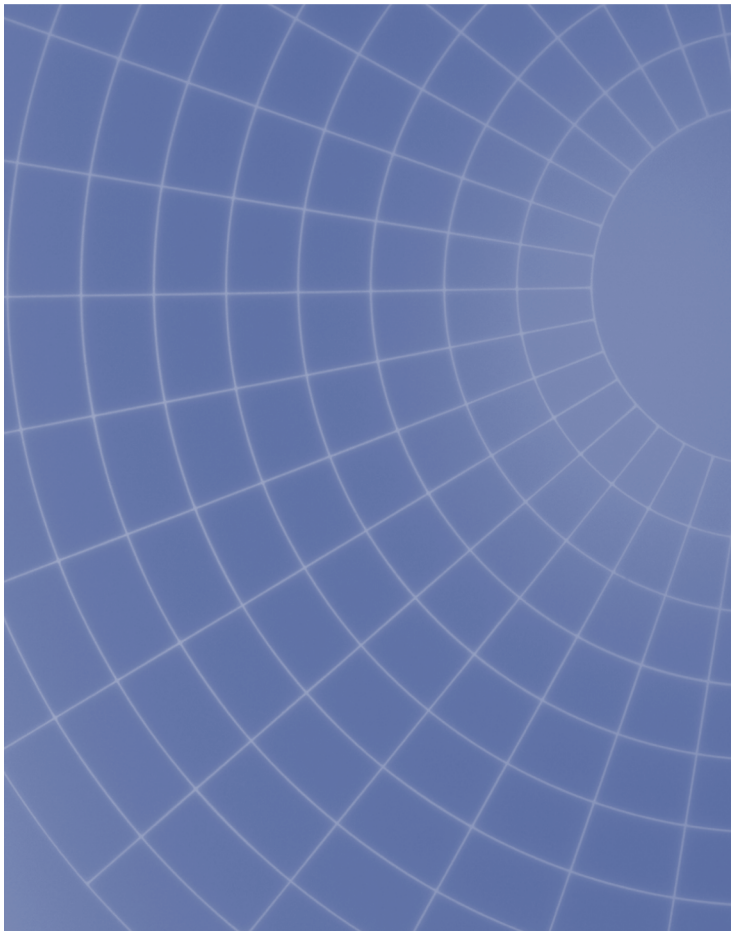
FDANEWS

Choosing the Best Device Sample Size for Verification and Validation



Copyright© 2014 by FDAnews. Print version ISBN: 978-1-60428-180-4. Digital version ISBN: 978-1-60428-181-1. Price: $397. All rights reserved. Photocopying or reproducing this report in any form, including electronic or facsimile transmission, scanning or electronic storage, is a violation of federal copyright law and is strictly prohibited without the publisher's express written permission.

This report may not be resold. FDAnews only sells its publications directly or through authorized resellers. Information concerning authorized resellers may be obtained from FDAnews, 300 N. Washington St., Suite 200, Falls Church, VA 22046-3431. Main telephone: (703) 538-7600. Toll free: (888) 838-5578.

While every effort has been made by FDAnews to ensure the accuracy of information in this report, this organization accepts no responsibility for errors or omissions. The report is sold as is, without warranty of any kind, either express or implied, respecting its contents, including but not limited to implied warranties for the report's quality, performance, merchantability, or fitness for any particular purpose. Neither FDAnews nor its dealers or distributors shall be liable to the purchaser or any other person or entity with respect to any liability, loss, or damage caused or alleged to be caused directly or indirectly by this report.

Choosing the Best Device Sample Size for Verification and Validation

Table of Contents

Statistical Techniques and Design Controls 3

Types of Applicable Problems 3

Confidence Level 4

Risk and Sample Size 5

Statistical Requirements for Verification and Validation 8

Types of Data Examined 9

Nonstatistical Techniques 10

Sources of Error 10

Statistical Justification 11

Determining Sample Size in Statistical Testing 14

Sample Collection Methods 15

Practical Examples 16

Appendices 22

A. FDA Design Control Guidance for Medical Device Manufacturers

B. NIST/SEMATECH e-Handbook of Statistical Methods

This report is largely drawn from materials presented during a recent FDAnews webinar led by Steven Walfish, president of Statistical Outsourcing Services. The consultancy provides statistical analysis and training to various industries, including the medical device manufacturing industry. Previously, Walfish was senior manager of biostatistics, nonclinical, at Human Genome Sciences and a senior associate at PricewaterhouseCoopers, where he specialized in the pharmaceutical industry.

Statistical Techniques and Design Controls

When looking at design verification and validation, the FDA expects devicemakers to answer two questions in a clear and detailed fashion:

- Did you make the product correctly?
- Did you make the right product?

Manufacturers may answer both questions through verification and validation testing, part of any design control requirements, using sample sizes appropriate for both the types of testing done and the types of product.

Design verification confirms that the design outputs meet the design input requirements. In other words, verification confirms that "you made it right," said Steven Walfish, president of Statistical Outsourcing Services.

Design validation confirms that the product fulfills its intended purpose or use and meets the user's needs. Validation confirms that a company "made the right thing," Walfish said, noting that a company would start the validation with a set of user requirements.

The use and justification of appropriate sample sizes is one of the requirements for effective verification and validation testing. This area receives a great deal of FDA attention during device approvals and inspections. From a statistical perspective, companies should figure out the best sample size to use for both types of testing. That helps "to get the level of comfort that we have, in fact, met the design inputs from our design output process, and that the final medical device does meet the intended use," Walfish said.

Types of Applicable Problems

Statistical methods can be applied to solve a diverse set of problems related to device verification and validation.

"There is a whole fleet of statistics that we can bring to bear during the validation and verification process," Walfish said. The correct sample size must be chosen for any of these activities, which include:

- Determining the usefulness of a limited number of test results of a product characteristic (hypothesis testing);
- Predicting the expected range of values for a product characteristic based on a limited number of test data (the confidence interval, or CI);
- Determining the number of tests required to provide adequate data to support a conclusion (the sample size);
- Comparing test data between two alternate designs in a design change (hypothesis testing);
- Predicting the amount of product that will fall within specifications (tolerance interval);
- Predicting system performance variation (analysis of variation, or ANOVA);

- Planning experiments to discover factors that influence a product's characteristic (design of experiments); and
- Determining the quantitative relationship between two or more variables (regression).

Understanding which statistical techniques to apply to which problems begins with what Walfish called the statistical building blocks. This requires a company to first answer a key question: What is the level of confidence the company wants to have for a particular device? Stated differently: How much risk is the company willing to live with as a result of its design and development process?

"We're not going to meet perfectly every design requirement or every user requirement," Walfish said. "There's going to be a certain risk that is going to occur."

The second thing companies must understand is the relationship between risk management and probabilistic models.

"A lot of people like to think of this — as some companies will call it — as a P1 value," Walfish said. This refers to the probability of a given failure occurring and, given that probability of occurrence, a determination of the severity of the impact on the patient if the failure does occur.

"That becomes your harm," he explained. "This is where we start to look at how you assess these things and get the right sample size to ensure you are not only meeting your user requirements, but also that you can assess the risk of putting the patient at any harm because of potential adverse consequences of that design."

Thus, the sample size not only helps device manufacturers meet their design verification and validation needs but also plays a key role in driving how they look at risk management, by relating risk to sample size.

With that in mind, companies must next examine the type of data they will be collecting and scrutinizing. Statistical data are generally categorized as either discrete or continuous. The root difference between the two is that discrete data are count data with a finite number of possible values, whereas continuous data have infinite possible outcomes. Continuous data are preferred for statistical exercises, such as developing sample sets for verification and validation testing, but are not always feasible.

With testing involving discrete data, a company will be doing simple pass/fail tests. Continuous data typically measure something about the output of a device, such as cycle times, voltages or pressures.

"Typically, very early on in design, we tend to set our specification around pass/fail attribute requirements," Walfish said. "So we count 50 cycles, and if after those 50 cycles it doesn't fail, we say that's a success and the device or component passed."

Confidence Level

The next step is for a company to look at how many devices, or units, it must test to provide sufficient confidence that zero failures in the sample can reasonably be interpreted to mean that the product meets the user requirements, including safety factors.

Two key factors influence this confidence level, which must be asserted through the use of statistics. Risk is the primary factor: A higher likelihood of harm to a user or patient requires a higher confidence

level. Related to sample size, this means that the sample size should be proportionate to the risk of the feature undergoing verification or validation. In other words, a greater risk potential calls for a larger sample size.

Conversely, if the potential for harm to a user or patient is less, the sample size necessary to achieve the desired confidence level will be smaller.

The second factor is variability, which covers variation from unit to unit or from batch to batch, as well as any variation in a company's measurement system. In short, where more variability is likely, a larger sample size is necessary to establish the desired confidence level. Conversely, where variability is small, sample sizes are likewise smaller.

Risk and Sample Size

It is impossible to discuss design controls without talking about risk management, as well. The ultimate goal of verification and validation testing is to show that there is a minimal risk that a device will not perform as intended. To prove this to the FDA's statistical satisfaction, companies must select sample sizes that will accurately reflect the real risk. To do this, devicemakers need to consider several factors. These include the product's overall risk profile.

"When you get to the verification and validation sample consideration piece, the most important thing is to tie it into your risk management," Walfish advised. "So when you are doing design verification and validation, your design history file and developing a risk plan, the sample size is ultimately going to be tied into that risk plan."

In a nutshell, a company can trade sample size for risk. The less risk a devicemaker is willing to take, the larger the sample size it will need. If more risk is acceptable, a smaller sample size will suffice.

Figure 1 shows a procedure that a device manufacturer might apply to determine whether to accept or reject a design based on the results of sample testing.

Figure 1

Risk and Sample Size

Confidence and reliability in a statistical assertion are expressed as Type I and Type II error.

Decision Based on Sample	Reality in Population		
		Accept	Reject
	Accept	Correct Decision	Type II Error (b) Consumer Risk
	Reject	Type I Error (a) Producer Risk	Correct Decision

"So we take a sample and, based on that sample, we either decide to accept or reject the design," Walfish explained. "Acceptable means the design meets the user's requirements. So if my sample says it's a good design, then I've made a correct design."

While that remains the goal of all verification and validation testing, in reality a devicemaker may find itself faced with a sample that shows an error. This can occur even if the design does, in fact, meet user requirements, Walfish noted. This would be a Type I error, or the "producer risk" noted in Figure 1.

This is where the concept of the CI comes into play, Walfish said. "When we say that we're 95 percent confident, we're saying that there's a 5 percent chance that we're going to reject the null hypothesis when it's true, or we're going to reject the design when, in reality, it's actually a good design."

CI changes are proportional to product variation and risk. The same applies to a Type II error, or risk to consumers; it is referred to as reliability.

"The most important situation is the Type II error, consumer risk," Walfish said. "This occurs when the sample size that we took — typically too small of a sample size — leads to a decision to accept the design when more testing would have led to a decision to reject it, and in fact, the consumers — the patients — are not going to get a device that doesn't meet their requirements."

As with the CI, a 95 percent reliability figure indicates that there is a 5 percent chance of a Type II error. Reliability is the more important of the two figures for device manufacturers. It also will receive more FDA scrutiny during an inspection. That is because reliability has a more direct bearing on the safety and efficacy of a device when it reaches the patient.

Accurate determination of appropriate confidence and reliability levels is critical to determining the correct sample size for the verification and validation testing of a specific device.

As a general rule of thumb, companies can look at the functionality of a particular device when determining appropriate levels of confidence and reliability at different risk levels. For instance, Walfish said, a confidence level of 95 percent and a reliability level of 90 percent would be appropriate for a low-risk device, such as a radiation shield for an X-ray machine. Such a device would need to be validated for its ability to shield an operator from extraneous radiation. This assumes that the failure mode and effects analysis (FMEA) and the product risk file do, in fact, identify the shield as a low-risk feature, Walfish emphasized.

If a product can be categorized as higher risk, those figures must change. In fact, devicemakers must be diligent in assuring that they have applied the correct risk category to any such product. For instance, for a medium-risk product, both the confidence and reliability figures would need to be at 95 percent, he suggested. Meanwhile, for a high-risk product, those figures would need to be 95 percent and 99 percent, respectively.

These figures are not carved in stone, Walfish emphasized, adding, "Every company, every organization is going to do this a little bit differently. I'm giving you a ballpark stick in the ground to work with. Every company has a different risk tolerance, and every company manufactures different risk and different classes of medical devices. If you're in the high-risk device world — Class III devices — you probably don't want to have anything less than 99 percent confidence and probably no less than 99 percent reliability, even for low-risk features."

With the risk-dependent confidence and reliability figures in place, a device manufacturer can determine the best sample size for its validation or verification testing by applying the following formula:

$$n = \ln(1 - \textit{Confidence)}/1\text{n}(\textit{Reliability}) = \ln(1 - 0.95)/1\text{n}(0.90) = 28$$

Using that formula, the sample size is 28 for a low-risk product with a confidence level of 95 percent and a reliability level of 90 percent. Walfish emphasized that that figure refers to the number of tests. The company would not have to necessarily test 28 devices; in fact, the number of individual devices tested could be as low as one, depending on variability among devices.

"If the variability of the device is small enough, I might be able to take a single device and test it 28 times and show that in all 28 times that I did that testing, my X-ray shield shielded the extraneous radiation," Walfish explained. "How do I do that? I take a Geiger counter, I put it by the shield, I shoot a beam, and I look at my Geiger counter, and if it's below a certain number, that's a pass. I do that 28 times; if all 28 times pass, then of course my validation and verification would pass."

To test using fewer devices, a company must be able to show there is minimal variability among devices; that will serve as its justification for not testing on 28 separate devices. That does not mean that a company must use only one device for the 28 tests. It could opt to use, for instance, seven devices, testing each four times.

The other important thing to note when using this approach in determining sample sizes, Walfish said, is that it assumes a straight pass/fail approach to the testing. This means that *all* 28 tests under the above example must "pass." If even just one test is a "fail," that means the device has failed verification or validation testing. As a result, the company must start over with a new series of 28 tests, after identifying and correcting any problems that contributed to the failure.

Statistical Requirements for Verification and Validation

When device manufacturers design verification and validation tests, particularly regarding choice of sample size, they must fully understand the requirements for statistical techniques, including how different techniques can affect the design control process. It is equally important that manufacturers be able to determine what is, and what is not, statistical in nature. In that way, they may understand what requirements lend themselves to statistics, and which are not statistical in nature, Walfish noted.

When applying statistics to these types of problems in the course of verification and validation testing, companies need to focus on the specific questions they need to answer. For instance, when a devicemaker undertakes hypothesis testing to see if a limited number of test results can accurately characterize a product feature, what the company really wants to determine is whether the sample being collected represents a good product characterization for the devices in question.

Likewise, when looking at the confidence level, the company is really trying to answer the question: How confident are we — 90 percent, 95 percent, 99 percent — that the true value of the data set of the sample population will fit inside this predetermined interval, which represents the likelihood that the device will perform as required.

"And when you have a hypothesis test and a confidence interval together, then we can come back and answer the question: 'How many samples do we need?'" Walfish explained. "My hypothesis test and my confidence interval are going to allow me to determine the appropriate sample size."

It is also important for devicemakers to remember that the key question may not necessarily be a design validation or verification question, he added. Sometimes a company just needs to do a hypothesis test to look at the differences between two designs, particularly in terms of whether certain design characteristics deliver better or worse performance, as discussed earlier in this report. This is another area where statistical analysis will prove useful.

Tolerance intervals are important to devicemakers. These figures indicate the ability of the company — and hence, of the user and the patient as well — to predict how much of a product will be within or outside specifications. For instance, Walfish said, if a manufacturer is 95 percent confident that 1 percent of its product could be out-of-spec, even though it doesn't observe any out-of-spec units during verification and validation, that company can still look at the variability and make a statement about what it estimates the out-of-spec, or out-of-tolerance, percentage of product will be.

Statistical techniques can be applied to predictions of system performance variation, as well. Here, companies would look at multiple factors simultaneously to see how they might affect the device design using ANOVA techniques.

Finally, statistics come into play during experimentation, particularly experimental design. Statistics help companies discover the factors that influence a product's characteristics, as well as the quantitative relationships between two or more variables during regression-type experiments, Walfish said.

Types of Data Examined

When devicemakers conduct validation and verification testing, they will look at two general types of data, as previously discussed. Continuous — or typical — data essentially apply to performance requirements. What companies must examine is part-to-part variability, a factor that is very important to a device's risk profile. In choosing a sample size for testing, companies must consider sources of variation, which can increase a device's risk.

For example, a tongue depressor represents a simple, low-risk medical device used frequently in doctors' offices. If a design requirement calls for the depressor to be flat, and not bowed, the company making this product must determine how much variability exists as part of its determining a sample size. Process variability and environmental impacts are among the sources of variability for this product, Walfish noted.

"But ultimately the question is: 'What is the risk to the patient if the physician was to get a slightly bowed tongue depressor?'" Walfish said. "It's probably a very low risk. So I'm not going to spend a lot of time in sample size and in the risk for this product."

However, a high-risk product, such as an implantable pacemaker, is another matter altogether, he said. If voltage is the specification to be tested, for instance, part-to-part variation in batteries and wiring, among other parts, can have an effect. The manufacturer has to take into account how those tolerances will affect the pacemaker's performance. And that in turn will affect the sample size for validation and verification testing.

Both verification and validation may require physical samples across lots, batches and operators to demonstrate that the specified requirements have been fulfilled. This must occur anytime a product has a high lot-to-lot variability.

"I used to work in capital equipment in the medical device industry, so you're making these $200,000 pieces of equipment. You're not making a lot of them during the validation and verification process. The assumption in that case is that the unit-to-unit variability is much less than the within-unit variability," Walfish said. "But if I'm making widgets, a plastic-injected molded part, and I'm making 30,000 of these things, it's probably going to be more important for me to be able to look at my verification and validation requirements across not just lots and batches, but across mold cavities and across different settings of my process. These are the things that are going to drive the variability of my process."

The second type of data that devicemakers may need to consider is discrete data, which typically involve functional requirements. As with continuous data, companies must consider sources of variation and risk in discrete data when determining sample size. However, individual parts or a process generally are not the sources of variation seen in discrete data. Rather, the variation stems from such things as interactions, timing, initial conditions, workflow, prior events, configurations, options and accessories.

"So if I'm going to look at a physical, visual characteristic of my medical device," Walfish said. "I don't know what causes that variability necessarily. I might need to go back and do some design experiments to understand what causes that variability because I don't have a variable output to be able to look at."

In some situations, Walfish explained, a company must test whether something works or not; he compared it to flicking a light switch: "If I turn on the switch in a room and the light goes on, I don't

have to test that 100 times. If it works once, it's always going to work. The design worked, in that it allowed the switch to turn on the light."

He also emphasized that validation and verification do not include process validation, but rather are intended simply to show that the design meets user requirements. In the light switch example above, the user wants the lights to come on when the switch is thrown. A comparable expectation in the device world might be when the user pushes a button, the device powers on. Another user requirement might be that a system boots up within one minute. These are all examples of discrete data.

In terms of sample size, devicemakers may use smaller numbers of physical samples, with repeated iterations to stress interactions across timing or users, for instance.

Going back again to the light switch example, Walfish said, "It's not as though I need to do that 30 times because the variability is almost nil between turning the switch on, or when I push a button today, later the same day and tomorrow. So we only have to worry about doing discrete data and statistical sample size for those characteristics that are discrete and have different sources of variability that we can quantify in the process."

Device companies also need to remember that, when dealing with discrete data, not all requirements are statistical in nature. Walfish stressed that companies must be able to provide the FDA with a documented rationale for when it considers a test nonstatistical versus statistical.

Nonstatistical Techniques

While the bulk of work developing sample sizes for verification and validation testing may center around statistical data and techniques, there are times when a nonstatistical approach — a simple pass/fail, it works or it does not work — is more appropriate. This type of testing can be done on a single unit of a product.

When companies opt to use a nonstatistical method, they must be able to justify that decision. This requires a clear understanding of when nonstatistical techniques are appropriate. Devicemakers must maintain written criteria for when it will consider a test to be nonstatistical versus statistical. These criteria can be as simple as a list of the types of tests that a company has deemed to be nonstatistical, Walfish said.

Companies should include the written criteria for statistical versus nonstatistical testing in the validation and verification protocol. The nonstatistical tests may sometimes be referred to as "unit tests," Walfish noted, referring to the ability to get the necessary information by testing just a single unit.

That justification comes down to risk. When the risk to the patient is lower, a nonstatistical unit test is more likely to be appropriate. Sometimes a nonstatistical test can be conducted simply to verify the existence of a feature, for instance. As long as the design includes that feature, and the test indicates that the feature is present, the user requirements have been met.

Sources of Error

All verification and validation testing must be concerned with variability, also known as error. There are, essentially, two sources of error.

The first source, which Walfish characterized as the most common and most-focused-on by devicemakers, is the variability of the device or component itself, or the sample statistic around the unit population. This is the standard deviation, and represents the part-to-part variability in the process.

The second source is measurement variability, which refers to variability in the measurement of a device characteristic or output due to the instrument used. The question asked here is: "How much variability will I get when I take the exact same unit and measure it repeatedly with the same instrument?" This source of error becomes particularly important when a company wants to look at variability within a part versus variability part-to-part.

"If my measurement system variability is very high, then I'm going to need to test the same unit more times than necessarily testing multiple units," Walfish said. "It becomes very important when we start to talk about how we're going to partition out our sample size."

Thus, companies must periodically perform a measurement system analysis (MSA). Walfish suggested that devicemakers should do so before undertaking any data collection and decision-making. MSAs also should be repeated periodically on a maintenance basis.

"Always, always, always get this measurement system under control," Walfish said. "Under the variability prior to doing any data collection and, more importantly, before you do any decision-making about the design."

Some key activities that could warrant an MSA include:

- Making product design decisions;
- Performing process validation;
- Improving phase in a black/green belt project; and
- Conducting a final product inspection.

Walfish placed extra emphasis on the first bullet point, saying, "If we're going to make a decision about the product design, we want to make sure that our measurement system is adequate, and if we've already done a measurement system analysis, that it's still valid, before we go about making any decisions about product design."

Finally, Walfish cautioned against focusing only on Type I errors — the so-called producer risk — when developing sampling plans.

"We focus on confidence levels. We focus in on doing hypothesis testing and getting 95 percent confidence. We focus in on doing a P value that's got to be less than 0.05," he said. "We focus, focus, focus on Type I errors, and it's a bad way to go around doing this, because a Type I error does not tell me the risk to a patient. All the Type I error tells me is what is the risk to us, the company."

While company risk is important, devicemakers also need to consider the FDA's priority, which is risk to the patient, or Type II errors. Thus, the agency would like to see sampling plans that also control for this type of risk.

Statistical Justification

Whenever a company comes up with a sample size for any validation or verification testing exercise, it must be able to justify that size in a way that will be acceptable to the FDA. A company should, for instance, have written criteria for determining when it considers a test to be statistical versus nonstatistical. This can be as simple as a list of the types of tests that it considers to be nonstatistical.

"In some cases, that discussion comes around risk," Walfish said. "You might make a statement that the risk to the patient is zero, so you're not going to do statistical justification. You're going to do a unit test and show that these are nonstatistical."

Companies also need to have written criteria for statistical testing. In some cases, they may rely on recognized standards for a particular statistical requirement, and these may specify a sample size and acceptance criteria. If this is the case, a company's statement that it is using a recognized standard may be considered a valid justification.

"For example, there is a standard for human factors," Walfish said. "So when we talk about sample size for user or usability testing, those sample sizes can be found in that particular standard document, which I think is HE75. A devicemaker can use that as its criteria."

He added that a company could test more or less than the standard, and use independent statistical justification as its rationale. The FDA allows a great deal of leeway for companies to use the approach that works best for any given individual situation.

"But if you don't want to come up with a statistical justification, it is sufficient to say that HE75, Annex A, requires this sample size for this level of confidence and this level of reliability," he said. "So ultimately, at the end of the day, you want to document these things, but sometimes standards exist that allow you to not have to document these particular tests."

When compiling a statistical justification, devicemakers must remember that there are two parts to every sample plan — the sample size and the acceptance criterion. Companies must justify both. The FDA expects that companies will have a rationale, in writing, for the sampling plan selected. The plan will be based on the risk to the patient. The confidence statements made about a product must be associated with passing a sampling plan.

For example, if a company is going to do a variables test on an attribute, by testing on continuous data, it might say that it is going to sample 60 units, with the mean value set between eight and 10. There must be a rationale to that acceptance criterion. The company must explain how it came up with all acceptance criteria as well as how it came up with the sample size.

"The rationale comes back to the risk to the patient," Walfish said. "You can say I am 95 percent confident with 90 percent reliability that I will meet the requirements. That says that I'm 95 percent confident that the true Type I error rate is no worse than 10 percent. So this has to be associated with a low-risk characteristic, because you're allowing yourself up to a 10 percent failure rate."

Q&As about Statistical Requirements for Verification and Validation

Question: *How is the sample size determination formula, $n = (Z_\alpha + Z_\beta)^2 S^2/\Delta^2$, different from the n = Zs over margin error squared?*

Answer: The Zs-squared over margin of error squared, or Zs over margin of error, the whole thing squared, only takes into account the Type I error. So, the Z is only the Type I error. What the first does is increase the sample size to effectively value both the Type I and Type II error. It will give a slightly larger sample size than the Zs over the meaningful difference.

Q: *In cases of process validation or relocation of manufacturing equipment, is $n = (Z_\alpha + Z_\beta)^2 S^2/\Delta^2$ appropriate?*

A: It's appropriate for everything. The only difference is that process validation or relocation will change the Type I, Type II errors because if a company has been making this product — relocation is a good example. The company is transferring production from one facility to another — it's been doing so for a while. It can probably run it at a lower confidence of reliability because it has a certain level of experience that hopefully will bear itself out.

Q: *When would a company use which sample size formulas?*

A: Three different sample size formulas have been discussed: The first — $n = (Z_{\alpha} + Z_{\beta})^2 S^2/\Delta^2$ — is for a mean value, which has to be within a certain spec. The sample size is the key value — $X + k * S < U$ — with the individual values being within a certain upper and lower specification. The natural log of one minus confidence over natural log of reliability ($X - k * S > L$). That's the attribute go/no-go, and that's when all of the tests must be found to be acceptable. So those are the three scenarios.

Q: *Is ANSI sampling used for manufacturing and not development?*

A: Don't use the ANSI Z1.4 through 1.9 for the purposes of validation — design validation or verification or process validation. Those standards are written for incoming inspection. There are switching rules associated with them, and that is not something seen in the design verification and validation activities.

Q: *Can process capability (Cpk) be used to size a sample size or criteria?*

A: Cpk can be related back to these things. Just multiply Cpk by 3; that would be the Z value. So, a company can say it has a Cpk of 1, that's the same thing as having a Z of 3. That's the same thing as having a reliability of about 0.27 percent. And then the company can use that information to come up with the sample size and the acceptance criteria that it needs to have for Cpk.

Q: *For calculating the sample size, which method is better, the first or the third?*

A: The first is used for continuous data where a company gets an actual reading, and the third for attribute where the only thing a company knows is it's either good or bad. So, if a company applies a force and it doesn't break, that's a Method 3. Or measure the force, and that's Method 1.

Q: *What is a source for the derivation on, for instance, the sample size for continuous data?*

A: The easiest one is $n = 1n(1 - Confidence)/1n(Reliability) = 1n(1 - 0.95)/1n(0 - 0.90) = 28$ at 95 percent confidence and 90 percent reliability; this is in almost every reliability textbook.

Q: *And how would the derivation for that come about?*

A: For the derivation of it, use the binomial distribution. It's just the binomial set X equals to zero.

Determining Sample Size in Statistical Testing

Once a devicemaker has determined the statistical requirements of its verification and validation activities, it is ready to determine the precise sample sizes needed for each phase of testing.

One very practical way of performing sample size calculations for continuous data is to use the idea of tolerance intervals — also known as the K value. Under this approach, a company would determine the confidence level and reliability and use the *ISO 16269-6* standard for tolerance levels (www.iso.org/obp/ui/#iso:std:iso:16269:-6:ed-2:v1:en) to find the value of K, as shown in the formulas below:

$$\overline{X} + k * S < U$$

$$\overline{X} - k * S > L$$

"Basically, this says that if my average, plus or minus K times the standard deviation, is within a certain upper and lower spec, I'm going to have a certain percent confidence that a certain percentage of individual values will be in the tolerance," Walfish explained.

He offered an example of a company that requires a 95 percent confidence and 95 percent reliability for a process with a mean of 10.0 and a standard deviation of 0.55, for instance, with a specification of 8.5 to 12.0.

"All we really have to do is focus on which specification the mean is closest to," Walfish explained. "That will give you the most conservative estimate of sample size, the largest sample size to meet this specification."

In this example, the mean, 10, is closest to the lower specification, so the company can solve for K thusly:

$$10 - k * 0.55 > 8.5,\ k = 2.727$$

Going to *ISO 16269-6*, the company can consult the table in the document's Appendix D, looking for 2.727 with 95 percent confidence and 95 percent reliability. There it would find an ISO-recommended sample size of 20 to 22. Walfish suggested that a sample size of 21 would be appropriate in this case.

"If all 21 units are between 8.5 and 12 when tested, then I can say that I'm 95 percent confident that 95 percent of the individual values will be within 8.5 and 12," he said. "If I need more reliability, I can do that. I'm going to keep the same K value of 2.727, but I'm going to go down a different column in that standard, and my sample size will be larger. In this example, your sample size gets somewhere around the high 30s in order to have 99 percent reliability."

Walfish noted that the National Institutes of Standards and Technology (NIST) engineering handbook (see Appendix B) also includes formulas for calculating K.

Another way to determine sample size for continuous data is via the standard hypothesis testing formula $n = (Z_\alpha + Z_\beta)^2\ S^2/\Delta^2$, where Z_α represents the Type I error, Z_β the Type II error, S the standard deviation and Δ the meaningful difference between the true population mean and what is observed in the sample. With the Δ value, the goal is to get as close as possible to the true population mean, so a small Δ is desirable.

In general, "the higher the sample size for continuous data, the closer the mean for my sample is to the true process mean," Walfish said.

For example, a company that makes a blood pressure cuff may need to be within ±2 mm of mercury of true pressure. Using the formula above, that means the Δ value is 2. The standard deviation (*S* in the formula above) represents variability between cuffs and within a cuff — e.g., in using the cuff to measure the blood pressure of the same person multiple times, how much variability in results can be expected, assuming that that individual's blood pressure does not change.

In this example, Walfish assumed a Type I error rate of 5 percent, meaning a 95 percent confidence level ($Z_{\alpha} = 1.96$), a Type II error rate of 1 percent for a 99 percent reliability figure ($Z_{\beta} = 2.326$) and unknown standard deviation and unknown Δ, but a requirement to detect a 0.5 standard deviation shift. It is assumed that this is a two-sided specification. Using the formula above, $n = (1.96 + 2.326)^2 (1)^2/0.5^2$, so the sample size, *n*, is 74.

"So this is a fairly robust method for knowing that there's only a 1 percent chance that I'm going to say that I'm within spec when in reality I'm not in spec," Walfish said. "That's a fairly good sampling plan for a mid- to high-risk decision."

Sample Collection Methods

Once a company has calculated a sample size for validation and verification testing, it must turn its attention to the process of collecting the samples, which the FDA also will scrutinize. Walfish characterized actual sample collection as "the hard part of the process."

"How do I know that I can take seven devices and do four replicates each, or if I can do four devices and do seven replicates each?" he asked rhetorically.

There are four general methods to collecting samples. The first and most common is known as simple random sampling, which involves randomly selecting the correct number of samples from the parts made. Each unit is selected independently, so that any and all samples are equally likely to be selected into making up sample size *n*.

"That makes a lot of sense if I'm going to make 30,000 injection-molded parts and I need to get a sample size of 300," Walfish said. "I'll randomly pick 300."

A company making a more complex device, such as an X-ray machine, rather than making a given number of machines or sub-assemblies, may want to be able to break up the testing into discrete points around a functional feature or features of the device.

"So I might take an X-ray tube sub-assembly and do some samples at a very low dose, some at a very high dose and some at a medium dose," Walfish said. "And say I have determined I have to do 28 tests; I'll make it 30 and do 10 tests each at the low, medium and high dose levels. If all 30 pass, then I'm going to say that the design passes."

A second and also common approach, particularly for high-volume devices, is systematic sampling. In such a case, a company designates that a sample will be pulled at regular intervals. For instance, Walfish said, a company might decide to pull for testing every 100th device manufactured. For a manufacturing run of 3,000 devices, that would provide a sample size of 30.

An advantage of this method is that it ensures that a devicemaker will subject to testing items from the beginning, middle and end of a production run on a particular production line.

"When I do that, I'm going to ensure that I'm going to get samples all across the process so that I get my best chance of testing the true functionality of the device and the design," Walfish said.

A third approach is known as composite sampling, in which a company takes parts and combines them into a single test. This approach can save costs, but does not provide complete information about the individual sampling units.

"Composite sampling doesn't really work very well in medical devices," Walfish cautioned. "But if you have a chemical process or an in vitro device, for instance, composite sampling can work in many cases."

Finally, the fourth approach is stratified random sampling. It ensures that a sample represents each of various predefined strata. This allows a company to either minimize the variability of the estimator or to minimize costs, Walfish said.

Practical Examples

Devicemakers may face some common scenarios that require particular approaches to verification and validation testing, as well as require different methods of choosing sample sizes. Walfish noted that he drew these examples from approaches that medical device manufacturers have actually used, and that have passed FDA muster.

A company may need to compare new and old designs with the goal of showing that the new design represents an improvement.

"In many cases, what happens is we're not actually making a brand-spanking-new design," Walfish said. "In reality, what we're doing is making an alteration to a current design, and we want to know if the change is an improvement, which typically means a reduction in the amount of the failure rate that the user or patient would see."

This sort of evaluation can be done using either discrete or continuous data to provide documented evidence of equivalence of the two designs or of an improvement in the newer design.

For example, the current design — designated Design A — has a 5 percent failure rate for, say, image quality. To check if a new design — Design B — has a lower failure rate for image quality, the manufacturer would have to have a 95 percent confidence that the new design reduces the failure rate by, for example, 1 percent.

Image quality could be considered a discrete data point. However, testing it as such may not be the most practical approach, as it would require a very large sample size.

The lower limit for calculations is 4 percent, the lower failure rate expected for Design B and the upper limit, represented by $\hat{p}$, is 5 percent, the failure rate for Design A. Solving for *n*, as in the formula below, a sample size of 1,285 would be required for a 1 percent reduction in failure rate.

$$n = \frac{Z_{\alpha}^{2} \times \left(\hat{p} \times (1 - \hat{p})\right)}{(Lower\ Limit - \hat{p})^{2}}$$

"No one's going to want to do that," Walfish said. "In most cases, we're not going to look at a criterion like reducing image quality. We're going to score the image quality and look for it as a continuous variable."

There are occasional circumstances where a sample size of just one may be an acceptable size, or even the most appropriate size, for testing a particular product. For instance, a patient table, such as those used in an operating suite or a doctor's office, must be verified and validated for its ability to hold a declared amount of weight. If the claim is that a given table can hold, for instance, 250 pounds, the manufacturer must show its documented rationale for why there is a negligible amount of variability among the tables it has constructed.

In one case, Walfish said, a company tested a single table. However, instead of testing the table with a load of 250 pounds, it did so with 800 pounds. The company argued that showing that the table can sustain a load of 800 pounds provided sufficient confidence that any table it produced would be able to meet the 250-pound requirement. The FDA accepted that rationale.

"That's not a statistical discussion; it's a scientific discussion," Walfish explained, also noting that the table had to undergo about 30 additional tests, many having to do with specific capabilities, including its ability to rotate. "*N* equals one and they did it through what is called accelerated testing."

Another feature that could lend itself to this sort of accelerated testing using a sample size of one might be that the instrument boots up when the power is turned on. This might not be appropriate, however, if the device is made with a specification or requirement that it boot up in less than, for instance, five seconds, Walfish cautioned.

"It depends on the variability from machine to machine," he explained. "But if we're asking if it's either present or not present — boots up or does not boot up — in that case, we could be looking at a sample size of one."

A third example might be the functionality of an alarm state for a given failure mode. In such a case, the manufacturer needs to demonstrate that the alarm will go off when that particular failure mode occurs. Like the weight capability for the operating table and the booting up question discussed previously, this represents a simple yes/no question.

For instance, if a device is supposed to alarm if a particular wire develops a kink, the test would involve kinking the wire and then determining if that action triggered the alarm. If it does, that is a pass. Likewise, if a plugged tube is supposed to cause an alarm, the company would test that by purposefully plugging the tube to determine whether the device alarmed as intended.

"You can do a greater-than-one sample size, but what I'm getting at is there are times when you only have to do one and show that for this failure mode there is an alarm that occurs," Walfish said, adding, "That does not mean that there won't be a sample size greater than one if you look at the total number of failure modes."

Walfish noted that many companies like to choose a sample size of three. When queried, they often will respond that they do so because the preamble to regulations mentions a minimum of three lots, or simply because they have always done it that way. However, a sample size of three "is not always the answer," he warned.

Walfish gave the example of a simple verification or validation test of a simple device attribute. If all three devices in the sample pass, the company is saying that it is 99 percent confident that the true error rate is less than 82 percent, or that it is 95 percent confident that the true error rate is less than 71 percent.

"Those are not good odds," he said. "Any time your liability is less than 50 percent, you might as well just do a sample size of one and flip a coin."

The problem, he explained, is that for continuous data, a 95 percent confidence level and 90 percent reliability level can only detect a shift greater than 1.7 standard deviations from the mean.

"Two would be 95 percent confidence and 1.7 is 90 percent, so we're really far out," he said. "So a sample size of three is not appropriate."

Even a very small increase — to a sample size of four, for instance — will increase the confidence figures to 99 percent confidence that the true error rate is less than 73 percent and 95 percent confidence that the rate is less than 60 percent. A sample size of four also reduces the standard deviation to 1.5, from 1.7.

"You can figure out that once you work your way up to five, six, seven and so on, you're going to start to see a rapid increase in the reliability side of things, a reduction of Type II errors," Walfish said. "And when you get around a sample size of about 30 — this magical 30 that people talk about — you start to see a leveling off in the impact on the reliability, so you have to make the sample size quite large after that to get small, incremental changes in reliability."

Q&As on Determining Sample Size in Statistical Testing

Question: *On nonstatistical examples and the sample size of one, is this a "common sense" discussion? And if not, using the surgical table example, it seems that there would have had to have been data first to show that there's very little variability from one table to another in order for the 800-pound test to really be valid for the product line. Do they actually provide data on very low variability for surgical tables, for example?*

Answer: Correct. This is not the company's first foray into surgical tables. It added some features, but one of the design requirements was that the tables hold at least a 250-pound force. The company said it knows from its designs that the tables can hold up to 800 pounds, so it opted to do an 800-pound test to show that it's sufficient and get that user requirement out of the way.

Q: *Is talk about risk about the probability event, the harm or a combination of the two?*

A: It's about the entire risk-management process. Ultimately, it comes down to the harm, but the harm is made up of probability statements themselves. So, what's the probability of occurrence, and then given that it occurred, what's the severity of that event? For instance, what's the probability that the catheter is going to break, and then if it does break, what is the ultimate severity of that? Put these two things together to get the harm. What companies really focus on from the sample size is the harm side of it.

But companies aren't really focusing on the harm. They end up focusing on the probability of occurrence, and that depends on their risk table; if their probability of occurrence is very high, but their severity is low, they'll take a different position than if the severity is high and the probability low.

Q: *So a low-risk feature on a Class II product may not be the same as a low-risk feature on a Class III product. What's the strategy for handling something like that?*

A: Typically, what differentiates a Class II from a Class III is that Class IIIs tend to be implantable devices, so they require medical intervention to either be inserted or removed. That's a general statement. It's true that there are Class III devices that are nonimplantable.

So for a Class III, a risk, even though it's a low risk of it occurring, if it requires medical intervention to resolve the issue, that becomes a higher risk than a low risk in a Class II that is a nonimplantable device. So companies want to have a little bit more confidence in reliability, even if it's low risk — meaning low probability of occurrence, low severity if it occurs — if medical intervention would be necessary, that puts the overall patient risk higher.

Q: *When determining risk to be mitigated in validation, should the acceptance criteria be the same for both challenge and nominal conditions?*

A: Yes. Don't make a distinction between the two. A lot of people do. So, if there's a challenge condition, it needs to perform similarly to nominal conditions.

Q: *How can later process controls and production or measurement of certain values in production be used to limit the sample sizes of verification and validation?*

A: If there's a certain level of control in place in the process, that can allow a company to reduce its Type I and Type II errors to get a small sample size. Many companies don't have that experience and functionality yet. But if a company has it, then it can reduce its sample size.

Q: *A company is doing a verification test on a stent deployment using, say, 30 samples and the stents have to deploy under a certain amount of force, so the company does 30. Then the company wants to validate the design and to do it in an animal model. But the company can only deploy one stent per animal and doesn't want to kill 30 animals, so how does the company justify a sampling size of, say, one?*

A: OK, so the company did 30, and all 30 deployed. Now if, in fact, the company is only looking at a go/no-go, it deployed or it didn't deploy, and didn't actually measure the force required to deploy, the company is going to be able to leverage that. That 30 give the company 95 percent confidence and 90 percent reliability.

Now, as for the validation piece, and the company is going to do it in an animal model, and it wants to have a reduced sample size — one — for ethical and cost reasons. It's just trying to show that in an animal model, it's going to get results that are similar or comparable to what it got during verification testing. And so now, it's no longer statistical justification. It's a scientific justification.

The problem is that usually, the FDA is not going to allow a company to get away with a sample size of one. It's going to usually require more than that or a simulation that a company can show that ties in its bench testing to the animal model.

Q: *If the company does three stent deployments in animals, how can it justify not doing 30, even though the company deployed 30?*

A: Everyone has that same fear: How do we justify it? It's not three that the company's passing; it's now zero failures in 33 deployments. That's sort of the idea of the stratification. It all comes back to the acceptance criteria and what is the risk the company is willing to take of having a late-stage failure. Because the company could do 15 simulated or 15 sub-components and say it's going to do 15 in the animal, but if it fails in the animal, it fails the entire design validation and verification. The company can put the two together, but the risk is that if it fails, it fails it all.

That's why most people say, 'Oh, I don't want to do that. I'd rather do 30, get enough confidence and reliability that this stuff is going to work, so that when I go into the animal model I'm not going to have any problems, and so do a smaller sample size.' A company can leverage its past work.

Q: *With regard to the idea that sample size does not imply unique units, does that mean that if a company is making spherical bearings — ball bearings — and its sample size is 30, does it need to measure 30 bearings?*

A: If the company is looking at, say, the diameter of a ball bearing, no. But if it's looking at the functionality of a ball bearing, and the functionality test doesn't destroy the ball bearing, it can show functionality multiple times. But most of the time when the company talks about functional testing, there's going to be some destruction to the ball bearing. But if it's just that it works, then yes, it's not an issue.

Q: *What about a real discrete component such as a ball bearing as to some other device that produces a signal. It could be a voltmeter. One turns it on, one measures the voltage on a battery, one gets a number. One turns it off, one turns it on. One can use that same digital voltmeter to measure that voltage source, and the assumption there is that all voltmeters are similar.*

A: That the variability between voltmeters is less than the variability within or between batteries, yes.

Q: *Is there a difference in sample size criteria between process verification and design verification?*

A: There's not a difference in sample size. There's a difference in the risk. Typically, for process validation, there doesn't have to be as high a level of confidence and reliability as there is in design verification and validation because typically a company has leveraged its experience in order to do that.

Q: *Is it true that if a company doesn't have a rationale, it can't pick a sample size? And is it correct that, since sample size should be proportionate to the risk to the business and the patient, if a company gets a certain number whether the scale is 1 to 100 or 1 to 1,000, it could assign a numeric range for low-, medium- and high-risk, and ultimately use those values however it desires to calculate its sample size?*

A: Yes, to both questions. That is correct. A company *should*, not *could*, assign a numeric range in that way. In other words, if it has a range of 1 to 100, any risk between 1 and 25 is low, 26 to 70 is medium and 71 to 100 is high. And set that across its entire product portfolio.

Q: *What about when a company runs a test and has data. The company did its sample size calculation. And it had one that failed. Can it go back and test additional specimens in order to get to its prescribed number?*

A: No, it can't do that. Say it tests 30 and it's supposed to accept on zero, and it gets one failure out of 30. Can it then go calculate how big of a sample size it would need to have with one failure to have the same level of confidence and reliability? No, because it can't continually retest until it passes. Because even if it says, OK, it'll exhaust the entire population but it passes, once it sees a failure, it has to start testing over again, supposedly with the idea that it has made some adjustments to the design. But it can't continually test unless it builds it in prior to testing. And there are statistical methods with double and sequential sampling plans. That's always a possibility.

Q: *What is the difference when a company predicts it or builds it into its protocol versus adding it on afterward?*

A: The difference is that if a company builds them into the protocol, there are stopping rules that are in place for failure. Typically, the way it works is that if there are zero failures, a company accepts. If there are two failures, it rejects. If there's one failure, the company then can take a second sample. So it

specifies how big the sample size will be and how many times it can sample. What a company doesn't want to do is have a sense of trying to work around the system.

Q: *So, it's more just showing that a company is not letting it look like it's working around the system. Because it's not actually changing the population.*

A: Correct. But ultimately, that's the other problem. So, at some point, the company is no longer doing sampling statistics. It's actually doing population statistics because it's sampling almost the entire population.

Q: *Could a company refer to a 510(k) in this same area to see what another company did under the Freedom of Information Act (FoIA) and get some sense of what's worked before in this context?*

A: Yes, if a company has a similar equivalent device and it's available via FoIA, as long as it's not redacted, a company can go for it. The FDA will not commit to a value. The agency will not say whether it will accept 95/90. It says that's each company's decision, based on the risk of the device.

Appendices

A. FDA Design Control Guidance for Medical Device Manufacturers

B. NIST/SEMATECH e-Handbook of Statistical Methods

Appendix A: FDA Design Control Guidance for Medical Device Manufacturers

Center for Devices and Radiological Health

DESIGN CONTROL GUIDANCE

FOR

MEDICAL DEVICE MANUFACTURERS

This Guidance relates to
FDA 21 CFR 820.30 and Sub-clause 4.4 of ISO 9001

March 11, 1997

FOREWORD

To ensure that good quality assurance practices are used for the design of medical devices and that they are consistent with quality system requirements worldwide, the Food and Drug Administration revised the Current Good Manufacturing Practice (CGMP) requirements by incorporating them into the Quality System Regulation, 21 CFR Part 820. An important component of the revision is the addition of design controls.

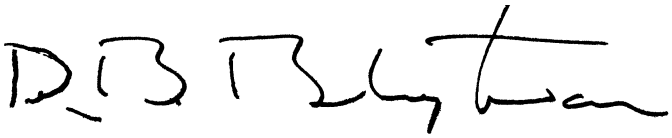
Because design controls must apply to a wide variety of devices, the regulation does not prescribe the practices that must be used. Instead, it establishes a framework that manufacturers must use when developing and implementing design controls. The framework provides manufacturers with the flexibility needed to develop design controls that both comply with the regulation and are most appropriate for their own design and development processes.

This guidance is intended to assist manufacturers in understanding the intent of the regulation. Design controls are based upon quality assurance and engineering principles. This guidance complements the regulation by describing its intent from a technical perspective using practical terms and examples.

Draft guidance was made publicly available in March, 1996. We appreciate the many comments, suggestions for improvement, and encouragement we received from industry, interested parties, and the Global Harmonization Task Force (GHTF) Study Group 3. The comments were systematically reviewed, and revisions made in response to those comments and suggestions are incorporated in this version. As experience is gained with the guidance, FDA will consider the need for additional revisions within the next six to eighteen months.

The Center publishes the results of its work in scientific journals and in its own technical reports. Through these reports, CDRH also provides assistance to industry and to the medical and healthcare professional communities in complying with the laws and regulations mandated by Congress. These reports are sold by the Government Printing Office (GPO) and by the National Technical Information Service (NTIS). Many reports, including this guidance document, are also available via Internet on the World Wide Web at www.fda.gov.

We welcome your comments and suggestions for future revisions.



D. Bruce Burlington, M.D.
Director
Center for Devices and Radiological Health

PREFACE

Effective implementation of design controls requires that the regulation and its intent be well understood. The Office of Compliance within CDRH is using several methods to assist manufacturers in developing this understanding. Methods include the use of presentations, teleconferences, practice audits, and written guidance.

Those persons in medical device companies charged with responsibility for developing, implementing, or applying design controls come from a wide variety of technical and non-technical backgrounds—engineering, business administration, life sciences, computer science, and the arts. Therefore, it is important that a tool be provided that conveys the intent of the regulation using practical terminology and examples. That is the purpose of his guidance.

The response of medical device manufacturers and other interested parties to the March, 1996 draft version of this guidance has significantly influenced this latest version. Most comments centered on the complaint that the guidance was too prescriptive. Therefore, it has been rewritten to be more pragmatic, focusing on principles rather than specific practices.

It is noteworthy that many comments offered suggestions for improving the guidance, and that the authors of the comments often acknowledged the value of design controls and the potential benefit of good guidance to the medical device industry, the public, and the FDA. Some comments even included examples of past experiences with the implementation of controls.

Finally, there are several people within CDRH that deserve recognition for their contributions to the development of this guidance. Al Taylor and Bill Midgette of the Office of Science and Technology led the development effort and served as co-chairs of the CDRH Design Control Guidance Team that reviewed the comments received last spring. Team members included Ashley Boulware, Bob Cangelosi, Andrew Lowrey, Deborah Lumbardo, Jack McCracken, Greg O'Connell, and Walter Scott. As the lead person within CDRH with responsibility for implementing the Quality System Regulation, Kim Trautman reviewed the guidance and coordinated its development with the many other concurrent and related activities. Their contributions are gratefully acknowledged.

FDA would also like to acknowledge the significant contributions made by the Global Harmonization Task Force (GHTF) Study Group 3. The Study Group reviewed and revised this guidance at multiple stages during its development. It is hoped that this cooperative effort will lead to this guidance being accepted as an internationally recognized guidance document through the GHTF later this year.

Lillian J. Gill
Director
Office of Compliance

ACKNOWLEDGEMENT

FDA wishes to acknowledge the contributions of the Global Harmonization Task Force (GHTF) Study Group 3 to the development of this guidance. As has been stated in the past, FDA is firmly committed to the international harmonization of standards and regulations governing medical devices. The GHTF was formed in 1992 to further this effort. The GHTF includes representatives of the Canadian Ministry of Health and Welfare; the Japanese Ministry of Health and Welfare; FDA; industry members from the European Union, Australia, Canada, Japan, and the United States; and a few delegates from observing countries.

Among other efforts, the GHTF Study Group 3 started developing guidance on the application of design controls to medical devices in the spring of 1995. Study Group 3 has recognized FDA's need to publish timely guidance on this topic in conjunction with promulgation of its new Quality System Regulation. The Study Group has therefore devoted considerable time and effort to combine its draft document with the FDA's efforts as well as to review and comment on FDA's subsequent revisions. FDA, for its part, delayed final release of its guidance pending final review by the Study Group. As a result, it is hoped that this document, with some minor editorial revisions to make the guidance global to several regulatory schemes, will be recognized through the GHTF as an international guidance document.

TABLE OF CONTENTS

FOREWORD i
PREFACE iii
ACKNOWLEDGEMENT iv
TABLE OF CONTENTS v
INTRODUCTION 1
SECTION A. GENERAL 7
SECTION B. DESIGN AND DEVELOPMENT PLANNING 9
SECTION C. DESIGN INPUT 13
SECTION D. DESIGN OUTPUT 19
SECTION E. DESIGN REVIEW 23
SECTION F. DESIGN VERIFICATION 29
SECTION G. DESIGN VALIDATION 33
SECTION H. DESIGN TRANSFER 37
SECTION I. DESIGN CHANGES 39
SECTION J. DESIGN HISTORY FILE (DHF) 43

INTRODUCTION

I. PURPOSE

This guidance is intended to assist manufacturers in understanding quality system requirements concerning design controls. Assistance is provided by interpreting the language of the quality systems requirements and explaining the underlying concepts in practical terms.

Design controls are an interrelated set of practices and procedures that are incorporated into the design and development process, i.e., a system of checks and balances. Design controls make systematic assessment of the design an integral part of development. As a result, deficiencies in design input requirements, and discrepancies between the proposed designs and requirements, are made evident and corrected earlier in the development process. Design controls increase the likelihood that the design transferred to production will translate into a device that is appropriate for its intended use.

In practice, design controls provide managers and designers with improved visibility of the design process. With improved visibility, managers are empowered to more effectively direct the design process—that is, to recognize problems earlier, make corrections, and adjust resource allocations. Designers benefit both by enhanced understanding of the degree of conformance of a design to user and patient needs, and by improved communications and coordination among all participants in the process.

The medical device industry encompasses a wide range of technologies and applications, ranging from simple hand tools to complex computer-controlled surgical machines, from implantable screws to artificial organs, from blood-glucose test strips to diagnostic imaging systems and laboratory test equipment. These devices are manufactured by companies varying in size and structure, methods of design and development, and methods of management. These factors significantly influence how design controls are actually applied. Given this diversity, this guidance does not suggest particular methods of implementation, and therefore, must not be used to assess compliance with the quality system requirements. Rather, the intent is to expand upon the distilled language of the quality system requirements with practical explanations and examples of design control principles. Armed with this basic knowledge, manufacturers can and should seek out technology-specific guidance on applying design controls to their particular situation.

When using this guidance, there could be a tendency to focus only on the time and effort required in developing and incorporating the controls into the design process. However, readers should keep in mind the intrinsic value of design controls as well. It is a well-established fact that the cost to correct design errors is lower when errors are detected early in the design and development process. Large and small companies that have achieved quality systems certification under ISO 9001 cite improvements in productivity, product quality, customer satisfaction, and company competitiveness. Additional benefits

are described in comments received from a quality assurance manager of a medical device firm regarding the value of a properly documented design control system:

> "...there are benefits to an organization and the quality improvement of an organization by having a written design control system. By defining this system on paper, a corporation allows all its employees to understand the requirements, the process, and expectations of design and how the quality of design is assured and perceived by the system. It also provides a baseline to review the system periodically for further improvements based on history, problems, and failures of the system (not the product)."

II. SCOPE

The guidance applies to the design of medical devices as well as the design of the associated manufacturing processes. The guidance is applicable to new designs as well as modifications or improvements to existing device designs. The guidance discusses subjects in the order in which they appear in FDA's Quality System regulation and is cross-referenced to International Organization for Standards (ISO) 9001:1994, *Quality Systems—Model for Quality Assurance in Design, Development, Production, Installation, and Servicing*, and the ISO draft international standard ISO/DIS 13485, *Quality Systems—Medical Devices—Particular Requirements for the Application of ISO 9001*, dated April 1996.

Design controls are a component of a comprehensive quality system that covers the life of a device. The assurance process is a total systems approach that extends from the development of device requirements through design, production, distribution, use, maintenance, and eventually, obsolescence. Design control begins with development and approval of design inputs, and includes the design of a device and the associated manufacturing processes.

Design control does not end with the transfer of a design to production. Design control applies to all changes to the device or manufacturing process design, including those occurring long after a device has been introduced to the market. This includes evolutionary changes such as performance enhancements as well as revolutionary changes such as corrective actions resulting from the analysis of failed product. The changes are part of a continuous, ongoing effort to design and develop a device that meets the needs of the user and/or patient. Thus, the design control process is revisited many times during the life of a product.

Some tools and techniques are described in the guidance. Although aspects of their utility are sometimes described, they are included in the guidance for illustrative purposes only. Including them does not mean that they are preferred. There may be alternative ways that are better suited to a particular manufacturer and design activity. The literature contains an abundance of information on tools and techniques. Such topics as project management, design review, process capability, and many others referred to in this guidance are

available in textbooks, periodicals, and journals. As a manufacturer applies design controls to a particular task, the appropriate tools and techniques used by competent personnel should be applied to meet the needs of the unique product or process for that manufacturer.

III. APPLICATION OF DESIGN CONTROLS

Design controls may be applied to any product development process. The simple example shown in Figure 1 illustrates the influence of design controls on a design process.

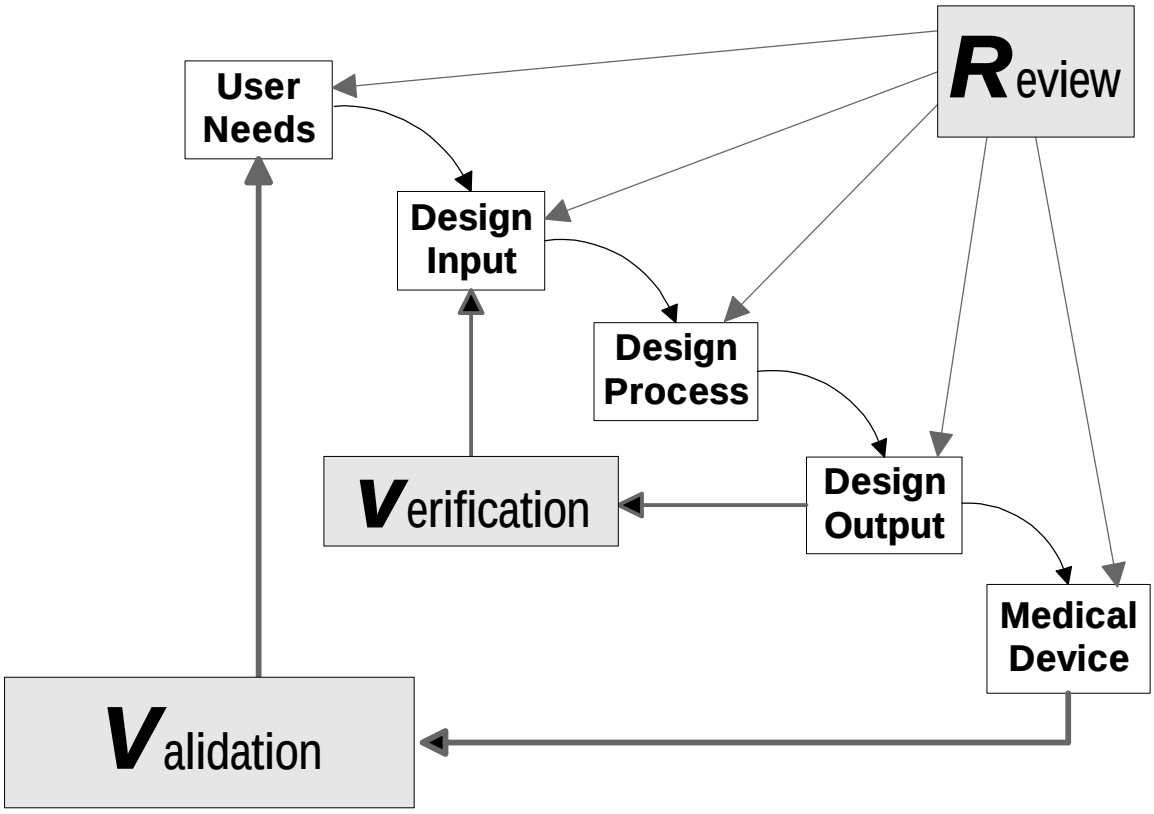


Figure 1 – Application of Design Controls to Waterfall Design Process (figure used with permission of Medical Devices Bureau, Health Canada)

The development process depicted in the example is a traditional waterfall model. The design proceeds in a logical sequence of phases or stages. Basically, requirements are developed, and a device is designed to meet those requirements. The design is then evaluated, transferred to production, and the device is manufactured. In practice, feedback paths would be required between each phase of the process and previous phases, representing the iterative nature of product development. However, this detail has been omitted from the figure to make the influence of the design controls on the design process more distinct.

The importance of the design input and verification of design outputs is illustrated by this example. When the design input has been reviewed and the design input requirements are determined to be acceptable, an iterative process of translating those requirements into a device design begins. The first step is conversion of the requirements into system or high-level specifications. Thus, these specifications are a design output. Upon verification that the high-level specifications conform to the design input requirements, they become the design input for the next step in the design process, and so on.

This basic technique is used repeatedly throughout the design process. Each design input is converted into a new design output; each output is verified as conforming to its input; and it then becomes the design input for another step in the design process. In this manner, the design input requirements are translated into a device design conforming to those requirements.

The importance of design reviews is also illustrated by the example. The design reviews are conducted at strategic points in the design process. For example, a review is conducted to assure that the design input requirements are adequate before they are converted into the design specifications. Another is used to assure that the device design is adequate before prototypes are produced for simulated use testing and clinical evaluation. Another, a validation review, is conducted prior to transfer of the design to production. Generally, they are used to provide assurance that an activity or phase has been completed in an acceptable manner, and that the next activity or phase can begin.

As the figure illustrates, design validation encompasses verification and extends the assessment to address whether devices produced in accordance with the design actually satisfy user needs and intended uses.

An analogy to automobile design and development may help to clarify these concepts. Fuel efficiency is a common design requirement. This requirement might be expressed as the number of miles-per-gallon of a particular grade of gasoline for a specified set of driving conditions. As the design of the car proceeds, the requirements, including the one for fuel efficiency, are converted into the many layers of system and subsystem specifications needed for design. As these various systems and subsystems are designed, design verification methods are used to establish conformance of each design to its own specifications. Because several specifications directly affect fuel efficiency, many of the verification activities help to provide confirmation that the overall design will meet the fuel efficiency requirement. This might include simulated road testing of prototypes or actual road testing. This is establishing by objective evidence that the design output conforms to the fuel efficiency requirement. However, these verification activities alone are not sufficient to validate the design. The design may be validated when a representative sample of users have driven production vehicles under a specified range of driving conditions and judged the fuel efficiency to be adequate. This is providing objective evidence that the particular requirement for a specific intended use can be consistently fulfilled.

CONCURRENT ENGINEERING. Although the waterfall model is a useful tool for introducing design controls, its usefulness in practice is limited. The model does apply to the development of some simpler devices. However, for more complex devices, a concurrent engineering model is more representative of the design processes in use in the industry.

In a traditional waterfall development scenario, the engineering department completes the product design and formally transfers the design to production. Subsequently, other departments or organizations develop processes to manufacture and service the product. Historically, there has frequently been a divergence between the intent of the designer and the reality of the factory floor, resulting in such undesirable outcomes as low manufacturing yields, rework or redesign of the product, or unexpectedly high cost to service the product.

One benefit of concurrent engineering is the involvement of production and service personnel throughout the design process, assuring the mutual optimization of the characteristics of a device and its related processes. While the primary motivations of concurrent engineering are shorter development time and reduced production cost, the practical result is often improved product quality.

Concurrent engineering encompasses a range of practices and techniques. From a design control standpoint, it is sufficient to note that concurrent engineering may blur the line between development and production. On the one hand, the concurrent engineering model properly emphasizes that the development of production processes is a design rather than a manufacturing activity. On the other hand, various components of a design may enter production before the design as a whole has been approved. Thus, concurrent engineering and other more complex models of development usually require a comprehensive matrix of reviews and approvals to ensure that each component and process design is validated prior to entering production, and the product as a whole is validated prior to design release.

RISK MANAGEMENT AND DESIGN CONTROLS. Risk management is the systematic application of management policies, procedures, and practices to the tasks of identifying, analyzing, controlling, and monitoring risk. It is intended to be a framework within which experience, insight, and judgment are applied to successfully manage risk. It is included in this guidance because of its effect on the design process.

Risk management begins with the development of the design input requirements. As the design evolves, new risks may become evident. To systematically identify and, when necessary, reduce these risks, the risk management process is integrated into the design process. In this way, unacceptable risks can be identified and managed earlier in the design process when changes are easier to make and less costly.

An example of this is an exposure control system for a general purpose x-ray system. The control function was allocated to software. Late in the development process, risk analysis of the system uncovered several failure modes that could result in overexposure to the

patient. Because the problem was not identified until the design was near completion, an expensive, independent, back-up timer had to be added to monitor exposure times.

THE QUALITY SYSTEM AND DESIGN CONTROLS. In addition to procedures and work instructions necessary for the implementation of design controls, policies and procedures may also be needed for other determinants of device quality that should be considered during the design process. The need for policies and procedures for these factors is dependent upon the types of devices manufactured by a company and the risks associated with their use. Management with executive responsibility has the responsibility for determining what is needed.

Example of topics for which policies and procedures may be appropriate are:

- risk management
- device reliability
- device durability
- device maintainability
- device serviceability
- human factors engineering
- software engineering
- use of standards
- configuration management
- compliance with regulatory requirements
- device evaluation (which may include third party product certification or approval)
- clinical evaluations
- document controls
- use of consultants
- use of subcontractors
- use of company historical data

SECTION A. GENERAL

I. REQUIREMENTS

§ 820.30(a) General.

(1) Each manufacturer of any class III or class II device, and the class I devices listed in paragraph (a) (2) of this section, shall establish and maintain procedures to control the design of the device in order to ensure that specified design requirements are met.

(2) The following class I devices are subject to design controls:

(i) Devices automated with computer software; and

(ii) The devices listed in the chart below.

Section	Device
868.6810	Catheter, Tracheobronchial Suction
878.4460	Glove, Surgeon's
880.6760	Restraint, Protective
892.5650	System, Applicator, Radionuclide, Manual
892.5740	Source, Radionuclide Teletherapy

II. DEFINITIONS

§ 820.3 (n) *Management with executive responsibility* means those senior employees of a manufacturer who have the authority to establish or make changes to the manufacturer's quality policy and quality system.

§ 820.3 (s) *Quality* means the totality of features and characteristics that bear on the ability of a device to satisfy fitness-for-use, including safety and performance.

§ 820.3 (v) *Quality system* means the organizational structure, responsibilities, procedures, processes, and resources for implementing quality management.

Cross reference to ISO 9001:1994 and ISO/DIS 13485 Section 4.4.1 General.

III. DISCUSSION AND POINTS TO CONSIDER

The essential quality aspects and the regulatory requirements, such as safety, performance, and dependability of a product (whether hardware, software, services, or processed materials) are established during the design and development phase. Deficient design can be a major cause of quality problems.

The context within which product design is to be carried out should be set by the manufacturer's senior management. It is their responsibility to establish a design and development plan which sets the targets to be met. This plan defines the constraints within which the design is to be implemented.

The quality system requirements do not dictate the types of design process that a manufacturer must use. Manufacturers should use processes best suited to their needs. However, whatever the processes may be, it is important that the design controls are applied in an appropriate manner. This guidance document contains examples of how this might be achieved in a variety of situations.

It is important to note that the design function may apply to various facets of the operation having differing styles and time scales. Such facets are related to products, including services and software, as well as to their manufacturing processes.

Senior management needs to decide how the design function is to be managed and by whom. Senior management should also ensure that internal policies are established for design issues such as:

- assessing new product ideas
- training and retraining of design managers and design staff
- use of consultants
- evaluation of the design process
- product evaluation, including third party product certification and approvals
- patenting or other means of design protection

It is for senior management to ensure that adequate resources are available to carry out the design in the required time. This may involve reinforcing the skills and equipment available internally and/or obtaining external resources.

SECTION B. DESIGN AND DEVELOPMENT PLANNING

I. REQUIREMENTS

§ 820.30(b) Design and development planning.

- Each manufacturer shall establish and maintain plans that describe or reference the design and development activities and define responsibility for implementation.
- The plans shall identify and describe the interfaces with different groups or activities that provide, or result in, input to the design and development process.
- The plans shall be reviewed, updated, and approved as design and development evolves.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 sections 4.4.2 Design and development planning and 4.4.3 Organizational and technical interfaces.

II. DISCUSSION AND POINTS TO CONSIDER

Design and development planning is needed to ensure that the design process is appropriately controlled and that device quality objectives are met. The plans must be consistent with the remainder of the design control requirements. The following elements would typically be addressed in the design and development plan or plans:

- Description of the goals and objectives of the design and development program; i.e., what is to be developed;
- Delineation of organizational responsibilities with respect to assuring quality during the design and development phase, to include interface with any contractors;
- Identification of the major tasks to be undertaken, deliverables for each task, and individual or organizational responsibilities (staff and resources) for completing each task;
- Scheduling of major tasks to meet overall program time constraints;
- Identification of major reviews and decision points;
- Selection of reviewers, the composition of review teams, and procedures to be followed by reviewers;
- Controls for design documentation;
- Notification activities.

Planning enables management to exercise greater control over the design and development process by clearly communicating policies, procedures, and goals to members of the

design and development team, and providing a basis for measuring conformance to quality system objectives.

Design activities should be specified at the level of detail necessary for carrying out the design process. The extent of design and development planning is dependent on the size of the developing organization and the size and complexity of the product to be developed. Some manufacturers may have documented policies and procedures which apply to all design and development activities. For each specific development program, such manufacturers may also prepare a plan which spells out the project-dependent elements in detail, and incorporates the general policies and procedures by reference. Other manufacturers may develop a comprehensive design and development plan which is specifically tailored to each individual project.

In summary, the form and organization of the planning documents are less important than their content. The following paragraphs discuss the key elements of design and development planning.

ORGANIZATIONAL RESPONSIBILITIES. The management responsibility section of the quality system requirements[1] requires management to establish a quality policy and implement an organizational structure to ensure quality. These are typically documented in a quality manual or similarly named document. In some cases, however, the design and development plan, rather than the quality manual, is the best vehicle for describing organizational responsibilities relative to design and development activities. The importance of defining responsibilities with clarity and without ambiguity should be recognized. When input to the design is from a variety of sources, their interrelationships and interfaces (as well as the pertinent responsibilities and authorities) should be defined, documented, coordinated, and controlled. This might be the case, for example, if a multidisciplinary product development team is assembled for a specific project, or if the team includes suppliers, contract manufacturers, users, outside consultants, or independent auditors.

TASK BREAKDOWN. The plan establishes, to the extent possible:

- The major tasks required to develop the product
- The time involved for each major task
- The resources and personnel required
- The allocation of responsibilities for completing each major task
- The prerequisite information necessary to start each major task and the interrelationship between tasks
- The form of each task output or deliverable
- Constraints, such as applicable codes, standards, and regulations

[1] § 820.20 of the FDA Quality System Regulation; section 4.1 of ISO 9001 and ISO/DIS 13485.

Tasks for all significant design activities, including verification and validation tasks, should be included in the design and development plan. For example, if clinical trials are anticipated, there may be tasks associated with appropriate regulatory requirements.

For complex projects, rough estimates may be provided initially, with the details left for the responsible organizations to develop. As development proceeds, the plan should evolve to incorporate more and better information.

The relationships between tasks should be presented in such a way that they are easily understood. It should be clear which tasks depend on others, and which tasks need to be performed concurrently. Planning should reflect the degree of perceived development risk; for example, tasks involving new technology or processes should be spelled out in greater detail, and perhaps be subjected to more reviews and checks, than tasks which are perceived as routine or straightforward.

The design and development plan may include a schedule showing starting and completion dates for each major task, project milestone, or key decision points. The method chosen and the detail will vary depending on the complexity of the project and the level of risk associated with the device. For small projects, the plan may consist of only a simple flow diagram or computer spreadsheet. For larger projects, there are a number of project management tools that are used to develop plans. Three of the most commonly used are the Program Evaluation and Review Technique (PERT), the Critical Path Method (CPM), and the Gantt chart. Software is available in many forms for these methods. When selecting these tools, be careful to choose one that best fits the needs of the project. Some of the software programs are far more complex than may be necessary.

Unless a manufacturer has experience with the same type of device, the plan will initially be limited in scope and detail. As work proceeds, the plan is refined. Lack of experience in planning often leads to optimistic schedules, but slippage may also occur for reasons beyond the control of planners, for example, personnel turnover, materiel shortage, or unexpected problems with a design element or process. Sometimes the schedule can be compressed by using additional resources, such as diverting staff or equipment from another project, hiring a contractor, or leasing equipment.

It is important that the schedule be updated to reflect current knowledge. At all times, the plan should be specified at a level of detail enabling management to make informed decisions, and provide confidence in meeting overall schedule and performance objectives. This is important because scheduling pressures have historically been a contributing factor in many design defects which caused injury. To the extent that good planning can prevent schedule pressures, the potential for design errors is reduced.

However, no amount of planning can eliminate all development risk. There is inherent conflict between the desire to maximize performance and the need to meet business objectives, including development deadlines. In some corporate cultures, impending deadlines create enormous pressure to cut corners. Planning helps to combat this dilemma by ensuring management awareness of pressure points. With awareness, decisions are

more likely to be made with appropriate oversight and consideration of all relevant factors. Thus, when concessions to the clock must be made, they can be justified and supported.

SECTION C. DESIGN INPUT

I. REQUIREMENTS

§ 820.30(c) Design input.

- Each manufacturer shall establish and maintain procedures to ensure that the design requirements relating to a device are appropriate and address the intended use of the device, including the needs of the user and patient.
- The procedures shall include a mechanism for addressing incomplete, ambiguous, or conflicting requirements.
- The design input requirements shall be documented and shall be reviewed and approved by designated individual(s).
- The approval, including the date and signature of the individual(s) approving the requirements, shall be documented.

Cross reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.4 Design input.

II. DEFINITIONS

§ 820.3(f) *Design input* means the physical and performance requirements of a device that are used as a basis for device design.

III. DISCUSSION AND POINTS TO CONSIDER

Design input is the starting point for product design. The requirements which form the design input establish a basis for performing subsequent design tasks and validating the design. Therefore, development of a solid foundation of requirements is the single most important design control activity.

Many medical device manufacturers have experience with the adverse effects that incomplete requirements can have on the design process. A frequent complaint of developers is that "there's never time to do it right, but there's always time to do it over." If essential requirements are not identified until validation, expensive redesign and rework may be necessary before a design can be released to production.

By comparison, the experience of companies that have designed devices using clear-cut, comprehensive sets of requirements is that rework and redesign are significantly reduced and product quality is improved. They know that the development of requirements for a medical device of even moderate complexity is a formidable, time-consuming task. They accept the investment in time and resources required to develop the requirements because they know the advantages to be gained in the long run.

Unfortunately, there are a number of common misconceptions regarding the meaning and practical application of the quality system requirements for design input. Many seem to arise from interpreting the requirements as a literal prescription, rather than a set of principles to be followed. In this guidance document, the focus is on explaining the principles and providing examples of how they may be applied in typical situations.

CONCEPT DOCUMENTS VERSUS DESIGN INPUT In some cases, the marketing staff, who maintain close contact with customers and users, determine a need for a new product, or enhancements to an existing product. Alternatively, the idea for a new product may evolve out of a research or clinical activity. In any case, the result is a concept document specifying some of the desired characteristics of the new product.

Some members of the medical device community view these marketing memoranda, or the equivalent, as the design input. However, that is not the intent of the quality system requirements. Such concept documents are rarely comprehensive, and should not be expected to be so. Rather, the intent of the quality system requirements is that the product conceptual description be elaborated, expanded, and transformed into a complete set of design input requirements which are written to an engineering level of detail.

This is an important concept. The use of qualitative terms in a concept document is both appropriate and practical. This is often not the case for a document to be used as a basis for design. Even the simplest of terms can have enormous design implications. For example, the term "must be portable" in a concept document raises questions in the minds of product developers about issues such as size and weight limitations, resistance to shock and vibration, the need for protection from moisture and corrosion, the capability of operating over a wide temperature range, and many others. Thus, a concept document may be the starting point for development, but it is not the design input requirement. This is a key principle—the design input requirements are the result of the first stage of the design control process.

RESEARCH AND DEVELOPMENT. Some manufacturers have difficulty in determining where research ends and development begins. Research activities may be undertaken in an effort to determine new business opportunities or basic characteristics for a new product. It may be reasonable to develop a rapid prototype to explore the feasibility of an idea or design approach, for example, prior to developing design input requirements. But manufacturers should avoid falling into the trap of equating the prototype design with a finished product design. Prototypes at this stage lack safety features and ancillary functions necessary for a finished product, and are developed under conditions which preclude adequate consideration of product variability due to manufacturing.

RESPONSIBILITY FOR DESIGN INPUT DEVELOPMENT. Regardless of who developed the initial product concept, product developers play a key role in developing the design input requirements. When presented with a set of important characteristics, it is the product developers who understand the auxiliary issues that must be addressed, as well as the level of detail necessary to design a product. Therefore, a second key principle is that

the product developer(s) ultimately bear responsibility for translating user and/or patient needs into a set of requirements which can be validated prior to implementation. While this is primarily an engineering function, the support or full participation of production and service personnel, key suppliers, etc., may be required to assure that the design input requirements are complete.

Care must be exercised in applying this principle. Effective development of design input requirements encompasses input from both the product developer as well as those representing the needs of the user, such as marketing. Terminology can be a problem. In some cases, the product conceptual description may be expressed in medical terms. Medical terminology is appropriate in requirements when the developers and reviewers are familiar with the language, but it is often preferable to translate the concepts into engineering terms at the requirements stage to minimize miscommunication with the development staff.

Another problem is incorrect assumptions. Product developers make incorrect assumptions about user needs, and marketing personnel make incorrect assumptions about the needs of the product designers. Incorrect assumptions can have serious consequences that may not be detected until late in the development process. Therefore, both product developers and those representing the user must take responsibility for critically examining proposed requirements, exploring stated and implied assumptions, and uncovering problems.

Some examples should clarify this point. A basic principle is that design input requirements should specify what the design is intended to do while carefully avoiding specific design solutions at this stage. For example, a concept document might dictate that the product be housed in a machined aluminum case. It would be prudent for product developers to explore why this type of housing was specified. Perhaps there is a valid reason—superior electrical shielding, mechanical strength, or reduced time to market as compared to a cast housing. Or perhaps machined aluminum was specified because a competitor's product is made that way, or simply because the user didn't think plastic would be strong enough.

Not all incorrect assumptions are made by users. Incorrect assumptions made by product developers may be equally damaging. Failure to understand the abuse to which a portable instrument would be subjected might result in the selection of housing materials inadequate for the intended use of the product.

There are occasions when it may be appropriate to specify part of the design solution in the design input requirements. For example, a manufacturer may want to share components or manufacturing processes across a family of products in order to realize economies of scale, or simply to help establish a corporate identity. In the case of a product upgrade, there may be clear consensus regarding the features to be retained. However, it is important to realize that every such design constraint reduces implementation flexibility and should therefore be documented and identified as a possible conflicting requirement for subsequent resolution.

SCOPE AND LEVEL OF DETAIL. Design input requirements must be comprehensive. This may be quite difficult for manufacturers who are implementing a system of design controls for the first time. Fortunately, the process gets easier with practice. It may be helpful to realize that design input requirements fall into three categories. Virtually every product will have requirements of all three types.

- Functional requirements specify what the device does, focusing on the operational capabilities of the device and processing of inputs and the resultant outputs.
- Performance requirements specify how much or how well the device must perform, addressing issues such as speed, strength, response times, accuracy, limits of operation, etc. This includes a quantitative characterization of the use environment, including, for example, temperature, humidity, shock, vibration, and electromagnetic compatibility. Requirements concerning device reliability and safety also fit into this category.
- Interface requirements specify characteristics of the device which are critical to compatibility with external systems; specifically, those characteristics which are mandated by external systems and outside the control of the developers. One interface which is important in every case is the user and/or patient interface.

What is the scope of the design input requirements development process and how much detail must be provided? The scope is dependent upon the complexity of a device and the risk associated with its use. For most medical devices, numerous requirements encompassing functions, performance, safety, and regulatory concerns are implied by the application. These implied requirements should be explicitly stated, in engineering terms, in the design input requirements.

Determining the appropriate level of detail requires experience. However, some general guidance is possible. The marketing literature contains product specifications, but these are superficial. The operator and service manuals may contain more detailed specifications and performance limits, but these also fall short of being comprehensive. Some insight as to what is necessary is provided by examining the requirements for a very common external interface. For the power requirements for AC-powered equipment, it is not sufficient to simply say that a unit shall be AC-powered. It is better to say that the unit shall be operable from AC power in North America, Europe, and Japan, but that is still insufficient detail to implement or validate the design. If one considers the situation just in North America, where the line voltage is typically 120 volts, many systems are specified to operate over the range of 108 to 132 volts. However, to account for the possibility of brownout, critical devices may be specified to operate from 95 to 132 volts or even wider ranges. Based on the intended use of the device, the manufacturer must choose appropriate performance limits.

There are many cases when it is impractical to establish every functional and performance characteristic at the design input stage. But in most cases, the form of the requirement can be determined, and the requirement can be stated with a to-be-determined (TBD) numerical value or a range of possible values. This makes it possible for reviewers to

assess whether the requirements completely characterize the intended use of the device, judge the impact of omissions, and track incomplete requirements to ensure resolution.

For complex designs, it is not uncommon for the design input stage to consume as much as thirty percent of the total project time. Unfortunately, some managers and developers have been trained to measure design progress in terms of hardware built, or lines of software code written. They fail to realize that building a solid foundation saves time during the implementation. Part of the solution is to structure the requirements documents and reviews such that tangible measures of progress are provided.

At the other extreme, many medical devices have very simple requirements. For example, many new devices are simply replacement parts for a product, or are kits of commodity items. Typically, only the packaging and labeling distinguishes these products from existing products. In such cases, there is no need to recreate the detailed design input requirements of the item. It is acceptable to simply cite the predecessor product documentation, add any new product information, and establish the unique packaging and labeling requirements.

ASSESSING DESIGN INPUT REQUIREMENTS FOR ADEQUACY. Eventually, the design input must be reviewed for adequacy. After review and approval, the design input becomes a controlled document. All future changes will be subject to the change control procedures, as discussed in Section I (Design Changes).

Any assessment of design input requirements boils down to a matter of judgment. As discussed in Section E (Design Review), it is important for the review team to be multidisciplinary and to have the appropriate authority. A number of criteria may be employed by the review team.

- Design input requirements should be unambiguous. That is, each requirement should be able to be verified by an objective method of analysis, inspection, or testing. For example, it is insufficient to state that a catheter must be able to withstand repeated flexing. A better requirement would state that the catheter should be formed into a 50 mm diameter coil and straightened out for a total of fifty times with no evidence of cracking or deformity. A qualified reviewer could then make a judgment whether this specified test method is representative of the conditions of use.
- Quantitative limits should be expressed with a measurement tolerance. For example, a diameter of 3.5 mm is an incomplete specification. If the diameter is specified as 3.500±0.005 mm, designers have a basis for determining how accurate the manufacturing processes have to be to produce compliant parts, and reviewers have a basis for determining whether the parts will be suitable for the intended use.
- The set of design input requirements for a product should be self-consistent. It is not unusual for requirements to conflict with one another or with a referenced industry standard due to a simple oversight. Such conflicts should be resolved early in the development process.

- The environment in which the product is intended to be used should be properly characterized. For example, manufacturers frequently make the mistake of specifying "laboratory" conditions for devices which are intended for use in the home. Yet, even within a single country, relative humidity in a home may range from 20 percent to 100 percent (condensing) due to climactic and seasonal variations. Household temperatures in many climates routinely exceed 40 °C during the hot season. Altitudes may exceed 3,000 m, and the resultant low atmospheric pressure may adversely affect some kinds of medical equipment. If environmental conditions are fully specified, a qualified reviewer can make a determination of whether the specified conditions are representative of the intended use.
- When industry standards are cited, the citations should be reviewed for completeness and relevance. For example, one medical device manufacturer claimed compliance with an industry standard covering mechanical shock and vibration. However, when the referenced standard was examined by a reviewer, it was found to prescribe only the method of testing, omitting any mention of pass/fail criteria. It was incumbent on the manufacturer in this case to specify appropriate performance limits for the device being tested, as well as the test method.

EVOLUTION OF THE DESIGN INPUT REQUIREMENTS. Large development projects often are implemented in stages. When this occurs, the design input requirements at each stage should be developed and reviewed following the principles set forth in this section. Fortunately, the initial set of requirements, covering the overall product, is by far the most difficult to develop. As the design proceeds, the output from the early stages forms the basis for the subsequent stages, and the information available to designers is inherently more extensive and detailed.

It is almost inevitable that verification activities will uncover discrepancies which result in changes to the design input requirements. There are two points to be made about this. One is that the change control process for design input requirements must be carefully managed. Often, a design change to correct one problem may create a new problem which must be addressed. Throughout the development process, it is important that any changes are documented and communicated to developers so that the total impact of the change can be determined. The change control process is crucial to device quality.

The second point is that extensive rework of the design input requirements suggests that the design input requirements may not be elaborated to a suitable level of detail, or insufficient resources are being devoted to defining and reviewing the requirements. Managers can use this insight to improve the design control process. From a design control perspective, the number of requirements changes made is less important than the thoroughness of the change control process.

SECTION D. DESIGN OUTPUT

I. REQUIREMENTS

§ 820.30(d) Design output.

- Each manufacturer shall establish and maintain procedures for defining and documenting design output in terms that allow an adequate evaluation of conformance to design input requirements.
- Design output procedures shall contain or make reference to acceptance criteria and shall ensure that those design outputs that are essential for the proper functioning of the device are identified.
- Design output shall be documented, reviewed, and approved before release.
- The approval, including the date and signature of the individual(s) approving the output, shall be documented.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.5 Design output.

II. DEFINITIONS

§ 820.3(g) *Design output* means the results of a design effort at each design phase and at the end of the total design effort. The finished design output is the basis for the device master record. The total finished design output consists of the device, its packaging and labeling, and the device master record.

§ 820.3(y) *Specification* means any requirement with which a product, process, service, or other activity must conform.

III. DISCUSSION AND POINTS TO CONSIDER

The quality system requirements for design output can be separated into two elements: Design output should be expressed in terms that allow adequate assessment of conformance to design input requirements and should identify the characteristics of the design that are crucial to the safety and proper functioning of the device. This raises two fundamental issues for developers:

- What constitutes design output?
- Are the form and content of the design output suitable?

The first issue is important because the typical development project produces voluminous records, some of which may not be categorized as design output. On the other hand, design output must be reasonably comprehensive to be effective. As a general rule, an item is design output if it is a work product, or deliverable item, of a design task listed in

the design and development plan, and the item defines, describes, or elaborates an element of the design implementation. Examples include block diagrams, flow charts, software high-level code, and system or subsystem design specifications. The design output in one stage is often part of the design input in subsequent stages.

Design output includes production specifications as well as descriptive materials which define and characterize the design.

PRODUCTION SPECIFICATIONS. Production specifications include drawings and documents used to procure components, fabricate, test, inspect, install, maintain, and service the device, such as the following:

- assembly drawings
- component and material specifications
- production and process specifications
- software machine code (e.g., diskette or master EPROM)
- work instructions
- quality assurance specifications and procedures
- installation and servicing procedures
- packaging and labeling specifications, including methods and processes used

In addition, as discussed in Section H (Design Transfer), production specifications may take on other forms. For example, some manufacturers produce assembly instructions on videotapes rather than written instructions. Similarly, a program diskette, used by a computer-aided milling machine to fabricate a part, would be considered a production specification. The videotape and the software on the program diskette are part of the device master record.

OTHER DESCRIPTIVE MATERIALS. Other design output items might be produced which are necessary to establish conformance to design input requirements, but are not used in its production. For example, for each part which is fabricated by computer-aided machine, there should be an assembly drawing which specifies the dimensions and characteristics of the part. It is a part of the design output because it establishes the basis for the machine tool program used to fabricate the part. Other examples of design output include the following:

- the results of risk analysis
- software source code
- results of verification activities
- biocompatibility test results
- bioburden test results

FORM AND CONTENT. Manufacturers must take steps to assure that the design output characterizes all important aspects of the design and is expressed in terms which

allow adequate verification and validation. Two basic mechanisms are available to manufacturers to accomplish these objectives.

- First, the manufacturer proactively can specify the form and content of design output at the planning stage. For some types of design output, form and content may be codified in a consensus standard which can be referenced. In other cases, a manufacturer could specify the desired characteristics, or even simply specify that the form and content of an existing document be followed.
- Second, form and content can be reviewed retroactively as a part of the design verification process. For example, the verification of design output could include assessing whether specified documentation standards have been adhered to.

As these examples illustrate, conformance with the quality system requirements concerning design output generally requires no "extra" effort on the part of the manufacturer, but simply the application of some common sense procedures during the planning, execution, and review of design tasks.

SECTION E. DESIGN REVIEW

I. REQUIREMENTS

§ 820.30(e) Design review.

- Each manufacturer shall establish and maintain procedures to ensure that formal documented reviews of the design results are planned and conducted at appropriate stages of the device's design development.
- The procedures shall ensure that participants at each design review include representatives of all functions concerned with the design stage being reviewed and an individual(s) who does not have direct responsibility for the design stage being reviewed, as well as any specialists needed.
- The results of a design review, including identification of the design, the date, and the individual(s) performing the review, shall be documented in the design history file (the DHF).

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.6 Design review.

II. DEFINITIONS

§ 820.3(h) *Design review* means a documented, comprehensive, systematic examination of a design to evaluate the adequacy of the design requirements, to evaluate the capability of the design to meet these requirements, and to identify problems.

III. DISCUSSION AND POINTS TO CONSIDER

In general, formal design reviews are intended to:

- provide a systematic assessment of design results, including the device design and the associated designs for production and support processes;
- provide feedback to designers on existing or emerging problems;
- assess project progress; and/or
- provide confirmation that the project is ready to move on to the next stage of development.

Many types of reviews occur during the course of developing a product. Reviews may have both an internal and external focus. The internal focus is on the feasibility of the design and the produceability of the design with respect to manufacturing and support

capabilities. The external focus is on the user requirements; that is, the device design is viewed from the perspective of the user.

The nature of reviews changes as the design progresses. During the initial stages, issues related to design input requirements will predominate. Next, the main function of the reviews may be to evaluate or confirm the choice of solutions being offered by the design team. Then, issues such as the choice of materials and the methods of manufacture become more important. During the final stages, issues related to the verification, validation, and production may predominate.

The term "review" is commonly used by manufacturers to describe a variety of design assessment activities. Most, but not all, of these activities meet the definition of formal design reviews. The following exceptions may help to clarify the distinguishing characteristics of design reviews.

- Each design document which constitutes the formal output, or deliverable, of a design task is normally subject to evaluation activities, sometimes referred to as informal peer review, supervisory review, or technical assessment. These activities, while they may be called reviews, are often better described as verification activities, because they are not intended to be comprehensive, definitive, and multidisciplinary in their scope. Rather, their purpose is to confirm that design output meets design input. Verification activities affect and add to the design output, and are themselves subject to subsequent design review.
- Developers may conduct routine or ad hoc meetings to discuss an issue, coordinate activities, or assess development progress. Decisions from such meetings may not require formal documentation; however, if a significant issue is resolved, this should be documented. If the outcome results in change to an approved design document, then applicable change control procedures should be followed, as discussed in Section I (Design Changes).

Control of the design review process is achieved by developing and implementing a formal design review program consistent with quality system requirements. The following issues should be addressed and documented in the design and development plan(s).

NUMBER AND TYPE OF REVIEWS. It is a well-accepted fact that the cost to correct design errors increases as the design nears completion, and the flexibility to implement an optimal solution decreases. When an error is discovered at the end of the development cycle, difficult decisions have to be made regarding an acceptable corrective action. When that corrective action is implemented in haste, the result is often an unintended consequence leading to a new problem. Thus, formal design reviews should be planned to detect problems early. A corollary is that planners should presume that problems will be detected, and allocate a reasonable amount of time to implement corrective actions. Typically, formal reviews are conducted at the end of each phase and at important milestones in the design process.

As discussed in Section C (Design Input), it is beneficial in almost every case to conduct a formal review of the design input requirements early in the development process. The number of reviews depends upon the complexity of the device.

- For a simple design, or a minor upgrade to an existing product, it might be appropriate to conduct a single review at the conclusion of the design process.
- For a product involving multiple subsystems, an early design task is to allocate the design input requirements among the various subsystems. For example, in a microprocessor-based system, designers must decide which functions will be performed by hardware and which by software. In another case, tolerance buildup from several components may combine to create a clearance problem. System designers must establish tolerance specifications for each component to meet the overall dimensional specification. In cases like these, a formal design review is a prudent step to ensure that all such system-level requirements have been allocated satisfactorily prior to engaging in detailed design of each subsystem.
- For complex systems, additional reviews are often built into the development plan. For example, engineering sketches may be developed for prototyping purposes prior to development of production drawings. Evaluation of the prototype would typically culminate in a formal design review. Similarly, software development commonly includes a high-level design phase, during which requirements are elaborated to a greater level of detail and algorithms are developed to implement key functions. A formal design review would typically be conducted to review this work prior to beginning detailed coding.

There are a number of approaches to conducting formal design reviews at the end of the design process. In some organizations, engineering essentially completes the design, tests an engineering prototype, and conducts a formal design review prior to turning the design over to manufacturing. In such cases, an additional review will be needed after the design has been validated using production devices.

In some instances, components having long lead times may enter production prior to completion of the overall device design. The primary motivation for early production is to reduce time to market. The manufacturer runs the business risk that the design review at the end of the design process will uncover a defect that must be corrected in production devices before any devices are distributed.

All of these approaches to scheduling formal design reviews are valid. What is important is that the manufacturer establish a reasonable rationale for the number and type of reviews, based on sound judgment.

SELECTION OF REVIEWERS. In determining who should participate in a formal design review, planners should consider the qualifications of reviewers, the types of expertise required to make an adequate assessment, and the independence of the reviewers. Each of these concerns is discussed briefly in the following paragraphs.

Qualifications. Formal design reviews should be conducted by person(s) having technical competence and experience at least comparable to the developers. For a small manufacturer, this may require that an outside consultant be retained to participate in the evaluation of the design.

A manufacturer will often employ one or more specialists to conduct certain types of specialized assessments which are beyond the capabilities of the designers. For example, a mechanical engineer may be retained to perform a structural analysis of a design, and perhaps conduct vibration testing to verify its performance under stress. Such specialists may be assigned to participate in the formal design review. Alternatively, they may be assigned to make an independent assessment and submit observations and recommendations to the reviewers. Either approach is valid.

Types of expertise required. Many medical device designs involve a number of technologies, such as electronics, mechanics, software, materials science, or pneumatics. In addition, a variety of clinical and manufacturing issues may influence the design. Manufacturers should carefully consider which interests should be represented at formal design reviews. Subtle distinctions in reviewer perspective may have dramatic impact on device quality. For example, the marketing department of a small manufacturer shared a new design with several surgeons on their advisory board. The surgeons all thought the design was terrific. Subsequently, the manufacturer invited two experienced operating room nurses to participate in the final design review. During the course of the review, it became apparent that while surgeons may be the customers, nurses are the primary users of the device, and no one up to that point had consulted with any nurses. The nurses at the design review didn't like some of the features of the design. After some further market survey, the manufacturer decided to make changes to the design to accommodate these concerns. It was unfortunate (and expensive) in this case that the user requirements were not considered until late in the development cycle, but the design review was ultimately very successful.

Independence. The formal design review should include at least one individual who does not have direct responsibility for the design stage under review. In a small company, complete independence is very difficult to obtain. Within the context of formal design reviews, the practical solution is simply to ensure a fresh perspective, based on the principle that those who are too close to the design may overlook design errors. Thus, reviewers will often be from the same organization as the developers, but they should not have been significantly involved in the activities under review. As discussed in the following section, the formal design review procedures play a large role in assuring independent and objective reviews.

DESIGN REVIEW PROCEDURES. The manufacturer should have documented formal design review procedures addressing the following:

- Evaluation of the design (including identification of concerns, i.e., issues and potential problems with the design)
- Resolution of concerns

- Implementation of corrective actions

Evaluation of the design. Many formal design reviews take the form of a meeting. At this meeting, the designer(s) may make presentations to explain the design implementation, and persons responsible for verification activities may present their findings to the reviewers. Reviewers may ask for clarification or additional information on any topic, and add their concerns to any raised by the presenters. This portion of the review is focused on finding problems, not resolving them.

There are many approaches to conducting design review meetings. In simple cases, the technical assessor and reviewer may be the same person, often a project manager or engineering supervisor, and the review meeting is a simple affair in the manager's office. For more elaborate reviews, detailed written procedures are desirable to ensure that all pertinent topics are discussed, conclusions accurately recorded, and action items documented and tracked.

There is a dangerous tendency for design review meetings to become adversarial affairs. The reputation of the designers tends to be linked to the number of discrepancies found, causing the designers to become defensive, while the reviewers score points by finding weaknesses in the design. The resulting contest can be counterproductive. An added complication is the presence of invited guests, often clinicians, who are expected to provide the user perspective. These reviewers are often very reluctant to ask probing questions, especially if they sense that they may become involved in a conflict where all the rules and relationships are not evident.

These difficulties can be avoided by stating the goals and ground rules for conducting the formal design review clearly at the outset. While the designers are in the best position to explain the best features of the design, they are also most likely to be aware of the design's weaknesses. If the designers and reviewers are encouraged to work together to systematically explore problems and find solutions, the resultant design will be improved and all parties will benefit from the process. Participants must be encouraged to ask questions, avoid making assumptions, and think critically. The focus must be on the design, not the participants.

Not all formal design reviews involve meetings. For extremely simple designs or design changes, it may be appropriate to specify a procedure in which review materials are distributed or circulated among the reviewers for independent assessment and approval. However, such a procedure negates the benefits of synergy and teamwork, and should be considered only in cases where the design issues are limited in scope and well defined.

Resolution of concerns. The reviewers consider concerns raised during the evaluation portion of the formal design review and decide on an appropriate disposition for each one. There is wide variation in the way companies implement decision-making processes. In some cases, the reviewers play an advisory role to the engineering manager or other company official, who directs the formal design review and ultimately selects a course of

action. In other cases, the reviewers are given limited or broad authority to make decisions and commit resources to resolve problems. The approach used should be documented.

In the real world, reviews often leave unresolved issues. Therefore, review procedures should include a process for resolving differences, and provide reviewers with enough leeway to make practical decisions while protecting the integrity of the process.

Implementation of corrective actions. Not all identified concerns result in corrective actions. The reviewers may decide that the issue is erroneous or immaterial. In most cases, however, resolution involves a design change, a requirements change, or a combination of the two. If the solution is evident, the reviewers may specify the appropriate corrective action; otherwise, an action item will be assigned to study the problem further. In any case, action items and corrective actions are normally tracked under the manufacturer's change control procedures.

RELATIONSHIP OF DESIGN REVIEW TO VERIFICATION AND VALIDATION. In practice, design review, verification, and validation overlap one another, and the relationship among them may be confusing. As a general rule, the sequence is: verification, review, validation, review.

In most cases, verification activities are completed prior to the design review, and the verification results are submitted to the reviewers along with the other design output to be reviewed. Alternatively, some verification activities may be treated as components of the design review, particularly if the verification activity is complex and requires multidisciplinary review.

Similarly, validation typically involves a variety of activities, including a determination that the appropriate verifications and reviews have been completed. Thus, at the conclusion of the validation effort, a review is usually warranted to assure that the validation is complete and adequate.

SECTION F. DESIGN VERIFICATION

I. REQUIREMENTS

§ 820.30(f) Design verification.

- Each manufacturer shall establish and maintain procedures for verifying the device design.
- Design verification shall confirm that the design output meets the design input requirements.
- The results of the design verification, including identification of the design, method(s), the date, and the individual(s) performing the verification, shall be documented in the Design History File.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.7 Design verification.

I. DEFINITIONS

§820.3(y) *Specification* means any requirement with which a product, process, service, or other activity must conform.

§ 820.3(z) *Validation* means confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled.

> **(1)** *Process Validation* means establishing by objective evidence that a process consistently produces a result or product meeting its predetermined specifications.
>
> **(2)** *Design Validation* means establishing by objective evidence that device specifications conform with user needs and intended use(s).

§820.3(aa) *Verification* means confirmation by examination and provision of objective evidence that specified requirements have been fulfilled.

III. DISCUSSION AND POINTS TO CONSIDER

Verification and validation are associated concepts with very important differences. Various organizations have different definitions for these terms. Medical device

manufacturers are encouraged to use the terminology of the quality system requirements in their internal procedures.

To illustrate the concepts, consider a building design analogy. In a typical scenario, the senior architect establishes the design input requirements and sketches the general appearance and construction of the building, but associates or contractors typically elaborate the details of the various mechanical systems. Verification is the process of checking at each stage whether the output conforms to requirements for that stage. For example: does the air conditioning system deliver the specified cooling capacity to each room? Is the roof rated to withstand so many newtons per square meter of wind loading? Is a fire alarm located within 50 meters of each location in the building?

At the same time, the architect has to keep in mind the broader question of whether the results are consistent with the ultimate user requirements. Does the air conditioning system keep the occupants comfortable throughout the building? Will the roof withstand weather extremes expected at the building site? Can the fire alarm be heard throughout the building? Those broader concerns are the essence of validation.

In the initial stages of design, verification is a key quality assurance technique. As the design effort progresses, verification activities become progressively more comprehensive. For example, heat or cooling delivery can be calculated and verified by the air conditioning designer, but the resultant air temperature can only be estimated. Occupant comfort is a function not only of delivered air temperature, but also humidity, heat radiation to or from nearby thermal masses, heat gain or loss through adjacent windows, etc. During the latter design phases, the interaction of these complex factors may be considered during verification of the design.

Validation follows successful verification, and ensures that each requirement for a particular use is fulfilled. Validation of user needs is possible only after the building is built. The air conditioning and fire alarm performance may be validated by testing and inspection, while the strength of the roof will probably be validated by some sort of analysis linked to building codes which are accepted as meeting the needs of the user—subject to possible confirmation during a subsequent severe storm.

Validation is the topic of Section G of this guidance document. The remainder of this section focuses on verification principles.

TYPES OF VERIFICATION ACTIVITIES. Verification activities are conducted at all stages and levels of device design. The basis of verification is a three-pronged approach involving tests, inspections, and analyses. Any approach which establishes conformance with a design input requirement is an acceptable means of verifying the design with respect to that requirement. In many cases, a variety of approaches are possible.

Complex designs require more and different types of verification activities. The nature of verification activities varies according to the type of design output. The intent of this guidance document is not to suggest or recommend verification techniques which should

be performed by device manufacturers. Rather, the manufacturer should select and apply appropriate verification techniques based on the generally accepted practices for the technologies employed in their products. Many of these practices are an integral part of the development process, and are routinely performed by developers. The objective of design controls is to ensure adequate oversight by making verification activities explicit and measuring the thoroughness of their execution. Following are a few examples of verification methods and activities.

- Worst case analysis of an assembly to verify that components are derated properly and not subject to overstress during handling and use.
- Thermal analysis of an assembly to assure that internal or surface temperatures do not exceed specified limits.
- Fault tree analysis of a process or design.
- Failure modes and effects analysis.
- Package integrity tests.
- Biocompatibility testing of materials.
- Bioburden testing of products to be sterilized.
- Comparison of a design to a previous product having an established history of successful use.

For some technologies, verification methods may be highly standardized. In other cases, the manufacturer may choose from a variety of applicable methods. In a few cases, the manufacturer must be creative in devising ways to verify a particular aspect of a design.

Some manufacturers erroneously equate production testing with verification. Whereas verification testing establishes conformance of design output with design input, the aim of production testing is to determine whether the unit under test has been correctly manufactured. In other words, production testing is designed to efficiently screen out manufacturing process errors and perhaps also to detect infant mortality failures. Typically, a small subset of functional and performance tests accomplish this objective with a high degree of accuracy. Therefore, production testing is rarely, if ever, comprehensive enough to verify the design. For example, a leakage test may be used during production to ensure that a hermetically-sealed enclosure was properly assembled. However, the leakage test may not be sensitive enough to detect long-term diffusion of gas through the packaging material. Permeability of the packaging material is an intrinsic property of the material rather than an assembly issue, and would likely be verified using a more specialized test than is used during production.

DOCUMENTATION OF VERIFICATION ACTIVITIES. Some verification methods result in a document by their nature. For example, a failure modes and effects analysis produces a table listing each system component, its postulated failure modes, and the effect of such failures on system operation.

Another self-documenting verification method is the traceability matrix. This method is particularly useful when the design input and output are both documents; it also has great utility in software development. In the most common form of the traceability matrix, the input requirements are enumerated in a table, and references are provided to each section in the output documents (or software modules) which address or satisfy each input requirement. The matrix can also be constructed "backwards," listing each feature in the design output and tracing which input requirement bears on that feature. This reverse approach is especially useful for detecting hidden assumptions. Hidden assumptions are dangerous because they often lead to overdesign, adding unnecessary cost and complexity to the design. In other cases, hidden assumptions turn out to be undocumented design input requirements which, once exposed, can be properly tracked and verified.

However, many verification activities are simply some sort of structured assessment of the design output relative to the design input. When this is the case, manufacturers may document completion of verification activities by linking these activities with the signoff procedures for documents. This may be accomplished by establishing a procedure whereby each design output document must be verified and signed by designated persons. The presence of the reviewers' signatures on the document signifies that the design output has been verified in accordance with the signoff procedure.

SECTION G. DESIGN VALIDATION

I. REQUIREMENTS

§ 820.30(g) Design validation.

- Each manufacturer shall establish and maintain procedures for validating the device design.
- Design validation shall be performed under defined operating conditions on initial production units, lots, or batches, or their equivalents.
- Design validation shall ensure that devices conform to defined user needs and intended uses and shall include testing of production units under actual or simulated use conditions.
- Design validation shall include software validation and risk analysis, where appropriate.
- The results of the design validation, including identification of the design, method(s), the date, and the individual(s) performing the validation, shall be documented in the Design History File.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.8 Design validation.

II. DEFINITIONS

§820.3(y) *Specification* means any requirement with which a product, process, service, or other activity must conform.

§ 820.3(z) *Validation* means confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled.

> **(1)** *Process Validation* means establishing by objective evidence that a process consistently produces a result or product meeting its predetermined specifications.
>
> **(2)** *Design Validation* means establishing by objective evidence that device specifications conform with user needs and intended use(s).

§820.3(aa) *Verification* means confirmation by examination and provision of objective evidence that specified requirements have been fulfilled.

III. DISCUSSION AND POINTS TO CONSIDER

Whereas verification is a detailed examination of aspects of a design at various stages in the development, design validation is a cumulative summation of all efforts to assure that the design will conform with user needs and intended use(s), given expected variations in components, materials, manufacturing processes, and the use environment.

VALIDATION PLANNING. Planning for validation should begin early in the design process. The performance characteristics that are to be assessed should be identified, and validation methods and acceptance criteria should be established. For complex designs, a schedule of validation activities and organizational or individual responsibilities will facilitate maintaining control over the process. The validation plan should be reviewed for appropriateness, completeness, and to ensure that user needs and intended uses are addressed.

VALIDATION REVIEW. Validation may expose deficiencies in the original assumptions concerning user needs and intended uses. A formal review process should be used to resolve any such deficiencies. As with verification, the perception of a deficiency might be judged insignificant or erroneous, or a corrective action may be required.

VALIDATION METHODS. Many medical devices do not require clinical trials. However, all devices require clinical evaluation and should be tested in the actual or simulated use environment as a part of validation. This testing should involve devices which are manufactured using the same methods and procedures expected to be used for ongoing production. While testing is always a part of validation, additional validation methods are often used in conjunction with testing, including analysis and inspection methods, compilation of relevant scientific literature, provision of historical evidence that similar designs and/or materials are clinically safe, and full clinical investigations or clinical trials.

Some manufacturers have historically used their best assembly workers or skilled lab technicians to fabricate test articles, but this practice can obscure problems in the manufacturing process. It may be beneficial to ask the best workers to evaluate and critique the manufacturing process by trying it out, but pilot production should simulate as closely as possible the actual manufacturing conditions.

Validation should also address product packaging and labeling. These components of the design may have significant human factors implications, and may affect product performance in unexpected ways. For example, packaging materials have been known to cause electrostatic discharge (ESD) failures in electronic devices. If the unit under test is delivered to the test site in the test engineer's briefcase, the packaging problem may not become evident until after release to market.

Validation should include simulation of the expected environmental conditions, such as temperature, humidity, shock and vibration, corrosive atmospheres, etc. For some classes

of device, the environmental stresses encountered during shipment and installation far exceed those encountered during actual use, and should be addressed during validation.

Particular care should be taken to distinguish among customers, users, and patients to ensure that validation addresses the needs of all relevant parties. For a consumer device, the customer, user, and patient may all be the same person. At the other extreme, the person who buys the device may be different from the person who routinely uses it on patients in a clinical setting. Hospital administrators, biomedical engineers, health insurance underwriters, physicians, nurses, medical technicians, and patients have distinct and sometimes competing needs with respect to a device design.

VALIDATION DOCUMENTATION. Validation is a compilation of the results of all validation activities. For a complex design, the detailed results may be contained in a variety of separate documents and summarized in a validation report. Supporting information should be explicitly referenced in the validation report and either included as an appendix or available in the design history file.

SECTION H. DESIGN TRANSFER

I. REQUIREMENTS

§ 820.30(h) Design transfer.

- Each manufacturer shall establish and maintain procedures to ensure that the device design is correctly translated into production specifications.

Cross reference to ISO 9001:1994 and ISO/DIS 13485 section 4.2.3(c) Quality planning.

II. DISCUSSION AND POINTS TO CONSIDER

Production specifications must ensure that manufactured devices are repeatedly and reliably produced within product and process capabilities. If a manufactured device deviates outside those capabilities, performance may be compromised. Thus, the process of encapsulating knowledge about the device into production specifications is critical to device quality.

The level of detail necessary to accomplish this objective varies widely, based on the type of device, the relationship between the design and manufacturing organizations, and the knowledge, experience, and skills of production workers. In some cases, devices are produced by contract manufacturers who have no involvement in the development and little or no contact with the designers. At the other extreme, some devices are hand-crafted by skilled artisans with extensive knowledge about the use of the product.

One normally associates the term “production specifications” with written documents, such as assembly drawings, component procurement specifications, workmanship standards, manufacturing instructions, and inspection and test specifications. While these types of documents are widely employed in medical device production, other equally acceptable means of conveying design information exist, and manufacturers have the flexibility to employ these alternate means of communication as appropriate. For example, each of the following could constitute “production specifications” within the meaning of the quality system requirements:

- documentation (in electronic format as well as paper)
- training materials, e.g., manufacturing processes, test and inspection methods
- digital data files, e.g., programmable device files, master EPROM, computer-aided manufacturing (CAM) programming files
- manufacturing jigs and aids, e.g., molds, sample wiring harness to be duplicated

Historically, shortcomings in the production specifications tend to be manifested late in the product life cycle. When the design is new, there is often intensive interaction between the design and production teams, providing ample opportunity for undocumented information flow. Later, as production experience is gained, some decoupling often occurs between design and production teams. In addition, key personnel may leave, and their replacements may lack comparable training, experience, or institutional knowledge.

Particular care should be taken when the product involves new and unproved manufacturing processes, or established processes which are new to the manufacturer. It may not be possible to determine the adequacy of full-scale manufacturing on the basis of successfully building prototypes or models in a laboratory and testing these prototypes or models. The engineering feasibility and production feasibility may be different because the equipment, tools, personnel, operating procedures, supervision and motivation could be different when a manufacturer scales up for routine production.

No design team can anticipate all factors bearing on the success of the design, but procedures for design transfer should address at least the following basic elements.

- First, the design and development procedures should include a qualitative assessment of the completeness and adequacy of the production specifications.
- Second, the procedures should ensure that all documents and articles which constitute the production specifications are reviewed and approved.
- Third, the procedures should ensure that only approved specifications are used to manufacture production devices.

The first item in the preceding list may be addressed during design transfer. The second and third elements are among the basic principles of document control and configuration management. As long as the production specifications are traditional paper documents, there is ample information available to guide manufacturers in implementing suitable procedures. When the production specifications include non-traditional means, flexibility and creativity may be needed to achieve comparable rigor.

SECTION I. DESIGN CHANGES

I. REQUIREMENTS

§ 820.30(i) Design changes.

- Each manufacturer shall establish and maintain procedures for the identification, documentation, validation or where appropriate verification, review, and approval of design changes before their implementation.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.4.9 Design changes.

II. DISCUSSION AND POINTS TO CONSIDER

There are two principal administrative elements involved in controlling design changes:

- Document control—enumeration of design documents, and tracking their status and revision history. Throughout this section, the term "document" is used in an inclusive sense to mean all design documents, drawings, and other items of design input or output which characterize the design or some aspect of it.
- Change control—enumeration of deficiencies and corrective actions arising from verification and review of the design, and tracking their resolution prior to design transfer.

For a small development project, an adequate process for managing change involves little more than documenting the design change, performing appropriate verification and validation, and keeping records of reviews. The main objectives are ensuring that:

- corrective actions are tracked to completion;
- changes are implemented in such a manner that the original problem is resolved and no new problems are created; or if new problems are created, they are also tracked to resolution; and
- design documentation is updated to accurately reflect the revised design.

For projects involving more than two persons, coordination and communication of design changes become vitally important. In other words, manufacturers should take steps to avoid the common situation where, for example, Jon and Marie agree to a make a change but neglect to inform Pat of their decision.

Medical device manufacturers are usually quite comfortable with the processes of document control and change control with respect to managing manufacturing documents. The principles of these processes are reviewed in the following paragraphs. Subsequently, we will explore how these may be applied to design activities.

DOCUMENT CONTROL. The features of a manufacturing document control system typically include the following:

- Documents should be identified (i.e., named and numbered) in accordance with some logical scheme which links the documents to the product or component they describe or depict and illuminates the drawing hierarchy.
- A master list or index of documents should be maintained which presents a comprehensive overview of the documentation which collectively defines the product and/or process.
- Approval procedures should be prescribed which govern entry of documents into the document control system.
- A history of document revisions should be maintained.
- Procedures for distributing copies of controlled documents and tracking their location should be prescribed.
- Files of controlled documents should be periodically inventoried to ensure that the contents are up to date.
- A person or persons should be assigned specific responsibility to oversee and carry out these procedures. It is desirable that the document control system be administered by a person who is not directly involved with developing or using the documents. For a small manufacturer, document control might be a part-time job for a technician or clerical staff person. More typically, one or more librarians or full-time clerical or paraprofessional employees are required to administer the system.
- There should be a procedure for removal and deletion of obsolete documents.

CHANGE CONTROL. Manufacturing change control is usually implemented using a set of standardized procedures similar to the following:

- A change request might be originated by a developer, manager, reviewer, marketing representative, user, customer, quality assurance representative, or production personnel, and identifies a design problem which the requester believes should be corrected. Change requests are typically reviewed following the manufacturer's prescribed review process, and the request might be rejected, deferred, or accepted.
- If a change request is accepted and corrective action is straightforward, a change order might be issued on the spot to implement the change. The change order pertains to an explicitly identified document or group of documents, and specifies the detailed revision of the document content which will fix the identified problem.
- Often, the change request results in an assignment to developers to further study the problem and develop a suitable corrective action. If the change is extensive, wholesale revision of affected documents may be warranted in lieu of issuing change orders.
- Change requests and change orders should be communicated to all persons whose work might be impacted by the change.

- It may not be practical to immediately revise documents affected by a change order. Instead, the common practice is to distribute and attach a copy of the change order to each controlled copy of the original document.
- Change control procedures should incorporate review and assessment of the impact of the design change on the design input requirements and intended uses.
- A mechanism should be established to track all change requests and change orders to ensure proper disposition.
- Change control procedures are usually administered by the document control staff.

APPLICATION OF DOCUMENT AND CHANGE CONTROLS TO DESIGN. The design control system has to be concerned with the creation and revision of documents, as well as the management of finished documents. Additional mechanisms are required to provide needed flexibility while preserving the integrity of design documentation. These additional mechanisms are embodied in the procedures for review and approval of various documents.

It is important that the design change procedures always include re-verifying and re-validating the design. Fortunately, most design changes occur early in the design process, prior to extensive design validation. Thus, for most design changes, a simple inspection is all that is required. The later in the development cycle that the change occurs, the more important the validation review becomes. There are numerous cases when seemingly innocuous design changes made late in the design phase or following release of the design to market have had disastrous consequences.

For example, a manufacturer encountered problems in the field with a valve sticking in a ventilator due to moisture in the breathing circuit. The problem was resolved by slightly increasing the weight of the disc. Since the change was minor, minimal testing was performed to verify the change. Subsequently, when the revised valves entered production, significant numbers of valves began failing. Investigation revealed that the heavier disc was causing the valve cage to separate due to higher inertia. This failure mode was more serious than the original sticking problem, and resulted in a safety recall.

SECTION J. DESIGN HISTORY FILE (DHF)

I. REQUIREMENTS

§ 820.30(j) Design history file.

- Each manufacturer shall establish and maintain a DHF for each type of device.
- The DHF shall contain or reference the records necessary to demonstrate that the design was developed in accordance with the approved design plan and the requirements of this part.

Cross-reference to ISO 9001:1994 and ISO/DIS 13485 section 4.16 Control of quality records.

II. DEFINITIONS

§ 820.3(e) *Design history file* (DHF) means a compilation of records which describes the design history of a finished device.

III. DISCUSSION AND POINTS TO CONSIDER

There is no specific requirement in ISO 9001 or ISO 13485 for a design history file. However, in order to market a medical device in the United States, a manufacturer must comply with the U. S. Food and Drug Administration (FDA) quality system regulation, which requires a design history file. For this reason, some guidance is provided on the U. S. FDA design history file.

Other national regulations require some form of documentation and records. Product documentation required by Canada, Europe, and Japan contain certain elements of the U. S. FDA design history file requirements without requiring all the elements to be compiled in a file.

Virtually every section of the design control requirements specifies information which should be recorded. The compilation of these records is sometimes referred to as the design history file. Throughout this guidance document, suggestions are made when warranted as to the form and content of documents contained in the design history file.

The primary beneficiary of the device history file is the device manufacturer. For example, in one case, a microprocessor-controlled enteral feeding pump was reported to be behaving erratically in the field. Some of the symptoms pointed to software problems. But the manufacturer admitted that they did not possess a copy of the software source code for the product. The software had been developed by a contractor who had delivered only a master EPROM (memory chip) which was duplicated by the manufacturer to install the

software in each machine. The contractor had subsequently withdrawn following a contractual dispute, leaving the manufacturer with no rights to the source code developed by the contractor, and no practical way to maintain the software. For this and other reasons, the product was the subject of a mandatory recall and all known units were collected and destroyed.

This is admittedly an extreme case, but many similar cases have been documented in which the manufacturer lacked design information necessary to validate a design and maintain it throughout the product life cycle. This occurs for the most innocent of reasons—contracts expire, companies reorganize, employees move on to new projects or new jobs. Even when the designer is available, he or she may forget why a particular decision was made years, months, or even weeks before. Since design decisions often directly affect the well-being of device users and patients, it is to the manufacturer's benefit to maintain the knowledge base which forms a basis for the product design.

Except for small projects, it is unusual for all design history documents to be filed in a single location. For example, many design engineers maintain laboratory notebooks which are typically retained in the engineers' personal files. In addition, the design history may include memoranda and electronic mail correspondence which are stored at various physical locations. Quality system plans applicable to a development project may reside in the quality assurance department, while the chief engineer may be responsible for maintaining design and development plans. These diverse records need not be consolidated at a single location. The intent is simply that manufacturers have access to the information when it is needed. If a manufacturer has established procedures for multiple filing systems which together satisfy that intent, there is no need to create additional procedures or records.

As an example of the level of detail which may be entailed, some manufacturers have policies covering laboratory notebooks. Manufacturers typically find that without such written procedures, a breakdown in communications eventually occurs, resulting in a loss of control. These procedures might address the following points.

- Laboratory notebooks are the property of the manufacturer, not the individual.
- A separate notebook is to be maintained for each project, and surrendered to the engineering librarian at the conclusion of the engineer's active participation in the project.
- Laboratory notebooks are to be surrendered if the employee leaves the company.
- Product development supervisors shall review employees' laboratory notebooks at specified intervals to ensure that records are complete, accurate, and legible.

There are no requirements on the location or organization of the design history file. In some cases, especially for simple designs, the designer will assemble and maintain the entire design history file. For larger projects, a document control system will likely be established for design documents, and these files will likely be maintained in some central location, usually within the product development department.

Based on the structure (or lack thereof) of the product development organization, more or less extensive controls will be required. For example, company policy should state unequivocally that all design history documentation is the property of the manufacturer, not the employee or contractor. Design and development contracts should explicitly specify the manufacturer's right to design information and establish standards for the form and content of design documentation. Finally, certain basic design information may be maintained in a single project file in a specified location. This may include the following:

- Detailed design and development plan specifying design tasks and deliverables.
- Copies of approved design input documents and design output documents.
- Documentation of design reviews.
- Validation documentation.
- When applicable, copies of controlled design documents and change control records.

END OF DOCUMENT: DCGD.DOC (Microsoft Word 7.0) or DCGD.PDF (Adobe Acrobat format)

Appendix B: NIST/SEMATECH e-Handbook of Statistical Methods

1. Exploratory Data Analysis

This chapter presents the assumptions, principles, and techniques necessary to gain insight into data via EDA--exploratory data analysis.

1. EDA Introduction

1. What is EDA?
2. EDA vs Classical & Bayesian
3. EDA vs Summary
4. EDA Goals
5. The Role of Graphics
6. An EDA/Graphics Example
7. General Problem Categories

2. EDA Assumptions

1. Underlying Assumptions
2. Importance
3. Techniques for Testing Assumptions
4. Interpretation of 4-Plot
5. Consequences

3. EDA Techniques

1. Introduction
2. Analysis Questions
3. Graphical Techniques: Alphabetical
4. Graphical Techniques: By Problem Category
5. Quantitative Techniques
6. Probability Distributions

4. EDA Case Studies

1. Introduction
2. By Problem Category

Detailed Chapter Table of Contents
References
Dataplot Commands for EDA Techniques

1. Exploratory Data Analysis - Detailed Table of Contents [1.]

This chapter presents the assumptions, principles, and techniques necessary to gain insight into data via EDA--exploratory data analysis.

1. EDA Introduction [1.1.]
    1. What is EDA? [1.1.1.]
    2. How Does Exploratory Data Analysis differ from Classical Data Analysis? [1.1.2.]
        1. Model [1.1.2.1.]
        2. Focus [1.1.2.2.]
        3. Techniques [1.1.2.3.]
        4. Rigor [1.1.2.4.]
        5. Data Treatment [1.1.2.5.]
        6. Assumptions [1.1.2.6.]
    3. How Does Exploratory Data Analysis Differ from Summary Analysis? [1.1.3.]
    4. What are the EDA Goals? [1.1.4.]
    5. The Role of Graphics [1.1.5.]
    6. An EDA/Graphics Example [1.1.6.]
    7. General Problem Categories [1.1.7.]

2. EDA Assumptions [1.2.]
    1. Underlying Assumptions [1.2.1.]
    2. Importance [1.2.2.]
    3. Techniques for Testing Assumptions [1.2.3.]
    4. Interpretation of 4-Plot [1.2.4.]
    5. Consequences [1.2.5.]
        1. Consequences of Non-Randomness [1.2.5.1.]
        2. Consequences of Non-Fixed Location Parameter [1.2.5.2.]
        3. Consequences of Non-Fixed Variation Parameter [1.2.5.3.]
        4. Consequences Related to Distributional Assumptions [1.2.5.4.]

3. EDA Techniques [1.3.]
    1. Introduction [1.3.1.]
    2. Analysis Questions [1.3.2.]
    3. Graphical Techniques: Alphabetic [1.3.3.]
        1. Autocorrelation Plot [1.3.3.1.]
            1. Autocorrelation Plot: Random Data [1.3.3.1.1.]
            2. Autocorrelation Plot: Moderate Autocorrelation [1.3.3.1.2.]
            3. Autocorrelation Plot: Strong Autocorrelation and Autoregressive Model [1.3.3.1.3.]
            4. Autocorrelation Plot: Sinusoidal Model [1.3.3.1.4.]
        2. Bihistogram [1.3.3.2.]
        3. Block Plot [1.3.3.3.]
        4. Bootstrap Plot [1.3.3.4.]
        5. Box-Cox Linearity Plot [1.3.3.5.]
        6. Box-Cox Normality Plot [1.3.3.6.]
        7. Box Plot [1.3.3.7.]

8. Complex Demodulation Amplitude Plot [1.3.3.8.]
9. Complex Demodulation Phase Plot [1.3.3.9.]
10. Contour Plot [1.3.3.10.]
    1. DOE Contour Plot [1.3.3.10.1.]
11. DOE Scatter Plot [1.3.3.11.]
12. DOE Mean Plot [1.3.3.12.]
13. DOE Standard Deviation Plot [1.3.3.13.]
14. Histogram [1.3.3.14.]
    1. Histogram Interpretation: Normal [1.3.3.14.1.]
    2. Histogram Interpretation: Symmetric, Non-Normal, Short-Tailed [1.3.3.14.2.]
    3. Histogram Interpretation: Symmetric, Non-Normal, Long-Tailed [1.3.3.14.3.]
    4. Histogram Interpretation: Symmetric and Bimodal [1.3.3.14.4.]
    5. Histogram Interpretation: Bimodal Mixture of 2 Normals [1.3.3.14.5.]
    6. Histogram Interpretation: Skewed (Non-Normal) Right [1.3.3.14.6.]
    7. Histogram Interpretation: Skewed (Non-Symmetric) Left [1.3.3.14.7.]
    8. Histogram Interpretation: Symmetric with Outlier [1.3.3.14.8.]
15. Lag Plot [1.3.3.15.]
    1. Lag Plot: Random Data [1.3.3.15.1.]
    2. Lag Plot: Moderate Autocorrelation [1.3.3.15.2.]
    3. Lag Plot: Strong Autocorrelation and Autoregressive Model [1.3.3.15.3.]
    4. Lag Plot: Sinusoidal Models and Outliers [1.3.3.15.4.]
16. Linear Correlation Plot [1.3.3.16.]
17. Linear Intercept Plot [1.3.3.17.]
18. Linear Slope Plot [1.3.3.18.]
19. Linear Residual Standard Deviation Plot [1.3.3.19.]
20. Mean Plot [1.3.3.20.]
21. Normal Probability Plot [1.3.3.21.]
    1. Normal Probability Plot: Normally Distributed Data [1.3.3.21.1.]
    2. Normal Probability Plot: Data Have Short Tails [1.3.3.21.2.]
    3. Normal Probability Plot: Data Have Long Tails [1.3.3.21.3.]
    4. Normal Probability Plot: Data are Skewed Right [1.3.3.21.4.]
22. Probability Plot [1.3.3.22.]
23. Probability Plot Correlation Coefficient Plot [1.3.3.23.]
24. Quantile-Quantile Plot [1.3.3.24.]
25. Run-Sequence Plot [1.3.3.25.]
26. Scatter Plot [1.3.3.26.]
    1. Scatter Plot: No Relationship [1.3.3.26.1.]
    2. Scatter Plot: Strong Linear (positive correlation) Relationship [1.3.3.26.2.]
    3. Scatter Plot: Strong Linear (negative correlation) Relationship [1.3.3.26.3.]
    4. Scatter Plot: Exact Linear (positive correlation) Relationship [1.3.3.26.4.]
    5. Scatter Plot: Quadratic Relationship [1.3.3.26.5.]
    6. Scatter Plot: Exponential Relationship [1.3.3.26.6.]
    7. Scatter Plot: Sinusoidal Relationship (damped) [1.3.3.26.7.]
    8. Scatter Plot: Variation of Y Does Not Depend on X (homoscedastic) [1.3.3.26.8.]
    9. Scatter Plot: Variation of Y Does Depend on X (heteroscedastic) [1.3.3.26.9.]
    10. Scatter Plot: Outlier [1.3.3.26.10.]
    11. Scatterplot Matrix [1.3.3.26.11.]
    12. Conditioning Plot [1.3.3.26.12.]
27. Spectral Plot [1.3.3.27.]
    1. Spectral Plot: Random Data [1.3.3.27.1.]
    2. Spectral Plot: Strong Autocorrelation and Autoregressive Model [1.3.3.27.2.]
    3. Spectral Plot: Sinusoidal Model [1.3.3.27.3.]
28. Standard Deviation Plot [1.3.3.28.]
29. Star Plot [1.3.3.29.]

30. Weibull Plot [1.3.3.30.]
31. Youden Plot [1.3.3.31.]
    1. DOE Youden Plot [1.3.3.31.1.]
32. 4-Plot [1.3.3.32.]
33. 6-Plot [1.3.3.33.]

4. Graphical Techniques: By Problem Category [1.3.4.]
5. Quantitative Techniques [1.3.5.]
    1. Measures of Location [1.3.5.1.]
    2. Confidence Limits for the Mean [1.3.5.2.]
    3. Two-Sample *t*-Test for Equal Means [1.3.5.3.]
        1. Data Used for Two-Sample *t*-Test [1.3.5.3.1.]
    4. One-Factor ANOVA [1.3.5.4.]
    5. Multi-factor Analysis of Variance [1.3.5.5.]
    6. Measures of Scale [1.3.5.6.]
    7. Bartlett's Test [1.3.5.7.]
    8. Chi-Square Test for the Standard Deviation [1.3.5.8.]
        1. Data Used for Chi-Square Test for the Standard Deviation [1.3.5.8.1.]
    9. F-Test for Equality of Two Standard Deviations [1.3.5.9.]
    10. Levene Test for Equality of Variances [1.3.5.10.]
    11. Measures of Skewness and Kurtosis [1.3.5.11.]
    12. Autocorrelation [1.3.5.12.]
    13. Runs Test for Detecting Non-randomness [1.3.5.13.]
    14. Anderson-Darling Test [1.3.5.14.]
    15. Chi-Square Goodness-of-Fit Test [1.3.5.15.]
    16. Kolmogorov-Smirnov Goodness-of-Fit Test [1.3.5.16.]
    17. Grubbs' Test for Outliers [1.3.5.17.]
    18. Yates Analysis [1.3.5.18.]
        1. Defining Models and Prediction Equations [1.3.5.18.1.]
        2. Important Factors [1.3.5.18.2.]
6. Probability Distributions [1.3.6.]
    1. What is a Probability Distribution [1.3.6.1.]
    2. Related Distributions [1.3.6.2.]
    3. Families of Distributions [1.3.6.3.]
    4. Location and Scale Parameters [1.3.6.4.]
    5. Estimating the Parameters of a Distribution [1.3.6.5.]
        1. Method of Moments [1.3.6.5.1.]
        2. Maximum Likelihood [1.3.6.5.2.]
        3. Least Squares [1.3.6.5.3.]
        4. PPCC and Probability Plots [1.3.6.5.4.]
    6. Gallery of Distributions [1.3.6.6.]
        1. Normal Distribution [1.3.6.6.1.]
        2. Uniform Distribution [1.3.6.6.2.]
        3. Cauchy Distribution [1.3.6.6.3.]
        4. t Distribution [1.3.6.6.4.]
        5. F Distribution [1.3.6.6.5.]
        6. Chi-Square Distribution [1.3.6.6.6.]
        7. Exponential Distribution [1.3.6.6.7.]
        8. Weibull Distribution [1.3.6.6.8.]
        9. Lognormal Distribution [1.3.6.6.9.]
        10. Fatigue Life Distribution [1.3.6.6.10.]
        11. Gamma Distribution [1.3.6.6.11.]
        12. Double Exponential Distribution [1.3.6.6.12.]
        13. Power Normal Distribution [1.3.6.6.13.]
        14. Power Lognormal Distribution [1.3.6.6.14.]

15. Tukey-Lambda Distribution [1.3.6.6.15.]
16. Extreme Value Type I Distribution [1.3.6.6.16.]
17. Beta Distribution [1.3.6.6.17.]
18. Binomial Distribution [1.3.6.6.18.]
19. Poisson Distribution [1.3.6.6.19.]

7. Tables for Probability Distributions [1.3.6.7.]
    1. Cumulative Distribution Function of the Standard Normal Distribution [1.3.6.7.1.]
    2. Upper Critical Values of the Student's-t Distribution [1.3.6.7.2.]
    3. Upper Critical Values of the F Distribution [1.3.6.7.3.]
    4. Critical Values of the Chi-Square Distribution [1.3.6.7.4.]
    5. Critical Values of the t^* Distribution [1.3.6.7.5.]
    6. Critical Values of the Normal PPCC Distribution [1.3.6.7.6.]

4. EDA Case Studies [1.4.]
    1. Case Studies Introduction [1.4.1.]
    2. Case Studies [1.4.2.]
        1. Normal Random Numbers [1.4.2.1.]
            1. Background and Data [1.4.2.1.1.]
            2. Graphical Output and Interpretation [1.4.2.1.2.]
            3. Quantitative Output and Interpretation [1.4.2.1.3.]
            4. Work This Example Yourself [1.4.2.1.4.]
        2. Uniform Random Numbers [1.4.2.2.]
            1. Background and Data [1.4.2.2.1.]
            2. Graphical Output and Interpretation [1.4.2.2.2.]
            3. Quantitative Output and Interpretation [1.4.2.2.3.]
            4. Work This Example Yourself [1.4.2.2.4.]
        3. Random Walk [1.4.2.3.]
            1. Background and Data [1.4.2.3.1.]
            2. Test Underlying Assumptions [1.4.2.3.2.]
            3. Develop A Better Model [1.4.2.3.3.]
            4. Validate New Model [1.4.2.3.4.]
            5. Work This Example Yourself [1.4.2.3.5.]
        4. Josephson Junction Cryothermometry [1.4.2.4.]
            1. Background and Data [1.4.2.4.1.]
            2. Graphical Output and Interpretation [1.4.2.4.2.]
            3. Quantitative Output and Interpretation [1.4.2.4.3.]
            4. Work This Example Yourself [1.4.2.4.4.]
        5. Beam Deflections [1.4.2.5.]
            1. Background and Data [1.4.2.5.1.]
            2. Test Underlying Assumptions [1.4.2.5.2.]
            3. Develop a Better Model [1.4.2.5.3.]
            4. Validate New Model [1.4.2.5.4.]
            5. Work This Example Yourself [1.4.2.5.5.]
        6. Filter Transmittance [1.4.2.6.]
            1. Background and Data [1.4.2.6.1.]
            2. Graphical Output and Interpretation [1.4.2.6.2.]
            3. Quantitative Output and Interpretation [1.4.2.6.3.]
            4. Work This Example Yourself [1.4.2.6.4.]
        7. Standard Resistor [1.4.2.7.]
            1. Background and Data [1.4.2.7.1.]
            2. Graphical Output and Interpretation [1.4.2.7.2.]
            3. Quantitative Output and Interpretation [1.4.2.7.3.]
            4. Work This Example Yourself [1.4.2.7.4.]

8. Heat Flow Meter 1 [1.4.2.8.]
    1. Background and Data [1.4.2.8.1.]
    2. Graphical Output and Interpretation [1.4.2.8.2.]
    3. Quantitative Output and Interpretation [1.4.2.8.3.]
    4. Work This Example Yourself [1.4.2.8.4.]
9. Fatigue Life of Aluminum Alloy Specimens [1.4.2.9.]
    1. Background and Data [1.4.2.9.1.]
    2. Graphical Output and Interpretation [1.4.2.9.2.]
10. Ceramic Strength [1.4.2.10.]
    1. Background and Data [1.4.2.10.1.]
    2. Analysis of the Response Variable [1.4.2.10.2.]
    3. Analysis of the Batch Effect [1.4.2.10.3.]
    4. Analysis of the Lab Effect [1.4.2.10.4.]
    5. Analysis of Primary Factors [1.4.2.10.5.]
    6. Work This Example Yourself [1.4.2.10.6.]
3. References For Chapter 1: Exploratory Data Analysis [1.4.3.]

1.1. EDA Introduction

Summary

What is exploratory data analysis? How did it begin? How and where did it originate? How is it differentiated from other data analysis approaches, such as classical and Bayesian? Is EDA the same as statistical graphics? What role does statistical graphics play in EDA? Is statistical graphics identical to EDA?

These questions and related questions are dealt with in this section. This section answers these questions and provides the necessary frame of reference for EDA assumptions, principles, and techniques.

Table of Contents for Section 1

1. What is EDA?
2. EDA versus Classical and Bayesian
    1. Models
    2. Focus
    3. Techniques
    4. Rigor
    5. Data Treatment
    6. Assumptions
3. EDA vs Summary
4. EDA Goals
5. The Role of Graphics
6. An EDA/Graphics Example
7. General Problem Categories

1.1.1. What is EDA?

Approach Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

Focus The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

Philosophy EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics per se.

History The seminal work in EDA is Exploratory Data Analysis, Tukey, (1977). Over the years it has benefitted from other noteworthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), The ABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.

Techniques Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on

graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.

2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.

3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

Data Analysis Approaches

EDA is a data analysis approach. What other data analysis approaches exist and how does EDA differ from these other approaches? Three popular data analysis approaches are:

1. Classical
2. Exploratory (EDA)
3. Bayesian

Paradigms for Analysis Techniques

These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

For classical analysis, the sequence is

Problem => Data => Model => Analysis => Conclusions

For EDA, the sequence is

Problem => Data => Analysis => Model => Conclusions

For Bayesian, the sequence is

Problem => Data => Model => Prior Distribution => Analysis => Conclusions

Method of dealing with underlying model for the data distinguishes the 3 approaches

Thus for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model. For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate. Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected

data to jointly make inferences and/or test assumptions about the model parameters.

In the real world, data analysts freely mix elements of all of the above three approaches (and other approaches). The above distinctions were made to emphasize the major differences among the three approaches.

Further discussion of the distinction between the classical and EDA approaches

Focusing on EDA versus classical, these two approaches differ as follows:

1. Models
2. Focus
3. Techniques
4. Rigor
5. Data Treatment
6. Assumptions

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.1. Model

Classical	The classical approach imposes models (both deterministic and probabilistic) on the data. Deterministic models include, for example, regression models and analysis of variance (ANOVA) models. The most common probabilistic model assumes that the errors about the deterministic model are normally distributed--this assumption affects the validity of the ANOVA F tests.
Exploratory	The Exploratory Data Analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.2. Focus

Classical	The two approaches differ substantially in focus. For classical analysis, the focus is on the model--estimating parameters of the model and generating predicted values from the model.
Exploratory	For exploratory data analysis, the focus is on the data--its structure, outliers, and models suggested by the data.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.3. Techniques

Classical	Classical techniques are generally quantitative in nature. They include ANOVA, t tests, chi-squared tests, and F tests.
Exploratory	EDA techniques are generally graphical. They include scatter plots, character plots, box plots, histograms, bihistograms, probability plots, residual plots, and mean plots.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.4. Rigor

Classical	Classical techniques serve as the probabilistic foundation of science and engineering; the most important characteristic of classical techniques is that they are rigorous, formal, and "objective".
Exploratory	EDA techniques do not share in that rigor or formality. EDA techniques make up for that lack of rigor by being very suggestive, indicative, and insightful about what the appropriate model should be. EDA techniques are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.5. Data Treatment

Classical	Classical estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers ("estimates"). This is both a virtue and a vice. The virtue is that these few numbers focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other characteristics (skewness, tail length, autocorrelation, etc.) of the same population. In this sense there is a loss of information due to this "filtering" process.
Exploratory	The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.

1. Exploratory Data Analysis
1.1. EDA Introduction
1.1.2. How Does Exploratory Data Analysis differ from Classical Data Analysis?

1.1.2.6. Assumptions

Classical	The "good news" of the classical approach is that tests based on classical techniques are usually very sensitive--that is, if a true shift in location, say, has occurred, such tests frequently have the power to detect such a shift and to conclude that such a shift is "statistically significant". The "bad news" is that classical tests depend on underlying assumptions (e.g., normality), and hence the validity of the test conclusions becomes dependent on the validity of the underlying assumptions. Worse yet, the exact underlying assumptions may be unknown to the analyst, or if known, untested. Thus the validity of the scientific conclusions becomes intrinsically linked to the validity of the underlying assumptions. In practice, if such assumptions are unknown or untested, the validity of the scientific conclusions becomes suspect.
Exploratory	Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions.

1.1.3. How Does Exploratory Data Analysis Differ from Summary Analysis?

Summary A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table.

Exploratory In contrast, EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data. There is an archival role in the research and manufacturing world for summary statistics, but there is an enormously larger role for the EDA approach.

1.1.4. What are the EDA Goals?

Primary and Secondary Goals

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Insight into the Data

Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data. Graphics are irreplaceable--there are no quantitative analogues that will give the same insight as well-chosen graphics.

To get a "feel" for the data, it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

1.1.5. The Role of Graphics

Quantitative/ Graphical

Statistics and data analysis procedures can broadly be split into two parts:

- quantitative
- graphical

Quantitative

Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

- hypothesis testing
- analysis of variance
- point estimates and confidence intervals
- least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis.

Graphical

On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. These include:

- scatter plots
- histograms
- probability plots
- residual plots
- box plots
- block plots

EDA Approach Relies Heavily on Graphical Techniques

The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of

- testing assumptions
- model selection
- model validation
- estimator selection
- relationship identification
- factor effect determination

- outlier detection

If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data.

1.1.6. An EDA/Graphics Example

Anscombe Example

A simple, classic (Anscombe) example of the central role that graphics play in terms of providing insight into a data set starts with the following data set:

Data

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

Summary Statistics

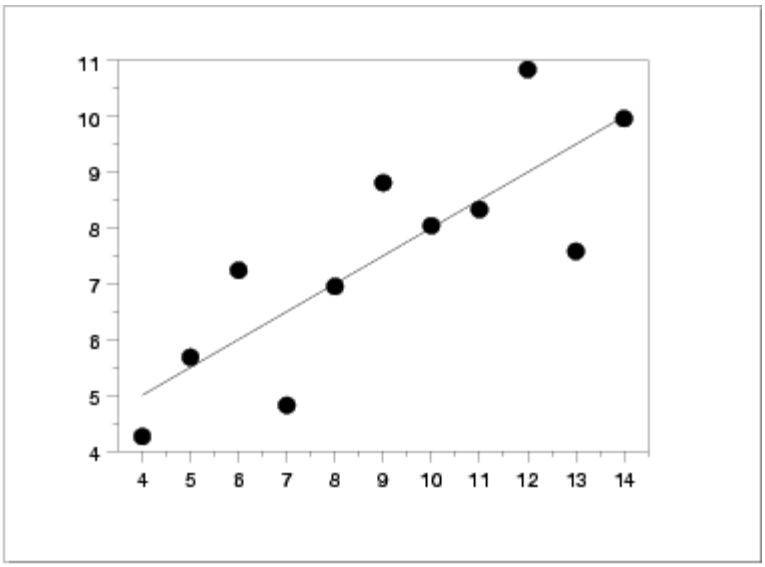
If the goal of the analysis is to compute summary statistics plus determine the best linear fit for Y as a function of X, the results might be given as:

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

The above quantitative analysis, although valuable, gives us only limited insight into the data.

Scatter Plot

In contrast, the following simple scatter plot of the data



suggests the following:

1. The data set "behaves like" a linear curve with some scatter;
2. there is no justification for a more complicated model (e.g., quadratic);
3. there are no outliers;
4. the vertical spread of the data appears to be of equal height irrespective of the *X*-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

Three Additional Data Sets

This kind of characterization for the data serves as the core for getting insight/feel for the data. Such insight/feel does not come from the quantitative statistics; on the contrary, calculations of quantitative statistics such as intercept and slope should be subsequent to the characterization and will make sense only if the characterization is true. To illustrate the loss of information that results when the graphics insight step is skipped, consider the following three data sets [Anscombe data sets 2, 3, and 4]:

X2	Y2	X3	Y3	X4	Y4
10.00	9.14	10.00	7.46	8.00	6.58
8.00	8.14	8.00	6.77	8.00	5.76
13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.77	9.00	7.11	8.00	8.84
11.00	9.26	11.00	7.81	8.00	8.47
14.00	8.10	14.00	8.84	8.00	7.04
6.00	6.13	6.00	6.08	8.00	5.25
4.00	3.10	4.00	5.39	19.00	12.50
12.00	9.13	12.00	8.15	8.00	5.56
7.00	7.26	7.00	6.42	8.00	7.91
5.00	4.74	5.00	5.73	8.00	6.89

Quantitative Statistics for Data Set 2

A quantitative analysis on data set 2 yields

N = 11
Mean of *X* = 9.0
Mean of *Y* = 7.5
Intercept = 3

Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

which is identical to the analysis for data set 1. One might naively assume that the two data sets are "equivalent" since that is what the statistics tell us; but what do the statistics not tell us?

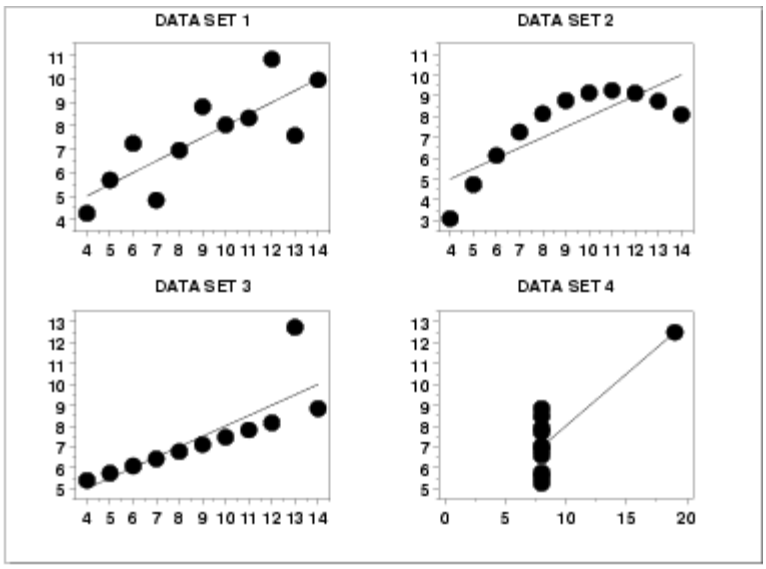
Quantitative Statistics for Data Sets 3 and 4

Remarkably, a quantitative analysis on data sets 3 and 4 also yields

N = 11
Mean of X = 9.0
Mean of Y = 7.5
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.236
Correlation = 0.816 (0.817 for data set 4)

which implies that in some quantitative sense, all four of the data sets are "equivalent". In fact, the four data sets are far from "equivalent" and a scatter plot of each data set, which would be step 1 of any EDA approach, would tell us that immediately.

Scatter Plots



Interpretation of Scatter Plots

Conclusions from the scatter plots are:

1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Importance

These points are exactly the substance that provide and

of Exploratory Analysis

define "insight" and "feel" for a data set. They are the goals and the fruits of an open exploratory data analysis (EDA) approach to the data. Quantitative statistics are not wrong per se, but they are incomplete. They are incomplete because they are numeric **summaries** which in the summarization operation do a good job of focusing on a particular aspect of the data (e.g., location, intercept, slope, degree of relatedness, etc.) by judiciously reducing the data to a few numbers. Doing so also **filters** the data, necessarily omitting and screening out other sometimes crucial information in the focusing operation. Quantitative statistics focus but also filter; and filtering is exactly what makes the quantitative approach incomplete at best and misleading at worst.

The estimated intercepts (= 3) and slopes (= 0.5) for data sets 2, 3, and 4 are misleading because the estimation is done in the context of an assumed linear model and that linearity assumption is the fatal flaw in this analysis.

The EDA approach of deliberately postponing the model selection until further along in the analysis has many rewards, not the least of which is the ultimate convergence to a much-improved model and the formulation of valid and supportable scientific and engineering conclusions.

1.1.7. General Problem Categories

Problem Classification

The following table is a convenient way to classify EDA problems.

Univariate and Control

UNIVARIATE	CONTROL
Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: 1. A number (the estimated constant in the model). 2. An estimate of uncertainty for the constant. 3. An estimate of the distribution for the error. Techniques: • 4-Plot • Probability Plot • PPCC Plot	Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: A "yes" or "no" to the question "Is the system out of control? ". Techniques: • Control Charts

Comparative and Screening

COMPARATIVE	SCREENING
Data: A single response variable and k independent variables ($Y, X_1, X_2, \ldots, X_k$), primary focus is on	Data: A single response variable and k independent variables ($Y, X_1, X_2, \ldots, X_k$).

	one (the primary factor) of these independent variables. Model: $y = f(x_1, x_2, ..., x_k) +$ error Output: A "yes" or "no" to the question "Is the primary factor significant?". Techniques: • Block Plot • Scatter Plot • Box Plot	Model: $y = f(x_1, x_2, ..., x_k) +$ error Output: 1. A ranked list (from most important to least important) of factors. 2. Best settings for the factors. 3. A good model/prediction equation relating Y to the factors. Techniques: • Block Plot • Probability Plot • Bihistogram
Optimization and Regression	OPTIMIZATION Data: A single response variable and k independent variables ($Y, X_1, X_2, ... , X_k$). Model: $y = f(x_1, x_2, ..., x_k) +$ error Output: Best settings for the factor variables. Techniques: • Block Plot • Least Squares Fitting • Contour Plot	REGRESSION Data: A single response variable and k independent variables ($Y, X_1, X_2, ... , X_k$). The independent variables can be continuous. Model: $y = f(x_1, x_2, ..., x_k) +$ error Output: A good model/prediction equation relating Y to the factors. Techniques: • Least Squares Fitting • Scatter Plot

	6-Plot

Time Series and Multivariate

TIME SERIES	MULTIVARIATE
Data: A column of time dependent numbers, *Y*. In addition, time is an indpendent variable. The time variable can be either explicit or implied. If the data are not equi-spaced, the time variable should be explicitly provided. Model: $y_t = f(t)$ + error The model can be either a time domain based or frequency domain based. Output: A good model/prediction equation relating *Y* to previous values of *Y*. Techniques: • Autocorrelation Plot • Spectrum • Complex Demodulation Amplitude Plot • Complex Demodulation Phase Plot • ARIMA Models	Data: *k* factor variables (X_1, X_2, ... , X_k). Model: The model is not explicit. Output: Identify underlying correlation structure in the data. Techniques: • Star Plot • Scatter Plot Matrix • Conditioning Plot • Profile Plot • Principal Components • Clustering • Discrimination/Classification Note that multivarate analysis is only covered lightly in this Handbook.

1.2. EDA Assumptions

Summary

The gamut of scientific and engineering experimentation is virtually limitless. In this sea of diversity is there any common basis that allows the analyst to systematically and validly arrive at supportable, repeatable research conclusions?

Fortunately, there is such a basis and it is rooted in the fact that every measurement process, however complicated, has certain underlying assumptions. This section deals with what those assumptions are, why they are important, how to go about testing them, and what the consequences are if the assumptions do not hold.

Table of Contents for Section 2

1. Underlying Assumptions
2. Importance
3. Testing Assumptions
4. Importance of Plots
5. Consequences

1.2.1. Underlying Assumptions

Assumptions Underlying a Measurement Process

There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

1. random drawings;
2. from a fixed distribution;
3. with the distribution having fixed location; and
4. with the distribution having fixed variation.

Univariate or Single Response Variable

The "fixed location" referred to in item 3 above differs for different problem types. The simplest problem type is univariate; that is, a single variable. For the univariate problem, the general model

response = deterministic component + random component

becomes

response = constant + error

Assumptions for Univariate Model

For this case, the "fixed location" is simply the unknown constant. We can thus imagine the process at hand to be operating under constant conditions that produce a single column of data with the properties that

- the data are uncorrelated with one another;
- the random component has a fixed distribution;
- the deterministic component consists of only a constant; and
- the random component has fixed variation.

Extrapolation to a Function of Many Variables

The universal power and importance of the univariate model is that it can easily be extended to the more general case where the deterministic component is not just a constant, but is in fact a function of many variables, and the engineering objective is to characterize and model the function.

Residuals

The key point is that regardless of how many factors there

Will Behave According to Univariate Assumptions

are, and regardless of how complicated the function is, if the engineer succeeds in choosing a good model, then the differences (residuals) between the raw response data and the predicted values from the fitted model should themselves behave like a univariate process. Furthermore, the residuals from this univariate process fit will behave like:

- random drawings;
- from a fixed distribution;
- with fixed location (namely, 0 in this case); and
- with fixed variation.

Validation of Model

Thus if the residuals from the fitted model do in fact behave like the ideal, then testing of underlying assumptions becomes a tool for the validation and quality of fit of the chosen model. On the other hand, if the residuals from the chosen fitted model violate one or more of the above univariate assumptions, then the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.

1.2.2. Importance

Predictability and Statistical Control	Predictability is an all-important goal in science and engineering. If the four underlying assumptions hold, then we have achieved probabilistic predictability--the ability to make probability statements not only about the process in the past, but also about the process in the future. In short, such processes are said to be "in statistical control".
Validity of Engineering Conclusions	Moreover, if the four assumptions are valid, then the process is amenable to the generation of valid scientific and engineering conclusions. If the four assumptions are not valid, then the process is drifting (with respect to location, variation, or distribution), unpredictable, and out of control. A simple characterization of such processes by a location estimate, a variation estimate, or a distribution "estimate" inevitably leads to engineering conclusions that are not valid, are not supportable (scientifically or legally), and which are not repeatable in the laboratory.

1.2.3. Techniques for Testing Assumptions

Testing Underlying Assumptions Helps Assure the Validity of Scientific and Engineering Conclusions

Because the validity of the final scientific/engineering conclusions is inextricably linked to the validity of the underlying univariate assumptions, it naturally follows that there is a real necessity that each and every one of the above four assumptions be routinely tested.

Four Techniques to Test Underlying Assumptions

The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:

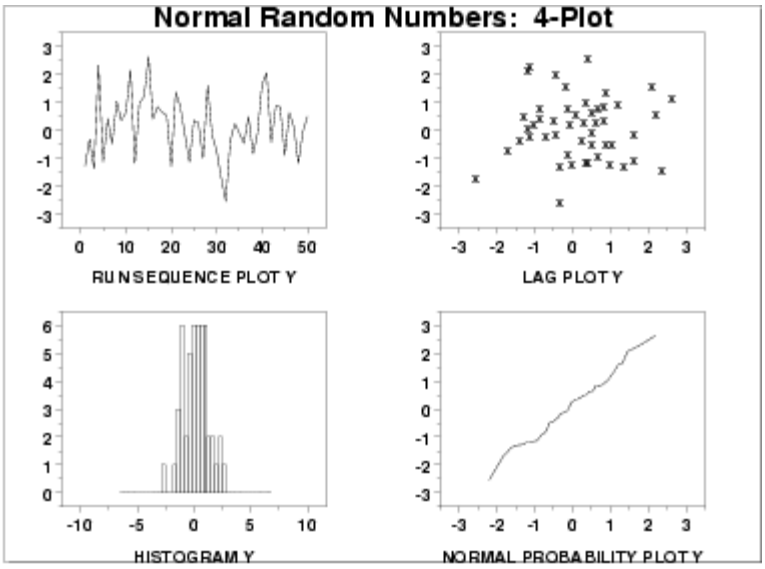
1. run sequence plot (Y_i versus i)
2. lag plot (Y_i versus Y_{i-1})
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y)

Plot on a Single Page for a Quick Characterization of the Data

The four EDA plots can be juxtaposed for a quick look at the characteristics of the data. The plots below are ordered as follows:

1. Run sequence plot - upper left
2. Lag plot - upper right
3. Histogram - lower left
4. Normal probability plot - lower right

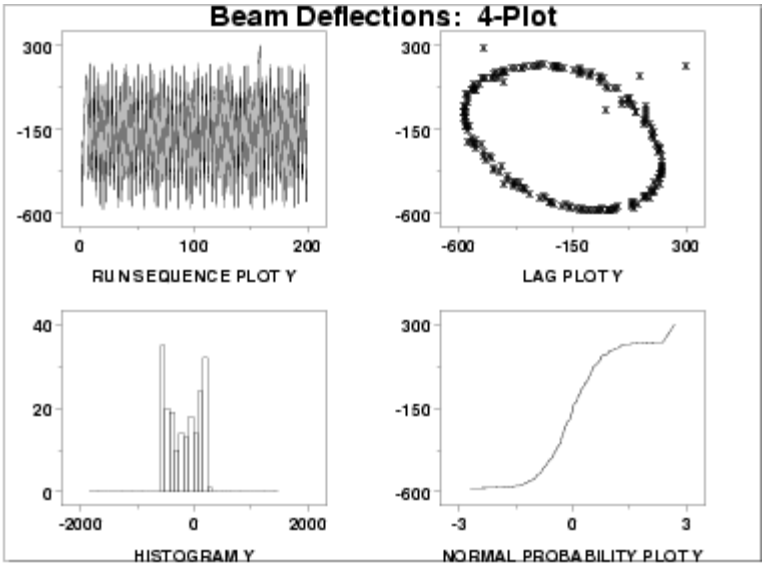
Sample Plot: Assumptions Hold



This 4-plot reveals a process that has fixed location, fixed variation, is random, apparently has a fixed approximately normal distribution, and has no outliers.

Sample Plot: Assumptions Do Not Hold

If one or more of the four underlying assumptions do not hold, then it will show up in the various plots as demonstrated in the following example.



This 4-plot reveals a process that has fixed location, fixed variation, is non-random (oscillatory), has a non-normal, U-shaped distribution, and has several outliers.

1.2.4. Interpretation of 4-Plot

Interpretation of EDA Plots: Flat and Equi-Banded, Random, Bell-Shaped, and Linear

The four EDA plots discussed on the previous page are used to test the underlying assumptions:

1. **Fixed Location:**
   If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.

2. **Fixed Variation:**
   If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.

3. **Randomness:**
   If the randomness assumption holds, then the lag plot will be structureless and random.

4. **Fixed Distribution:**
   If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then
   1. the histogram will be bell-shaped, and
   2. the normal probability plot will be linear.

Plots Utilized to Test the Assumptions

Conversely, the underlying assumptions are tested using the EDA plots:

- **Run Sequence Plot:**
  If the run sequence plot is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.

- **Lag Plot:**
  If the lag plot is structureless, then the randomness assumption holds.

- **Histogram:**
  If the histogram is bell-shaped, the underlying distribution is symmetric and perhaps approximately normal.

Normal Probability Plot:
If the normal probability plot is linear, the underlying distribution is approximately normal.

If all four of the assumptions hold, then the process is said definitionally to be "in statistical control".

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.2.5. Consequences

What If Assumptions Do Not Hold?	If some of the underlying assumptions do not hold, what can be done about it? What corrective actions can be taken? The positive way of approaching this is to view the testing of underlying assumptions as a framework for learning about the process. Assumption-testing promotes insight into important aspects of the process that may not have surfaced otherwise.
Primary Goal is Correct and Valid Scientific Conclusions	The primary goal is to have correct, validated, and complete scientific/engineering conclusions flowing from the analysis. This usually includes intermediate goals such as the derivation of a good-fitting model and the computation of realistic parameter estimates. It should always include the ultimate goal of an understanding and a "feel" for "what makes the process tick". There is no more powerful catalyst for discovery than the bringing together of an experienced/expert scientist/engineer and a data set ripe with intriguing "anomalies" and characteristics.
Consequences of Invalid Assumptions	The following sections discuss in more detail the consequences of invalid assumptions: 1. Consequences of non-randomness 2. Consequences of non-fixed location parameter 3. Consequences of non-fixed variation 4. Consequences related to distributional assumptions

1.2.5.1. Consequences of Non-Randomness

Randomness Assumption

There are four underlying assumptions:

1. randomness;
2. fixed location;
3. fixed variation; and
4. fixed distribution.

The randomness assumption is the most critical but the least tested.

Consequeces of Non-Randomness

If the randomness assumption does not hold, then

1. All of the usual statistical tests are invalid.
2. The calculated uncertainties for commonly used statistics become meaningless.
3. The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
4. The simple model: y = constant + error becomes invalid.
5. The parameter estimates become suspect and non-supportable.

Non-Randomness Due to Autocorrelation

One specific and common type of non-randomness is autocorrelation. Autocorrelation is the correlation between Y_t and Y_{t-k}, where k is an integer that defines the lag for the autocorrelation. That is, autocorrelation is a time dependent non-randomness. This means that the value of the current point is highly dependent on the previous point if $k = 1$ (or k points ago if k is not 1). Autocorrelation is typically detected via an autocorrelation plot or a lag plot.

If the data are not random due to autocorrelation, then

1. Adjacent data values may be related.
2. There may not be n independent snapshots of the phenomenon under study.
3. There may be undetected "junk"-outliers.
4. There may be undetected "information-rich"-outliers.

1.2.5.2. Consequences of Non-Fixed Location Parameter

Location Estimate

The usual estimate of location is the mean

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

from N measurements Y_1, Y_2, ... , Y_N.

Consequences of Non-Fixed Location

If the run sequence plot does not support the assumption of fixed location, then

1. The location may be drifting.
2. The single location estimate may be meaningless (if the process is drifting).
3. The choice of location estimator (e.g., the sample mean) may be sub-optimal.
4. The usual formula for the uncertainty of the mean:

$$s(\bar{Y}) = \frac{1}{\sqrt{N(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

   may be invalid and the numerical value optimistically small.
5. The location estimate may be poor.
6. The location estimate may be biased.

1.2.5.3. Consequences of Non-Fixed Variation Parameter

Variation Estimate

The usual estimate of variation is the standard deviation

$$s_Y = \frac{1}{\sqrt{(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

from N measurements $Y_1, Y_2, \ldots, Y_N$.

Consequences of Non-Fixed Variation

If the run sequence plot does not support the assumption of fixed variation, then

1. The variation may be drifting.
2. The single variation estimate may be meaningless (if the process variation is drifting).
3. The variation estimate may be poor.
4. The variation estimate may be biased.

1. Exploratory Data Analysis
1.2. EDA Assumptions
1.2.5. Consequences

1.2.5.4. Consequences Related to Distributional Assumptions

Distributional Analysis

Scientists and engineers routinely use the mean (average) to estimate the "middle" of a distribution. It is not so well known that the variability and the noisiness of the mean as a location estimator are intrinsically linked with the underlying distribution of the data. For certain distributions, the mean is a poor choice. For any given distribution, there exists an optimal choice-- that is, the estimator with minimum variability/noisiness. This optimal choice may be, for example, the median, the midrange, the midmean, the mean, or something else. The implication of this is to "estimate" the distribution first, and then--based on the distribution--choose the optimal estimator. The resulting engineering parameter estimators will have less variability than if this approach is not followed.

Case Studies

The airplane glass failure case study gives an example of determining an appropriate distribution and estimating the parameters of that distribution. The uniform random numbers case study gives an example of determining a more appropriate centrality parameter for a non-normal distribution.

Other consequences that flow from problems with distributional assumptions are:

Distribution

1. The distribution may be changing.
2. The single distribution estimate may be meaningless (if the process distribution is changing).
3. The distribution may be markedly non-normal.
4. The distribution may be unknown.
5. The true probability distribution for the error may remain unknown.

Model

1. The model may be changing.
2. The single model estimate may be meaningless.
3. The default model
   $$Y = \text{constant} + \text{error}$$
   may be invalid.
4. If the default model is insufficient, information about

a better model may remain undetected.
5. A poor deterministic model may be fit.
6. Information about an improved model may go undetected.

Process

1. The process may be out-of-control.
2. The process may be unpredictable.
3. The process may be un-modelable.

1. Exploratory Data Analysis

1.3. EDA Techniques

Summary

After you have collected a set of data, how do you do an exploratory data analysis? What techniques do you employ? What do the various techniques focus on? What conclusions can you expect to reach?

This section provides answers to these kinds of questions via a gallery of EDA techniques and a detailed description of each technique. The techniques are divided into graphical and quantitative techniques. For exploratory data analysis, the emphasis is primarily on the graphical techniques.

Table of Contents for Section 3

1. Introduction
2. Analysis Questions
3. Graphical Techniques: Alphabetical
4. Graphical Techniques: By Problem Category
5. Quantitative Techniques: Alphabetical
6. Probability Distributions

1.3.1. Introduction

Graphical and Quantitative Techniques

This section describes many techniques that are commonly used in exploratory and classical data analysis. This list is by no means meant to be exhaustive. Additional techniques (both graphical and quantitative) are discussed in the other chapters. Specifically, the product comparisons chapter has a much more detailed description of many classical statistical techniques.

EDA emphasizes graphical techniques while classical techniques emphasize quantitative techniques. In practice, an analyst typically uses a mixture of graphical and quantitative techniques. In this section, we have divided the descriptions into graphical and quantitative techniques. This is for organizational clarity and is not meant to discourage the use of both graphical and quantitiative techniques when analyzing data.

Use of Techniques Shown in Case Studies

This section emphasizes the techniques themselves; how the graph or test is defined, published references, and sample output. The use of the techniques to answer engineering questions is demonstrated in the case studies section. The case studies do not demonstrate all of the techniques.

Availability in Software

The sample plots and output in this section were generated with the Dataplot software program. Other general purpose statistical data analysis programs can generate most of the plots, intervals, and tests discussed here, or macros can be written to acheive the same result.

1.3.2. Analysis Questions

EDA Questions

Some common questions that exploratory data analysis is used to answer are:

1. What is a typical value?
2. What is the uncertainty for a typical value?
3. What is a good distributional fit for a set of numbers?
4. What is a percentile?
5. Does an engineering modification have an effect?
6. Does a factor have an effect?
7. What are the most important factors?
8. Are measurements coming from different laboratories equivalent?
9. What is the best function for relating a response variable to a set of factor variables?
10. What are the best settings for factors?
11. Can we separate signal from noise in time dependent data?
12. Can we extract any structure from multivariate data?
13. Does the data have outliers?

Analyst Should Identify Relevant Questions for his Engineering Problem

A critical early step in any analysis is to identify (for the engineering problem at hand) which of the above questions are relevant. That is, we need to identify which questions we want answered and which questions have no bearing on the problem at hand. After collecting such a set of questions, an equally important step, which is invaluable for maintaining focus, is to prioritize those questions in decreasing order of importance. EDA techniques are tied in with each of the questions. There are some EDA techniques (e.g., the scatter plot) that are broad-brushed and apply almost universally. On the other hand, there are a large number of EDA techniques that are specific and whose specificity is tied in with one of the above questions. Clearly if one chooses not to explicitly identify relevant questions, then one cannot take advantage of these question-specific EDA technqiues.

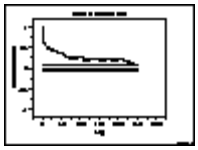
EDA Approach Emphasizes Graphics

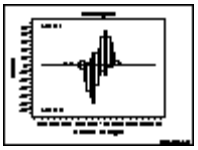
Most of these questions can be addressed by techniques discussed in this chapter. The process modeling and process improvement chapters also address many of the questions above. These questions are also relevant for the classical approach to statistics. What distinguishes the EDA approach is an emphasis on graphical techniques to gain insight as

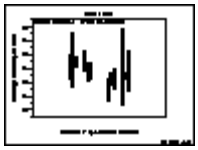
opposed to the classical approach of quantitative tests. Most data analysts will use a mix of graphical and classical quantitative techniques to address these problems.

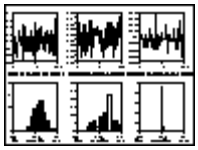
1.3.3. Graphical Techniques: Alphabetic

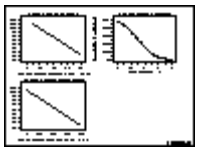
This section provides a gallery of some useful graphical techniques. The techniques are ordered alphabetically, so this section is not intended to be read in a sequential fashion. The use of most of these graphical techniques is demonstrated in the case studies in this chapter. A few of these graphical techniques are demonstrated in later chapters.

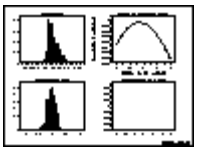

Autocorrelation Plot: 1.3.3.1

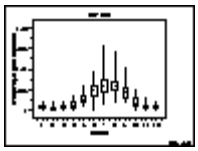

Bihistogram: 1.3.3.2

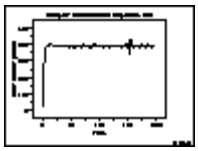

Block Plot: 1.3.3.3

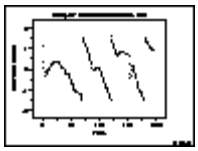

Bootstrap Plot: 1.3.3.4

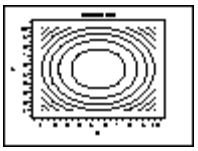

Box-Cox Linearity Plot: 1.3.3.5

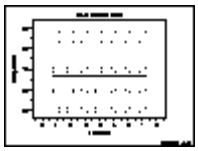

Box-Cox Normality Plot: 1.3.3.6

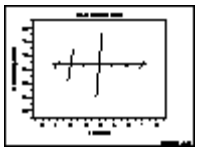

Box Plot: 1.3.3.7

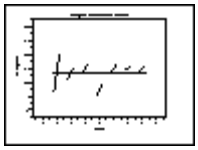

Complex Demodulation Amplitude Plot: 1.3.3.8

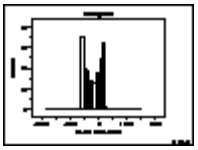

Complex Demodulation Phase Plot: 1.3.3.9

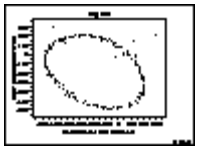

Contour Plot: 1.3.3.10

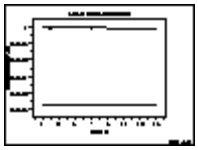

DOE Scatter Plot: 1.3.3.11

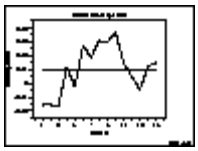

DOE Mean Plot: 1.3.3.12

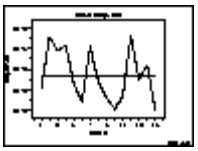

DOE Standard Deviation Plot: 1.3.3.13

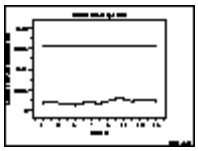

Histogram: 1.3.3.14

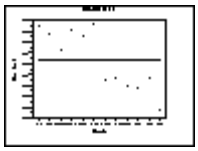

Lag Plot: 1.3.3.15

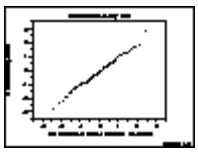

Linear Correlation Plot: 1.3.3.16

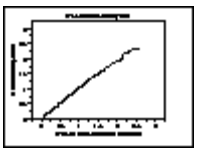

Linear Intercept Plot: 1.3.3.17

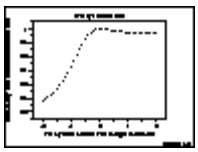

Linear Slope Plot: 1.3.3.18

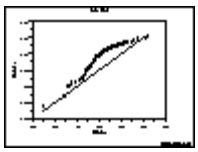

Linear Residual Standard Deviation Plot: 1.3.3.19

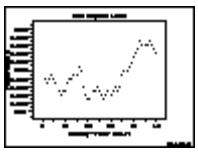

Mean Plot: 1.3.3.20

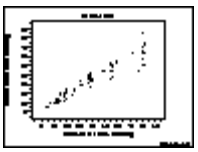

Normal Probability Plot: 1.3.3.21

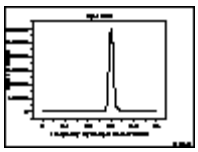

Probability Plot: 1.3.3.22

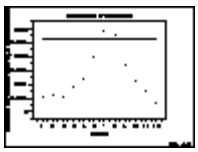

Probability Plot Correlation Coefficient Plot: 1.3.3.23

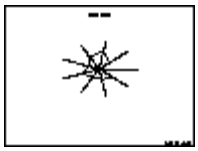

Quantile-Quantile Plot: 1.3.3.24

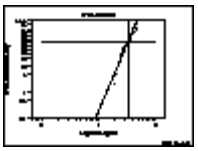

Run Sequence Plot: 1.3.3.25

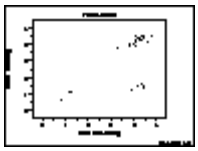

Scatter Plot: 1.3.3.26

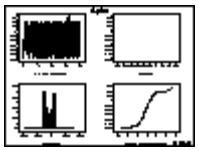

Spectrum: 1.3.3.27

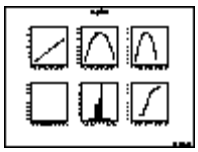

Standard Deviation Plot: 1.3.3.28


Star Plot: 1.3.3.29


Weibull Plot: 1.3.3.30


Youden Plot: 1.3.3.31


4-Plot: 1.3.3.32


6-Plot: 1.3.3.33

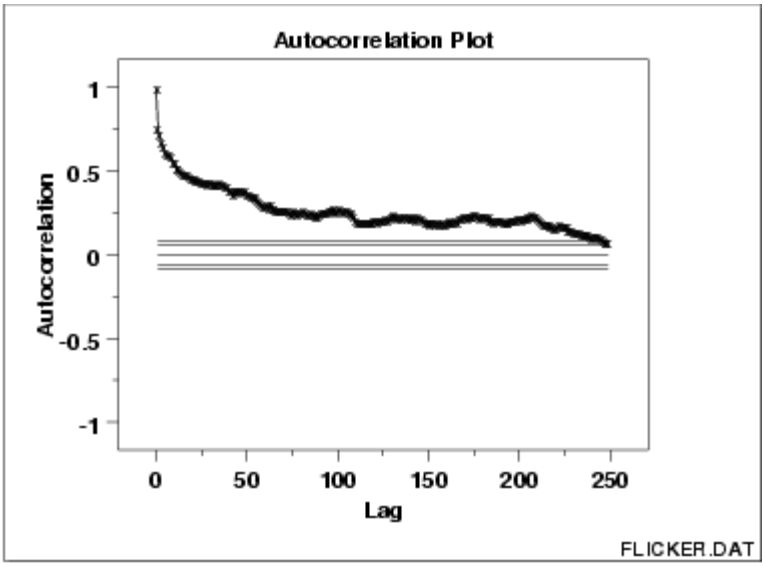
1.3.3.1. Autocorrelation Plot

Purpose: Check Randomness

Autocorrelation plots (Box and Jenkins, pp. 28-32) are a commonly-used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

In addition, autocorrelation plots are used in the model identification stage for Box-Jenkins autoregressive, moving average time series models.

Sample Plot: Autocorrelations should be near-zero for randomness. Such is not the case in this example and thus the randomness assumption fails



This sample autocorrelation plot shows that the time series is not random, but rather has a high degree of autocorrelation between adjacent and near-adjacent observations.

Definition: r(h) versus h

Autocorrelation plots are formed by

- Vertical axis: Autocorrelation coefficient

$$R_h = C_h/C_0$$

where C_h is the autocovariance function

$$C_h = \frac{1}{N}\sum_{t=1}^{N-h}(Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

and C_0 is the variance function

$$C_0 = \frac{\sum_{t=1}^{N}(Y_t - \bar{Y})^2}{N}$$

Note--R_h is between -1 and +1.

Note--Some sources may use the following formula for the autocovariance function

$$C_h = \frac{1}{N-h}\sum_{t=1}^{N-h}(Y_t - \bar{Y})(Y_{t+h} - \bar{Y})$$

Although this definition has less bias, the (1/*N*) formulation has some desirable statistical properties and is the form most commonly used in the statistics literature. See pages 20 and 49-50 in Chatfield for details.

- Horizontal axis: Time lag *h* (*h* = 1, 2, 3, ...)
- The above line also contains several horizontal reference lines. The middle line is at zero. The other four lines are 95 % and 99 % confidence bands. Note that there are two distinct formulas for generating the confidence bands.

    1. If the autocorrelation plot is being used to test for randomness (i.e., there is no time dependence in the data), the following formula is recommended:

        $$\pm\frac{z_{1-\alpha/2}}{\sqrt{N}}$$

        where *N* is the sample size, *z* is the cumulative distribution function of the standard normal distribution and α is the significance level. In this case, the confidence bands have fixed width that depends on the sample size. This is the formula that was used to generate the confidence bands in the above plot.

    2. Autocorrelation plots are also used in the model identification stage for fitting ARIMA models. In this case, a moving average model is assumed for the data and the following confidence bands should be generated:

$$\pm z_{1-\alpha/2}\sqrt{\frac{1}{N}(1+2\sum_{i=1}^{k}y_i^2)}$$

where *k* is the lag, *N* is the sample size, *z* is the cumulative distribution function of the standard normal distribution and α is the significance level. In this case, the confidence bands increase as the lag increases.

Questions

The autocorrelation plot can provide answers to the following questions:

1. Are the data random?
2. Is an observation related to an adjacent observation?
3. Is an observation related to an observation twice-removed? (etc.)
4. Is the observed time series white noise?
5. Is the observed time series sinusoidal?
6. Is the observed time series autoregressive?
7. What is an appropriate model for the observed time series?
8. Is the model

   Y = constant + error

   valid and sufficient?

9. Is the formula $s_{\bar{Y}} = s/\sqrt{N}$ valid?

Importance: Ensure validity of engineering conclusions

Randomness (along with fixed model, fixed variation, and fixed distribution) is one of the four assumptions that typically underlie all measurement processes. The randomness assumption is critically important for the following three reasons:

1. Most standard statistical tests depend on randomness. The validity of the test conclusions is directly linked to the validity of the randomness assumption.

2. Many commonly-used statistical formulae depend on the randomness assumption, the most common formula being the formula for determining the standard deviation of the sample mean:

   $$s_{\bar{Y}} = s/\sqrt{N}$$

   where s is the standard deviation of the data. Although heavily used, the results from using this formula are of no value unless the randomness

assumption holds.

3. For univariate data, the default model is

   Y = constant + error

   If the data are not random, this model is incorrect and invalid, and the estimates for the parameters (such as the constant) become nonsensical and invalid.

In short, if the analyst does not check for randomness, then the validity of many of the statistical conclusions becomes suspect. The autocorrelation plot is an excellent way of checking for such randomness.

Examples

Examples of the autocorrelation plot for several common situations are given in the following pages.

1. Random (= White Noise)
2. Weak autocorrelation
3. Strong autocorrelation and autoregressive model
4. Sinusoidal model

Related Techniques

Partial Autocorrelation Plot
Lag Plot
Spectral Plot
Seasonal Subseries Plot

Case Study

The autocorrelation plot is demonstrated in the beam deflection data case study.

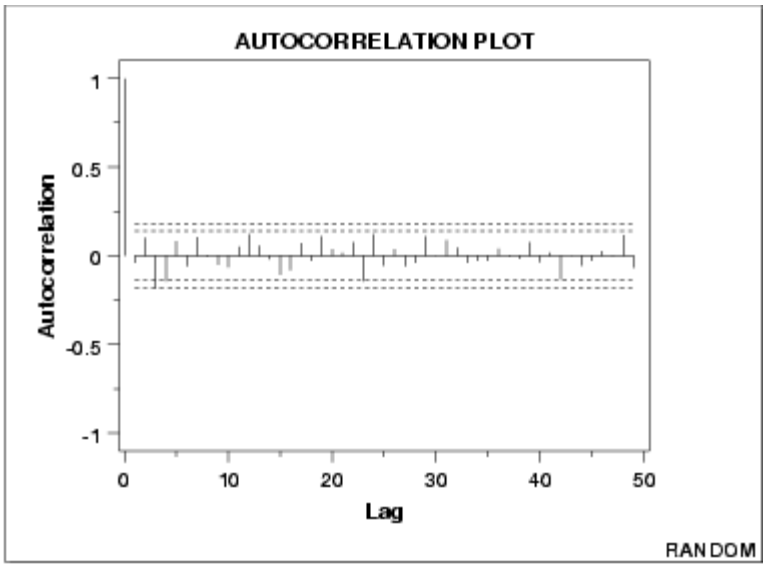
Software

Autocorrelation plots are available in most general purpose statistical software programs.

1.3.3.1.1. Autocorrelation Plot: Random Data

Autocorrelation Plot

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from this plot.

1. There are no significant autocorrelations.
2. The data are random.

Discussion

Note that with the exception of lag 0, which is always 1 by definition, almost all of the autocorrelations fall within the 95% confidence limits. In addition, there is no apparent pattern (such as the first twenty-five being positive and the second twenty-five being negative). This is the abscence of a pattern we expect to see if the data are in fact random.

A few lags slightly outside the 95% and 99% confidence limits do not neccessarily indicate non-randomness. For a 95% confidence interval, we might expect about one out of twenty lags to be statistically significant due to random fluctuations.

There is no associative ability to infer from a current value Y_i as to what the next value Y_{i+1} will be. Such non-association is the essense of randomness. In short, adjacent

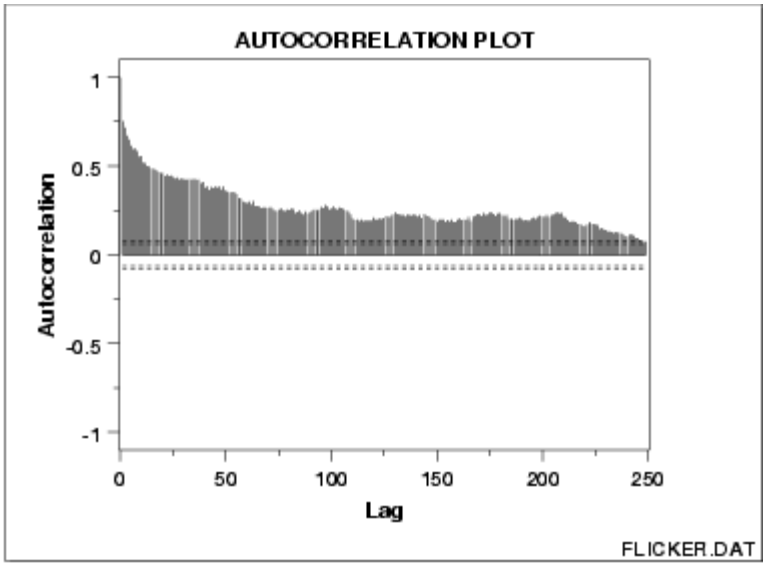
observations do not "co-relate", so we call this the "no autocorrelation" case.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.3.1.2. Autocorrelation Plot: Moderate Autocorrelation

Autocorrelation Plot

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from this plot.

1. The data come from an underlying autoregressive model with moderate positive autocorrelation.

Discussion

The plot starts with a moderately high autocorrelation at lag 1 (approximately 0.75) that gradually decreases. The decreasing autocorrelation is generally linear, but with significant noise. Such a pattern is the autocorrelation plot signature of "moderate autocorrelation", which in turn provides moderate predictability if modeled properly.

Recommended Next Step

The next step would be to estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be performed by using least squares linear regression or by fitting a Box-Jenkins autoregressive (AR) model.

The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit non-randomness, the residuals after fitting $\mathbf{Y}_i$ against $\mathbf{Y}_{i-1}$ should result in random residuals. Assessing whether or not the proposed model in fact sufficiently removed the randomness is discussed in detail in the Process Modeling chapter.

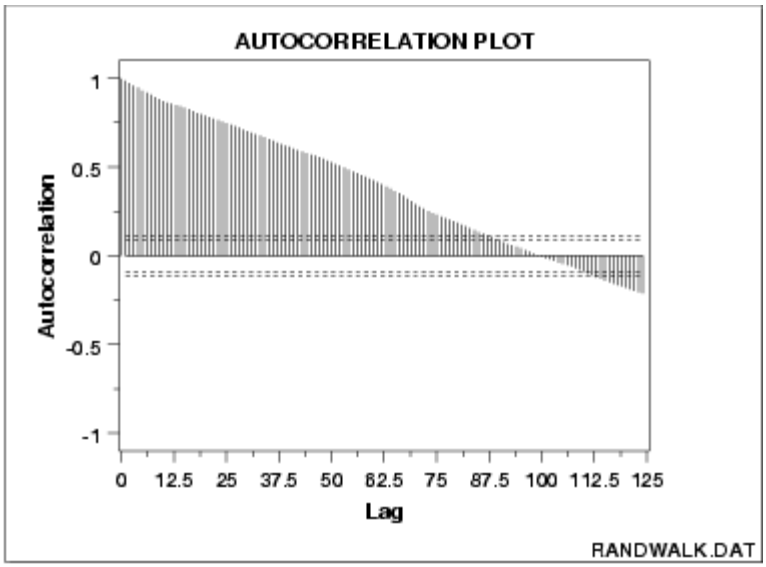
The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1.3.3.1.3. Autocorrelation Plot: Strong Autocorrelation and Autoregressive Model

Autocorrelation Plot for Strong Autocorrelation

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from the above plot.

1. The data come from an underlying autoregressive model with strong positive autocorrelation.

Discussion

The plot starts with a high autocorrelation at lag 1 (only slightly less than 1) that slowly declines. It continues decreasing until it becomes negative and starts showing an incresing negative autocorrelation. The decreasing autocorrelation is generally linear with little noise. Such a pattern is the autocorrelation plot signature of "strong autocorrelation", which in turn provides high predictability if modeled properly.

Recommended Next Step

The next step would be to estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be performed by using least squares linear regression or by fitting a Box-Jenkins autoregressive (AR) model.

The randomness assumption for least squares fitting applies to the residuals of the model. That is, even though the original data exhibit non-randomness, the residuals after fitting $\mathbf{Y_i}$ against $\mathbf{Y_{i-1}}$ should result in random residuals. Assessing whether or not the proposed model in fact sufficiently removed the randomness is discussed in detail in the Process Modeling chapter.

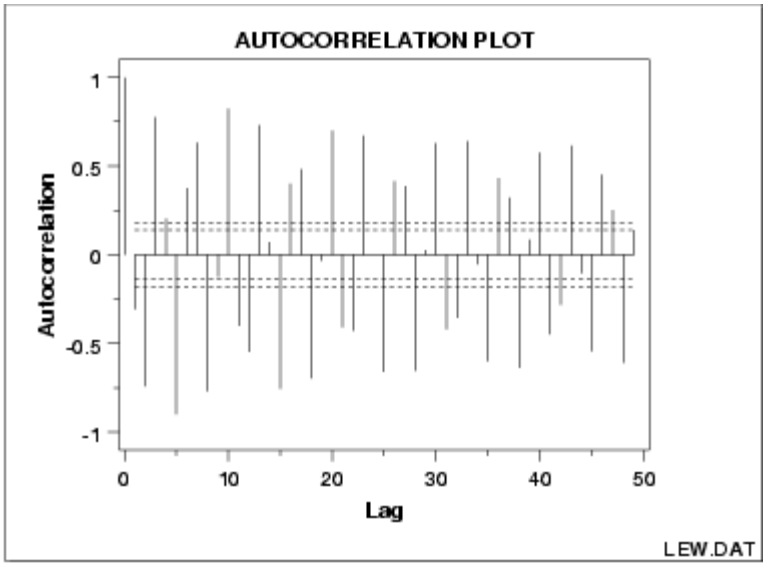
The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1.3.3.1.4. Autocorrelation Plot: Sinusoidal Model

Autocorrelation Plot for Sinusoidal Model

The following is a sample autocorrelation plot.



Conclusions

We can make the following conclusions from the above plot.

1. The data come from an underlying sinusoidal model.

Discussion

The plot exhibits an alternating sequence of positive and negative spikes. These spikes are not decaying to zero. Such a pattern is the autocorrelation plot signature of a sinusoidal model.

Recommended Next Step

The beam deflection case study gives an example of modeling a sinusoidal model.

1.3.3.2. Bihistogram

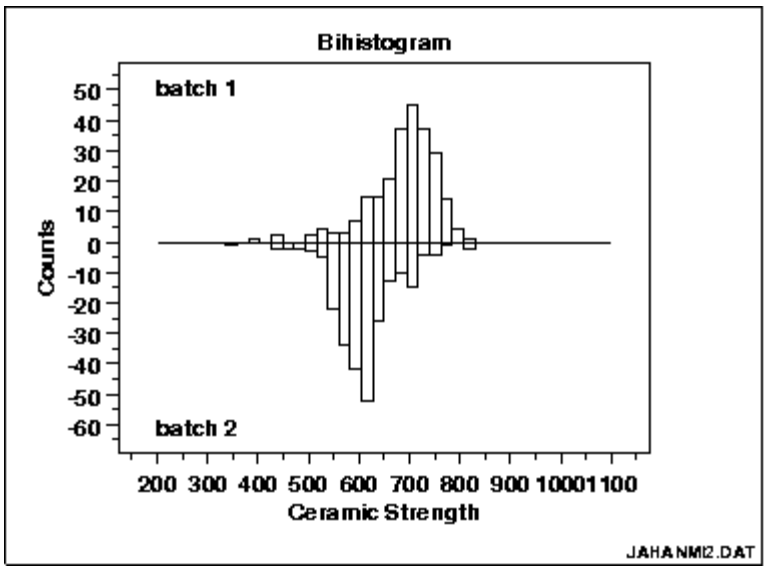
Purpose: Check for a change in location, variation, or distribution

The bihistogram is an EDA tool for assessing whether a before-versus-after engineering modification has caused a change in

- location;
- variation; or
- distribution.

It is a graphical alternative to the two-sample t-test. The bihistogram can be more powerful than the t-test in that all of the distributional features (location, scale, skewness, outliers) are evident on a single plot. It is also based on the common and well-understood histogram.

Sample Plot: This bihistogram reveals that there is a significant difference in ceramic breaking strength between batch 1 (above) and batch 2 (below)



From the above bihistogram, we can see that batch 1 is centered at a ceramic strength value of approximately 725 while batch 2 is centered at a ceramic strength value of approximately 625. That indicates that these batches are displaced by about 100 strength units. Thus the batch factor has a significant effect on the location (typical value) for strength and hence batch is said to be "significant" or to "have an effect". We thus see graphically and convincingly what a t-test or analysis of variance would indicate quantitatively.

With respect to variation, note that the spread (variation) of the above-axis batch 1 histogram does not appear to be that much different from the below-axis batch 2 histogram. With respect to distributional shape, note that the batch 1 histogram is skewed left while the batch 2 histogram is more symmetric with even a hint of a slight skewness to the right.

Thus the bihistogram reveals that there is a clear difference between the batches with respect to location and distribution, but not in regard to variation. Comparing batch 1 and batch 2, we also note that batch 1 is the "better batch" due to its 100-unit higher average strength (around 725).

Definition: Two adjoined histograms

Bihistograms are formed by vertically juxtaposing two histograms:

- Above the axis: Histogram of the response variable for condition 1
- Below the axis: Histogram of the response variable for condition 2

Questions

The bihistogram can provide answers to the following questions:

1. Is a (2-level) factor significant?
2. Does a (2-level) factor have an effect?
3. Does the location change between the 2 subgroups?
4. Does the variation change between the 2 subgroups?
5. Does the distributional shape change between subgroups?
6. Are there any outliers?

Importance: Checks 3 out of the 4 underlying assumptions of a measurement process

The bihistogram is an important EDA tool for determining if a factor "has an effect". Since the bihistogram provides insight into the validity of three (location, variation, and distribution) out of the four (missing only randomness) underlying assumptions in a measurement process, it is an especially valuable tool. Because of the dual (above/below) nature of the plot, the bihistogram is restricted to assessing factors that have only two levels. However, this is very common in the before-versus-after character of many scientific and engineering experiments.

Related Techniques

t test (for shift in location)
F test (for shift in variation)
Kolmogorov-Smirnov test (for shift in distribution)
Quantile-quantile plot (for shift in location and distribution)

Case Study

The bihistogram is demonstrated in the ceramic strength data case study.

Software

The bihistogram is not widely available in general purpose

statistical software programs. Bihistograms can be generated using Dataplot and R software.

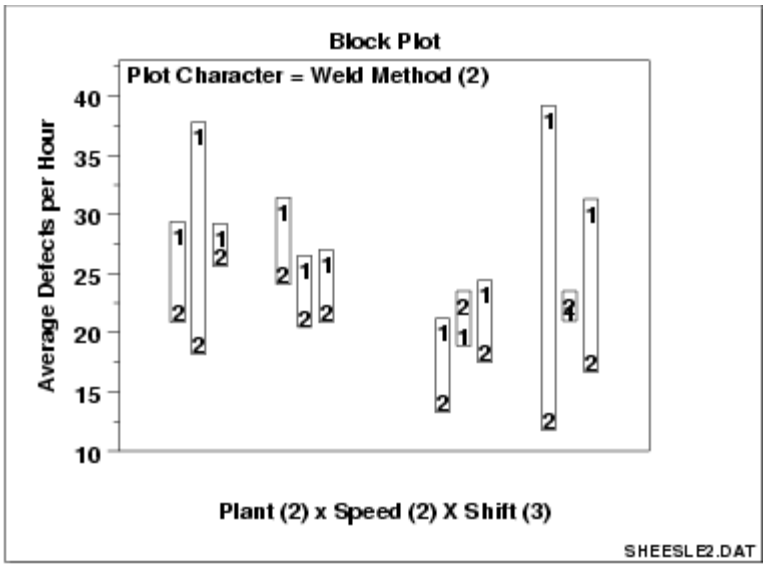
1.3.3.3. Block Plot

Purpose: Check to determine if a factor of interest has an effect robust over all other factors

The block plot (Filliben 1993) is an EDA tool for assessing whether the factor of interest (the primary factor) has a statistically significant effect on the response, and whether that conclusion about the primary factor effect is valid robustly over all other nuisance or secondary factors in the experiment.

It replaces the analysis of variance test with a less assumption-dependent binomial test and should be routinely used whenever we are trying to robustly decide whether a primary factor has an effect.

Sample Plot: Weld method 2 is lower (better) than weld method 1 in 10 of 12 cases



This block plot reveals that in 10 of the 12 cases (bars), weld method 2 is lower (better) than weld method 1. From a binomial point of view, weld method is statistically significant.

Definition

Block Plots are formed as follows:

- Vertical axis: Response variable Y
- Horizontal axis: All combinations of all levels of all nuisance (secondary) factors X1, X2, ...
- Plot Character: Levels of the primary factor XP

Discussion: Primary factor is denoted by plot character: within-bar plot character.

Average number of defective lead wires per hour from a study with four factors,

1. weld strength (2 levels)
2. plant (2 levels)
3. speed (2 levels)
4. shift (3 levels)

are shown in the plot above. Weld strength is the primary factor and the other three factors are nuisance factors. The 12 distinct positions along the horizontal axis correspond to all possible combinations of the three nuisance factors, i.e., 12 = 2 plants x 2 speeds x 3 shifts. These 12 conditions provide the framework for assessing whether any conclusions about the 2 levels of the primary factor (weld method) can truly be called "general conclusions". If we find that one weld method setting does better (smaller average defects per hour) than the other weld method setting for all or most of these 12 nuisance factor combinations, then the conclusion is in fact general and robust.

Ordering along the horizontal axis

In the above chart, the ordering along the horizontal axis is as follows:

- The left 6 bars are from plant 1 and the right 6 bars are from plant 2.
- The first 3 bars are from speed 1, the next 3 bars are from speed 2, the next 3 bars are from speed 1, and the last 3 bars are from speed 2.
- Bars 1, 4, 7, and 10 are from the first shift, bars 2, 5, 8, and 11 are from the second shift, and bars 3, 6, 9, and 12 are from the third shift.

Setting 2 is better than setting 1 in 10 out of 12 cases

In the block plot for the first bar (plant 1, speed 1, shift 1), weld method 1 yields about 28 defects per hour while weld method 2 yields about 22 defects per hour--hence the difference for this combination is about 6 defects per hour and weld method 2 is seen to be better (smaller number of defects per hour).

Is "weld method 2 is better than weld method 1" a general conclusion?

For the second bar (plant 1, speed 1, shift 2), weld method 1 is about 37 while weld method 2 is only about 18. Thus weld method 2 is again seen to be better than weld method 1. Similarly for bar 3 (plant 1, speed 1, shift 3), we see weld method 2 is smaller than weld method 1. Scanning over all of the 12 bars, we see that weld method 2 is smaller than weld method 1 in 10 of the 12 cases, which is highly suggestive of a robust weld method effect.

An event

What is the chance of 10 out of 12 happening by chance?

with chance probability of only 2%

This is probabilistically equivalent to testing whether a coin is fair by flipping it and getting 10 heads in 12 tosses. The chance (from the binomial distribution) of getting 10 (or more extreme: 11, 12) heads in 12 flips of a fair coin is about 2%. Such low-probability events are usually rejected as untenable and in practice we would conclude that there is a difference in weld methods.

Advantage: Graphical and binomial

The advantages of the block plot are as follows:

- A quantitative procedure (analysis of variance) is replaced by a graphical procedure.
- An F-test (analysis of variance) is replaced with a binomial test, which requires fewer assumptions.

Questions

The block plot can provide answers to the following questions:

1. Is the factor of interest significant?
2. Does the factor of interest have an effect?
3. Does the location change between levels of the primary factor?
4. Has the process improved?
5. What is the best setting (= level) of the primary factor?
6. How much of an average improvement can we expect with this best setting of the primary factor?
7. Is there an interaction between the primary factor and one or more nuisance factors?
8. Does the effect of the primary factor change depending on the setting of some nuisance factor?
9. Are there any outliers?

Importance: Robustly checks the significance of the factor of interest

The block plot is a graphical technique that pointedly focuses on whether or not the primary factor conclusions are in fact robustly general. This question is fundamentally different from the generic multi-factor experiment question where the analyst asks, "What factors are important and what factors are not" (a screening problem)? Global data analysis techniques, such as analysis of variance, can potentially be improved by local, focused data analysis techniques that take advantage of this difference.

Related Techniques

t test (for shift in location for exactly 2 levels)
ANOVA (for shift in location for 2 or more levels)
Bihistogram (for shift in location, variation, and distribution for exactly 2 levels).

Case Study

The block plot is demonstrated in the ceramic strength data case study.

Software

Block plots are not currently available in most general

purpose statistical software programs. However they can be generated using Dataplot and, with some programming, R software.

1.3.3.4. Bootstrap Plot

Purpose: Estimate uncertainty

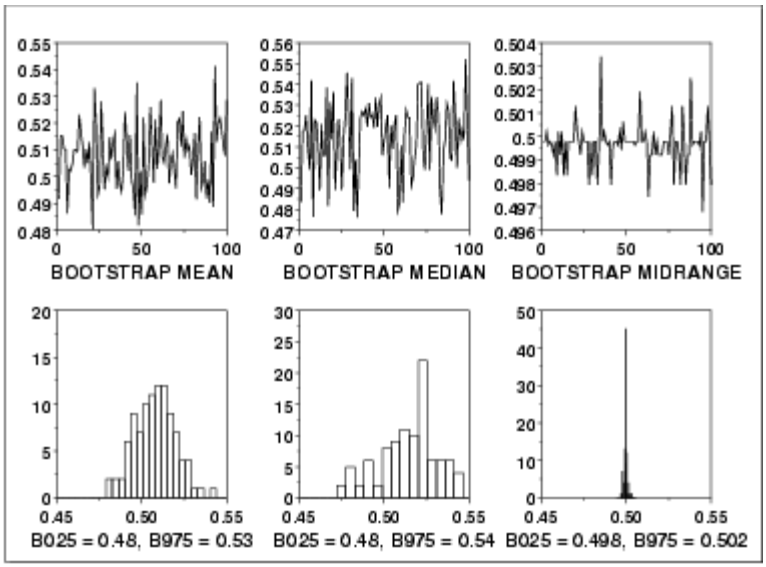
The bootstrap (Efron and Gong) plot is used to estimate the uncertainty of a statistic.

Generate subsamples with replacement

To generate a bootstrap uncertainty estimate for a given statistic from a set of data, a subsample of a size less than or equal to the size of the data set is generated from the data, and the statistic is calculated. This subsample is generated *with replacement* so that any data point can be sampled multiple times or not sampled at all. This process is repeated for many subsamples, typically between 500 and 1000. The computed values for the statistic form an estimate of the sampling distribution of the statistic.

For example, to estimate the uncertainty of the median from a dataset with 50 elements, we generate a subsample of 50 elements and calculate the median. This is repeated at least 500 times so that we have at least 500 values for the median. Although the number of bootstrap samples to use is somewhat arbitrary, 500 subsamples is usually sufficient. To calculate a 90% confidence interval for the median, the sample medians are sorted into ascending order and the value of the 25th median (assuming exactly 500 subsamples were taken) is the lower confidence limit while the value of the 475th median (assuming exactly 500 subsamples were taken) is the upper confidence limit.

Sample Plot:



This bootstrap plot was generated from 500 uniform random numbers. Bootstrap plots and corresponding histograms were generated for the mean, median, and mid-range. The histograms for the corresponding statistics clearly show that for uniform random numbers the mid-range has the smallest variance and is, therefore, a superior location estimator to the mean or the median.

Definition

The bootstrap plot is formed by:

- Vertical axis: Computed value of the desired statistic for a given subsample.
- Horizontal axis: Subsample number.

The bootstrap plot is simply the computed value of the statistic versus the subsample number. That is, the bootstrap plot generates the values for the desired statistic. This is usually immediately followed by a histogram or some other distributional plot to show the location and variation of the sampling distribution of the statistic.

Questions

The bootstrap plot is used to answer the following questions:

- What does the sampling distribution for the statistic look like?
- What is a 95% confidence interval for the statistic?
- Which statistic has a sampling distribution with the smallest variance? That is, which statistic generates the narrowest confidence interval?

Importance

The most common uncertainty calculation is generating a confidence interval for the mean. In this case, the uncertainty formula can be derived mathematically. However, there are many situations in which the uncertainty formulas are mathematically intractable. The bootstrap provides a method for calculating the uncertainty in these cases.

Cautuion on use of the bootstrap	The bootstrap is not appropriate for all distributions and statistics (Efron and Tibrashani). For example, because of the shape of the uniform distribution, the bootstrap is not appropriate for estimating the distribution of statistics that are heavily dependent on the tails, such as the range.
Related Techniques	Histogram Jackknife The jacknife is a technique that is closely related to the bootstrap. The jackknife is beyond the scope of this handbook. See the Efron and Gong article for a discussion of the jackknife.
Case Study	The bootstrap plot is demonstrated in the uniform random numbers case study.
Software	The bootstrap is becoming more common in general purpose statistical software programs. However, it is still not supported in many of these programs. Both R software and Dataplot support a bootstrap capability.

1.3.3.5. Box-Cox Linearity Plot

Purpose: Find the transformation of the X variable that maximizes the correlation between a Y and an X variable

When performing a linear fit of Y against X, an appropriate transformation of X can often significantly improve the fit. The Box-Cox transformation (Box and Cox, 1964) is a particularly useful family of transformations. It is defined as:

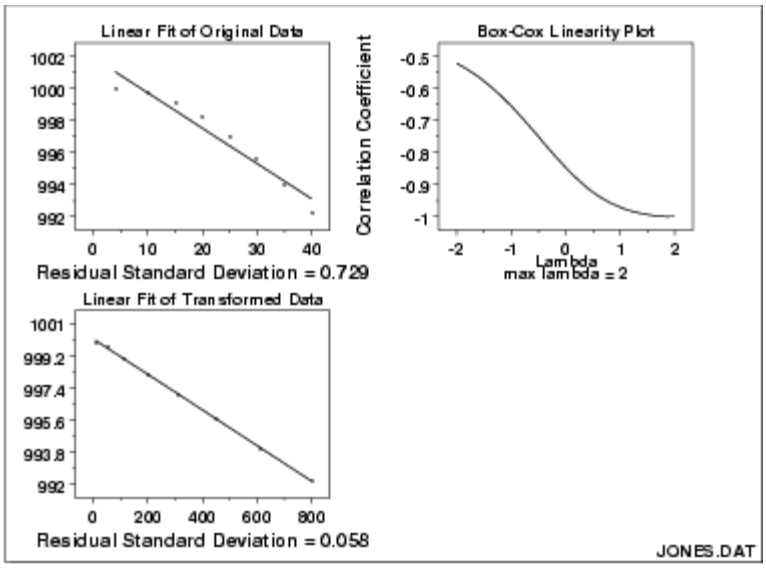
$$T(X) = (X^{\lambda} - 1)/\lambda$$

where X is the variable being transformed and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the above formula.

The Box-Cox linearity plot is a plot of the correlation between Y and the transformed X for given values of λ. That is, λ is the coordinate for the horizontal axis variable and the value of the correlation between Y and the transformed X is the coordinate for the vertical axis of the plot. The value of λ corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for λ.

Transforming X is used to improve the fit. The Box-Cox transformation applied to Y can be used as the basis for meeting the error assumptions. That case is not covered here. See page 225 of (Draper and Smith, 1981) or page 77 of (Ryan, 1997) for a discussion of this case.

Sample Plot



The plot of the original data with the predicted values from a linear fit indicate that a quadratic fit might be preferable. The Box-Cox linearity plot shows a value of $\lambda = 2.0$. The plot of the transformed data with the predicted values from a linear fit with the transformed data shows a better fit (verified by the significant reduction in the residual standard deviation).

Definition

Box-Cox linearity plots are formed by

- Vertical axis: Correlation coefficient from the transformed X and Y
- Horizontal axis: Value for λ

Questions

The Box-Cox linearity plot can provide answers to the following questions:

1. Would a suitable transformation improve my fit?
2. What is the optimal value of the transformation parameter?

Importance: Find a suitable transformation

Transformations can often significantly improve a fit. The Box-Cox linearity plot provides a convenient way to find a suitable transformation without engaging in a lot of trial and error fitting.

Related Techniques

Linear Regression
Box-Cox Normality Plot

Case Study

The Box-Cox linearity plot is demonstrated in the Alaska pipeline data case study.

Software

Box-Cox linearity plots are not a standard part of most general purpose statistical software programs. However, the underlying technique is based on a transformation and

computing a correlation coefficient. So if a statistical program supports these capabilities, writing a macro for a Box-Cox linearity plot should be feasible.

1.3.3.6. Box-Cox Normality Plot

Purpose: Find transformation to normalize data

Many statistical tests and intervals are based on the assumption of normality. The assumption of normality often leads to tests that are simple, mathematically tractable, and powerful compared to tests that do not make the normality assumption. Unfortunately, many real data sets are in fact not approximately normal. However, an appropriate transformation of a data set can often yield a data set that does follow approximately a normal distribution. This increases the applicability and usefulness of statistical techniques based on the normality assumption.

The Box-Cox transformation is a particulary useful family of transformations. It is defined as:

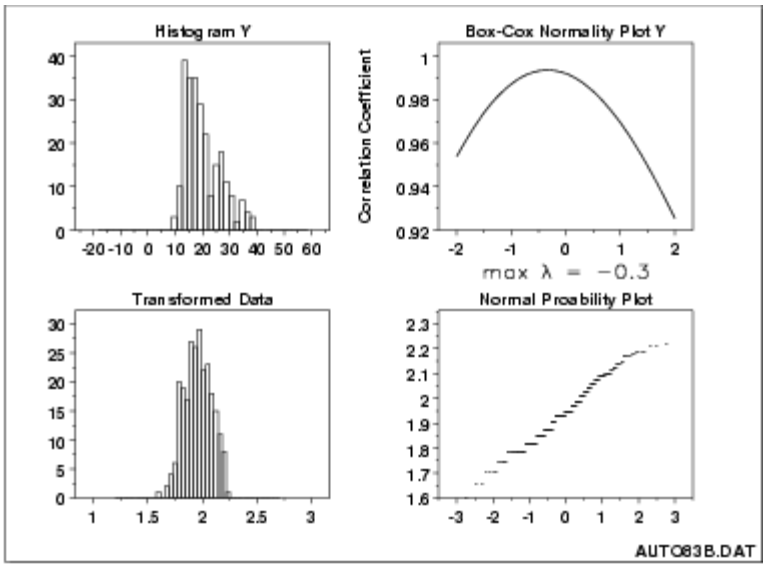
$$T(Y) = (Y^{\lambda} - 1)/\lambda$$

where Y is the response variable and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the above formula.

Given a particular transformation such as the Box-Cox transformation defined above, it is helpful to define a measure of the normality of the resulting transformation. One measure is to compute the correlation coefficient of a normal probability plot. The correlation is computed between the vertical and horizontal axis variables of the probability plot and is a convenient measure of the linearity of the probability plot (the more linear the probability plot, the better a normal distribution fits the data).

The Box-Cox normality plot is a plot of these correlation coefficients for various values of the λ parameter. The value of λ corresponding to the maximum correlation on the plot is then the optimal choice for λ.

Sample Plot



The histogram in the upper left-hand corner shows a data set that has significant right skewness (and so does not follow a normal distribution). The Box-Cox normality plot shows that the maximum value of the correlation coefficient is at λ = -0.3. The histogram of the data after applying the Box-Cox transformation with λ = -0.3 shows a data set for which the normality assumption is reasonable. This is verified with a normal probability plot of the transformed data.

Definition

Box-Cox normality plots are formed by:

- Vertical axis: Correlation coefficient from the normal probability plot after applying Box-Cox transformation
- Horizontal axis: Value for λ

Questions

The Box-Cox normality plot can provide answers to the following questions:

1. Is there a transformation that will normalize my data?
2. What is the optimal value of the transformation parameter?

Importance: Normalization Improves Validity of Tests

Normality assumptions are critical for many univariate intervals and hypothesis tests. It is important to test the normality assumption. If the data are in fact clearly not normal, the Box-Cox normality plot can often be used to find a transformation that will approximately normalize the data.

Related Techniques

Normal Probability Plot
Box-Cox Linearity Plot

Software

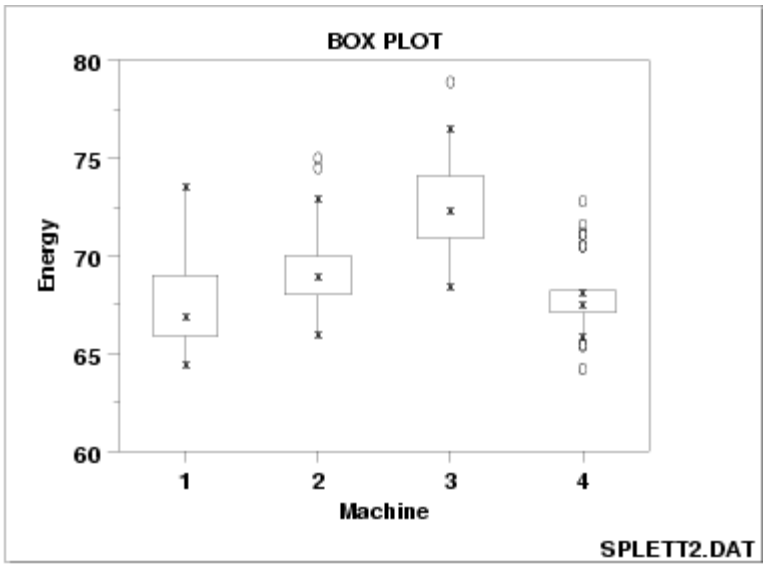
Box-Cox normality plots are not a standard part of most general purpose statistical software programs. However, the underlying technique is based on a normal probability plot and computing a correlation coefficient. So if a statistical program supports these capabilities, writing a macro for a Box-Cox normality plot should be feasible.

1.3.3.7. Box Plot

Purpose: Check location and variation shifts

Box plots (Chambers 1983) are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

Sample Plot: This box plot reveals that machine has a significant effect on energy with respect to location and possibly variation



This box plot, comparing four machines for energy output, shows that machine has a significant effect on energy with respect to both location and variation. Machine 3 has the highest energy response (about 72.5); machine 4 has the least variable energy response with about 50% of its readings being within 1 energy unit.

Definition

Box plots are formed by

Vertical axis: Response variable
Horizontal axis: The factor of interest

More specifically, we

1. Calculate the median and the quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile).

2. Plot a symbol at the median (or draw a line) and draw a box (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.

3. Draw a line from the lower quartile to the minimum point and another line from the upper quartile to the maximum point. Typically a symbol is drawn at these minimum and maximum points, although this is optional.

Thus the box plot identifies the middle 50% of the data, the median, and the extreme points.

Single or multiple box plots can be drawn

A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. For a single box plot, the width of the box is arbitrary. For multiple box plots, the width of the box plot can be set proportional to the number of points in the given group or sample (some software implementations of the box plot simply set all the boxes to the same width).

Box plots with fences

There is a useful variation of the box plot that more specifically identifies outliers. To create this variation:

1. Calculate the median and the lower and upper quartiles.

2. Plot a symbol at the median and draw a box between the lower and upper quartiles.

3. Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.

4. Calculate the following points:

   L1 = lower quartile - 1.5*IQ
   L2 = lower quartile - 3.0*IQ
   U1 = upper quartile + 1.5*IQ
   U2 = upper quartile + 3.0*IQ

5. The line from the lower quartile to the minimum is now drawn from the lower quartile to the smallest point that is greater than L1. Likewise, the line from the upper quartile to the maximum is now drawn to the largest point smaller than U1.

6. Points between L1 and L2 or between U1 and U2 are drawn as small circles. Points less than L2 or greater than U2 are drawn as large circles.

Questions

The box plot can provide answers to the following questions:

1. Is a factor significant?
2. Does the location differ between subgroups?
3. Does the variation differ between subgroups?
4. Are there any outliers?

Importance: *Check the significance of a factor*	The box plot is an important EDA tool for determining if a factor has a significant effect on the response with respect to either location or variation. The box plot is also an effective tool for summarizing large quantities of information.
Related Techniques	Mean Plot Analysis of Variance
Case Study	The box plot is demonstrated in the ceramic strength data case study.
Software	Box plots are available in most general purpose statistical software programs.

1.3.3.8. Complex Demodulation Amplitude Plot

Purpose: Detect Changing Amplitude in Sinusoidal Models

In the frequency analysis of time series models, a common model is the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is the amplitude, ϕ is the phase shift, and ω is the dominant frequency. In the above model, α and ϕ are constant, that is they do not vary with time, t_i.

The complex demodulation amplitude plot (Granger, 1964) is used to determine if the assumption of constant amplitude is justifiable. If the slope of the complex demodulation amplitude plot is not zero, then the above model is typically replaced with the model:

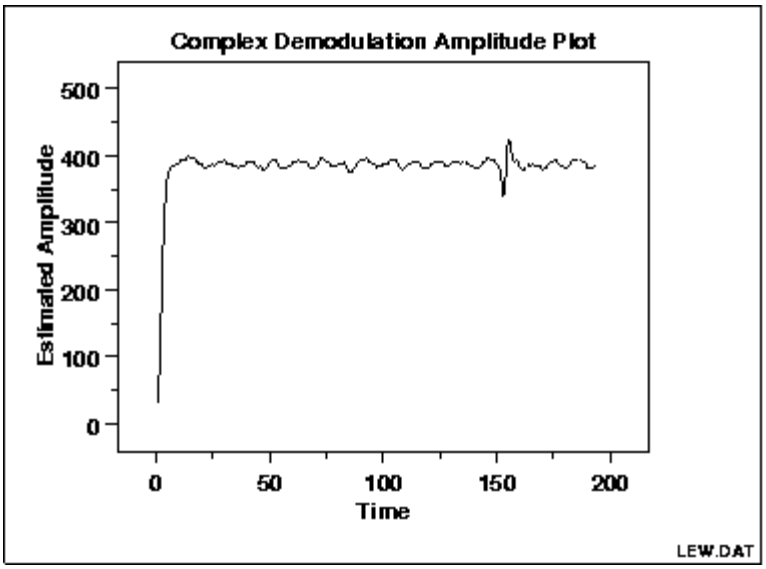
$$Y_i = C + \alpha_i \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

where $\hat{\alpha}_i$ is some type of linear model fit with standard least squares. The most common case is a linear fit, that is the model becomes

$$Y_i = C + \left(B_0 + B_1 * t_i\right) \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

Quadratic models are sometimes used. Higher order models are relatively rare.

Sample Plot:



This complex demodulation amplitude plot shows that:

- the amplitude is fixed at approximately 390;
- there is a start-up effect; and
- there is a change in amplitude at around x = 160 that should be investigated for an outlier.

Definition: The complex demodulation amplitude plot is formed by:

- Vertical axis: Amplitude
- Horizontal axis: Time

The mathematical computations for determining the amplitude are beyond the scope of the Handbook. Consult Granger (Granger, 1964) for details.

Questions The complex demodulation amplitude plot answers the following questions:

1. Does the amplitude change over time?
2. Are there any outliers that need to be investigated?
3. Is the amplitude different at the beginning of the series (i.e., is there a start-up effect)?

Importance: Assumption Checking As stated previously, in the frequency analysis of time series models, a common model is the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is assumed to be constant, that is it does not vary with time. It is important to check whether or not this assumption is reasonable.

The complex demodulation amplitude plot can be used to verify this assumption. If the slope of this plot is essentially zero, then the assumption of constant amplitude is justified. If

it is not, α should be replaced with some type of time-varying model. The most common cases are linear ($B_0 + B_1*t$) and quadratic ($B_0 + B_1*t + B_2*t^2$).

Related Techniques
Spectral Plot
Complex Demodulation Phase Plot
Non-Linear Fitting

Case Study
The complex demodulation amplitude plot is demonstrated in the beam deflection data case study.

Software
Complex demodulation amplitude plots are available in some, but not most, general purpose statistical software programs.

1.3.3.9. Complex Demodulation Phase Plot

Purpose: Improve the estimate of frequency in sinusoidal time series models

As stated previously, in the frequency analysis of time series models, a common model is the sinusoidal model:

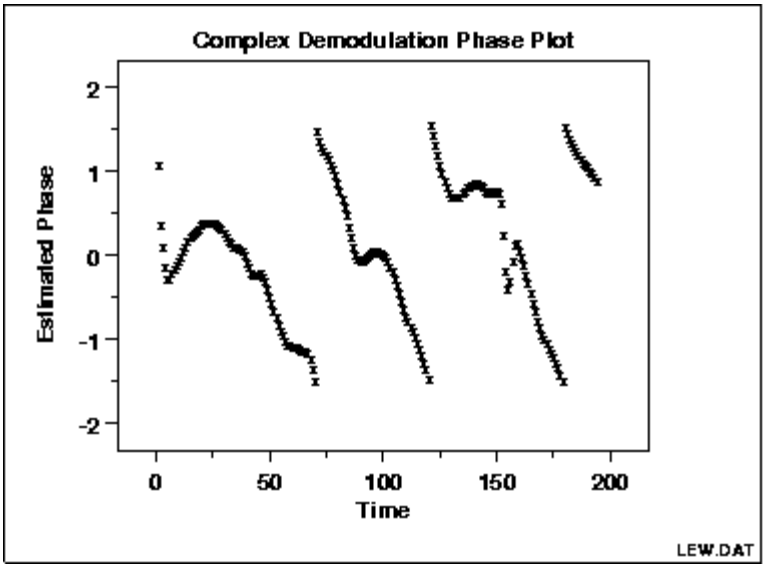
$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

In this equation, α is the amplitude, ϕ is the phase shift, and ω is the dominant frequency. In the above model, α and ϕ are constant, that is they do not vary with time t_i.

The complex demodulation phase plot (Granger, 1964) is used to improve the estimate of the frequency (i.e., ω) in this model.

If the complex demodulation phase plot shows lines sloping from left to right, then the estimate of the frequency should be increased. If it shows lines sloping right to left, then the frequency should be decreased. If there is essentially zero slope, then the frequency estimate does not need to be modified.

Sample Plot:



This complex demodulation phase plot shows that:

the specified demodulation frequency is incorrect;
- the demodulation frequency should be increased.

Definition

The complex demodulation phase plot is formed by:

- Vertical axis: Phase
- Horizontal axis: Time

The mathematical computations for the phase plot are beyond the scope of the Handbook. Consult Granger (Granger, 1964) for details.

Questions

The complex demodulation phase plot answers the following question:

Is the specified demodulation frequency correct?

Importance of a Good Initial Estimate for the Frequency

The non-linear fitting for the sinusoidal model:

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

is usually quite sensitive to the choice of good starting values. The initial estimate of the frequency, ω, is obtained from a spectral plot. The complex demodulation phase plot is used to assess whether this estimate is adequate, and if it is not, whether it should be increased or decreased. Using the complex demodulation phase plot with the spectral plot can significantly improve the quality of the non-linear fits obtained.

Related Techniques

Spectral Plot
Complex Demodulation Phase Plot
Non-Linear Fitting

Case Study

The complex demodulation amplitude plot is demonstrated in the beam deflection data case study.

Software

Complex demodulation phase plots are available in some, but not most, general purpose statistical software programs.

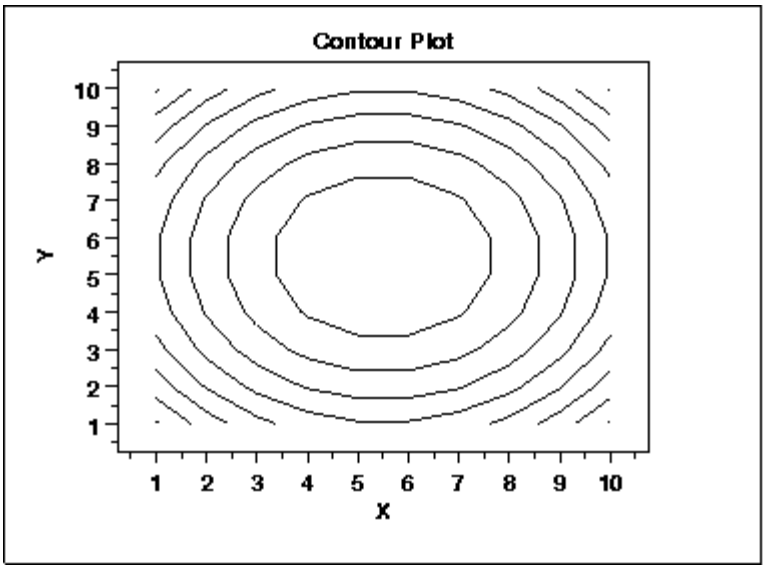
1.3.3.10. Contour Plot

Purpose: Display 3-d surface on 2-d plot

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant *z* slices, called contours, on a 2-dimensional format. That is, given a value for *z*, lines are drawn for connecting the (x,y) coordinates where that *z* value occurs.

The contour plot is an alternative to a 3-D surface plot.

Sample Plot:



This contour plot shows that the surface is symmetric and peaks in the center.

Definition

The contour plot is formed by:

- Vertical axis: Independent variable 2
- Horizontal axis: Independent variable 1
- Lines: iso-response values

The independent variables are usually restricted to a regular grid. The actual techniques for determining the correct iso-response values are rather complex and are almost always computer generated.

An additional variable may be required to specify the Z values for drawing the iso-lines. Some software packages require explicit values. Other software packages will determine them automatically.

If the data (or function) do not form a regular grid, you typically need to perform a 2-D interpolation to form a regular grid.

Questions

The contour plot is used to answer the question

How does Z change as a function of X and Y?

Importance: Visualizing 3-dimensional data

For univariate data, a run sequence plot and a histogram are considered necessary first steps in understanding the data. For 2-dimensional data, a scatter plot is a necessary first step in understanding the data.

In a similar manner, 3-dimensional data should be plotted. Small data sets, such as result from designed experiments, can typically be represented by block plots, DOE mean plots, and the like ("DOE" stands for "Design of Experiments"). For large data sets, a contour plot or a 3-D surface plot should be considered a necessary first step in understanding the data.

DOE Contour Plot

The DOE contour plot is a specialized contour plot used in the design of experiments. In particular, it is useful for full and fractional designs.

Related Techniques

3-D Plot

Software

Contour plots are available in most general purpose statistical software programs. They are also available in many general purpose graphics and mathematics programs. These programs vary widely in the capabilities for the contour plots they generate. Many provide just a basic contour plot over a rectangular grid while others permit color filled or shaded contours.

Most statistical software programs that support design of experiments will provide a DOE contour plot capability.

1.3.3.10.1. DOE Contour Plot

DOE Contour Plot: Introduction

The DOE contour plot is a specialized contour plot used in the analysis of full and fractional experimental designs. These designs often have a low level, coded as "-1" or "-", and a high level, coded as "+1" or "+" for each factor. In addition, there can optionally be one or more center points. Center points are at the mid-point between the low and high level for each factor and are coded as "0".

The DOE contour plot is generated for two factors. Typically, this would be the two most important factors as determined by previous analyses (e.g., through the use of the DOE mean plots and an analysis of variance). If more than two factors are important, you may want to generate a series of DOE contour plots, each of which is drawn for two of these factors. You can also generate a matrix of all pairwise DOE contour plots for a number of important factors (similar to the scatter plot matrix for scatter plots).

The typical application of the DOE contour plot is in determining settings that will maximize (or minimize) the response variable. It can also be helpful in determining settings that result in the response variable hitting a pre-determined target value. The DOE contour plot plays a useful role in determining the settings for the next iteration of the experiment. That is, the initial experiment is typically a fractional factorial design with a fairly large number of factors. After the most important factors are determined, the DOE contour plot can be used to help define settings for a full factorial or response surface design based on a smaller number of factors.

Construction of DOE Contour Plot

The following are the primary steps in the construction of the DOE contour plot.

1. The *x* and *y* axes of the plot represent the values of the first and second factor (independent) variables.

2. The four vertex points are drawn. The vertex points are (-1,-1), (-1,1), (1,1), (1,-1). At each vertex point, the average of all the response values at that vertex point is printed.

3. Similarly, if there are center points, a point is drawn at (0,0) and the average of the response values at the center points is printed.

4. The **linear** DOE contour plot assumes the model:

$$Y = \mu + \beta_1 \cdot U_1 + \beta_2 \cdot U_2 + \beta_{12} \cdot U_1 \cdot U_2$$

where μ is the overall mean of the response variable. The values of β_1, β_2, β_{12}, and μ are estimated from the vertex points using least squares estimation.

In order to generate a single contour line, we need a value for *Y*, say Y . Next, we

solve for U_2 in terms of U_1 and, after doing the algebra, we have the equation:

$$U_2 = \frac{(Y_0 - \mu) - \beta_1 \cdot U_1}{\beta_2 + \beta_{12} \cdot U_1}$$

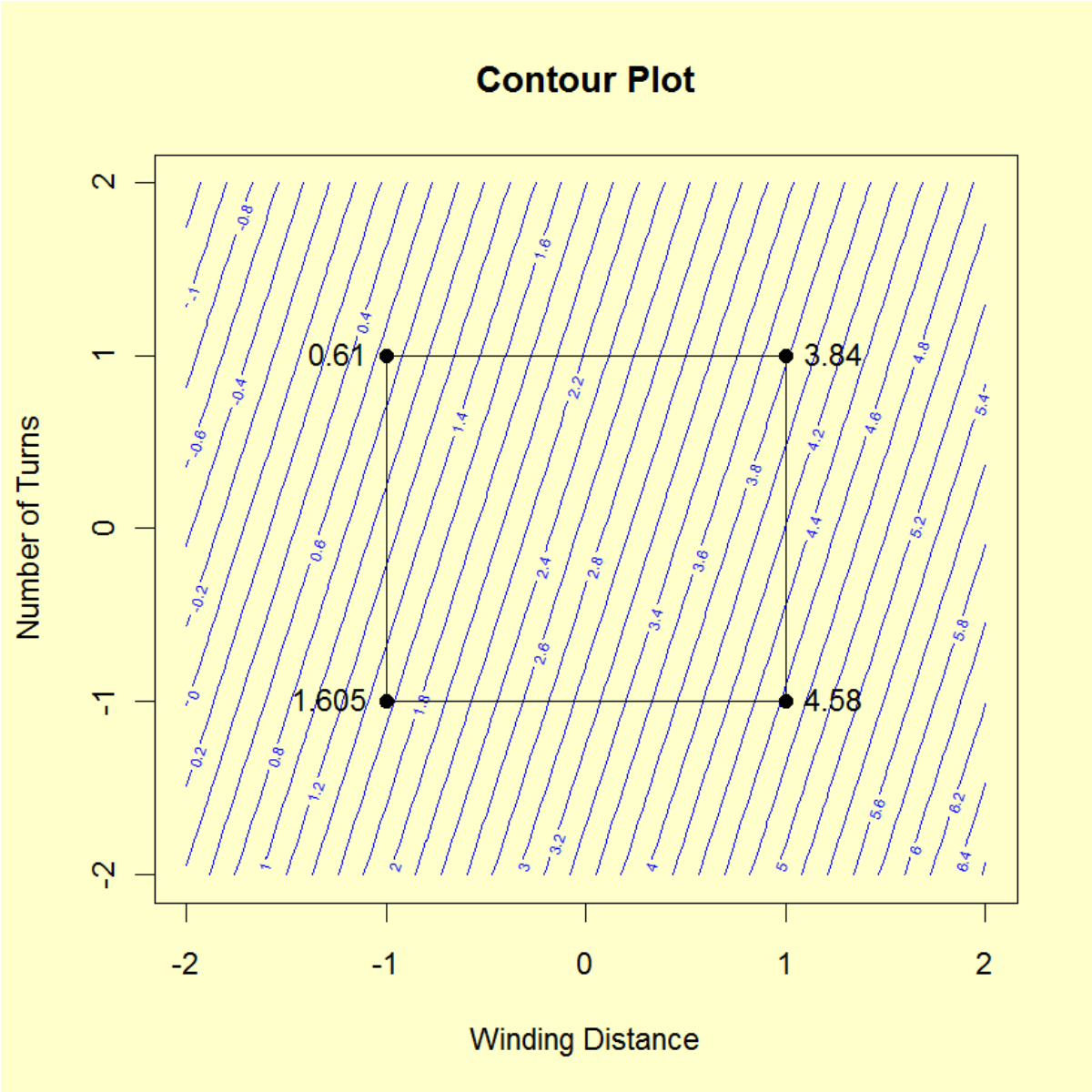
We generate a sequence of points for U_1 in the range -2 to 2 and compute the corresponding values of U_2. These points constitute a single contour line corresponding to $Y = Y_0$.

The user specifies the target values for which contour lines will be generated.

The above algorithm assumes a linear model for the design. DOE contour plots can also be generated for the case in which we assume a quadratic model for the design. The algebra for solving for U_2 in terms of U_1 becomes more complicated, but the fundamental idea is the same. Quadratic models are needed for the case when the average for the center points does not fall in the range defined by the vertex point (i.e., there is curvature).

Sample DOE Contour Plot

The following is a DOE contour plot for the data used in the Eddy current case study. The analysis in that case study demonstrated that X1 and X2 were the most important factors.



Interpretation of the Sample DOE Contour Plot

From the above DOE contour plot we can derive the following information.

1. Interaction significance;
2. Best (data) setting for these two dominant factors;

Interaction Significance

Note the appearance of the contour plot. If the contour curves are linear, then that implies that the interaction term is not significant; if the contour curves have considerable curvature, then that implies that the interaction term is large and important. In our case, the contour curves do not have considerable curvature, and so we conclude that the X1*X2 term is not significant.

Best Settings

To determine the best factor settings for the already-run experiment, we first must define what "best" means. For the Eddy current data set used to generate this DOE contour plot, "best" means to **maximize** (rather than minimize or hit a target) the response. Hence from the contour plot we determine the best settings for the two dominant factors by simply scanning the four vertices and choosing the vertex with the **largest** value (= average response). In this case, it is (X1 = +1, X2 = +1).

As for factor X3, the contour plot provides no best setting information, and so we would resort to other tools: the main effects plot, the interaction effects matrix, or the ordered data to determine optimal X3 settings.

Case Study The Eddy current case study demonstrates the use of the DOE contour plot in the context of the analysis of a full factorial design.

Software DOE Contour plots are available in many statistical software programs that analyze data from designed experiments.

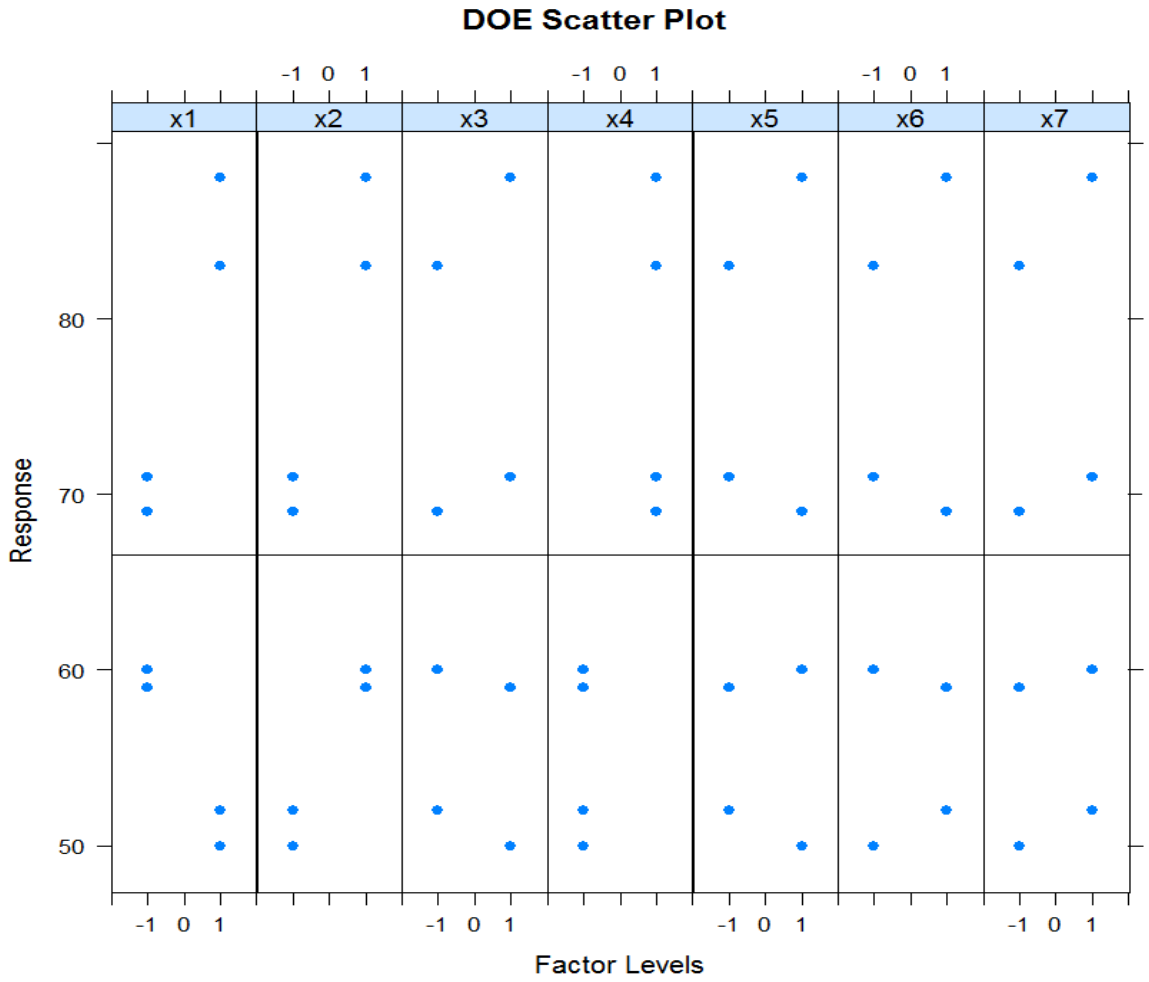
1.3.3.11. DOE Scatter Plot

Purpose: Determine Important Factors with Respect to Location and Scale

The DOE scatter plot shows the response values for each level of each factor (i.e., independent) variable. This graphically shows how the location and scale vary for both within a factor variable and between different factor variables. This graphically shows which are the important factors and can help provide a ranked list of important factors from a designed experiment. The DOE scatter plot is a complement to the traditional analyis of variance of designed experiments.

DOE scatter plots are typically used in conjunction with the DOE mean plot and the DOE standard deviation plot. The DOE mean plot replaces the raw response values with mean response values while the DOE standard deviation plot replaces the raw response values with the standard deviation of the response values. There is value in generating all 3 of these plots. The DOE mean and standard deviation plots are useful in that the summary measures of location and spread stand out (they can sometimes get lost with the raw plot). However, the raw data points can reveal subtleties, such as the presence of outliers, that might get lost with the summary statistics.

Sample Plot: Factors 4, 2, 3, and 7 are the Important Factors.



Description of the Plot

For this sample plot, there are seven factors and each factor has two levels. For each factor, we define a distinct *x* coordinate for each level of the factor. For example, for factor 1, level 1 is coded as 0.8 and level 2 is coded as 1.2. The *y* coordinate is simply the value of the response variable. The solid horizontal line is drawn at the overall mean of the response variable. The vertical dotted lines are added for clarity.

Although the plot can be drawn with an arbitrary number of levels for a factor, it is really only useful when there are two or three levels for a factor.

Conclusions

This sample DOE scatter plot shows that:

1. there does not appear to be any outliers;
2. the levels of factors 2 and 4 show distinct location differences; and
3. the levels of factor 1 show distinct scale differences.

Definition: Response Values Versus Factor Variables

DOE scatter plots are formed by:

- Vertical axis: Value of the response variable
- Horizontal axis: Factor variable (with each level of the factor coded with a slightly offset *x* coordinate)

Questions

The DOE scatter plot can be used to answer the following questions:

1. Which factors are important with respect to location and scale?
2. Are there outliers?

Importance: Identify Important Factors with Respect to Location and Scale

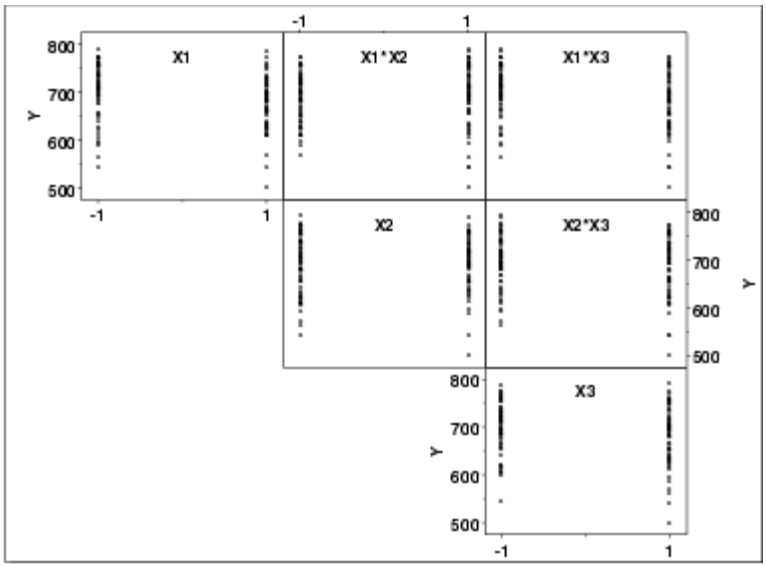
The goal of many designed experiments is to determine which factors are important with respect to location and scale. A ranked list of the important factors is also often of interest. DOE scatter, mean, and standard deviation plots show this graphically. The DOE scatter plot additionally shows if outliers may potentially be distorting the results.

DOE scatter plots were designed primarily for analyzing designed experiments. However, they are useful for any type of multi-factor data (i.e., a response variable with two or more factor variables having a small number of distinct levels) whether or not the data were generated from a designed experiment.

Extension for Interaction Effects

Using the concept of the scatterplot matrix, the DOE scatter plot can be extended to display first order interaction effects.

Specifically, if there are *k* factors, we create a matrix of plots with *k* rows and *k* columns. On the diagonal, the plot is simply a DOE scatter plot with a single factor. For the off-diagonal plots, we multiply the values of $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$. For the common 2-level designs (i.e., each factor has two levels) the values are typically coded as -1 and 1, so the multiplied values are also -1 and 1. We then generate a DOE scatter plot for this interaction variable. This plot is called a DOE interaction effects plot and an example is shown below.



Interpretation of the DOE Interaction Effects Plot

We can first examine the diagonal elements for the main effects. These diagonal plots show a great deal of overlap between the levels for all three factors. This indicates that location and scale effects will be relatively small.

We can then examine the off-diagonal plots for the first order interaction effects. For example, the plot in the first row and second column is the interaction between factors X1 and X2. As with the main effect plots, no clear patterns are evident.

Related Techniques

DOE mean plot
DOE standard deviation plot
Block plot
Box plot
Analysis of variance

Case Study The DOE scatter plot is demonstrated in the ceramic strength data case study.

Software DOE scatter plots are available in some general purpose statistical software programs, although the format may vary somewhat between these programs. They are essentially just scatter plots with the X variable defined in a particular way, so it should be feasible to write macros for DOE scatter plots in most statistical software programs.

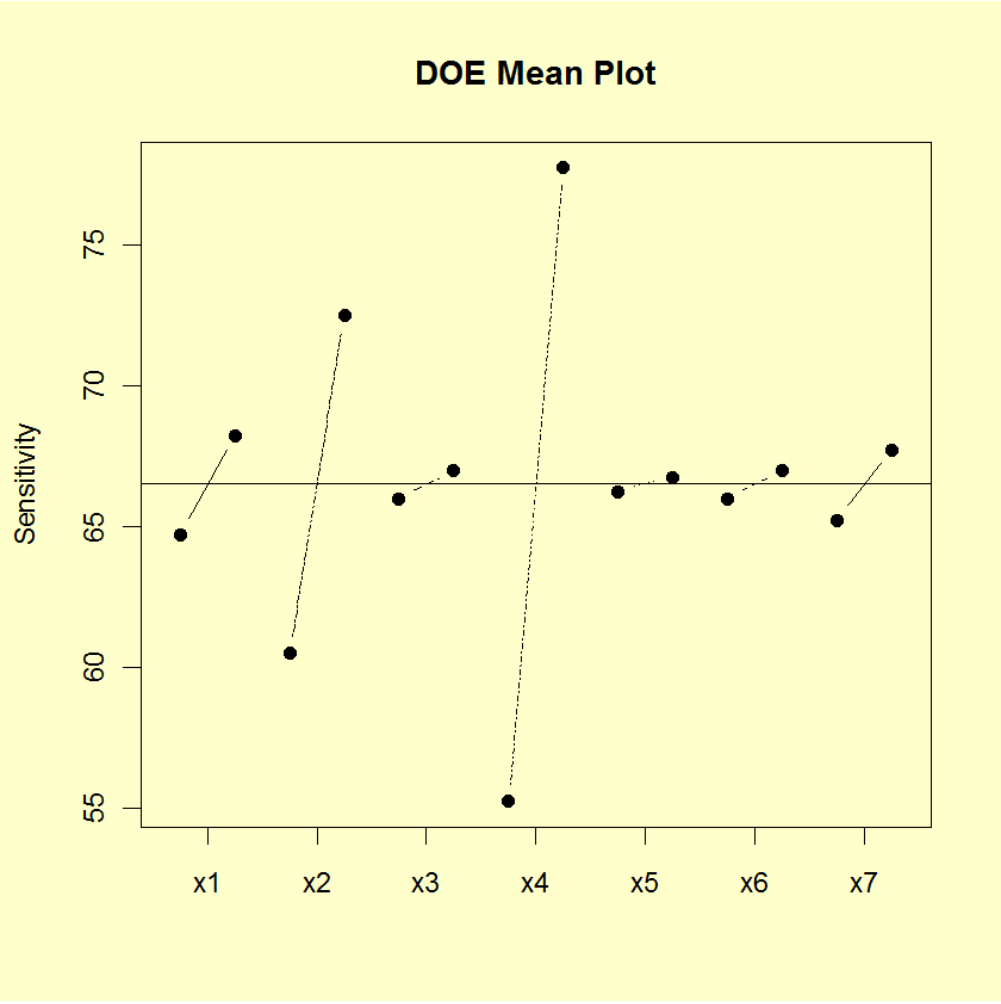
1.3.3.12. DOE Mean Plot

Purpose: Detect Important Factors With Respect to Location

The DOE mean plot is appropriate for analyzing data from a designed experiment, with respect to important factors, where the factors are at two or more levels. The plot shows mean values for the two or more levels of each factor plotted by factor. The means for a single factor are connected by a straight line. The DOE mean plot is a complement to the traditional analysis of variance of designed experiments.

This plot is typically generated for the mean. However, it can be generated for other location statistics such as the median.

Sample Plot: Factors 4, 2, and 1 Are the Most Important Factors



This sample DOE mean plot shows that:

1. factor 4 is the most important;
2. factor 2 is the second most important;
3. factor 1 is the third most important;
4. factor 7 is the fourth most important;
5. factor 6 is the fifth most important;
6. factors 3 and 5 are relatively unimportant.

In summary, factors 4, 2, and 1 seem to be clearly important, factors 3 and 5 seem to be clearly unimportant, and factors 6 and 7 are borderline factors whose inclusion in any subsequent models will be determined by further analyses.

Definition: Mean Response Versus Factor Variables

DOE mean plots are formed by:

- Vertical axis: Mean of the response variable for each level of the factor
- Horizontal axis: Factor variable

Questions

The DOE mean plot can be used to answer the following questions:

1. Which factors are important? The DOE mean plot does not provide a definitive answer to this question, but it does help categorize factors as "clearly important", "clearly not important", and "borderline importance".
2. What is the ranking list of the important factors?

Importance: Determine Significant Factors

The goal of many designed experiments is to determine which factors are significant. A ranked order listing of the important factors is also often of interest. The DOE mean plot is ideally suited for answering these types of questions and we recommend its routine use in analyzing designed experiments.

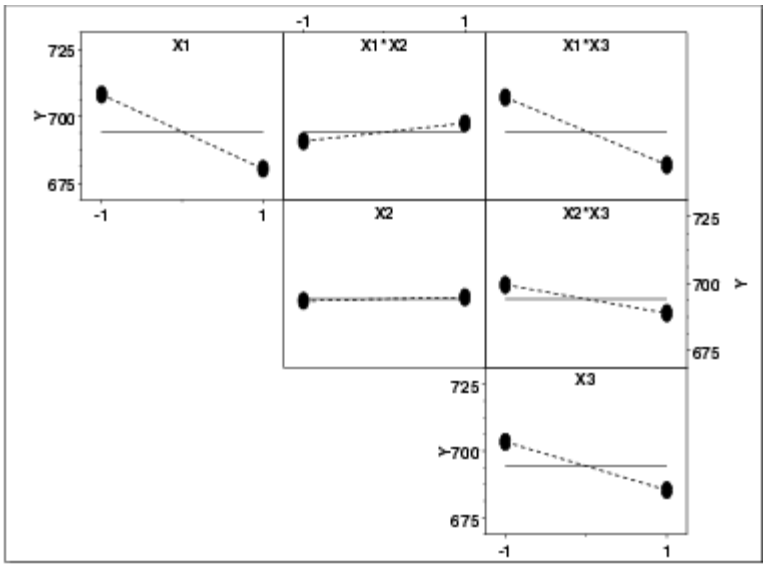
Extension for Interaction Effects

Using the concept of the scatter plot matrix, the DOE mean plot can be extended to display first-order interaction effects.

Specifically, if there are k factors, we create a matrix of plots with k rows and k columns. On the diagonal, the plot is simply a DOE mean plot with a single factor. For the off-diagonal plots, measurements at each level of the interaction are plotted versus level, where level is X_i times X_j and X_i is the code for the ith main effect level and X_j is the code for the jth main effect.

For the common 2-level designs (i.e., each factor has two levels) the values are typically coded as -1 and 1, so the multiplied values are also -1 and 1. We then generate a DOE mean plot for this interaction variable. This plot is called a DOE interaction effects plot and an example is shown below.

DOE Interaction Effects Plot



This plot shows that the most significant factor is X1 and the most significant interaction is between X1 and X3.

Related Techniques

DOE scatter plot
DOE standard deviation plot
Block plot
Box plot
Analysis of variance

Case Study

The DOE mean plot and the DOE interaction effects plot are demonstrated in the ceramic strength data case study.

Software

DOE mean plots are available in some general purpose statistical software programs, although the format may vary somewhat between these programs. It may be feasible to write macros for DOE mean plots in some statistical software programs that do not support this plot directly.

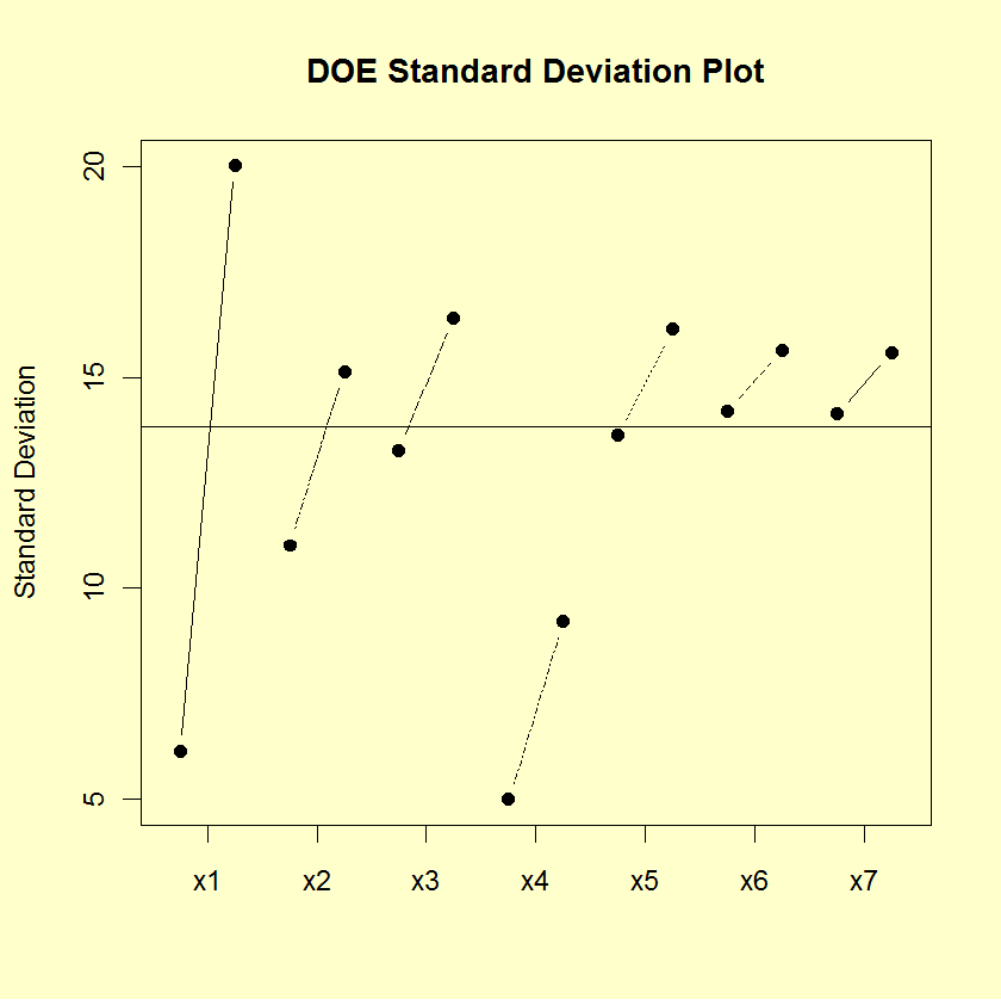
1.3.3.13. DOE Standard Deviation Plot

Purpose: Detect Important Factors With Respect to Scale

The DOE standard deviation plot is appropriate for analyzing data from a designed experiment, with respect to important factors, where the factors are at two or more levels and there are repeated values at each level. The plot shows standard deviation values for the two or more levels of each factor plotted by factor. The standard deviations for a single factor are connected by a straight line. The DOE standard deviation plot is a complement to the traditional analysis of variance of designed experiments.

This plot is typically generated for the standard deviation. However, it can also be generated for other scale statistics such as the range, the median absolute deviation, or the average absolute deviation.

Sample Plot



This sample DOE standard deviation plot shows that:

1. factor 1 has the greatest difference in standard deviations between factor levels;
2. factor 4 has a significantly lower average standard deviation than the average standard deviations of other factors (but the level 1 standard deviation for factor 1 is about the same as the level 1 standard deviation for factor 4);
3. for all factors, the level 1 standard deviation is smaller than the level 2 standard deviation.

Definition: Response Standard Deviations Versus Factor Variables

DOE standard deviation plots are formed by:

- Vertical axis: Standard deviation of the response variable for each level of the factor
- Horizontal axis: Factor variable

Questions

The DOE standard deviation plot can be used to answer the following questions:

1. How do the standard deviations vary across factors?
2. How do the standard deviations vary within a factor?
3. Which are the most important factors with respect to scale?
4. What is the ranked list of the important factors with respect to scale?

Importance: Assess Variability

The goal with many designed experiments is to determine which factors are significant. This is usually determined from the means of the factor levels (which can be conveniently shown with a DOE mean plot). A secondary goal is to assess the variability of the responses both within a factor and between factors. The DOE standard deviation plot is a convenient way to do this.

Related Techniques

DOE scatter plot
DOE mean plot
Block plot
Box plot
Analysis of variance

Case Study

The DOE standard deviation plot is demonstrated in the ceramic strength data case study.

Software

DOE standard deviation plots are not available in most general purpose statistical software programs. It may be feasible to write macros for DOE standard deviation plots in some statistical software programs that do not support them directly.

1.3.3.14. Histogram

Purpose: Summarize a Univariate Data Set

The purpose of a histogram (Chambers) is to graphically summarize the distribution of a univariate data set.

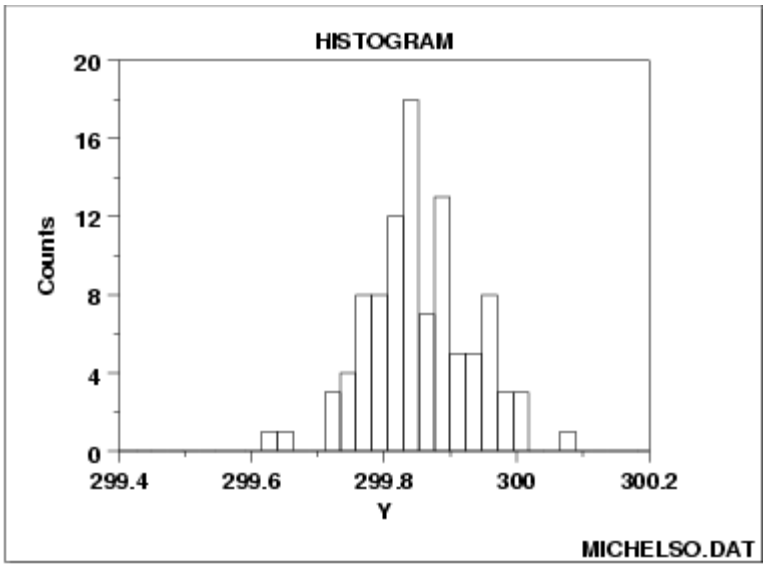
The histogram graphically shows the following:

1. center (i.e., the location) of the data;
2. spread (i.e., the scale) of the data;
3. skewness of the data;
4. presence of outliers; and
5. presence of multiple modes in the data.

These features provide strong indications of the proper distributional model for the data. The probability plot or a goodness-of-fit test can be used to verify the distributional model.

The examples section shows the appearance of a number of common features revealed by histograms.

Sample Plot



Definition

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin are counted. That is

Vertical axis: Frequency (i.e., counts for each bin)
- Horizontal axis: Response variable

The classes can either be defined arbitrarily by the user or via some systematic rule. A number of theoretically derived rules have been proposed by Scott (Scott 1992).

The cumulative histogram is a variation of the histogram in which the vertical axis gives not just the counts for a single bin, but rather gives the counts for that bin plus all bins for smaller values of the response variable.

Both the histogram and cumulative histogram have an additional variant whereby the counts are replaced by the normalized counts. The names for these variants are the relative histogram and the relative cumulative histogram.

There are two common ways to normalize the counts.

1. The normalized count is the count in a class divided by the total number of observations. In this case the relative counts are normalized to sum to one (or 100 if a percentage scale is used). This is the intuitive case where the height of the histogram bar represents the proportion of the data in each class.

2. The normalized count is the count in the class divided by the number of observations times the class width. For this normalization, the area (or integral) under the histogram is equal to one. From a probabilistic point of view, this normalization results in a relative histogram that is most akin to the probability density function and a relative cumulative histogram that is most akin to the cumulative distribution function. If you want to overlay a probability density or cumulative distribution function on top of the histogram, use this normalization. Although this normalization is less intuitive (relative frequencies greater than 1 are quite permissible), it is the appropriate normalization if you are using the histogram to model a probability density function.

Questions

The histogram can be used to answer the following questions:

1. What kind of population distribution do the data come from?
2. Where are the data located?
3. How spread out are the data?
4. Are the data symmetric or skewed?
5. Are there outliers in the data?

Examples

1. Normal
2. Symmetric, Non-Normal, Short-Tailed

3. Symmetric, Non-Normal, Long-Tailed
4. Symmetric and Bimodal
5. Bimodal Mixture of 2 Normals
6. Skewed (Non-Symmetric) Right
7. Skewed (Non-Symmetric) Left
8. Symmetric with Outlier

Related Techniques
Box plot
Probability plot

The techniques below are not discussed in the Handbook. However, they are similar in purpose to the histogram. Additional information on them is contained in the Chambers and Scott references.

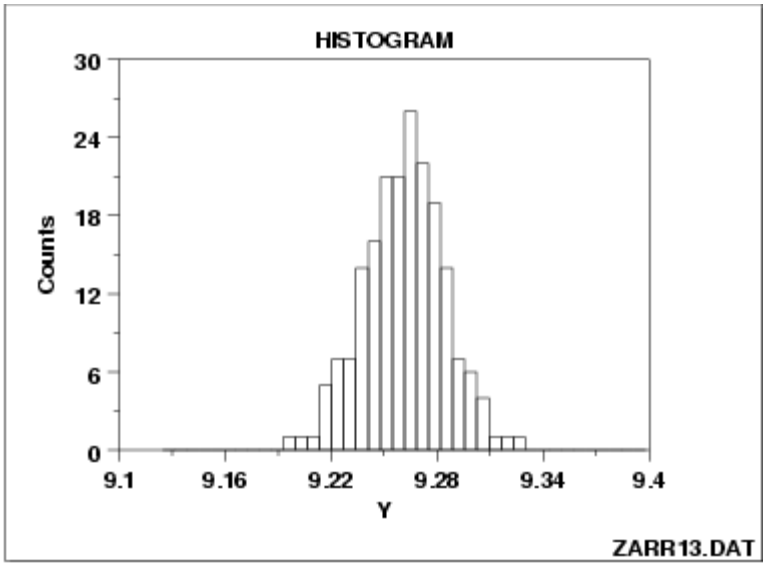
Frequency Plot
Stem and Leaf Plot
Density Trace

Case Study
The histogram is demonstrated in the heat flow meter data case study.

Software
Histograms are available in most general purpose statistical software programs. They are also supported in most general purpose charting, spreadsheet, and business graphics programs.

1.3.3.14.1. Histogram Interpretation: Normal

Symmetric, Moderate-Tailed Histogram



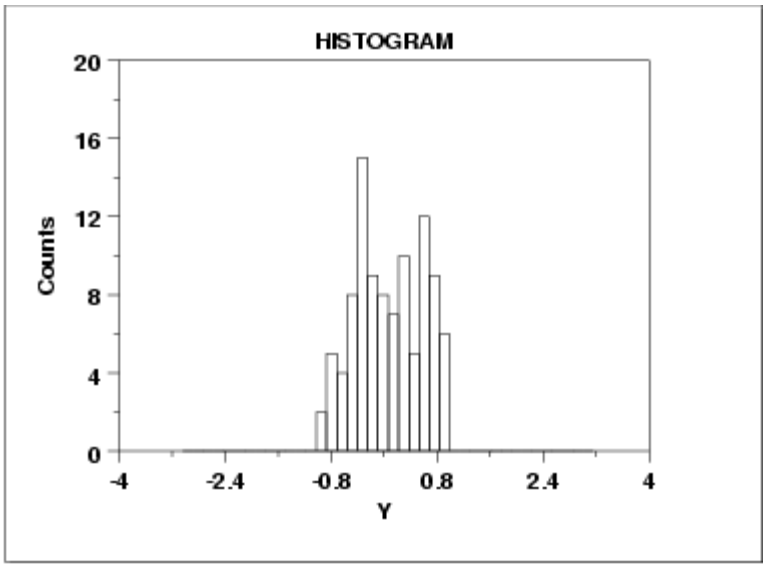
Note the classical bell-shaped, symmetric histogram with most of the frequency counts bunched in the middle and with the counts dying off out in the tails. From a physical science/engineering point of view, the normal distribution is that distribution which occurs most often in nature (due in part to the central limit theorem).

Recommended Next Step

If the histogram indicates a symmetric, moderate tailed distribution, then the recommended next step is to do a normal probability plot to confirm approximate normality. If the normal probability plot is linear, then the normal distribution is a good model for the data.

1.3.3.14.2. Histogram Interpretation: Symmetric, Non-Normal, Short-Tailed

Symmetric, Short-Tailed Histogram



Description of What Short-Tailed Means

For a symmetric distribution, the "body" of a distribution refers to the "center" of the distribution--commonly that region of the distribution where most of the probability resides--the "fat" part of the distribution. The "tail" of a distribution refers to the extreme regions of the distribution--both left and right. The "tail length" of a distribution is a term that indicates how fast these extremes approach zero.

For a short-tailed distribution, the tails approach zero very fast. Such distributions commonly have a truncated ("sawed-off") look. The classical short-tailed distribution is the uniform (rectangular) distribution in which the probability is constant over a given range and then drops to zero everywhere else--we would speak of this as having no tails, or extremely short tails.

For a moderate-tailed distribution, the tails decline to zero in a moderate fashion. The classical moderate-tailed distribution is the normal (Gaussian) distribution.

For a long-tailed distribution, the tails decline to zero very slowly--and hence one is apt to see probability a long way from the body of the distribution. The classical long-tailed distribution is the Cauchy distribution.

In terms of tail length, the histogram shown above would be characteristic of a "short-tailed" distribution.

The optimal (unbiased and most precise) estimator for location for the center of a distribution is heavily dependent on the tail length of the distribution. The common choice of taking N observations and using the calculated sample mean as the best estimate for the center of the distribution is a good choice for the normal distribution (moderate tailed), a poor choice for the uniform distribution (short tailed), and a horrible choice for the Cauchy distribution (long tailed). Although for the normal distribution the sample mean is as precise an estimator as we can get, for the uniform and Cauchy distributions, the sample mean is not the best estimator.

For the uniform distribution, the midrange

midrange = (smallest + largest) / 2

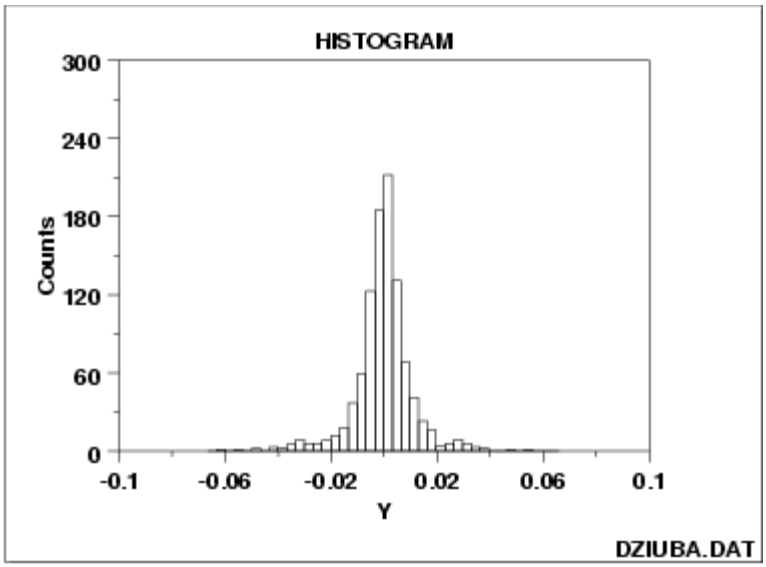
is the best estimator of location. For a Cauchy distribution, the median is the best estimator of location.

Recommended Next Step

If the histogram indicates a symmetric, short-tailed distribution, the recommended next step is to generate a uniform probability plot. If the uniform probability plot is linear, then the uniform distribution is an appropriate model for the data.

1.3.3.14.3. Histogram Interpretation: Symmetric, Non-Normal, Long-Tailed

Symmetric, Long-Tailed Histogram



Description of Long-Tailed

The previous example contains a discussion of the distinction between short-tailed, moderate-tailed, and long-tailed distributions.

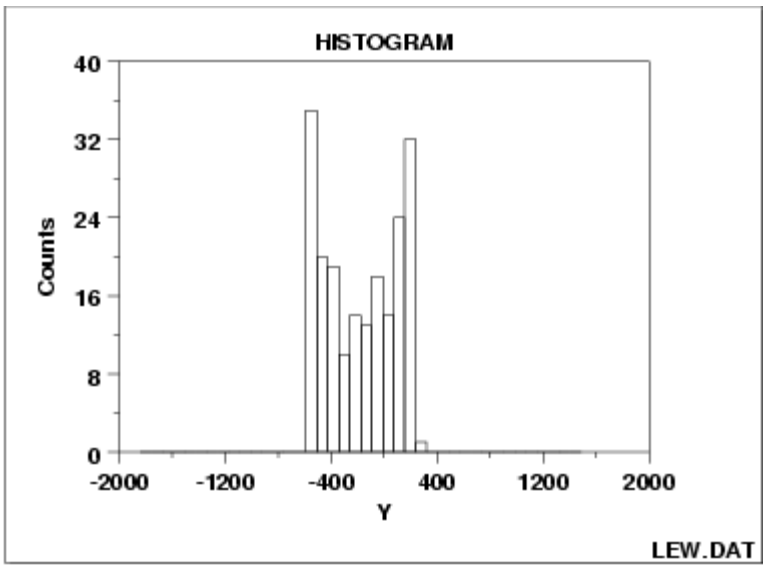
In terms of tail length, the histogram shown above would be characteristic of a "long-tailed" distribution.

Recommended Next Step

If the histogram indicates a symmetric, long tailed distribution, the recommended next step is to do a Cauchy probability plot. If the Cauchy probability plot is linear, then the Cauchy distribution is an appropriate model for the data. Alternatively, a Tukey Lambda PPCC plot may provide insight into a suitable distributional model for the data.

1.3.3.14.4. Histogram Interpretation: Symmetric and Bimodal

Symmetric, Bimodal Histogram



Description of Bimodal

The mode of a distribution is that value which is most frequently occurring or has the largest probability of occurrence. The sample mode occurs at the peak of the histogram.

For many phenomena, it is quite common for the distribution of the response values to cluster around a single mode (unimodal) and then distribute themselves with lesser frequency out into the tails. The normal distribution is the classic example of a unimodal distribution.

The histogram shown above illustrates data from a bimodal (2 peak) distribution. The histogram serves as a tool for diagnosing problems such as bimodality. Questioning the underlying reason for distributional non-unimodality frequently leads to greater insight and improved deterministic modeling of the phenomenon under study. For example, for the data presented above, the bimodal histogram is caused by sinusoidality in the data.

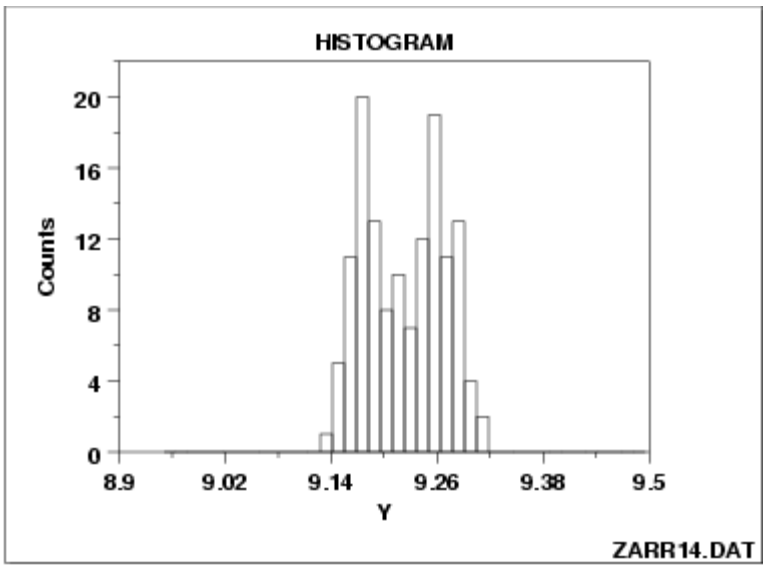
Recommended Next Step

If the histogram indicates a symmetric, bimodal distribution, the recommended next steps are to:

1. Do a run sequence plot or a scatter plot to check for sinusoidality.
2. Do a lag plot to check for sinusoidality. If the lag plot is elliptical, then the data are sinusoidal.
3. If the data are sinusoidal, then a spectral plot is used to graphically estimate the underlying sinusoidal frequency.
4. If the data are not sinusoidal, then a Tukey Lambda PPCC plot may determine the best-fit symmetric distribution for the data.
5. The data may be fit with a mixture of two distributions. A common approach to this case is to fit a mixture of 2 normal or lognormal distributions. Further discussion of fitting mixtures of distributions is beyond the scope of this Handbook.

1.3.3.14.5. Histogram Interpretation: Bimodal Mixture of 2 Normals

Histogram from Mixture of 2 Normal Distributions



Discussion of Unimodal and Bimodal

The histogram shown above illustrates data from a bimodal (2 peak) distribution.

In contrast to the previous example, this example illustrates bimodality due not to an underlying deterministic model, but bimodality due to a mixture of probability models. In this case, each of the modes appears to have a rough bell-shaped component. One could easily imagine the above histogram being generated by a process consisting of two normal distributions with the same standard deviation but with two different locations (one centered at approximately 9.17 and the other centered at approximately 9.26). If this is the case, then the research challenge is to determine physically why there are two similar but separate sub-processes.

Recommended Next Steps

If the histogram indicates that the data might be appropriately fit with a mixture of two normal distributions, the recommended next step is:

Fit the normal mixture model using either least squares or maximum likelihood. The general normal mixing model is

$$M = p\phi_1 + (1-p)\phi_2$$

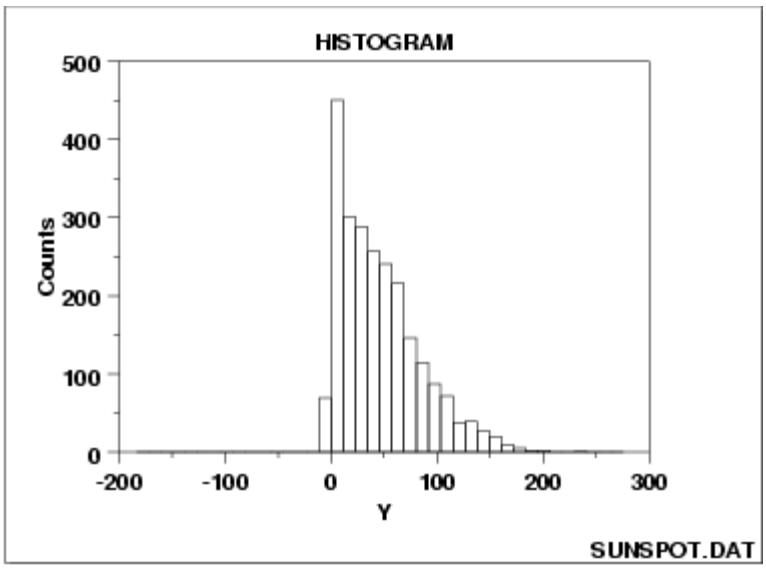
where p is the mixing proportion (between 0 and 1) and ϕ_1 and ϕ_2 are normal probability density functions with location and scale parameters μ_1, σ_1, μ_2, and σ_2, respectively. That is, there are 5 parameters to estimate in the fit.

Whether maximum likelihood or least squares is used, the quality of the fit is sensitive to good starting values. For the mixture of two normals, the histogram can be used to provide initial estimates for the location and scale parameters of the two normal distributions.

Both Dataplot code and R code can be used to fit a mixture of two normals.

1.3.3.14.6. Histogram Interpretation: Skewed (Non-Normal) Right

Right-Skewed Histogram



Discussion of Skewness

A symmetric distribution is one in which the 2 "halves" of the histogram appear as mirror-images of one another. A skewed (non-symmetric) distribution is a distribution in which there is no such mirror-imaging.

For skewed distributions, it is quite common to have one tail of the distribution considerably longer or drawn out relative to the other tail. A "skewed right" distribution is one in which the tail is on the right side. A "skewed left" distribution is one in which the tail is on the left side. The above histogram is for a distribution that is skewed right.

Skewed distributions bring a certain philosophical complexity to the very process of estimating a "typical value" for the distribution. To be specific, suppose that the analyst has a collection of 100 values randomly drawn from a distribution, and wishes to summarize these 100 observations by a "typical value". What does typical value mean? If the distribution is symmetric, the typical value is unambiguous-- it is a well-defined center of the distribution. For example, for a bell-shaped symmetric distribution, a center point is identical to that value at the peak of the distribution.

For a skewed distribution, however, there is no "center" in the usual sense of the word. Be that as it may, several "typical value" metrics are often used for skewed distributions. The first metric is the mode of the distribution. Unfortunately, for severely-skewed distributions, the mode may be at or near the left or right tail of the data and so it seems not to be a good representative of the center of the distribution. As a second choice, one could conceptually argue that the mean (the point on the horizontal axis where the distributiuon would balance) would serve well as the typical value. As a third choice, others may argue that the median (that value on the horizontal axis which has exactly 50% of the data to the left (and also to the right) would serve as a good typical value.

For symmetric distributions, the conceptual problem disappears because at the population level the mode, mean, and median are identical. For skewed distributions, however, these 3 metrics are markedly different. In practice, for skewed distributions the most commonly reported typical value is the mean; the next most common is the median; the least common is the mode. Because each of these 3 metrics reflects a different aspect of "centerness", it is recommended that the analyst report at least 2 (mean and median), and preferably all 3 (mean, median, and mode) in summarizing and characterizing a data set.

Some Causes for Skewed Data

Skewed data often occur due to lower or upper bounds on the data. That is, data that have a lower bound are often skewed right while data that have an upper bound are often skewed left. Skewness can also result from start-up effects. For example, in reliability applications some processes may have a large number of initial failures that could cause left skewness. On the other hand, a reliability process could have a long start-up period where failures are rare resulting in right-skewed data.

Data collected in scientific and engineering applications often have a lower bound of zero. For example, failure data must be non-negative. Many measurement processes generate only positive data. Time to occurence and size are common measurements that cannot be less than zero.

Recommended Next Steps

If the histogram indicates a right-skewed data set, the recommended next steps are to:

1. Quantitatively summarize the data by computing and reporting the sample mean, the sample median, and the sample mode.

2. Determine the best-fit distribution (skewed-right)

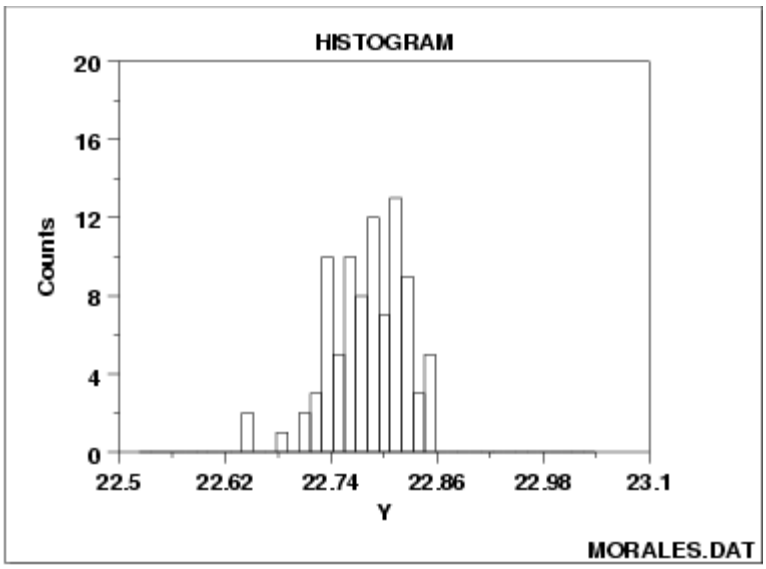
from the

- Weibull family (for the maximum)
- Gamma family
- Chi-square family
- Lognormal family
- Power lognormal family

3. Consider a normalizing transformation such as the Box-Cox transformation.

1.3.3.14.7. Histogram Interpretation: Skewed (Non-Symmetric) Left

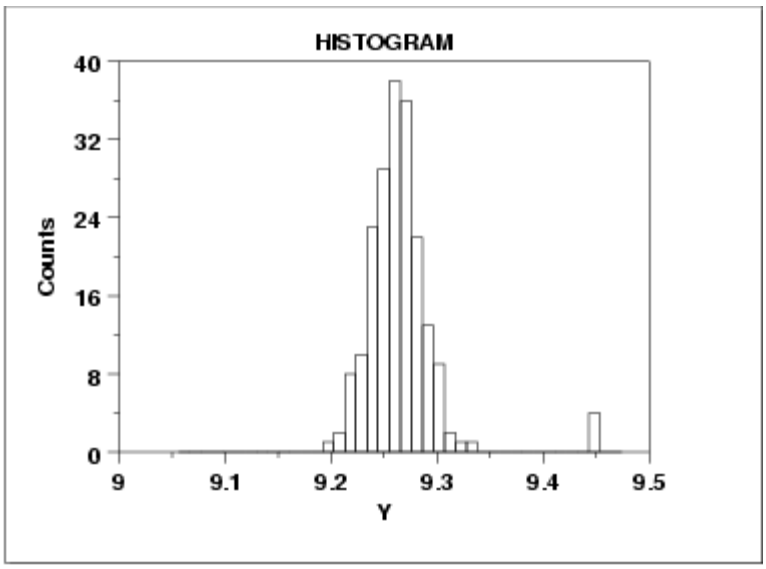
Skewed Left Histogram



The issues for skewed left data are similar to those for skewed right data.

1.3.3.14.8. Histogram Interpretation: Symmetric with Outlier

Symmetric Histogram with Outlier



Discussion of Outliers

A symmetric distribution is one in which the 2 "halves" of the histogram appear as mirror-images of one another. The above example is symmetric with the exception of outlying data near Y = 4.5.

An outlier is a data point that comes from a distribution different (in location, scale, or distributional form) from the bulk of the data. In the real world, outliers have a range of causes, from as simple as

1. operator blunders
2. equipment failures
3. day-to-day effects
4. batch-to-batch differences
5. anomalous input conditions
6. warm-up effects

to more subtle causes such as

1. A change in settings of factors that (knowingly or unknowingly) affect the response.

2. Nature is trying to tell us something.

Outliers Should be Investigated

All outliers should be taken seriously and should be investigated thoroughly for explanations. Automatic outlier-rejection schemes (such as throw out all data beyond 4 sample standard deviations from the sample mean) are particularly dangerous.

The classic case of automatic outlier rejection becoming automatic information rejection was the South Pole ozone depletion problem. Ozone depletion over the South Pole would have been detected years earlier except for the fact that the satellite data recording the low ozone readings had outlier-rejection code that automatically screened out the "outliers" (that is, the low ozone readings) before the analysis was conducted. Such inadvertent (and incorrect) purging went on for years. It was not until ground-based South Pole readings started detecting low ozone readings that someone decided to double-check as to why the satellite had not picked up this fact--it had, but it had gotten thrown out!

The best attitude is that outliers are our "friends", outliers are trying to tell us something, and we should not stop until we are comfortable in the explanation for each outlier.

Recommended Next Steps

If the histogram shows the presence of outliers, the recommended next steps are:

1. Graphically check for outliers (in the commonly encountered normal case) by generating a box plot. In general, box plots are a much better graphical tool for detecting outliers than are histograms.

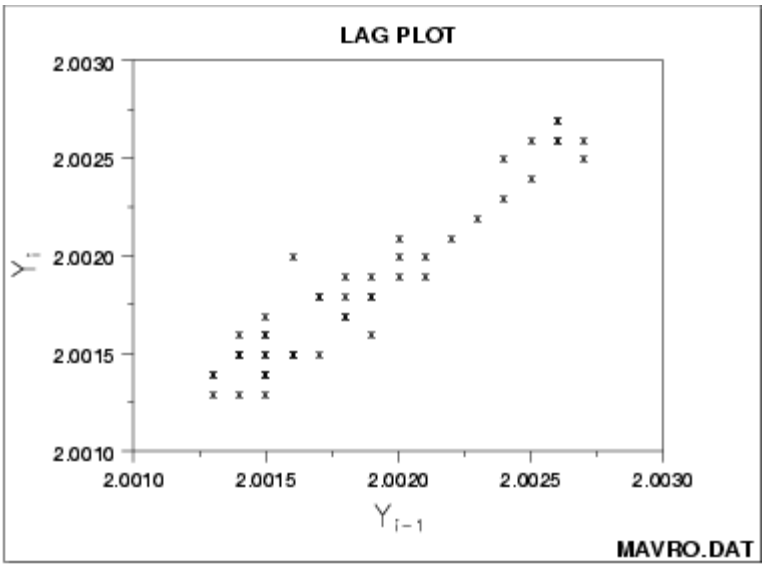
2. Quantitatively check for outliers (in the commonly encountered normal case) by carrying out Grubbs test which indicates how many sample standard deviations away from the sample mean are the data in question. Large values indicate outliers.

1.3.3.15. Lag Plot

Purpose: Check for randomness

A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random. Several common patterns for lag plots are shown in the examples below.

Sample Plot



This sample lag plot exhibits a linear pattern. This shows that the data are strongly non-random and further suggests that an autoregressive model might be appropriate.

Definition

A lag is a fixed time displacement. For example, given a data set Y_1, Y_2 ..., Y_n, Y_2 and Y_7 have lag 5 since 7 - 2 = 5. Lag plots can be generated for any arbitrary lag, although the most commonly used lag is 1.

A plot of lag 1 is a plot of the values of Y_i versus Y_{i-1}

- Vertical axis: Y_i for all *i*
- Horizontal axis: Y_{i-1} for all *i*

Questions

Lag plots can provide answers to the following questions:

1. Are the data random?
2. Is there serial correlation in the data?
3. What is a suitable model for the data?
4. Are there outliers in the data?

Importance Inasmuch as randomness is an underlying assumption for most statistical estimation and testing techniques, the lag plot should be a routine tool for researchers.

Examples

- Random (White Noise)
- Weak autocorrelation
- Strong autocorrelation and autoregressive model
- Sinusoidal model and outliers

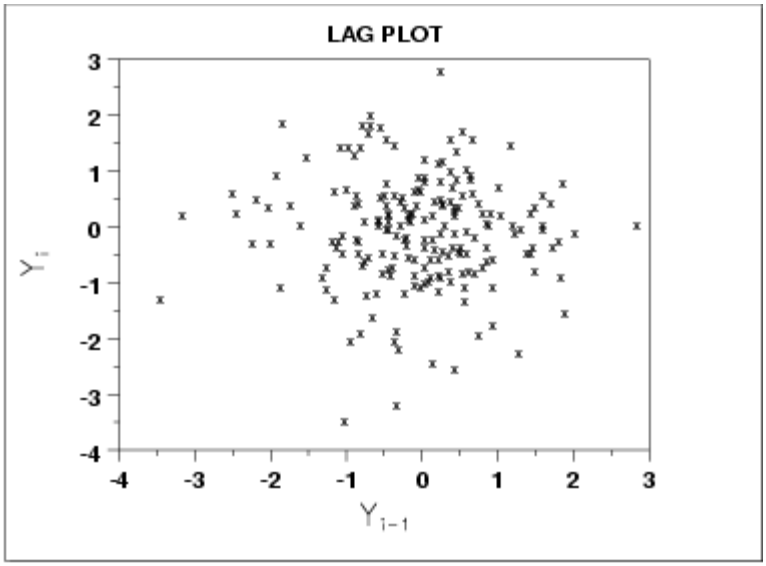
Related Techniques
Autocorrelation Plot
Spectrum
Runs Test

Case Study The lag plot is demonstrated in the beam deflection data case study.

Software Lag plots are not directly available in most general purpose statistical software programs. Since the lag plot is essentially a scatter plot with the 2 variables properly lagged, it should be feasible to write a macro for the lag plot in most statistical programs.

1.3.3.15.1. Lag Plot: Random Data

Lag Plot



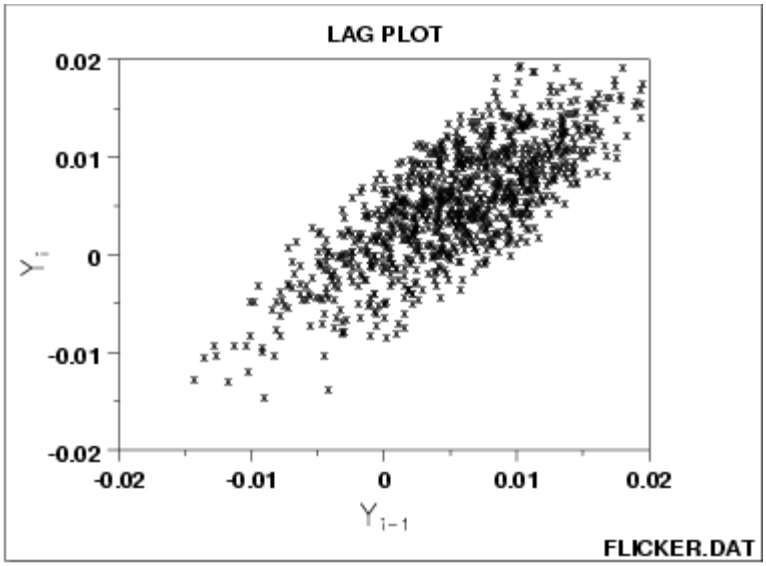
Conclusions We can make the following conclusions based on the above plot.

1. The data are random.
2. The data exhibit no autocorrelation.
3. The data contain no outliers.

Discussion The lag plot shown above is for lag = 1. Note the absence of structure. One cannot infer, from a current value Y_{i-1}, the next value Y_i. Thus for a known value Y_{i-1} on the horizontal axis (say, Y_{i-1} = +0.5), the Y_i-th value could be virtually anything (from Y_i = -2.5 to Y_i = +1.5). Such non-association is the essence of randomness.

1.3.3.15.2. Lag Plot: Moderate Autocorrelation

Lag Plot



Conclusions

We can make the conclusions based on the above plot.

1. The data are from an underlying autoregressive model with moderate positive autocorrelation
2. The data contain no outliers.

Discussion

In the plot above for lag = 1, note how the points tend to cluster (albeit noisily) along the diagonal. Such clustering is the lag plot signature of moderate autocorrelation.

If the process were completely random, knowledge of a current observation (say $Y_{i-1} = 0$) would yield virtually no knowledge about the next observation Y_i. If the process has moderate autocorrelation, as above, and if $Y_{i-1} = 0$, then the range of possible values for Y_i is seen to be restricted to a smaller range (.01 to +.01). This suggests prediction is possible using an autoregressive model.

Recommended Next Step

Estimate the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Since Y and Y are precisely the axes of the lag plot,

Y_i Y_{i-1}

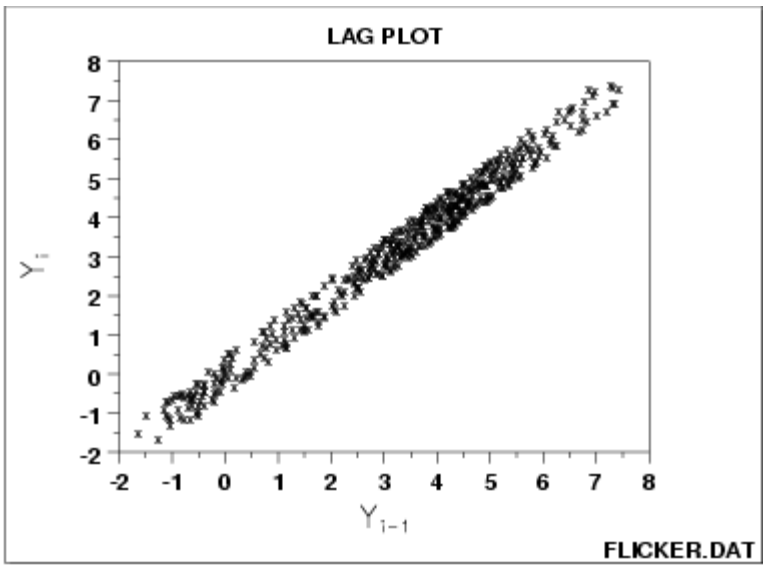
such estimation is a linear regression straight from the lag plot.

The residual standard deviation for the autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

1.3.3.15.3. Lag Plot: Strong Autocorrelation and Autoregressive Model

Lag Plot



Conclusions

We can make the following conclusions based on the above plot.

1. The data come from an underlying autoregressive model with strong positive autocorrelation
2. The data contain no outliers.

Discussion

Note the tight clustering of points along the diagonal. This is the lag plot signature of a process with strong positive autocorrelation. Such processes are highly non-random--there is strong association between an observation and a succeeding observation. In short, if you know Y_{i-1} you can make a strong guess as to what Y_i will be.

If the above process were completely random, the plot would have a shotgun pattern, and knowledge of a current observation (say $Y_{i-1} = 3$) would yield virtually no knowledge about the next observation Y_i (it could here be anywhere from -2 to +8). On the other hand, if the process had strong autocorrelation, as seen above, and if $Y_{i-1} = 3$, then the range of possible values for Y is seen to be

i

restricted to a smaller range (2 to 4)--still wide, but an improvement nonetheless (relative to -2 to +8) in predictive power.

Recommended Next Step

When the lag plot shows a strongly autoregressive pattern and only successive observations appear to be correlated, the next steps are to:

1. Extimate the parameters for the autoregressive model:

   $$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

   Since Y_i and Y_{i-1} are precisely the axes of the lag plot, such estimation is a linear regression straight from the lag plot.

   The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

   $$Y_i = A_0 + E_i$$

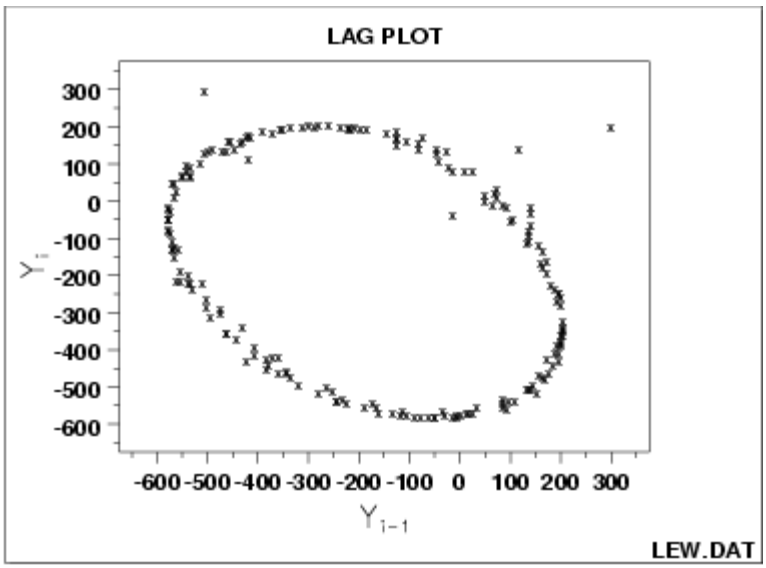
2. Reexamine the system to arrive at an explanation for the strong autocorrelation. Is it due to the

   1. phenomenon under study; or
   2. drifting in the environment; or
   3. contamination from the data acquisition system?

   Sometimes the source of the problem is contamination and carry-over from the data acquisition system where the system does not have time to electronically recover before collecting the next data point. If this is the case, then consider slowing down the sampling rate to achieve randomness.

1.3.3.15.4. Lag Plot: Sinusoidal Models and Outliers

Lag Plot



Conclusions

We can make the following conclusions based on the above plot.

1. The data come from an underlying single-cycle sinusoidal model.
2. The data contain three outliers.

Discussion

In the plot above for lag = 1, note the tight elliptical clustering of points. Processes with a single-cycle sinusoidal model will have such elliptical lag plots.

Consequences of Ignoring Cyclical Pattern

If one were to naively assume that the above process came from the null model

$$Y_i = A_0 + E_i$$

and then estimate the constant by the sample mean, then the analysis would suffer because

1. the sample mean would be biased and meaningless;
2. the confidence limits would be meaningless and optimistically small.

The proper model

$$Y_i = C + \alpha \sin(2\pi\omega t_i + \phi) + E_i$$

(where α is the amplitude, ω is the frequency--between 0 and .5 cycles per observation--, and ϕ is the phase) can be fit by standard non-linear least squares, to estimate the coefficients and their uncertainties.

The lag plot is also of value in outlier detection. Note in the above plot that there appears to be 4 points lying off the ellipse. However, in a lag plot, each point in the original data set Y shows up twice in the lag plot--once as Y_i and once as Y_{i-1}. Hence the outlier in the upper left at Y_i = 300 is the same raw data value that appears on the far right at Y_{i-1} = 300. Thus (-500,300) and (300,200) are due to the same outlier, namely the 158th data point: 300. The correct value for this 158th point should be approximately -300 and so it appears that a sign got dropped in the data collection. The other two points lying off the ellipse, at roughly (100,100) and at (0,-50), are caused by two faulty data values: the third data point of -15 should be about +125 and the fourth data point of +141 should be about -50, respectively. Hence the 4 apparent lag plot outliers are traceable to 3 actual outliers in the original run sequence: at points 4 (-15), 5 (141) and 158 (300). In retrospect, only one of these (point 158 (= 300)) is an obvious outlier in the run sequence plot.

Unexpected Value of EDA

Frequently a technique (e.g., the lag plot) is constructed to check one aspect (e.g., randomness) which it does well. Along the way, the technique also highlights some other anomaly of the data (namely, that there are 3 outliers). Such outlier identification and removal is extremely important for detecting irregularities in the data collection system, and also for arriving at a "purified" data set for modeling. The lag plot plays an important role in such outlier identification.

Recommended Next Step

When the lag plot indicates a sinusoidal model with possible outliers, the recommended next steps are:

1. Do a spectral plot to obtain an initial estimate of the frequency of the underlying cycle. This will be helpful as a starting value for the subsequent non-linear fitting.
2. Omit the outliers.
3. Carry out a non-linear fit of the model to the 197 points.

$$Y_i = C + \alpha \sin(2\pi\omega t_i + \phi) + E_i$$

1.3.3.16. Linear Correlation Plot

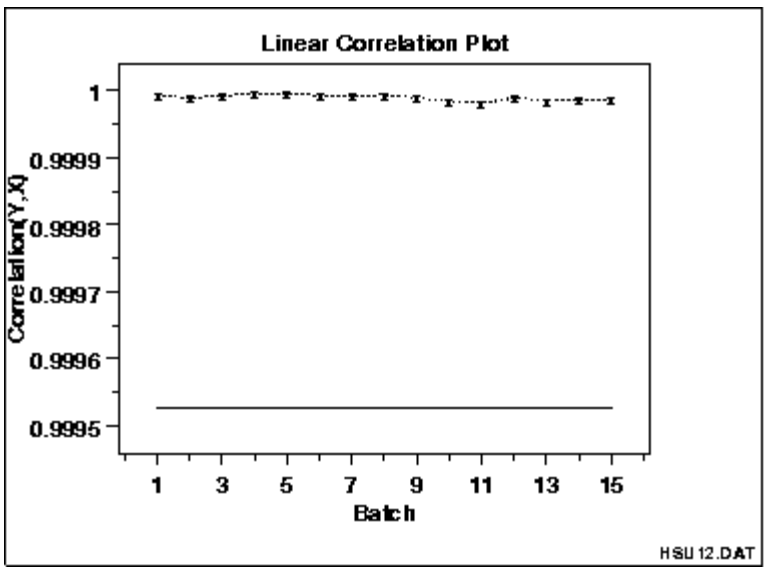
Purpose: Detect changes in correlation between groups

Linear correlation plots are used to assess whether or not correlations are consistent across groups. That is, if your data is in groups, you may want to know if a single correlation can be used across all the groups or whether separate correlations are required for each group.

Linear correlation plots are often used in conjunction with linear slope, linear intercept, and linear residual standard deviation plots. A linear correlation plot could be generated intially to see if linear fitting would be a fruitful direction. If the correlations are high, this implies it is worthwhile to continue with the linear slope, intercept, and residual standard deviation plots. If the correlations are weak, a different model needs to be pursued.

In some cases, you might not have groups. Instead you may have different data sets and you want to know if the same correlation can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear slope plot as for groups.

Sample Plot



This linear correlation plot shows that the correlations are high for all groups. This implies that linear fits could

provide a good model for each of these groups.

Definition: Group Correlations Versus Group ID

Linear correlation plots are formed by:

- Vertical axis: Group correlations
- Horizontal axis: Group identifier

A reference line is plotted at the correlation between the full data sets.

Questions

The linear correlation plot can be used to answer the following questions.

1. Are there linear relationships across groups?
2. Are the strength of the linear relationships relatively constant across the groups?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear correlation plots help answer this question in the context of linear fitting.

Related Techniques

Linear Intercept Plot
Linear Slope Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study

The linear correlation plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear correlation plot. However, if the statistical program can generate correlations over a group, it should be feasible to write a macro to generate this plot.

1.3.3.17. Linear Intercept Plot

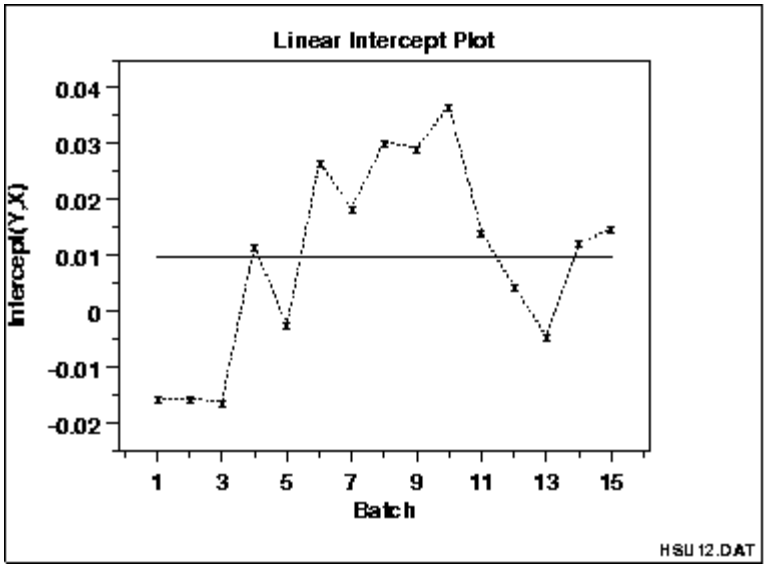
Purpose: Detect changes in linear intercepts between groups

Linear intercept plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

Linear intercept plots are typically used in conjunction with linear slope and linear residual standard deviation plots.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear intercept plot as for groups.

Sample Plot



This linear intercept plot shows that there is a shift in intercepts. Specifically, the first three intercepts are lower than the intercepts for the other groups. Note that these are small differences in the intercepts.

Definition: Group Intercepts Versus

Linear intercept plots are formed by:

- Vertical axis: Group intercepts from linear fits
- Horizontal axis: Group identifier

Group ID

A reference line is plotted at the intercept from a linear fit using all the data.

Questions

The linear intercept plot can be used to answer the following questions.

1. Is the intercept from linear fits relatively constant across groups?
2. If the intercepts vary across groups, is there a discernible pattern?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear intercept plots help answer this question in the context of linear fitting.

Related Techniques

Linear Correlation Plot
Linear Slope Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study

The linear intercept plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear intercept plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot.

1.3.3.18. Linear Slope Plot

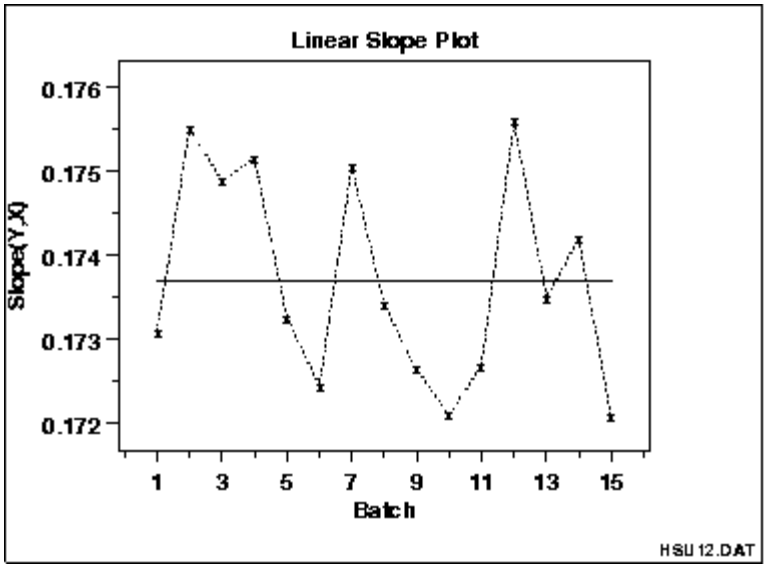
Purpose: Detect changes in linear slopes between groups

Linear slope plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

Linear slope plots are typically used in conjunction with linear intercept and linear residual standard deviation plots.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear slope plot as for groups.

Sample Plot



This linear slope plot shows that the slopes are about 0.174 (plus or minus 0.002) for all groups. There does not appear to be a pattern in the variation of the slopes. This implies that a single fit may be adequate.

Definition: Group Slopes Versus Group ID

Linear slope plots are formed by:

- Vertical axis: Group slopes from linear fits
- Horizontal axis: Group identifier

A reference line is plotted at the slope from a linear fit using all the data.

Questions

The linear slope plot can be used to answer the following questions.

1. Do you get the same slope across groups for linear fits?
2. If the slopes differ, is there a discernible pattern in the slopes?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear slope plots help answer this question in the context of linear fitting.

Related Techniques

Linear Intercept Plot
Linear Correlation Plot
Linear Residual Standard Deviation Plot
Linear Fitting

Case Study

The linear slope plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear slope plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot.

1.3.3.19. Linear Residual Standard Deviation Plot

Purpose: Detect Changes in Linear Residual Standard Deviation Between Groups

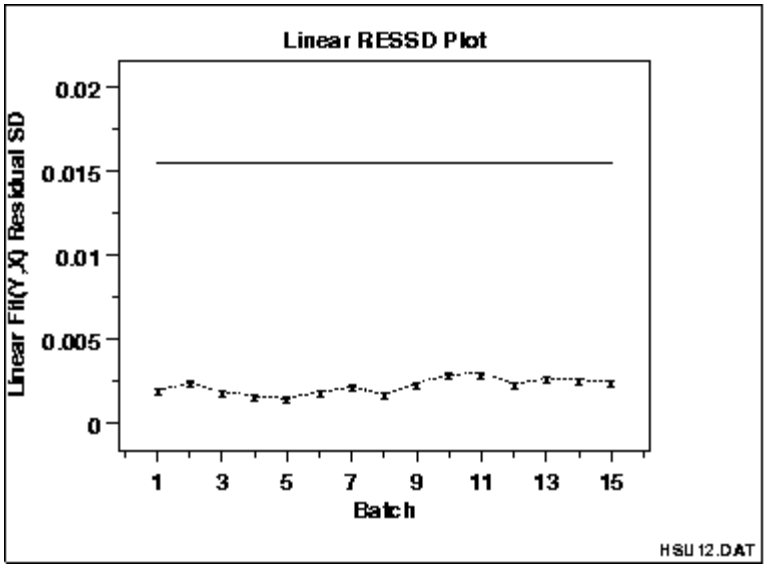
Linear residual standard deviation (RESSD) plots are used to graphically assess whether or not linear fits are consistent across groups. That is, if your data have groups, you may want to know if a single fit can be used across all the groups or whether separate fits are required for each group.

The residual standard deviation is a goodness-of-fit measure. That is, the smaller the residual standard deviation, the closer is the fit to the data.

Linear RESSD plots are typically used in conjunction with linear intercept and linear slope plots. The linear intercept and slope plots convey whether or not the fits are consistent across groups while the linear RESSD plot conveys whether the adequacy of the fit is consistent across groups.

In some cases you might not have groups. Instead, you have different data sets and you want to know if the same fit can be adequately applied to each of the data sets. In this case, simply think of each distinct data set as a group and apply the linear RESSD plot as for groups.

Sample Plot



This linear RESSD plot shows that the residual standard deviations from a linear fit are about 0.0025 for all the groups.

Definition: Group Residual Standard Deviation Versus Group ID

Linear RESSD plots are formed by:

- Vertical axis: Group residual standard deviations from linear fits
- Horizontal axis: Group identifier

A reference line is plotted at the residual standard deviation from a linear fit using all the data. This reference line will typically be much greater than any of the individual residual standard deviations.

Questions

The linear RESSD plot can be used to answer the following questions.

1. Is the residual standard deviation from a linear fit constant across groups?
2. If the residual standard deviations vary, is there a discernible pattern across the groups?

Importance: Checking Group Homogeneity

For grouped data, it may be important to know whether the different groups are homogeneous (i.e., similar) or heterogeneous (i.e., different). Linear RESSD plots help answer this question in the context of linear fitting.

Related Techniques

Linear Intercept Plot
Linear Slope Plot
Linear Correlation Plot
Linear Fitting

Case Study

The linear residual standard deviation plot is demonstrated in the Alaska pipeline data case study.

Software

Most general purpose statistical software programs do not support a linear residual standard deviation plot. However, if the statistical program can generate linear fits over a group, it should be feasible to write a macro to generate this plot.

1.3.3.20. Mean Plot

Purpose: Detect changes in location between groups

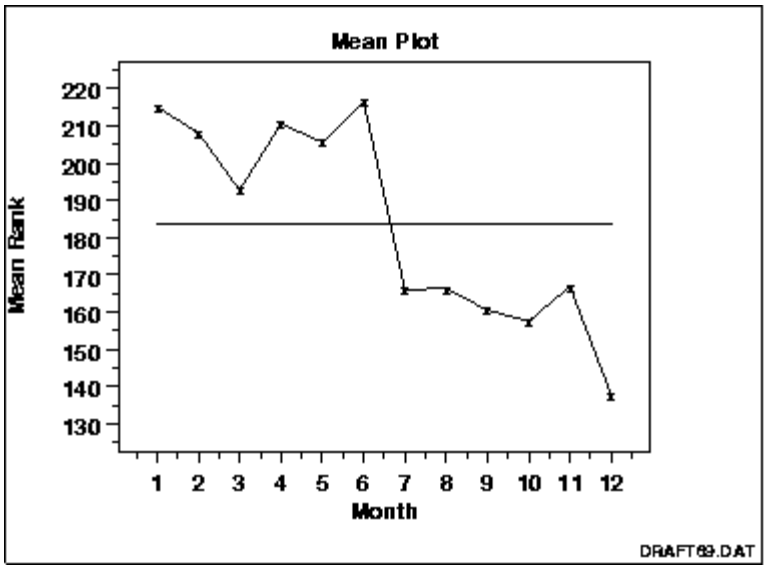
Mean plots are used to see if the mean varies between different groups of the data. The grouping is determined by the analyst. In most cases, the data set contains a specific grouping variable. For example, the groups may be the levels of a factor variable. In the sample plot below, the months of the year provide the grouping.

Mean plots can be used with ungrouped data to determine if the mean is changing over time. In this case, the data are split into an arbitrary number of equal-sized groups. For example, a data series with 400 points can be divided into 10 groups of 40 points each. A mean plot can then be generated with these groups to see if the mean is increasing or decreasing over time.

Although the mean is the most commonly used measure of location, the same concept applies to other measures of location. For example, instead of plotting the mean of each group, the median or the trimmed mean might be plotted instead. This might be done if there were significant outliers in the data and a more robust measure of location than the mean was desired.

Mean plots are typically used in conjunction with standard deviation plots. The mean plot checks for shifts in location while the standard deviation plot checks for shifts in scale.

Sample Plot



This sample mean plot shows a shift of location after the 6th month.

Definition: Group Means Versus Group ID

Mean plots are formed by:

- Vertical axis: Group mean
- Horizontal axis: Group identifier

A reference line is plotted at the overall mean.

Questions

The mean plot can be used to answer the following questions.

1. Are there any shifts in location?
2. What is the magnitude of the shifts in location?
3. Is there a distinct pattern in the shifts in location?

Importance: Checking Assumptions

A common assumption in 1-factor analyses is that of constant location. That is, the location is the same for different levels of the factor variable. The mean plot provides a graphical check for that assumption. A common assumption for univariate data is that the location is constant. By grouping the data into equal intervals, the mean plot can provide a graphical test of this assumption.

Related Techniques

Standard Deviation Plot
DOE Mean Plot
Box Plot

Software

Most general purpose statistical software programs do not support a mean plot. However, if the statistical program can generate the mean over a group, it should be feasible to write a macro to generate this plot.

1.3.3.21. Normal Probability Plot

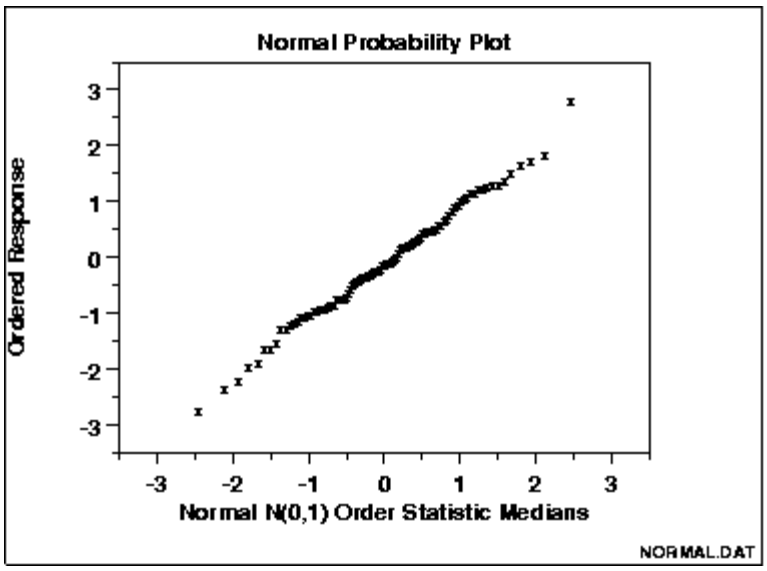
Purpose: Check If Data Are Approximately Normally Distributed

The normal probability plot (Chambers 1983) is a graphical technique for assessing whether or not a data set is approximately normally distributed.

The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

The normal probability plot is a special case of the probability plot. We cover the normal probability plot separately due to its importance in many applications.

Sample Plot



The points on this plot form a nearly linear pattern, which indicates that the normal distribution is a good model for this data set.

Definition: Ordered Response Values Versus Normal Order Statistic Medians

The normal probability plot is formed by:

- Vertical axis: Ordered response values
- Horizontal axis: Normal order statistic medians

The observations are plotted as a function of the corresponding normal order statistic medians which are

defined as:

$$N(i) = G(U(i))$$

where U(i) are the uniform order statistic medians (defined below) and G is the percent point function of the normal distribution. The percent point function is the inverse of the cumulative distribution function (probability that x is less than or equal to some value). That is, given a probability, we want the corresponding x of the cumulative distribution function.

The uniform order statistic medians are defined as:

$$U(i) = 1 - U(n) \text{ for } i = 1$$
$$U(i) = (i - 0.3175)/(n + 0.365) \text{ for } i = 2, 3, ..., n-1$$
$$U(i) = 0.5^{(1/n)} \text{ for } i = n$$

In addition, a straight line can be fit to the points and added as a reference line. The further the points vary from this line, the greater the indication of departures from normality.

Probability plots for distributions other than the normal are computed in exactly the same way. The normal percent point function (the G) is simply replaced by the percent point function of the desired distribution. That is, a probability plot can easily be generated for any distribution for which you have the percent point function.

One advantage of this method of computing probability plots is that the intercept and slope estimates of the fitted line are in fact estimates for the location and scale parameters of the distribution. Although this is not too important for the normal distribution since the location and scale are estimated by the mean and standard deviation, respectively, it can be useful for many other distributions.

The correlation coefficient of the points on the normal probability plot can be compared to a table of critical values to provide a formal test of the hypothesis that the data come from a normal distribution.

Questions

The normal probability plot is used to answer the following questions.

1. Are the data normally distributed?
2. What is the nature of the departure from normality (data skewed, shorter than expected tails, longer than expected tails)?

Importance: Check

The underlying assumptions for a measurement process are that the data should behave like:

Normality Assumption

1. random drawings;
2. from a fixed distribution;
3. with fixed location;
4. with fixed scale.

Probability plots are used to assess the assumption of a fixed distribution. In particular, most statistical models are of the form:

response = deterministic + random

where the deterministic part is the fit and the random part is error. This error component in most common statistical models is specifically assumed to be normally distributed with fixed location and scale. This is the most frequent application of normal probability plots. That is, a model is fit and a normal probability plot is generated for the residuals from the fitted model. If the residuals from the fitted model are not normally distributed, then one of the major assumptions of the model has been violated.

Examples

1. Data are normally distributed
2. Data have short tails
3. Data have fat tails
4. Data are skewed right

Related Techniques

Histogram
Probability plots for other distributions (e.g., Weibull)
Probability plot correlation coefficient plot (PPCC plot)
Anderson-Darling Goodness-of-Fit Test
Chi-Square Goodness-of-Fit Test
Kolmogorov-Smirnov Goodness-of-Fit Test

Case Study

The normal probability plot is demonstrated in the heat flow meter data case study.

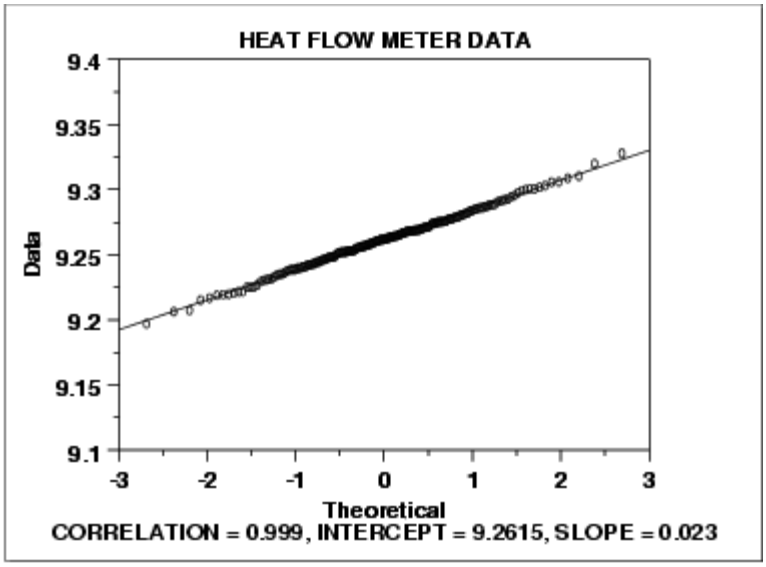
Software

Most general purpose statistical software programs can generate a normal probability plot.

1.3.3.21.1. Normal Probability Plot: Normally Distributed Data

Normal Probability Plot

The following normal probability plot is from the heat flow meter data.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a strongly linear pattern. There are only minor deviations from the line fit to the points on the probability plot.
2. The normal distribution appears to be a good model for these data.

Discussion

Visually, the probability plot shows a strongly linear pattern. This is verified by the correlation coefficient of 0.9989 of the line fit to the probability plot. The fact that the points in the lower and upper extremes of the plot do not deviate significantly from the straight-line pattern indicates that there are not any significant outliers (relative to a normal distribution).

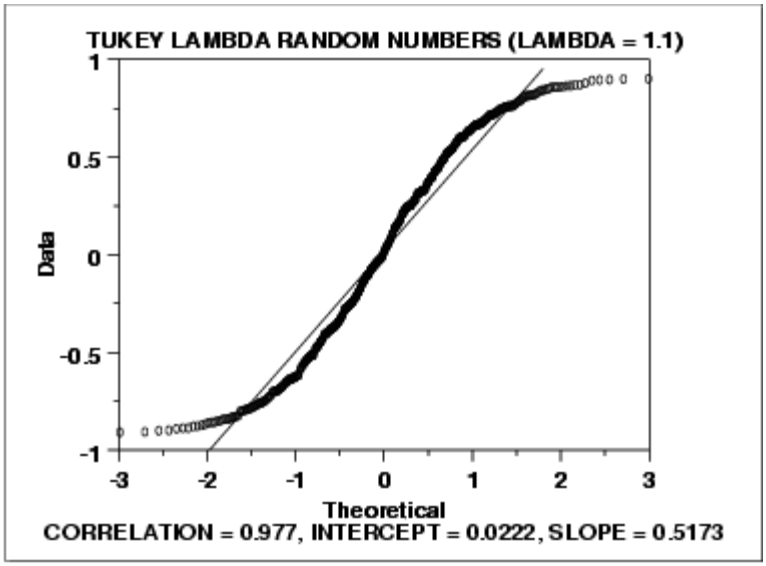
In this case, we can quite reasonably conclude that the normal distribution provides an excellent model for the data. The intercept and slope of the fitted line give estimates of

9.26 and 0.023 for the location and scale parameters of the fitted normal distribution.

1.3.3.21.2. Normal Probability Plot: Data Have Short Tails

Normal Probability Plot for Data with Short Tails

The following is a normal probability plot for 500 random numbers generated from a Tukey-Lambda distribution with the λ parameter equal to 1.1.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a non-linear pattern.
2. The normal distribution is not a good model for these data.

Discussion

For data with short tails relative to the normal distribution, the non-linearity of the normal probability plot shows up in two ways. First, the middle of the data shows an S-like pattern. This is common for both short and long tails. Second, the first few and the last few points show a marked departure from the reference fitted line. In comparing this plot to the long tail example in the next section, the important difference is the direction of the departure from the fitted line for the first few and last few points. For short tails, the first few points show increasing departure from the fitted line *above* the line and last few points show increasing departure from the fitted line *below* the line. For long tails,

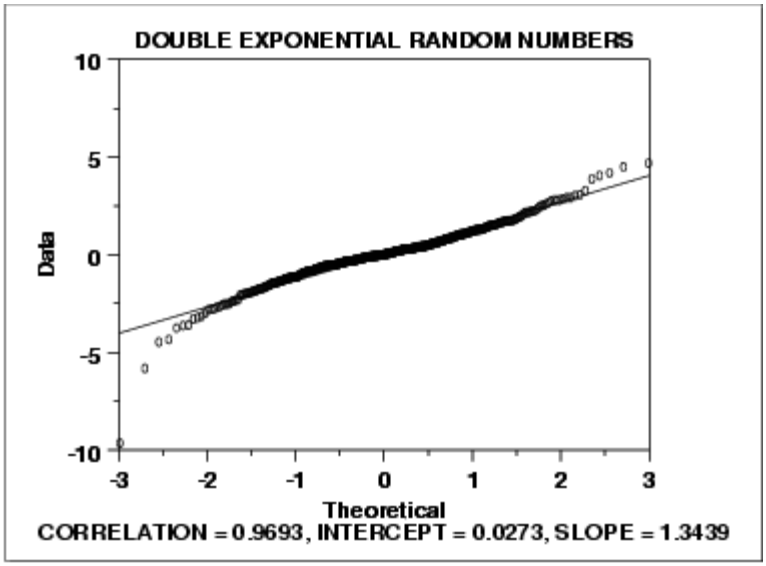
this pattern is reversed.

In this case, we can reasonably conclude that the normal distribution does not provide an adequate fit for this data set. For probability plots that indicate short-tailed distributions, the next step might be to generate a Tukey Lambda PPCC plot. The Tukey Lambda PPCC plot can often be helpful in identifying an appropriate distributional family.

1.3.3.21.3. Normal Probability Plot: Data Have Long Tails

Normal Probability Plot for Data with Long Tails

The following is a normal probability plot of 500 numbers generated from a double exponential distribution. The double exponential distribution is symmetric, but relative to the normal it declines rapidly and has longer tails.



Conclusions

We can make the following conclusions from the above plot.

1. The normal probability plot shows a reasonably linear pattern in the center of the data. However, the tails, particularly the lower tail, show departures from the fitted line.
2. A distribution other than the normal distribution would be a good model for these data.

Discussion

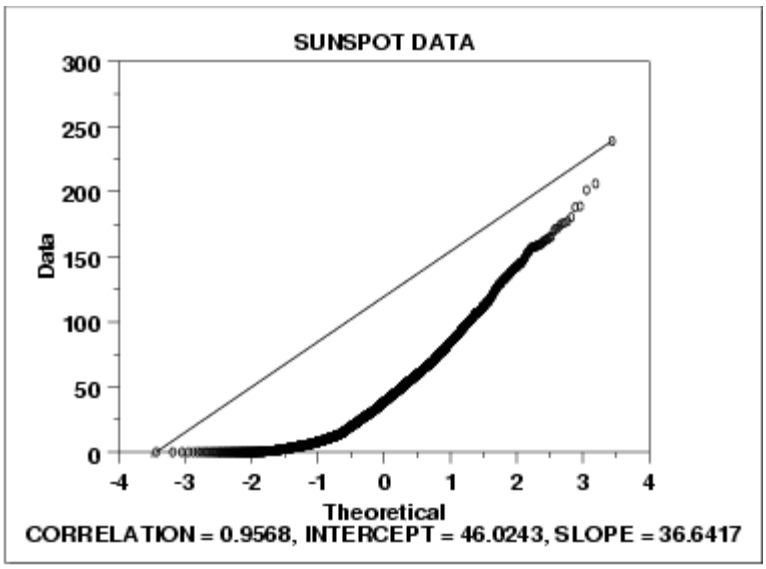
For data with long tails relative to the normal distribution, the non-linearity of the normal probability plot can show up in two ways. First, the middle of the data may show an S-like pattern. This is common for both short and long tails. In this particular case, the S pattern in the middle is fairly mild. Second, the first few and the last few points show marked departure from the reference fitted line. In the plot above, this is most noticeable for the first few data points. In

comparing this plot to the short-tail example in the previous section, the important difference is the direction of the departure from the fitted line for the first few and the last few points. For long tails, the first few points show increasing departure from the fitted line *below* the line and last few points show increasing departure from the fitted line *above* the line. For short tails, this pattern is reversed.

In this case we can reasonably conclude that the normal distribution can be improved upon as a model for these data. For probability plots that indicate long-tailed distributions, the next step might be to generate a Tukey Lambda PPCC plot. The Tukey Lambda PPCC plot can often be helpful in identifying an appropriate distributional family.

1.3.3.21.4. Normal Probability Plot: Data are Skewed Right

Normal Probability Plot for Data that are Skewed Right



Conclusions We can make the following conclusions from the above plot.

1. The normal probability plot shows a strongly non-linear pattern. Specifically, it shows a quadratic pattern in which all the points are below a reference line drawn between the first and last points.
2. The normal distribution is not a good model for these data.

Discussion This quadratic pattern in the normal probability plot is the signature of a significantly right-skewed data set. Similarly, if all the points on the normal probability plot fell above the reference line connecting the first and last points, that would be the signature pattern for a significantly left-skewed data set.

In this case we can quite reasonably conclude that we need to model these data with a right skewed distribution such as the Weibull or lognormal.

1.3.3.22. Probability Plot

Purpose: Check If Data Follow a Given Distribution

The probability plot (Chambers 1983) is a graphical technique for assessing whether or not a data set follows a given distribution such as the normal or Weibull.

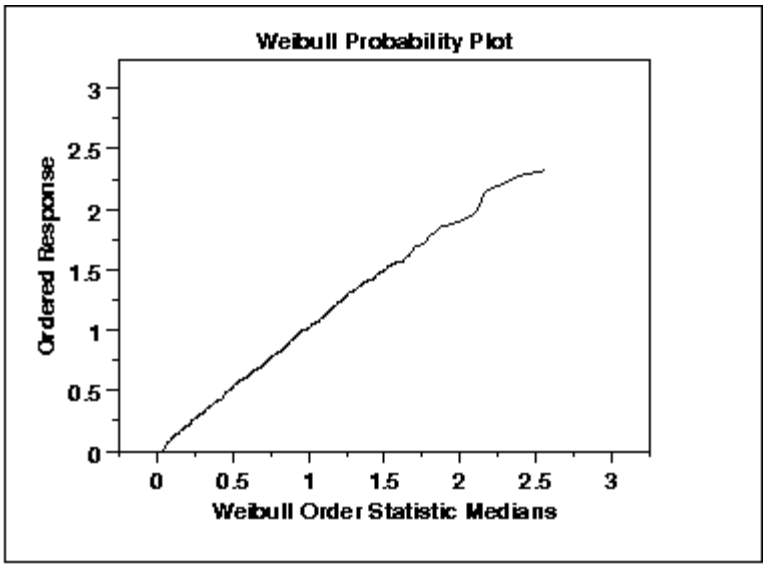
The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. Departures from this straight line indicate departures from the specified distribution.

The correlation coefficient associated with the linear fit to the data in the probability plot is a measure of the goodness of the fit. Estimates of the location and scale parameters of the distribution are given by the intercept and slope. Probability plots can be generated for several competing distributions to see which provides the best fit, and the probability plot generating the highest correlation coefficient is the best choice since it generates the straightest probability plot.

For distributions with shape parameters (not counting location and scale parameters), the shape parameters must be known in order to generate the probability plot. For distributions with a single shape parameter, the probability plot correlation coefficient (PPCC) plot provides an excellent method for estimating the shape parameter.

We cover the special case of the normal probability plot separately due to its importance in many statistical applications.

Sample Plot



This data is a set of 500 Weibull random numbers with a shape parameter = 2, location parameter = 0, and scale parameter = 1. The Weibull probability plot indicates that the Weibull distribution does in fact fit these data well.

Definition: Ordered Response Values Versus Order Statistic Medians for the Given Distribution

The probability plot is formed by:

- Vertical axis: Ordered response values
- Horizontal axis: Order statistic medians for the given distribution

The order statistic medians are defined as:

N(i) = G(U(i))

where the U(i) are the uniform order statistic medians (defined below) and G is the percent point function for the desired distribution. The percent point function is the inverse of the cumulative distribution function (probability that *x* is less than or equal to some value). That is, given a probability, we want the corresponding *x* of the cumulative distribution function.

The uniform order statistic medians are defined as:

m(i) = 1 - m(n) for i = 1
m(i) = (i - 0.3175)/(n + 0.365) for i = 2, 3, ..., n-1
m(i) = 0.5**(1/n) for i = n

In addition, a straight line can be fit to the points and added as a reference line. The further the points vary from this line, the greater the indication of a departure from the specified distribution.

This definition implies that a probability plot can be easily generated for any distribution for which the percent point function can be computed.

One advantage of this method of computing proability plots is that the intercept and slope estimates of the fitted line are in fact estimates for the location and scale parameters of the distribution. Although this is not too important for the normal distribution (the location and scale are estimated by the mean and standard deviation, respectively), it can be useful for many other distributions.

Questions

The probability plot is used to answer the following questions:

- Does a given distribution, such as the Weibull, provide a good fit to my data?
- What distribution best fits my data?
- What are good estimates for the location and scale parameters of the chosen distribution?

Importance: Check distributional assumption

The discussion for the normal probability plot covers the use of probability plots for checking the fixed distribution assumption.

Some statistical models assume data have come from a population with a specific type of distribution. For example, in reliability applications, the Weibull, lognormal, and exponential are commonly used distributional models. Probability plots can be useful for checking this distributional assumption.

Related Techniques

Histogram
Probability Plot Correlation Coefficient (PPCC) Plot
Hazard Plot
Quantile-Quantile Plot
Anderson-Darling Goodness of Fit
Chi-Square Goodness of Fit
Kolmogorov-Smirnov Goodness of Fit

Case Study

The probability plot is demonstrated in the uniform random numbers case study.

Software

Most general purpose statistical software programs support probability plots for at least a few common distributions.

1.3.3.23. Probability Plot Correlation Coefficient Plot

Purpose: Graphical Technique for Finding the Shape Parameter of a Distributional Family that Best Fits a Data Set

The probability plot correlation coefficient (PPCC) plot (Filliben 1975) is a graphical technique for identifying the shape parameter for a distributional family that best describes the data set. This technique is appropriate for families, such as the Weibull, that are defined by a single shape parameter and location and scale parameters, and it is not appropriate for distributions, such as the normal, that are defined only by location and scale parameters.

The PPCC plot is generated as follows. For a series of values for the shape parameter, the correlation coefficient is computed for the probability plot associated with a given value of the shape parameter. These correlation coefficients are plotted against their corresponding shape parameters. The maximum correlation coefficient corresponds to the optimal value of the shape parameter. For better precision, two iterations of the PPCC plot can be generated; the first is for finding the right neighborhood and the second is for fine tuning the estimate.

The PPCC plot is used first to find a good value of the shape parameter. The probability plot is then generated to find estimates of the location and scale parameters and in addition to provide a graphical assessment of the adequacy of the distributional fit.

Compare Distributions

In addition to finding a good choice for estimating the shape parameter of a given distribution, the PPCC plot can be useful in deciding which distributional family is most appropriate. For example, given a set of reliabilty data, you might generate PPCC plots for a Weibull, lognormal, gamma, and inverse Gaussian distributions, and possibly others, on a single page. This one page would show the best value for the shape parameter for several distributions and would additionally indicate which of these distributional families provides the best fit (as measured by the maximum probability plot correlation coefficient). That is, if the maximum PPCC value for the Weibull is 0.99 and only 0.94 for the lognormal, then we could reasonably

conclude that the Weibull family is the better choice.

Tukey-Lambda PPCC Plot for Symmetric Distributions

The Tukey Lambda PPCC plot, with shape parameter λ, is particularly useful for symmetric distributions. It indicates whether a distribution is short or long tailed and it can further indicate several common distributions. Specifically,

1. λ = -1: distribution is approximately Cauchy
2. λ = 0: distribution is exactly logistic
3. λ = 0.14: distribution is approximately normal
4. λ = 0.5: distribution is U-shaped
5. λ = 1: distribution is exactly uniform

If the Tukey Lambda PPCC plot gives a maximum value near 0.14, we can reasonably conclude that the normal distribution is a good model for the data. If the maximum value is less than 0.14, a long-tailed distribution such as the double exponential or logistic would be a better choice. If the maximum value is near -1, this implies the selection of very long-tailed distribution, such as the Cauchy. If the maximum value is greater than 0.14, this implies a short-tailed distribution such as the Beta or uniform.

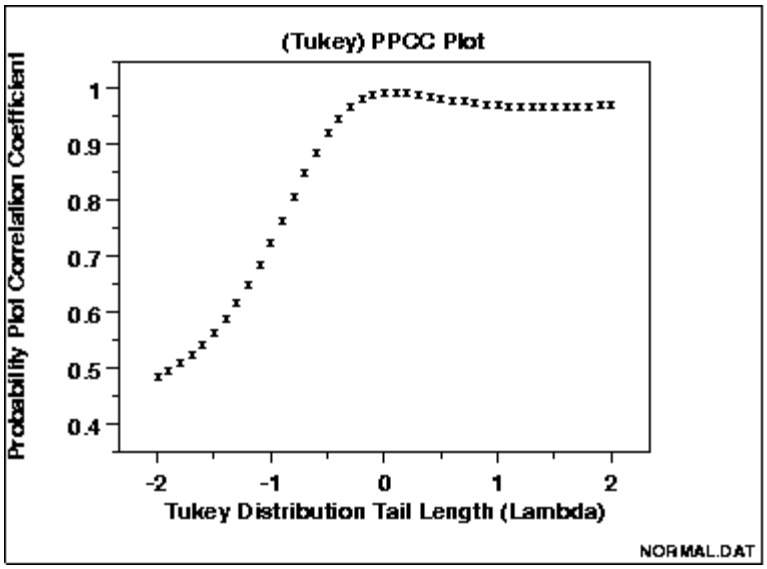
The Tukey-Lambda PPCC plot is used to suggest an appropriate distribution. You should follow-up with PPCC and probability plots of the appropriate alternatives.

Use Judgement When Selecting An Appropriate Distributional Family

When comparing distributional models, do not simply choose the one with the maximum PPCC value. In many cases, several distributional fits provide comparable PPCC values. For example, a lognormal and Weibull may both fit a given set of reliability data quite well. Typically, we would consider the complexity of the distribution. That is, a simpler distribution with a marginally smaller PPCC value may be preferred over a more complex distribution. Likewise, there may be theoretical justification in terms of the underlying scientific model for preferring a distribution with a marginally smaller PPCC value in some cases. In other cases, we may not need to know if the distributional model is optimal, only that it is adequate for our purposes. That is, we may be able to use techniques designed for normally distributed data even if other distributions fit the data somewhat better.

Sample Plot

The following is a PPCC plot of 100 normal random numbers. The maximum value of the correlation coefficient = 0.997 at λ = 0.099.



This PPCC plot shows that:

1. the best-fit symmetric distribution is nearly normal;
2. the data are not long tailed;
3. the sample mean would be an appropriate estimator of location.

We can follow-up this PPCC plot with a normal probability plot to verify the normality model for the data.

Definition: The PPCC plot is formed by:

- Vertical axis: Probability plot correlation coefficient;
- Horizontal axis: Value of shape parameter.

Questions The PPCC plot answers the following questions:

1. What is the best-fit member within a distributional family?
2. Does the best-fit member provide a good fit (in terms of generating a probability plot with a high correlation coefficient)?
3. Does this distributional family provide a good fit compared to other distributions?
4. How sensitive is the choice of the shape parameter?

Importance Many statistical analyses are based on distributional assumptions about the population from which the data have been obtained. However, distributional families can have radically different shapes depending on the value of the shape parameter. Therefore, finding a reasonable choice for the shape parameter is a necessary step in the analysis. In many analyses, finding a good distributional model for the data is the primary focus of the analysis. In both of these cases, the PPCC plot is a valuable tool.

Related Techniques	Probability Plot Maximum Likelihood Estimation Least Squares Estimation Method of Moments Estimation
Software	PPCC plots are currently not available in most common general purpose statistical software programs. However, the underlying technique is based on probability plots and correlation coefficients, so it should be possible to write macros for PPCC plots in statistical programs that support these capabilities. Dataplot supports PPCC plots.

1.3.3.24. Quantile-Quantile Plot

Purpose: Check If Two Data Sets Can Be Fit With the Same Distribution

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

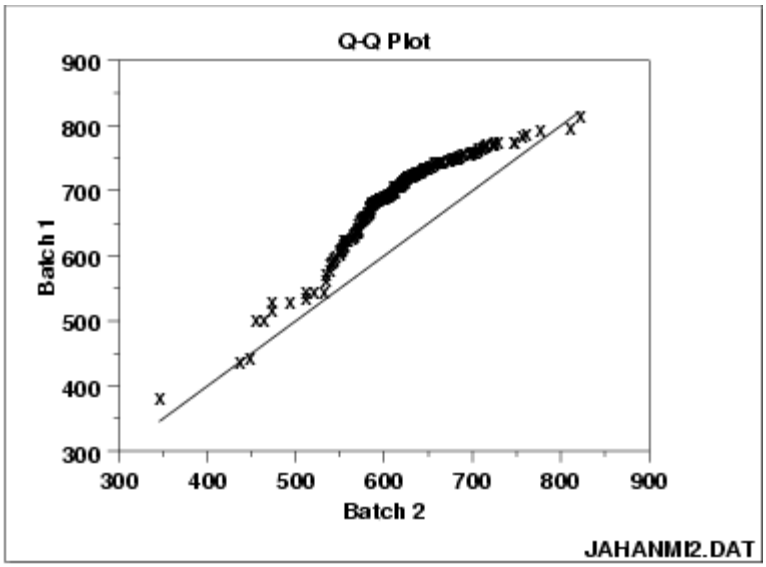
The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.

2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

Sample Plot



This q-q plot shows that

1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Definition: Quantiles for Data Set 1 Versus Quantiles of Data Set 2

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

Questions

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

Importance: Check for Common

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data

Distribution	sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.
Related Techniques	Bihistogram T Test F Test 2-Sample Chi-Square Test 2-Sample Kolmogorov-Smirnov Test
Case Study	The quantile-quantile plot is demonstrated in the ceramic strength data case study.
Software	Q-Q plots are available in some general purpose statistical software programs. If the number of data points in the two samples are equal, it should be relatively easy to write a macro in statistical programs that do not support the q-q plot. If the number of points are not equal, writing a macro for a q-q plot may be difficult.

1.3.3.25. Run-Sequence Plot

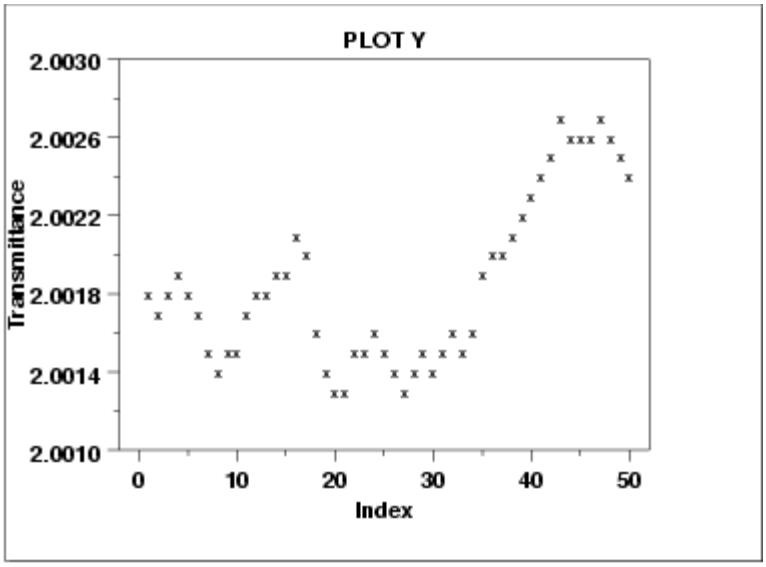
Purpose: Check for Shifts in Location and Scale and Outliers

Run sequence plots (Chambers 1983) are an easy way to graphically summarize a univariate data set. A common assumption of univariate data sets is that they behave like:

1. random drawings;
2. from a fixed distribution;
3. with a common location; and
4. with a common scale.

With run sequence plots, shifts in location and scale are typically quite evident. Also, outliers can easily be detected.

Sample Plot: Last Third of Data Shows a Shift of Location



This sample run sequence plot shows that the location shifts up for the last third of the data.

Definition: y(i) Versus i

Run sequence plots are formed by:

- Vertical axis: Response variable Y(i)
- Horizontal axis: Index i (i = 1, 2, 3, ...)

Questions

The run sequence plot can be used to answer the following questions

1. Are there any shifts in location?
2. Are there any shifts in variation?
3. Are there any outliers?

The run sequence plot can also give the analyst an excellent feel for the data.

Importance: Check Univariate Assumptions

For univariate data, the default model is

$$Y = \text{constant} + \text{error}$$

where the error is assumed to be random, from a fixed distribution, and with constant location and scale. The validity of this model depends on the validity of these assumptions. The run sequence plot is useful for checking for constant location and scale.

Even for more complex models, the assumptions on the error term are still often the same. That is, a run sequence plot of the residuals (even from very complex models) is still vital for checking for outliers and for detecting shifts in location and scale.

Related Techniques

Scatter Plot
Histogram
Autocorrelation Plot
Lag Plot

Case Study

The run sequence plot is demonstrated in the Filter transmittance data case study.

Software

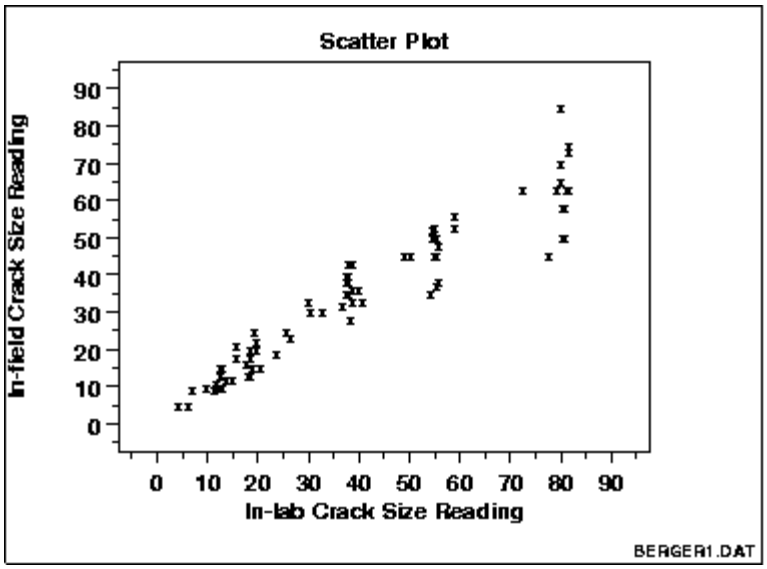
Run sequence plots are available in most general purpose statistical software programs.

1.3.3.26. Scatter Plot

Purpose: Check for Relationship

A scatter plot (Chambers 1983) reveals relationships or association between two variables. Such relationships manifest themselves by any non-random structure in the plot. Various common types of patterns are demonstrated in the examples.

Sample Plot: Linear Relationship Between Variables Y and X



This sample plot reveals a linear relationship between the two variables indicating that a linear regression model might be appropriate.

Definition: Y Versus X

A scatter plot is a plot of the values of Y versus the corresponding values of X:

- Vertical axis: variable Y--usually the response variable
- Horizontal axis: variable X--usually some variable we suspect may ber related to the response

Questions

Scatter plots can provide answers to the following questions:

1. Are variables X and Y related?
2. Are variables X and Y linearly related?
3. Are variables X and Y non-linearly related?
4. Does the variation in Y change depending on X?
5. Are there outliers?

Examples

1. No relationship
2. Strong linear (positive correlation)
3. Strong linear (negative correlation)
4. Exact linear (positive correlation)
5. Quadratic relationship
6. Exponential relationship
7. Sinusoidal relationship (damped)
8. Variation of Y doesn't depend on X (homoscedastic)
9. Variation of Y does depend on X (heteroscedastic)
10. Outlier

Combining Scatter Plots

Scatter plots can also be combined in multiple plots per page to help understand higher-level structure in data sets with more than two variables.

The scatterplot matrix generates all pairwise scatter plots on a single page. The conditioning plot, also called a co-plot or subset plot, generates scatter plots of Y versus X dependent on the value of a third variable.

Causality Is Not Proved By Association

The scatter plot uncovers relationships in data. "Relationships" means that there is some structured association (linear, quadratic, etc.) between X and Y. Note, however, that even though

causality implies association

association does NOT imply causality.

Scatter plots are a useful diagnostic tool for determining association, but if such association exists, the plot may or may not suggest an underlying cause-and-effect mechanism. A scatter plot can never "prove" cause and effect--it is ultimately only the researcher (relying on the underlying science/engineering) who can conclude that causality actually exists.

Appearance

The most popular rendition of a scatter plot is

1. some plot character (e.g., X) at the data points, and
2. no line connecting data points.

Other scatter plot format variants include

1. an optional plot character (e.g, X) at the data points, but
2. a solid line connecting data points.

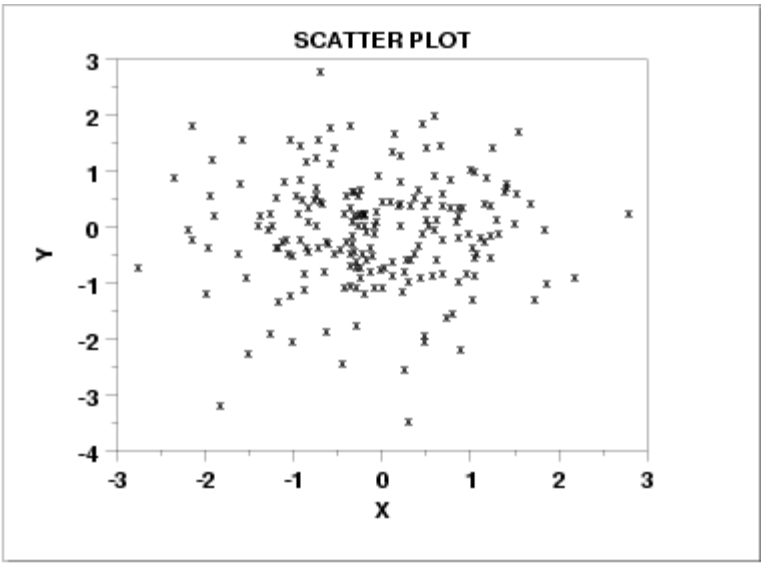
In both cases, the resulting plot is referred to as a scatter plot, although the former (discrete and disconnected) is the author's personal preference since nothing makes it onto the screen except the data--there are no interpolative artifacts to

bias the interpretation.

Related Techniques	Run Sequence Plot Box Plot Block Plot
Case Study	The scatter plot is demonstrated in the load cell calibration data case study.
Software	Scatter plots are a fundamental technique that should be available in any general purpose statistical software program. Scatter plots are also available in most graphics and spreadsheet programs as well.

1.3.3.26.1. Scatter Plot: No Relationship

Scatter Plot with No Relationship

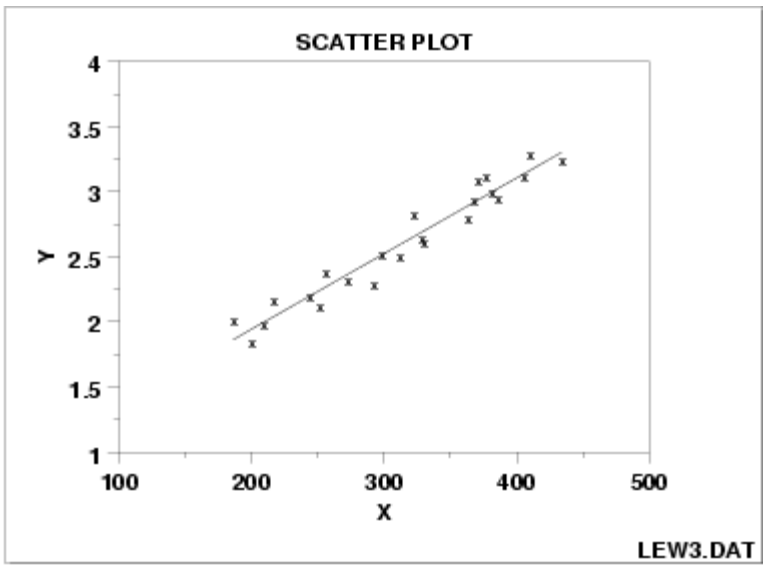


Discussion

Note in the plot above how for a given value of X (say X = 0.5), the corresponding values of Y range all over the place from Y = -2 to Y = +2. The same is true for other values of X. This lack of predictablility in determining Y from a given value of X, and the associated amorphous, non-structured appearance of the scatter plot leads to the summary conclusion: no relationship.

1.3.3.26.2. Scatter Plot: Strong Linear (positive correlation) Relationship

Scatter Plot Showing Strong Positive Linear Correlation

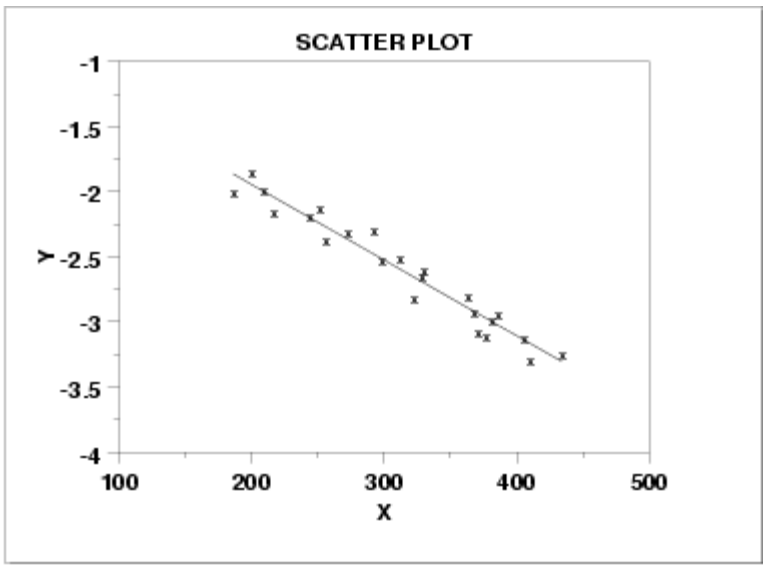


Discussion

Note in the plot above how a straight line comfortably fits through the data; hence a linear relationship exists. The scatter about the line is quite small, so there is a strong linear relationship. The slope of the line is positive (small values of *X* correspond to small values of *Y*; large values of *X* correspond to large values of *Y*), so there is a positive co-relation (that is, a positive correlation) between *X* and *Y*.

1.3.3.26.3. Scatter Plot: Strong Linear (negative correlation) Relationship

Scatter Plot Showing a Strong Negative Correlation

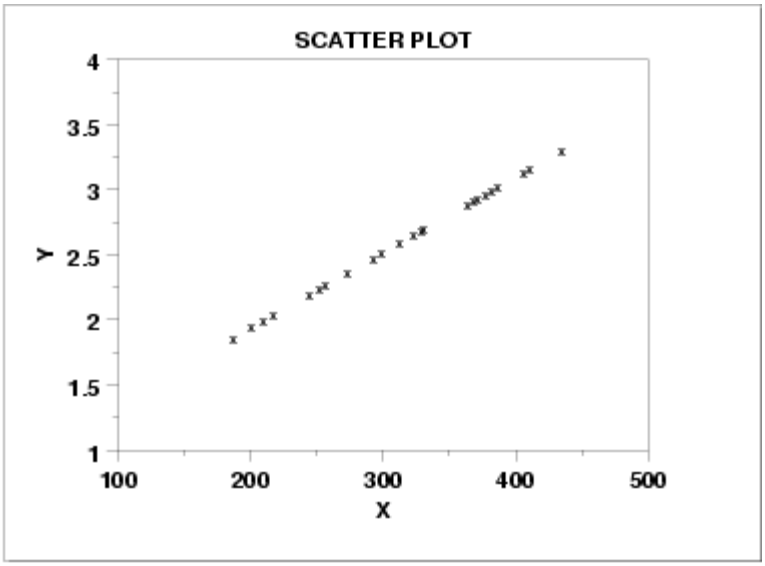


Discussion

Note in the plot above how a straight line comfortably fits through the data; hence there is a linear relationship. The scatter about the line is quite small, so there is a strong linear relationship. The slope of the line is negative (small values of X correspond to large values of Y; large values of X correspond to small values of Y), so there is a negative co-relation (that is, a negative correlation) between X and Y.

1.3.3.26.4. Scatter Plot: Exact Linear (positive correlation) Relationship

Scatter Plot Showing an Exact Linear Relationship

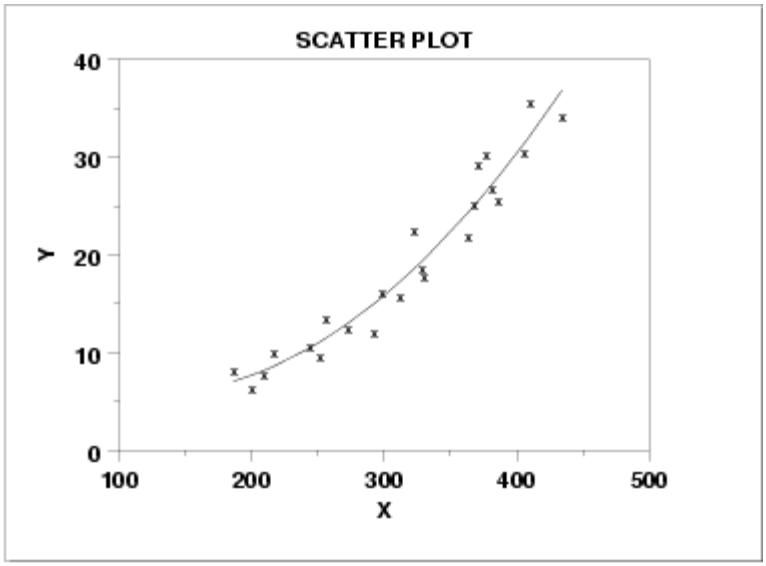


Discussion

Note in the plot above how a straight line comfortably fits through the data; hence there is a linear relationship. The scatter about the line is zero--there is perfect predictability between *X* and *Y*), so there is an exact linear relationship. The slope of the line is positive (small values of *X* correspond to small values of *Y*; large values of *X* correspond to large values of *Y*), so there is a positive co-relation (that is, a positive correlation) between *X* and *Y*.

1.3.3.26.5. Scatter Plot: Quadratic Relationship

Scatter Plot Showing Quadratic Relationship



Discussion

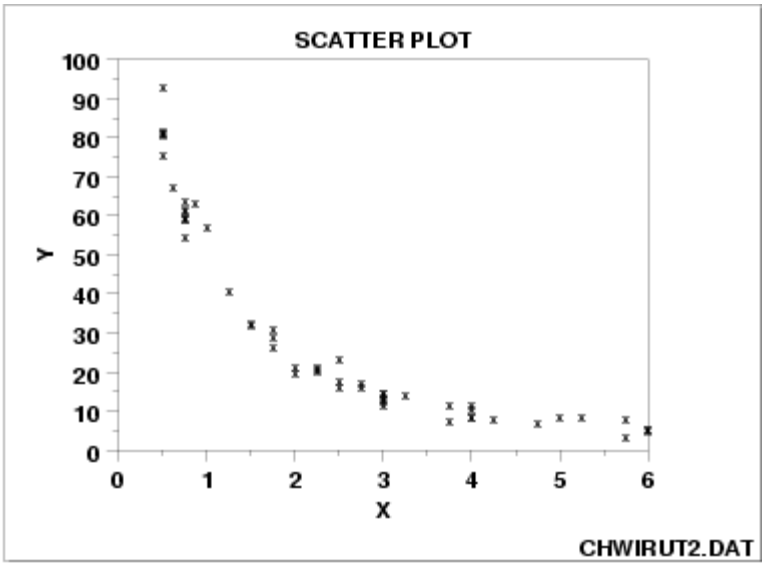
Note in the plot above how no imaginable simple straight line could ever adequately describe the relationship between *X* and *Y*--a curved (or curvilinear, or non-linear) function is needed. The simplest such curvilinear function is a quadratic model

$$Y_i = A + BX_i + CX_i^2 + E_i$$

for some A, B, and C. Many other curvilinear functions are possible, but the data analysis principle of parsimony suggests that we try fitting a quadratic function first.

1.3.3.26.6. Scatter Plot: Exponential Relationship

Scatter Plot Showing Exponential Relationship



Discussion

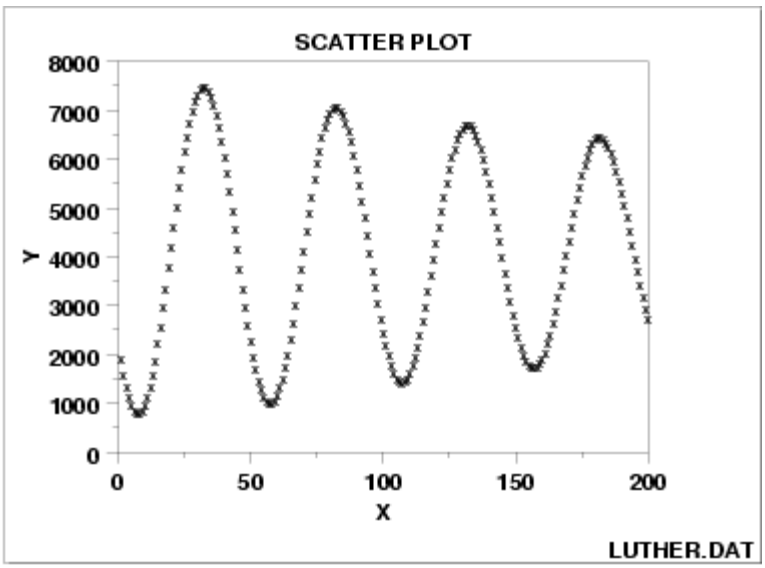
Note that a simple straight line is grossly inadequate in describing the relationship between *X* and *Y*. A quadratic model would prove lacking, especially for large values of *X*. In this example, the large values of *X* correspond to nearly constant values of *Y*, and so a non-linear function beyond the quadratic is needed. Among the many other non-linear functions available, one of the simpler ones is the exponential model

$$Y_i = A + Be^{CX_i} + E_i$$

for some A, B, and C. In this case, an exponential function would, in fact, fit well, and so one is led to the summary conclusion of an exponential relationship.

1.3.3.26.7. Scatter Plot: Sinusoidal Relationship (damped)

Scatter Plot Showing a Sinusoidal Relationship



Discussion

The complex relationship between *X* and *Y* appears to be basically oscillatory, and so one is naturally drawn to the trigonometric sinusoidal model:

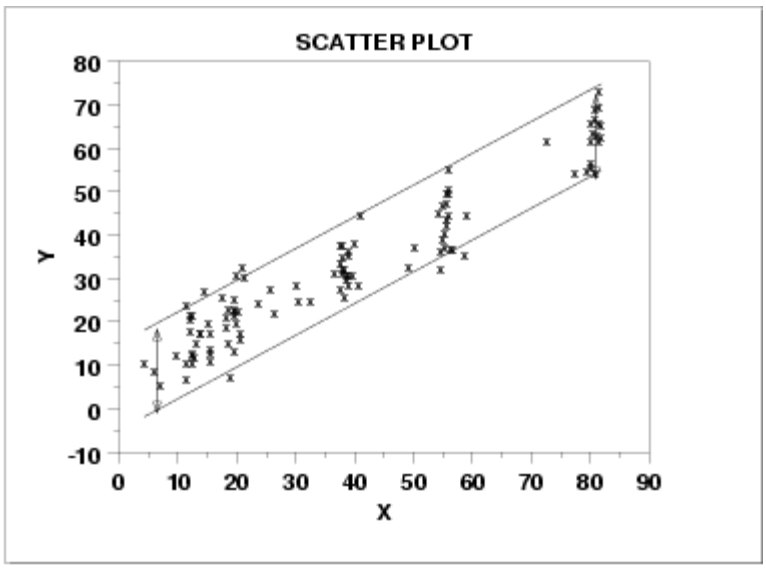
$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

Closer inspection of the scatter plot reveals that the amount of swing (the amplitude α in the model) does not appear to be constant but rather is decreasing (damping) as *X* gets large. We thus would be led to the conclusion: damped sinusoidal relationship, with the simplest corresponding model being

$$Y_i = C + \left(B_0 + B_1 * t_i\right) \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

1.3.3.26.8. Scatter Plot: Variation of Y Does Not Depend on X (homoscedastic)

Scatter Plot Showing Homoscedastic Variability



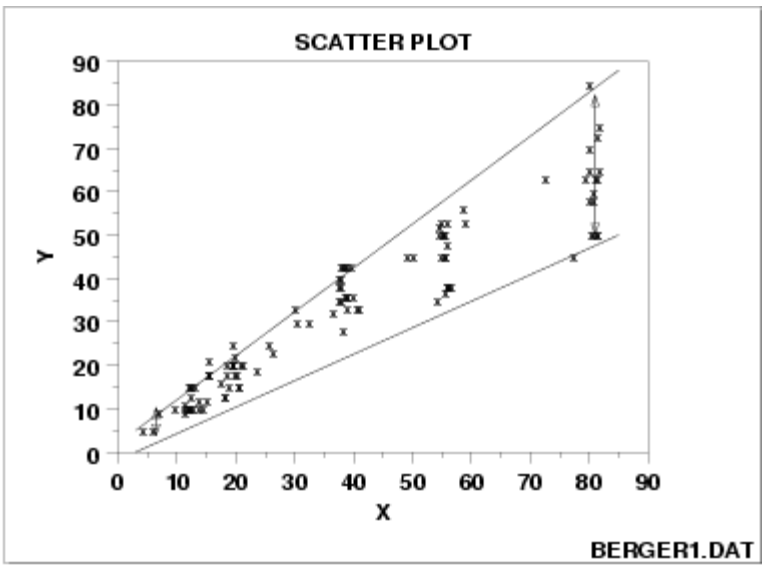
Discussion

This scatter plot reveals a linear relationship between *X* and *Y*: for a given value of *X*, the predicted value of *Y* will fall on a line. The plot further reveals that the variation in *Y* about the predicted value is about the same (+- 10 units), regardless of the value of *X*. Statistically, this is referred to as homoscedasticity. Such homoscedasticity is very important as it is an underlying assumption for regression, and its violation leads to parameter estimates with inflated variances. If the data are homoscedastic, then the usual regression estimates can be used. If the data are not homoscedastic, then the estimates can be improved using weighting procedures as shown in the next example.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.3. Graphical Techniques: Alphabetic
1.3.3.26. Scatter Plot

1.3.3.26.9. Scatter Plot: Variation of Y Does Depend on X (heteroscedastic)

Scatter Plot Showing Heteroscedastic Variability



Discussion

This scatter plot reveals an approximate linear relationship between *X* and *Y*, but more importantly, it reveals a statistical condition referred to as heteroscedasticity (that is, nonconstant variation in *Y* over the values of *X*). For a heteroscedastic data set, the variation in *Y* differs depending on the value of *X*. In this example, small values of *X* yield small scatter in *Y* while large values of *X* result in large scatter in *Y*.

Heteroscedasticity complicates the analysis somewhat, but its effects can be overcome by:

1. proper weighting of the data with noisier data being weighted less, or by

2. performing a *Y* variable transformation to achieve homoscedasticity. The Box-Cox normality plot can help determine a suitable transformation.

Impact of Ignoring Unequal

Fortunately, unweighted regression analyses on heteroscedastic data produce estimates of the coefficients that are unbiased. However, the coefficients will not be as

Variability in the Data

precise as they would be with proper weighting.

Note further that if heteroscedasticity does exist, it is frequently useful to plot and model the local variation $var(Y_i|X_i)$ as a function of *X*, as in $var(Y_i|X_i) = g(X_i)$. This modeling has two advantages:

1. it provides additional insight and understanding as to how the response *Y* relates to *X*; and

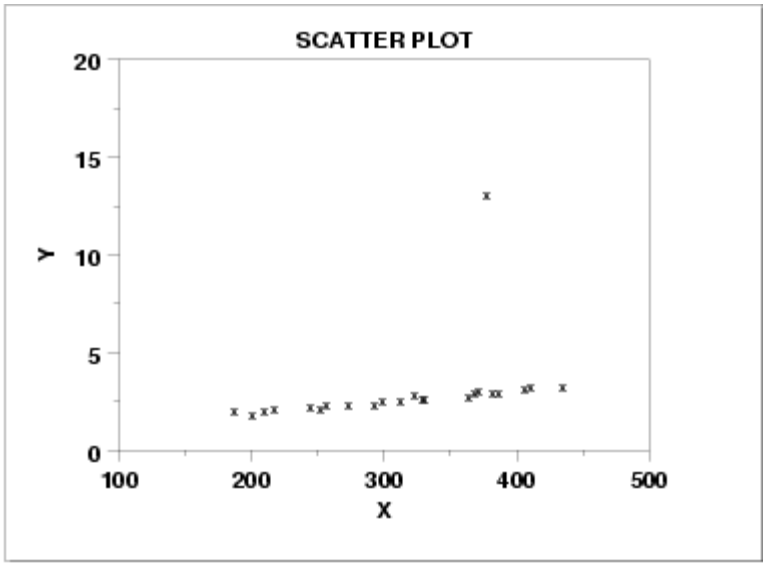
2. it provides a convenient means of forming weights for a weighted regression by simply using

$$w_i = W(Y_i|X_i) = \frac{1}{Var(Y_i|X_i)} = \frac{1}{g(X_i)}$$

The topic of non-constant variation is discussed in some detail in the process modeling chapter.

1.3.3.26.10. Scatter Plot: Outlier

Scatter Plot Showing Outliers



Discussion The scatter plot here reveals

1. a basic linear relationship between *X* and *Y* for most of the data, and
2. a single outlier (at *X* = 375).

An outlier is defined as a data point that emanates from a different model than do the rest of the data. The data here appear to come from a linear model with a given slope and variation except for the outlier which appears to have been generated from some other model.

Outlier detection is important for effective modeling. Outliers should be excluded from such model fitting. If all the data here are included in a linear regression, then the fitted model will be poor virtually everywhere. If the outlier is omitted from the fitting process, then the resulting fit will be excellent almost everywhere (for all points except the outlying point).

1.3.3.26.11. Scatterplot Matrix

Purpose: Check Pairwise Relationships Between Variables

Given a set of variables $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_k}$, the scatterplot matrix contains all the pairwise scatter plots of the variables on a single page in a matrix format. That is, if there are k variables, the scatterplot matrix will have k rows and k columns and the ith row and jth column of this matrix is a plot of $\boldsymbol{X_i}$ versus $\boldsymbol{X_j}$.

Although the basic concept of the scatterplot matrix is simple, there are numerous alternatives in the details of the plots.

1. The diagonal plot is simply a 45-degree line since we are plotting $\boldsymbol{X_i}$ versus $\boldsymbol{X_i}$. Although this has some usefulness in terms of showing the univariate distribution of the variable, other alternatives are common. Some users prefer to use the diagonal to print the variable label. Another alternative is to plot the univariate histogram on the diagonal. Alternatively, we could simply leave the diagonal blank.

2. Since $\boldsymbol{X_i}$ versus $\boldsymbol{X_j}$ is equivalent to $\boldsymbol{X_j}$ versus $\boldsymbol{X_i}$ with the axes reversed, some prefer to omit the plots below the diagonal.

3. It can be helpful to overlay some type of fitted curve on the scatter plot. Although a linear or quadratic fit can be used, the most common alternative is to overlay a lowess curve.

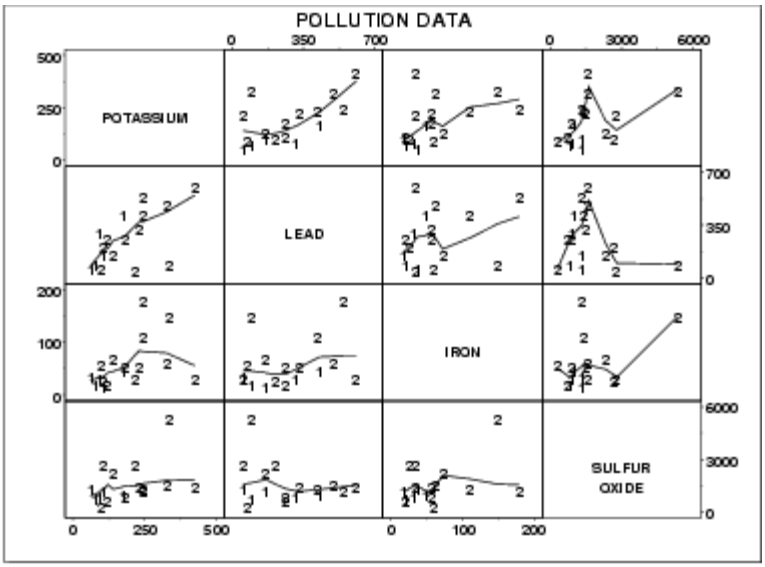
4. Due to the potentially large number of plots, it can be somewhat tricky to provide the axes labels in a way that is both informative and visually pleasing. One alternative that seems to work well is to provide axis labels on alternating rows and columns. That is, row one will have tic marks and axis labels on the left vertical axis for the first plot only while row two will have the tic marks and axis labels for the right vertical axis for the last plot in the row only. This alternating pattern continues for the remaining rows.

A similar pattern is used for the columns and the horizontal axes labels. Another alternative is to put the minimum and maximum scale value in the diagonal plot with the variable name.

5. Some analysts prefer to connect the scatter plots. Others prefer to leave a little gap between each plot.

6. Although this plot type is most commonly used for scatter plots, the basic concept is both simple and powerful and extends easily to other plot formats that involve pairwise plots such as the quantile-quantile plot and the bihistogram.

Sample Plot



This sample plot was generated from pollution data collected by NIST chemist Lloyd Currie.

There are a number of ways to view this plot. If we are primarily interested in a particular variable, we can scan the row and column for that variable. If we are interested in finding the strongest relationship, we can scan all the plots and then determine which variables are related.

Definition

Given *k* variables, scatter plot matrices are formed by creating *k* rows and *k* columns. Each row and column defines a single scatter plot

The individual plot for row *i* and column *j* is defined as

- Vertical axis: Variable $\boldsymbol{X_i}$
- Horizontal axis: Variable $\boldsymbol{X_j}$

Questions

The scatterplot matrix can provide answers to the following questions:

1. Are there pairwise relationships between the variables?
2. If there are relationships, what is the nature of these relationships?
3. Are there outliers in the data?
4. Is there clustering by groups in the data?

Linking and Brushing

The scatterplot matrix serves as the foundation for the concepts of linking and brushing.

By linking, we mean showing how a point, or set of points, behaves in each of the plots. This is accomplished by highlighting these points in some fashion. For example, the highlighted points could be drawn as a filled circle while the remaining points could be drawn as unfilled circles. A typical application of this would be to show how an outlier shows up in each of the individual pairwise plots. Brushing extends this concept a bit further. In brushing, the points to be highlighted are interactively selected by a mouse and the scatterplot matrix is dynamically updated (ideally in real time). That is, we can select a rectangular region of points in one plot and see how those points are reflected in the other plots. Brushing is discussed in detail by Becker, Cleveland, and Wilks in the paper *"Dynamic Graphics for Data Analysis"* (Cleveland and McGill, 1988).

Related Techniques

Star plot
Scatter plot
Conditioning plot
Locally weighted least squares

Software

Scatterplot matrices are becoming increasingly common in general purpose statistical software programs. If a software program does not generate scatterplot matrices, but it does provide multiple plots per page and scatter plots, it should be possible to write a macro to generate a scatterplot matrix. Brushing is available in a few of the general purpose statistical software programs that emphasize graphical approaches.

1.3.3.26.12. Conditioning Plot

Purpose: Check pairwise relationship between two variables conditional on a third variable

A conditioning plot, also known as a coplot or subset plot, is a plot of two variables conditional on the value of a third variable (called the conditioning variable). The conditioning variable may be either a variable that takes on only a few discrete values or a continuous variable that is divided into a limited number of subsets.

One limitation of the scatterplot matrix is that it cannot show interaction effects with another variable. This is the strength of the conditioning plot. It is also useful for displaying scatter plots for groups in the data. Although these groups can also be plotted on a single plot with different plot symbols, it can often be visually easier to distinguish the groups using the conditioning plot.

Although the basic concept of the conditioning plot matrix is simple, there are numerous alternatives in the details of the plots.

1. It can be helpful to overlay some type of fitted curve on the scatter plot. Although a linear or quadratic fit can be used, the most common alternative is to overlay a lowess curve.

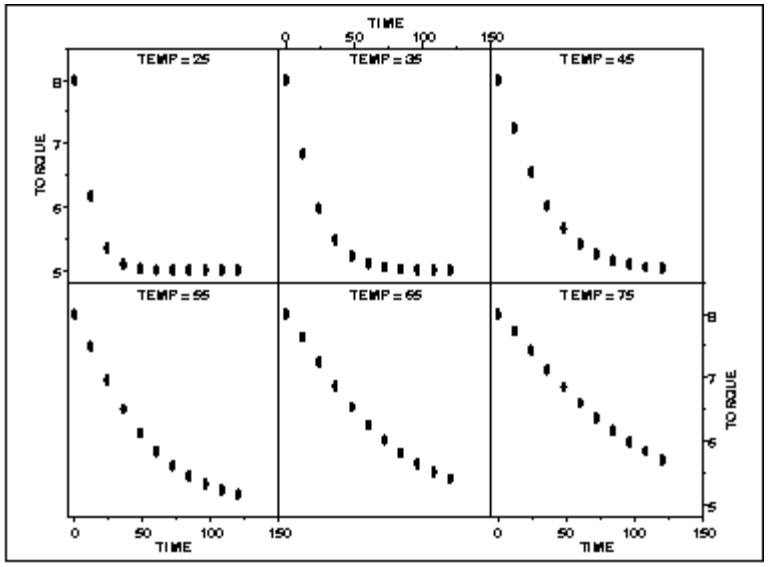
2. Due to the potentially large number of plots, it can be somewhat tricky to provide the axis labels in a way that is both informative and visually pleasing. One alternative that seems to work well is to provide axis labels on alternating rows and columns. That is, row one will have tic marks and axis labels on the left vertical axis for the first plot only while row two will have the tic marks and axis labels for the right vertical axis for the last plot in the row only. This alternating pattern continues for the remaining rows. A similar pattern is used for the columns and the horizontal axis labels. Note that this approach only works if the axes limits are fixed to common values for all of the plots.

3. Some analysts prefer to connect the scatter plots. Others prefer to leave a little gap between each plot. Alternatively, each plot can have its own labeling with

the plots not connected.

4. Although this plot type is most commonly used for scatter plots, the basic concept is both simple and powerful and extends easily to other plot formats.

Sample Plot



In this case, temperature has six distinct values. We plot torque versus time for each of these temperatures. This example is discussed in more detail in the process modeling chapter.

Definition

Given the variables ***X***, ***Y***, and ***Z***, the conditioning plot is formed by dividing the values of ***Z*** into *k* groups. There are several ways that these groups may be formed. There may be a natural grouping of the data, the data may be divided into several equal sized groups, the grouping may be determined by clusters in the data, and so on. The page will be divided into *n* rows and *c* columns where $nc \geq k$. Each row and column defines a single scatter plot.

The individual plot for row *i* and column *j* is defined as

- Vertical axis: Variable ***Y***
- Horizontal axis: Variable ***X***

where only the points in the group corresponding to the *i*th row and *j*th column are used.

Questions

The conditioning plot can provide answers to the following questions:

1. Is there a relationship between two variables?
2. If there is a relationship, does the nature of the relationship depend on the value of a third variable?
3. Are groups in the data similar?
4. Are there outliers in the data?

Related Techniques	Scatter plot Scatterplot matrix Locally weighted least squares
Software	Scatter plot matrices are becoming increasingly common in general purpose statistical software programs, including. If a software program does not generate conditioning plots, but it does provide multiple plots per page and scatter plots, it should be possible to write a macro to generate a conditioning plot.

1.3.3.27. Spectral Plot

Purpose: Examine Cyclic Structure

A spectral plot (Jenkins and Watts 1968 or Bloomfield 1976) is a graphical technique for examining cyclic structure in the frequency domain. It is a smoothed Fourier transform of the autocovariance function.

The frequency is measured in cycles per unit time where unit time is defined to be the distance between 2 points. A frequency of 0 corresponds to an infinite cycle while a frequency of 0.5 corresponds to a cycle of 2 data points. Equi-spaced time series are inherently limited to detecting frequencies between 0 and 0.5.

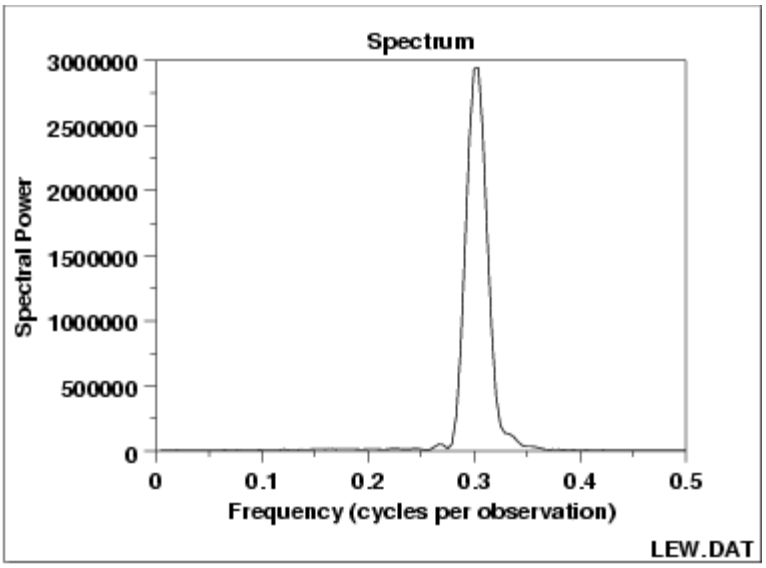
Trends should typically be removed from the time series before applying the spectral plot. Trends can be detected from a run sequence plot. Trends are typically removed by differencing the series or by fitting a straight line (or some other polynomial curve) and applying the spectral analysis to the residuals.

Spectral plots are often used to find a starting value for the frequency, ω, in the sinusoidal model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

See the beam deflection case study for an example of this.

Sample Plot



This spectral plot shows one dominant frequency of approximately 0.3 cycles per observation.

Definition: Variance Versus Frequency

The spectral plot is formed by:

- Vertical axis: Smoothed variance (power)
- Horizontal axis: Frequency (cycles per observation)

The computations for generating the smoothed variances can be involved and are not discussed further here. The details can be found in the Jenkins and Bloomfield references and in most texts that discuss the frequency analysis of time series.

Questions

The spectral plot can be used to answer the following questions:

1. How many cyclic components are there?
2. Is there a dominant cyclic frequency?
3. If there is a dominant cyclic frequency, what is it?

Importance Check Cyclic Behavior of Time Series

The spectral plot is the primary technique for assessing the cyclic nature of univariate time series in the frequency domain. It is almost always the second plot (after a run sequence plot) generated in a frequency domain analysis of a time series.

Examples

1. Random (= White Noise)
2. Strong autocorrelation and autoregressive model
3. Sinusoidal model

Related Techniques

Autocorrelation Plot
Complex Demodulation Amplitude Plot
Complex Demodulation Phase Plot

Case Study

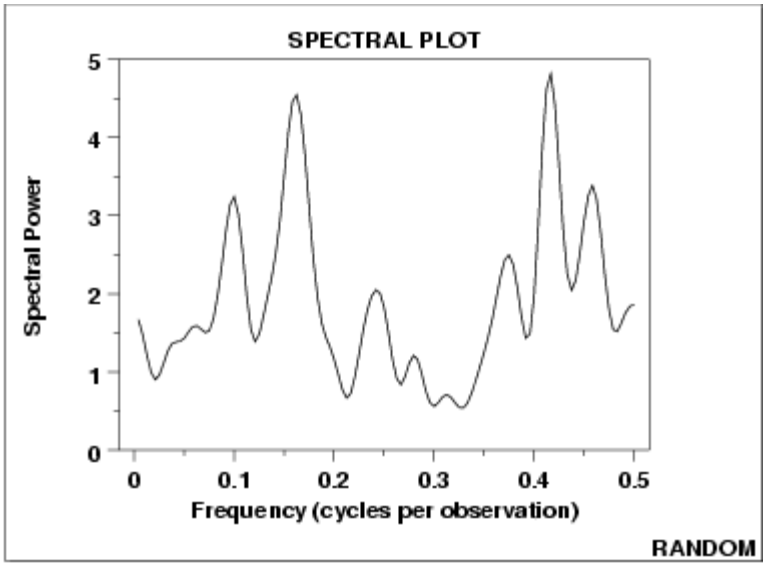
The spectral plot is demonstrated in the beam deflection data case study.

Software

Spectral plots are a fundamental technique in the frequency analysis of time series. They are available in many general purpose statistical software programs.

1.3.3.27.1. Spectral Plot: Random Data

Spectral Plot of 200 Normal Random Numbers



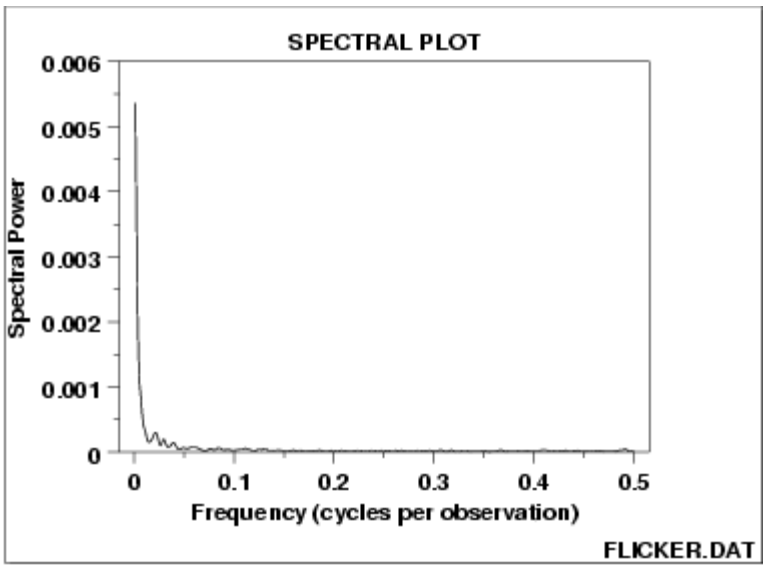
Conclusions We can make the following conclusions from the above plot.

1. There are no dominant peaks.
2. There is no identifiable pattern in the spectrum.
3. The data are random.

Discussion For random data, the spectral plot should show no dominant peaks or distinct pattern in the spectrum. For the sample plot above, there are no clearly dominant peaks and the peaks seem to fluctuate at random. This type of appearance of the spectral plot indicates that there are no significant cyclic patterns in the data.

1.3.3.27.2. Spectral Plot: Strong Autocorrelation and Autoregressive Model

Spectral Plot for Random Walk Data



Conclusions

We can make the following conclusions from the above plot.

1. Strong dominant peak near zero.
2. Peak decays rapidly towards zero.
3. An autoregressive model is an appropriate model.

Discussion

This spectral plot starts with a dominant peak near zero and rapidly decays to zero. This is the spectral plot signature of a process with strong positive autocorrelation. Such processes are highly non-random in that there is high association between an observation and a succeeding observation. In short, if you know Y_i you can make a strong guess as to what Y_{i+1} will be.

Recommended Next Step

The next step would be to determine the parameters for the autoregressive model:

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Such estimation can be done by linear regression or by

fitting a Box-Jenkins autoregressive (AR) model.

The residual standard deviation for this autoregressive model will be much smaller than the residual standard deviation for the default model

$$Y_i = A_0 + E_i$$

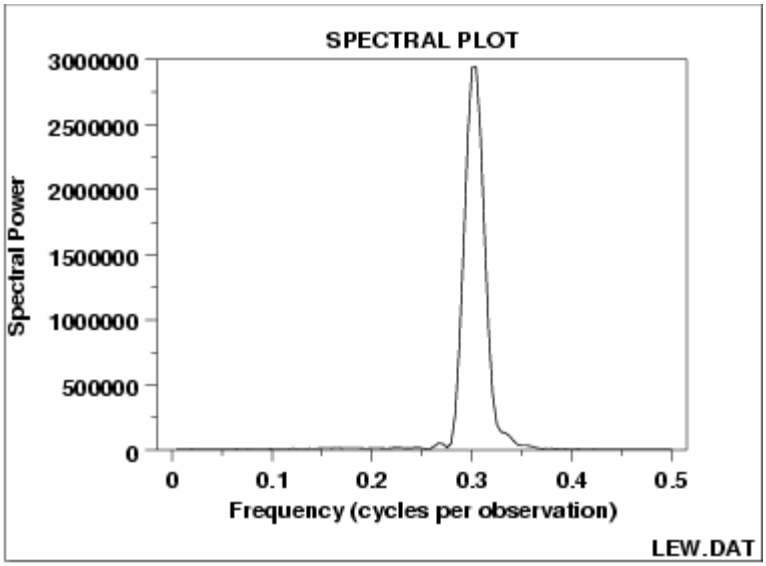
Then the system should be reexamined to find an explanation for the strong autocorrelation. Is it due to the

1. phenomenon under study; or
2. drifting in the environment; or
3. contamination from the data acquisition system (DAS)?

Oftentimes the source of the problem is item (3) above where contamination and carry-over from the data acquisition system result because the DAS does not have time to electronically recover before collecting the next data point. If this is the case, then consider slowing down the sampling rate to re-achieve randomness.

1.3.3.27.3. Spectral Plot: Sinusoidal Model

Spectral Plot for Sinusoidal Model



Conclusions

We can make the following conclusions from the above plot.

1. There is a single dominant peak at approximately 0.3.
2. There is an underlying single-cycle sinusoidal model.

Discussion

This spectral plot shows a single dominant frequency. This indicates that a single-cycle sinusoidal model might be appropriate.

If one were to naively assume that the data represented by the graph could be fit by the model

$$Y_i = A_0 + E_i$$

and then estimate the constant by the sample mean, the analysis would be incorrect because

- the sample mean is biased;
- the confidence interval for the mean, which is valid only for random data, is meaningless and too small.

On the other hand, the choice of the proper model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

where α is the amplitude, ω is the frequency (between 0 and .5 cycles per observation), and ϕ is the phase can be fit by non-linear least squares. The beam deflection data case study demonstrates fitting this type of model.

Recommended Next Steps

The recommended next steps are to:

1. Estimate the frequency from the spectral plot. This will be helpful as a starting value for the subsequent non-linear fitting. A complex demodulation phase plot can be used to fine tune the estimate of the frequency before performing the non-linear fit.

2. Do a complex demodulation amplitude plot to obtain an initial estimate of the amplitude and to determine if a constant amplitude is justified.

3. Carry out a non-linear fit of the model

$$Y_i = C + \alpha \sin\left(2\pi\omega t_i + \phi\right) + E_i$$

1.3.3.28. Standard Deviation Plot

Purpose: Detect Changes in Scale Between Groups

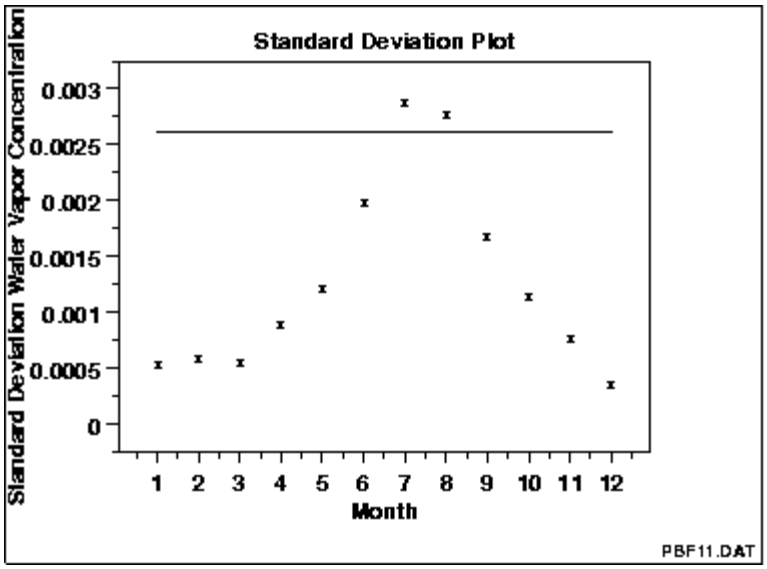
Standard deviation plots are used to see if the standard deviation varies between different groups of the data. The grouping is determined by the analyst. In most cases, the data provide a specific grouping variable. For example, the groups may be the levels of a factor variable. In the sample plot below, the months of the year provide the grouping.

Standard deviation plots can be used with ungrouped data to determine if the standard deviation is changing over time. In this case, the data are broken into an arbitrary number of equal-sized groups. For example, a data series with 400 points can be divided into 10 groups of 40 points each. A standard deviation plot can then be generated with these groups to see if the standard deviation is increasing or decreasing over time.

Although the standard deviation is the most commonly used measure of scale, the same concept applies to other measures of scale. For example, instead of plotting the standard deviation of each group, the median absolute deviation or the average absolute deviation might be plotted instead. This might be done if there were significant outliers in the data and a more robust measure of scale than the standard deviation was desired.

Standard deviation plots are typically used in conjunction with mean plots. The mean plot would be used to check for shifts in location while the standard deviation plot would be used to check for shifts in scale.

Sample Plot



This sample standard deviation plot shows

1. there is a shift in variation;
2. greatest variation is during the summer months.

Definition: Group Standard Deviations Versus Group ID

Standard deviation plots are formed by:

- Vertical axis: Group standard deviations
- Horizontal axis: Group identifier

A reference line is plotted at the overall standard deviation.

Questions

The standard deviation plot can be used to answer the following questions.

1. Are there any shifts in variation?
2. What is the magnitude of the shifts in variation?
3. Is there a distinct pattern in the shifts in variation?

Importance: Checking Assumptions

A common assumption in 1-factor analyses is that of equal variances. That is, the variance is the same for different levels of the factor variable. The standard deviation plot provides a graphical check for that assumption. A common assumption for univariate data is that the variance is constant. By grouping the data into equi-sized intervals, the standard deviation plot can provide a graphical test of this assumption.

Related Techniques

Mean Plot
DOE Standard Deviation Plot

Software

Most general purpose statistical software programs do not support a standard deviation plot. However, if the statistical program can generate the standard deviation for a group, it should be feasible to write a macro to generate this plot.

1.3.3.29. Star Plot

Purpose: Display Multivariate Data

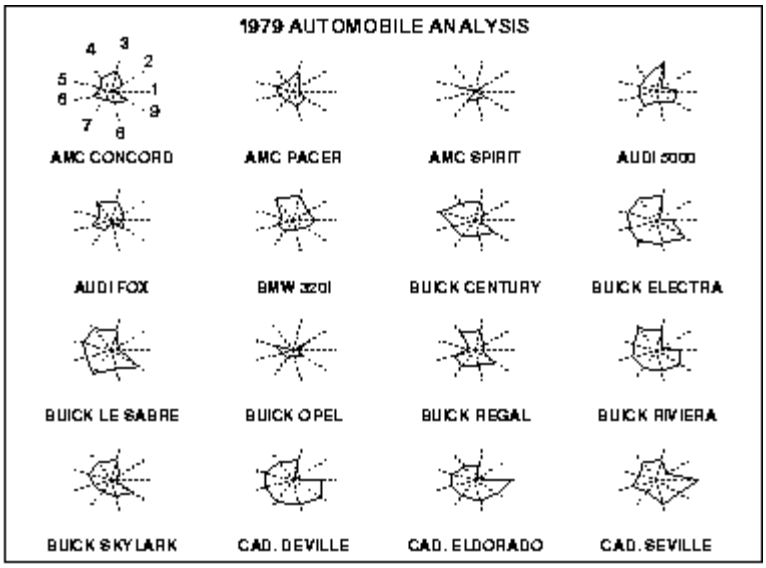
The star plot (Chambers 1983) is a method of displaying multivariate data. Each star represents a single observation. Typically, star plots are generated in a multi-plot format with many stars on each page and each star representing one observation.

Star plots are used to examine the relative values for a single data point (e.g., point 3 is large for variables 2 and 4, small for variables 1, 3, 5, and 6) and to locate similar points or dissimilar points.

Sample Plot

The plot below contains the star plots of 16 cars. The data file actually contains 74 cars, but we restrict the plot to what can reasonably be shown on one page. The variable list for the sample star plot is

1 Price
2 Mileage (MPG)
3 1978 Repair Record (1 = Worst, 5 = Best)
4 1977 Repair Record (1 = Worst, 5 = Best)
5 Headroom
6 Rear Seat Room
7 Trunk Space
8 Weight
9 Length



We can look at these plots individually or we can use them to identify clusters of cars with similar features. For example, we can look at the star plot of the Cadillac Seville and see that it is one of the most expensive cars, gets below average (but not among the worst) gas mileage, has an average repair record, and has average-to-above-average roominess and size. We can then compare the Cadillac models (the last three plots) with the AMC models (the first three plots). This comparison shows distinct patterns. The AMC models tend to be inexpensive, have below average gas mileage, and are small in both height and weight and in roominess. The Cadillac models are expensive, have poor gas mileage, and are large in both size and roominess.

Definition

The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables. The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance and the origin of the name of this plot.

Questions

The star plot can be used to answer the following questions:

1. What variables are dominant for a given observation?
2. Which observations are most similar, i.e., are there clusters of observations?
3. Are there outliers?

Weakness in Technique

Star plots are helpful for small-to-moderate-sized multivariate data sets. Their primary weakness is that their effectiveness is limited to data sets with less than a few hundred points. After that, they tend to be overwhelming.

Graphical techniques suited for large data sets are discussed

by Scott.

Related Techniques

Alternative ways to plot multivariate data are discussed in Chambers, du Toit, and Everitt.

Software

Star plots are available in some general purpose statistical software progams.

1.3.3.30. Weibull Plot

Purpose: Graphical Check To See If Data Come From a Population That Would Be Fit by a Weibull Distribution

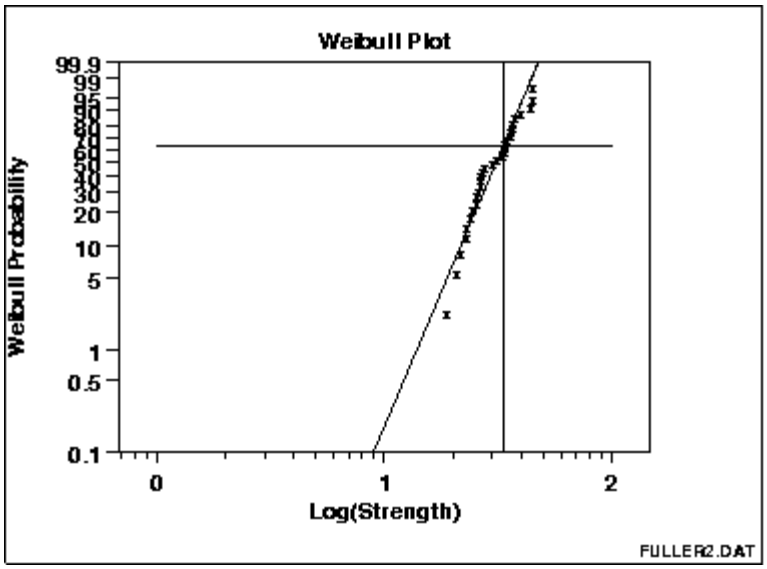
The Weibull plot (Nelson 1982) is a graphical technique for determining if a data set comes from a population that would logically be fit by a 2-parameter Weibull distribution (the location is assumed to be zero).

The Weibull plot has special scales that are designed so that if the data do in fact follow a Weibull distribution, the points will be linear (or nearly linear). The least squares fit of this line yields estimates for the shape and scale parameters of the Weibull distribution (the location is assumed to be zero).

Specifically, the shape parameter is the reciprocal of the slope of the fitted line and the scale parameter is the exponent of the intercept of the fitted line.

The Weibull distribution also has the property that the scale parameter falls at the 63.2% point irrespective of the value of the shape parameter. The plot shows a horizontal line at this 63.2% point and a vertical line where the horizontal line intersects the least squares fitted line. This vertical line shows the value of scale parameter.

Sample Plot



This Weibull plot shows that:

1. the assumption of a Weibull distribution is

reasonable;
2. the scale parameter estimate is computed to be 33.32;
3. the shape parameter estimate is computed to be 5.28; and
4. there are no outliers.

Note that the values on the x-axis ("0", "1", and "2") are the exponents. These actually denote the value $10^0 = 1$, $10^1 = 10$, and $10^2 = 100$.

Definition: Weibull Cumulative Probability Versus LN(Ordered Response)

The Weibull plot is formed by:

- Vertical axis: Weibull cumulative probability expressed as a percentage
- Horizontal axis: ordered failure times (in a LOG10 scale)

The vertical scale is $\ln(-\ln(1-p))$ where $p=(i-0.3)/(n+0.4)$ and i is the rank of the observation. This scale is chosen in order to linearize the resulting plot for Weibull data.

Questions

The Weibull plot can be used to answer the following questions:

1. Do the data follow a 2-parameter Weibull distribution?
2. What is the best estimate of the shape parameter for the 2-parameter Weibull distribution?
3. What is the best estimate of the scale (= variation) parameter for the 2-parameter Weibull distribution?

Importance: Check Distributional Assumptions

Many statistical analyses, particularly in the field of reliability, are based on the assumption that the data follow a Weibull distribution. If the analysis assumes the data follow a Weibull distribution, it is important to verify this assumption and, if verified, find good estimates of the Weibull parameters.

Related Techniques

Weibull Probability Plot
Weibull PPCC Plot
Weibull Hazard Plot

The Weibull probability plot (in conjunction with the Weibull PPCC plot), the Weibull hazard plot, and the Weibull plot are all similar techniques that can be used for assessing the adequacy of the Weibull distribution as a model for the data, and additionally providing estimation for the shape, scale, or location parameters.

The Weibull hazard plot and Weibull plot are designed to handle censored data (which the Weibull probability plot does not).

Case Study The Weibull plot is demonstrated in the fatigue life of aluminum alloy specimens case study.

Software Weibull plots are generally available in statistical software programs that are designed to analyze reliability data.

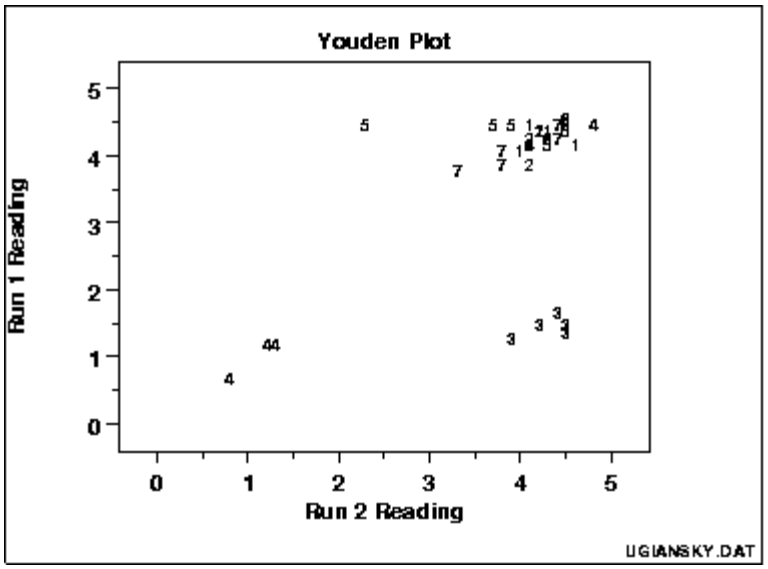
1.3.3.31. Youden Plot

Purpose: Interlab Comparisons

Youden plots are a graphical technique for analyzing interlab data when each lab has made two runs on the same product or one run on two different products.

The Youden plot is a simple but effective method for comparing both the within-laboratory variability and the between-laboratory variability.

Sample Plot



This plot shows:

1. Not all labs are equivalent.
2. Lab 4 is biased low.
3. Lab 3 has within-lab variability problems.
4. Lab 5 has an outlying run.

Definition: Response 1 Versus Response 2 Coded by Lab

Youden plots are formed by:

1. Vertical axis: Response variable 1 (i.e., run 1 or product 1 response value)
2. Horizontal axis: Response variable 2 (i.e., run 2 or product 2 response value)

In addition, the plot symbol is the lab id (typically an integer from 1 to *k* where *k* is the number of labs).

Sometimes a 45-degree reference line is drawn. Ideally, a lab generating two runs of the same product should produce reasonably similar results. Departures from this reference line indicate inconsistency from the lab. If two different products are being tested, then a 45-degree line may not be appropriate. However, if the labs are consistent, the points should lie near some fitted straight line.

Questions

The Youden plot can be used to answer the following questions:

1. Are all labs equivalent?
2. What labs have between-lab problems (reproducibility)?
3. What labs have within-lab problems (repeatability)?
4. What labs are outliers?

Importance

In interlaboratory studies or in comparing two runs from the same lab, it is useful to know if consistent results are generated. Youden plots should be a routine plot for analyzing this type of data.

DOE Youden Plot

The DOE Youden plot is a specialized Youden plot used in the design of experiments. In particular, it is useful for full and fractional designs.

Related Techniques

Scatter Plot

Software

The Youden plot is essentially a scatter plot, so it should be feasible to write a macro for a Youden plot in any general purpose statistical program that supports scatter plots.

1.3.3.31.1. DOE Youden Plot

DOE Youden Plot: Introduction

The DOE (Design of Experiments) Youden plot is a specialized Youden plot used in the analysis of full and fractional experiment designs. In particular, it is used in conjunction with the Yates algorithm. These designs may have a low level, coded as "-1" or "-", and a high level, coded as "+1" or "+", for each factor. In addition, there can optionally be one or more center points. Center points are at the midpoint between the low and high levels for each factor and are coded as "0".

The Yates agorithm and the the DOE Youden plot only use the "-1" and "+1" points. The Yates agorithm is used to estimate factor effects. The DOE Youden plot can be used to help determine the approriate model to based on the effect estimates from the Yates algorithm.

Construction of DOE Youden Plot

The following are the primary steps in the construction of the DOE Youden plot.

1. For a given factor or interaction term, compute the mean of the response variable for the low level of the factor and for the high level of the factor. Any center points are omitted from the computation.

2. Plot the point where the y-coordinate is the mean for the high level of the factor and the x-coordinate is the mean for the low level of the factor. The character used for the plot point should identify the factor or interaction term (e.g., "1" for factor 1, "13" for the interaction between factors 1 and 3).

3. Repeat steps 1 and 2 for each factor and interaction term of the data.

The high and low values of the interaction terms are obtained by multiplying the corresponding values of the main level factors. For example, the interaction term X_{13} is obtained by multiplying the values for X_1 with the corresponding values of X_3. Since the values for X_1 and X_3 are either "-1" or "+1", the resulting values for X_{13} are also either "-1" or "+1".

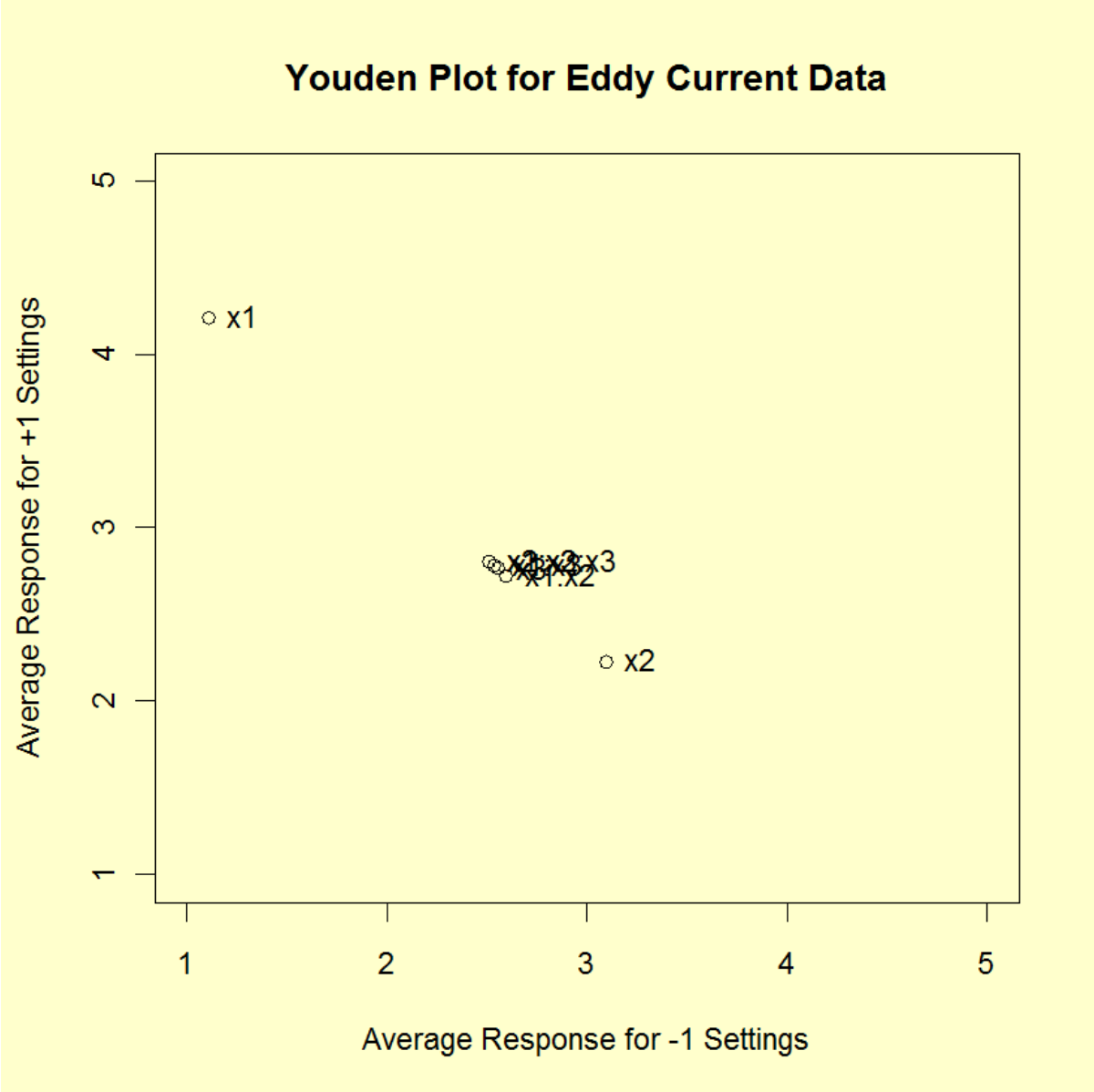
In summary, the DOE Youden plot is a plot of the mean of the response variable for the high level of a factor or interaction term against the mean of the response variable for the low level of that factor or interaction term.

For unimportant factors and interaction terms, these mean values should be nearly the same. For important factors and interaction terms, these mean values should be quite different. So the interpretation of the plot is that unimportant factors should be clustered together near the grand mean. Points that stand apart from this cluster identify important factors that should be included in the model.

Sample DOE

The following is a DOE Youden plot for the data used in the Eddy current case study. The

Youden Plot analysis in that case study demonstrated that X1 and X2 were the most important factors.



Interpretation of the Sample DOE Youden Plot

From the above DOE Youden plot, we see that factors 1 and 2 stand out from the others. That is, the mean response values for the low and high levels of factor 1 and factor 2 are quite different. For factor 3 and the 2 and 3-term interactions, the mean response values for the low and high levels are similar.

We would conclude from this plot that factors 1 and 2 are important and should be included in our final model while the remaining factors and interactions should be omitted from the final model.

Case Study

The Eddy current case study demonstrates the use of the DOE Youden plot in the context of the analysis of a full factorial design.

Software

DOE Youden plots are not typically available as built-in plots in statistical software programs. However, it should be relatively straightforward to write a macro to generate this plot in most general purpose statistical software programs.

1.3.3.32. 4-Plot

Purpose: Check Underlying Statistical Assumptions

The 4-plot is a collection of 4 specific EDA graphical techniques whose purpose is to test the assumptions that underlie most measurement processes. A 4-plot consists of a

1. run sequence plot;
2. lag plot;
3. histogram;
4. normal probability plot.

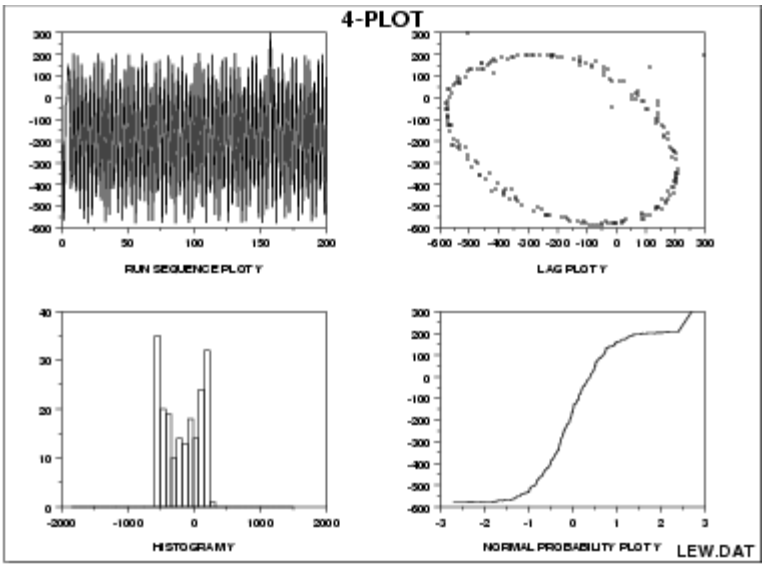
If the 4 underlying assumptions of a typical measurement process hold, then the above 4 plots will have a characteristic appearance (see the normal random numbers case study below); if any of the underlying assumptions fail to hold, then it will be revealed by an anomalous appearance in one or more of the plots. Several commonly encountered situations are demonstrated in the case studies below.

Although the 4-plot has an obvious use for univariate and time series data, its usefulness extends far beyond that. Many statistical models of the form

$$Y_i = f(X_1, ..., X_k) + E_i$$

have the same underlying assumptions for the error term. That is, no matter how complicated the functional fit, the assumptions on the underlying error term are still the same. The 4-plot can and should be routinely applied to the residuals when fitting models regardless of whether the model is simple or complicated.

Sample Plot: Process Has Fixed Location, Fixed Variation, Non-Random (Oscillatory), Non-Normal U-Shaped



Distribution, and Has 3 Outliers.

This 4-plot reveals the following:

1. the fixed location assumption is justified as shown by the run sequence plot in the upper left corner.
2. the fixed variation assumption is justified as shown by the run sequence plot in the upper left corner.
3. the randomness assumption is violated as shown by the non-random (oscillatory) lag plot in the upper right corner.
4. the assumption of a common, normal distribution is violated as shown by the histogram in the lower left corner and the normal probability plot in the lower right corner. The distribution is non-normal and is a U-shaped distribution.
5. there are several outliers apparent in the lag plot in the upper right corner.

Definition: 1. Run Sequence Plot; 2. Lag Plot; 3. Histogram; 4. Normal Probability Plot

The 4-plot consists of the following:

1. Run sequence plot to test fixed location and variation.
   - Vertically: Y_i
   - Horizontally: i
2. Lag Plot to test randomness.
   - Vertically: Y_i
   - Horizontally: Y_{i-1}
3. Histogram to test (normal) distribution.
   - Vertically: Counts
   - Horizontally: Y
4. Normal probability plot to test normal distribution.
   - Vertically: Ordered Y_i
   - Horizontally: Theoretical values from a normal N(0,1) distribution for ordered Y_i

Questions

4-plots can provide answers to many questions:

1. Is the process in-control, stable, and predictable?
2. Is the process drifting with respect to location?
3. Is the process drifting with respect to variation?
4. Are the data random?
5. Is an observation related to an adjacent observation?
6. If the data are a time series, is is white noise?
7. If the data are a time series and not white noise, is it sinusoidal, autoregressive, etc.?
8. If the data are non-random, what is a better model?
9. Does the process follow a normal distribution?
10. If non-normal, what distribution does the process follow?
11. Is the model

    $$Y_i = A_0 + E_i$$

    valid and sufficient?

12. If the default model is insufficient, what is a better model?
13. Is the formula $s_{\bar{Y}} = s/\sqrt{N}$ valid?
14. Is the sample mean a good estimator of the process location?
15. If not, what would be a better estimator?
16. Are there any outliers?

Importance: Testing Underlying Assumptions Helps Ensure the Validity of the Final Scientific and Engineering Conclusions

There are 4 assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

1. random drawings;
2. from a fixed distribution;
3. with that distribution having a fixed location; and
4. with that distribution having fixed variation.

Predictability is an all-important goal in science and engineering. If the above 4 assumptions hold, then we have achieved probabilistic predictability--the ability to make probability statements not only about the process in the past, but also about the process in the future. In short, such processes are said to be "statistically in control". If the 4 assumptions do not hold, then we have a process that is drifting (with respect to location, variation, or distribution), is unpredictable, and is out of control. A simple characterization of such processes by a location estimate, a variation estimate, or a distribution "estimate" inevitably leads to optimistic and grossly invalid engineering conclusions.

Inasmuch as the validity of the final scientific and engineering conclusions is inextricably linked to the

validity of these same 4 underlying assumptions, it naturally follows that there is a real necessity for all 4 assumptions to be routinely tested. The 4-plot (run sequence plot, lag plot, histogram, and normal probability plot) is seen as a simple, efficient, and powerful way of carrying out this routine checking.

Interpretation: Flat, Equi-Banded, Random, Bell-Shaped, and Linear

Of the 4 underlying assumptions:

1. If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
2. If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be approximately the same over the entire horizontal axis.
3. If the randomness assumption holds, then the lag plot will be structureless and random.
4. If the fixed distribution assumption holds (in particular, if the fixed normal distribution assumption holds), then the histogram will be bell-shaped and the normal probability plot will be approximatelylinear.

If all 4 of the assumptions hold, then the process is "statistically in control". In practice, many processes fall short of achieving this ideal.

Related Techniques

Run Sequence Plot
Lag Plot
Histogram
Normal Probability Plot

Autocorrelation Plot
Spectral Plot
PPCC Plot

Case Studies

The 4-plot is used in most of the case studies in this chapter:

1. Normal random numbers (the ideal)
2. Uniform random numbers
3. Random walk
4. Josephson junction cryothermometry
5. Beam deflections
6. Filter transmittance
7. Standard resistor
8. Heat flow meter 1

Software

It should be feasible to write a macro for the 4-plot in any general purpose statistical software program that supports the capability for multiple plots per page and supports the underlying plot techniques.

1.3.3.33. 6-Plot

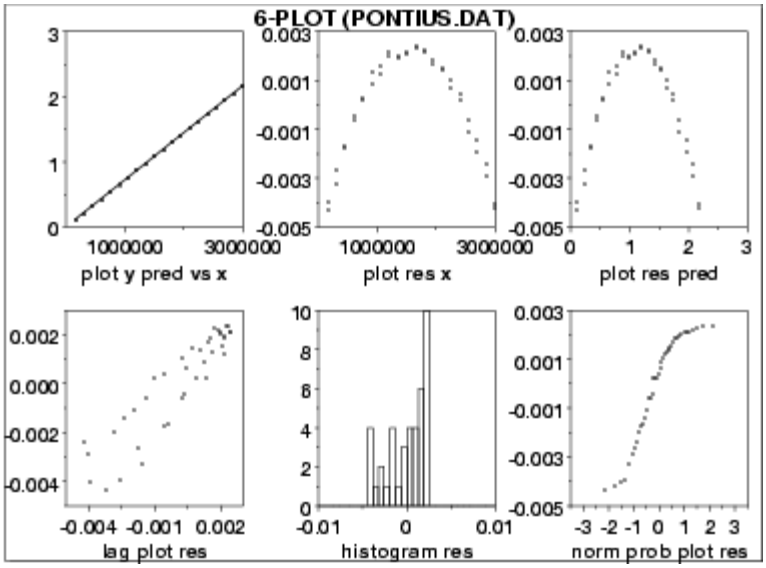
Purpose: Graphical Model Validation

The 6-plot is a collection of 6 specific graphical techniques whose purpose is to assess the validity of a Y versus X fit. The fit can be a linear fit, a non-linear fit, a LOWESS (locally weighted least squares) fit, a spline fit, or any other fit utilizing a single independent variable.

The 6 plots are:

1. Scatter plot of the response and predicted values versus the independent variable;
2. Scatter plot of the residuals versus the independent variable;
3. Scatter plot of the residuals versus the predicted values;
4. Lag plot of the residuals;
5. Histogram of the residuals;
6. Normal probability plot of the residuals.

Sample Plot



This 6-plot, which followed a linear fit, shows that the linear model is not adequate. It suggests that a quadratic model would be a better model.

Definition: 6

The 6-plot consists of the following:

Component Plots

1. Response and predicted values
    - Vertical axis: Response variable, predicted values
    - Horizontal axis: Independent variable
2. Residuals versus independent variable
    - Vertical axis: Residuals
    - Horizontal axis: Independent variable
3. Residuals versus predicted values
    - Vertical axis: Residuals
    - Horizontal axis: Predicted values
4. Lag plot of residuals
    - Vertical axis: RES(I)
    - Horizontal axis: RES(I-1)
5. Histogram of residuals
    - Vertical axis: Counts
    - Horizontal axis: Residual values
6. Normal probability plot of residuals
    - Vertical axis: Ordered residuals
    - Horizontal axis: Theoretical values from a normal N(0,1) distribution for ordered residuals

Questions

The 6-plot can be used to answer the following questions:

1. Are the residuals approximately normally distributed with a fixed location and scale?
2. Are there outliers?
3. Is the fit adequate?
4. Do the residuals suggest a better fit?

Importance: Validating Model

A model involving a response variable and a single independent variable has the form:

$$Y_i = f(X_i) + E_i$$

where Y is the response variable, X is the independent variable, *f* is the linear or non-linear fit function, and E is the random component. For a good model, the error component should behave like:

1. random drawings (i.e., independent);
2. from a fixed distribution;
3. with fixed location; and
4. with fixed variation.

In addition, for fitting models it is usually further assumed that the fixed distribution is normal and the fixed location is zero. For a good model the fixed variation should be as small as possible. A necessary component of fitting models is to verify these assumptions for the error component and to assess whether the variation for the error component is sufficiently small. The histogram, lag plot, and normal probability plot are used to verify the fixed distribution,

location, and variation assumptions on the error component. The plot of the response variable and the predicted values versus the independent variable is used to assess whether the variation is sufficiently small. The plots of the residuals versus the independent variable and the predicted values is used to assess the independence assumption.

Assessing the validity and quality of the fit in terms of the above assumptions is an absolutely vital part of the model-fitting process. No fit should be considered complete without an adequate model validation step.

Related Techniques

Linear Least Squares
Non-Linear Least Squares
Scatter Plot
Run Sequence Plot
Lag Plot
Normal Probability Plot
Histogram

Case Study

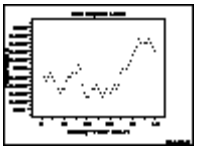
The 6-plot is used in the Alaska pipeline data case study.

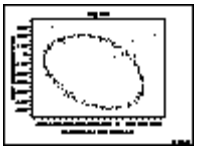
Software

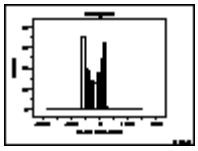
It should be feasible to write a macro for the 6-plot in any general purpose statistical software program that supports the capability for multiple plots per page and supports the underlying plot techniques.

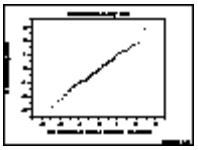
1.3.4. Graphical Techniques: By Problem Category

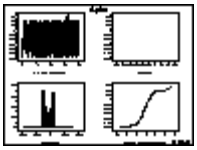
Univariate
y = c + e

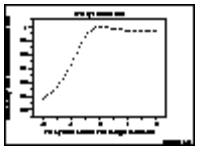

Run Sequence Plot: 1.3.3.25

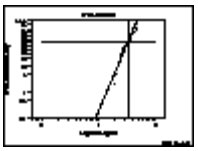

Lag Plot: 1.3.3.15

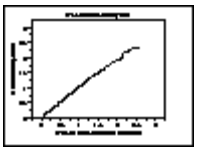

Histogram: 1.3.3.14

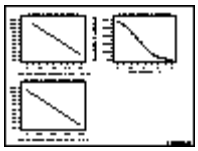

Normal Probability Plot: 1.3.3.21

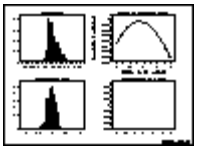

4-Plot: 1.3.3.32

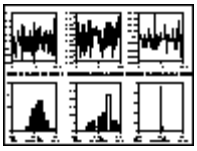

PPCC Plot: 1.3.3.23


Weibull Plot: 1.3.3.30

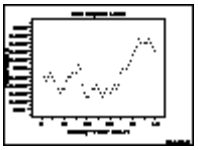

Probability Plot: 1.3.3.22

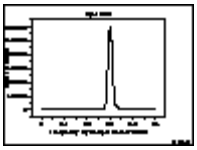

Box-Cox Linearity Plot: 1.3.3.5

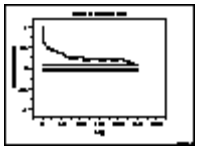

Box-Cox Normality Plot: 1.3.3.6

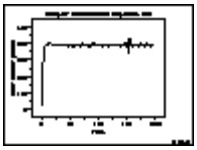

Bootstrap Plot: 1.3.3.4

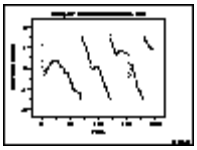
Time Series
y = f(t) + e


Run Sequence Plot: 1.3.3.25


Spectral Plot: 1.3.3.27

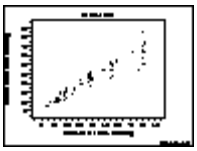

Autocorrelation Plot: 1.3.3.1

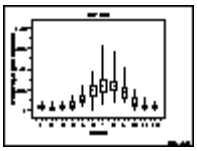

Complex Demodulation Amplitude Plot: 1.3.3.8

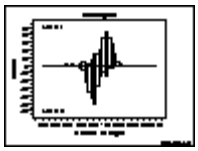

Complex Demodulation Phase Plot: 1.3.3.9

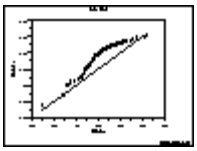
---

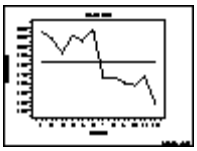
1 Factor
y = f(x) + e

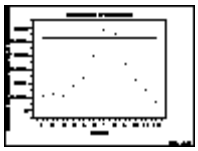

Scatter Plot: 1.3.3.26


Box Plot: 1.3.3.7


Bihistogram: 1.3.3.2

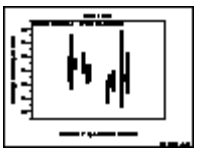

Quantile-Quantile Plot: 1.3.3.24


Mean Plot: 1.3.3.20

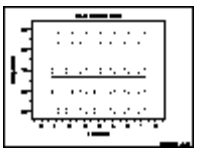

Standard Deviation Plot: 1.3.3.28

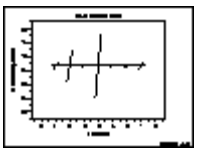
Multi-Factor/Comparative

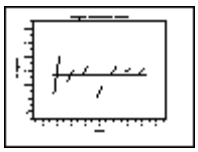
y = f(xp, x1,x2,...,xk) + e

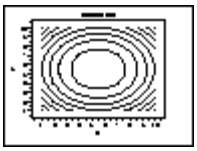

Block Plot: 1.3.3.3

Multi-Factor/Screening
y = f(x1,x2,x3,...,xk) + e

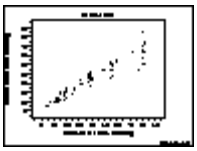

DOE Scatter Plot: 1.3.3.11

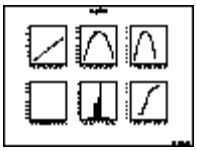

DOE Mean Plot: 1.3.3.12

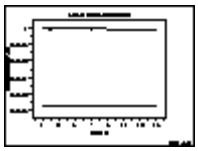

DOE Standard Deviation Plot: 1.3.3.13

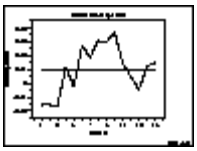

Contour Plot: 1.3.3.10

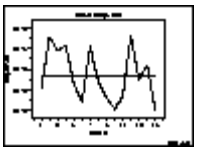
Regression
y = f(x1,x2,x3,...,xk) + e

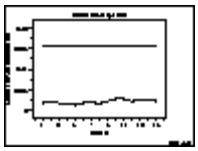

Scatter Plot: 1.3.3.26


6-Plot: 1.3.3.33

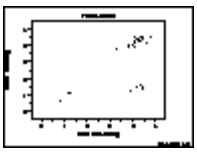

Linear Correlation Plot: 1.3.3.16


Linear Intercept Plot: 1.3.3.17

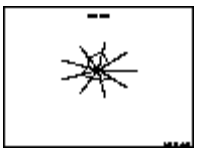

Linear Slope Plot: 1.3.3.18


Linear Residual Standard Deviation Plot:1.3.3.19

Interlab
(y1,y2) = f(x) + e


Youden Plot: 1.3.3.31

Multivariate
(y1,y2,...,yp)


Star Plot: 1.3.3.29

1.3.5. Quantitative Techniques

Confirmatory Statistics

The techniques discussed in this section are classical statistical methods as opposed to EDA techniques. EDA and classical techniques are not mutually exclusive and can be used in a complementary fashion. For example, the analysis can start with some simple graphical techniques such as the 4-plot followed by the classical confirmatory methods discussed herein to provide more rigorous statements about the conclusions. If the classical methods yield different conclusions than the graphical analysis, then some effort should be invested to explain why. Often this is an indication that some of the assumptions of the classical techniques are violated.

Many of the quantitative techniques fall into two broad categories:

1. Interval estimation
2. Hypothesis tests

Interval Estimates

It is common in statistics to estimate a parameter from a sample of data. The value of the parameter using all of the possible data, not just the sample data, is called the population parameter or true value of the parameter. An estimate of the true parameter value is made using the sample data. This is called a point estimate or a sample estimate.

For example, the most commonly used measure of location is the mean. The population, or true, mean is the sum of all the members of the given population divided by the number of members in the population. As it is typically impractical to measure every member of the population, a random sample is drawn from the population. The sample mean is calculated by summing the values in the sample and dividing by the number of values in the sample. This sample mean is then used as the point estimate of the population mean.

Interval estimates expand on point estimates by incorporating the uncertainty of the point estimate. In the example for the mean above, different samples from the same population will generate different values for the

sample mean. An interval estimate quantifies this uncertainty in the sample estimate by computing lower and upper values of an interval which will, with a given level of confidence (i.e., probability), contain the population parameter.

Hypothesis Tests

Hypothesis tests also address the uncertainty of the sample estimate. However, instead of providing an interval, a hypothesis test attempts to refute a specific claim about a population parameter based on the sample data. For example, the hypothesis might be one of the following:

- the population mean is equal to 10
- the population standard deviation is equal to 5
- the means from two populations are equal
- the standard deviations from 5 populations are equal

To reject a hypothesis is to conclude that it is false. However, to accept a hypothesis does not mean that it is true, only that we do not have evidence to believe otherwise. Thus hypothesis tests are usually stated in terms of both a condition that is doubted (null hypothesis) and a condition that is believed (alternative hypothesis).

A common format for a hypothesis test is:

H_0:	A statement of the null hypothesis, e.g., two population means are equal.
H_a:	A statement of the alternative hypothesis, e.g., two population means are not equal.
Test Statistic:	The test statistic is based on the specific hypothesis test.
Significance Level:	The significance level, α, defines the sensitivity of the test. A value of $\alpha = 0.05$ means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true. This is also called the type I error. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are commonly used. The probability of rejecting the null hypothesis when it is in fact false is called the power of the test and is denoted by $1 - \beta$. Its complement, the probability of accepting the null hypothesis when the alternative hypothesis is, in fact, true (type II error), is called β and can only be computed for a specific alternative hypothesis.
Critical Region:	The critical region encompasses those values of the test statistic that lead to a rejection of the null hypothesis. Based on the distribution of the test statistic and the significance level,

a cut-off value for the test statistic is computed. Values either above or below or both (depending on the direction of the test) this cut-off define the critical region.

Practical Versus Statistical Significance

It is important to distinguish between statistical significance and practical significance. Statistical significance simply means that we reject the null hypothesis. The ability of the test to detect differences that lead to rejection of the null hypothesis depends on the sample size. For example, for a particularly large sample, the test may reject the null hypothesis that two process means are equivalent. However, in practice the difference between the two means may be relatively small to the point of having no real engineering significance. Similarly, if the sample size is small, a difference that is large in engineering terms may not lead to rejection of the null hypothesis. The analyst should not just blindly apply the tests, but should combine engineering judgement with statistical analysis.

Bootstrap Uncertainty Estimates

In some cases, it is possible to mathematically derive appropriate uncertainty intervals. This is particularly true for intervals based on the assumption of a normal distribution. However, there are many cases in which it is not possible to mathematically derive the uncertainty. In these cases, the bootstrap provides a method for empirically determining an appropriate interval.

Table of Contents

Some of the more common classical quantitative techniques are listed below. This list of quantitative techniques is by no means meant to be exhaustive. Additional discussions of classical statistical techniques are contained in the product comparisons chapter.

- Location
    1. Measures of Location
    2. Confidence Limits for the Mean and One Sample t-Test
    3. Two Sample t-Test for Equal Means
    4. One Factor Analysis of Variance
    5. Multi-Factor Analysis of Variance
- Scale (or variability or spread)
    1. Measures of Scale
    2. Bartlett's Test
    3. Chi-Square Test
    4. F-Test
    5. Levene Test
- Skewness and Kurtosis
    1. Measures of Skewness and Kurtosis
- Randomness
    1. Autocorrelation
    2. Runs Test

Distributional Measures

1. Anderson-Darling Test
2. Chi-Square Goodness-of-Fit Test
3. Kolmogorov-Smirnov Test

- Outliers
  1. Detection of Outliers
  2. Grubbs Test
  3. Tietjen-Moore Test
  4. Generalized Extreme Deviate Test
- 2-Level Factorial Designs
  1. Yates Algorithm

1.3.5.1. Measures of Location

Location

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

Definition of Location

The first step is to define what we mean by a typical value. For univariate data, there are three common definitions:

1. mean - the mean is the sum of the data points divided by the number of data points. That is,

$$\bar{Y} = \sum_{i=1}^{N} Y_i / N$$

   The mean is that value that is most commonly referred to as the average. We will use the term average as a synonym for the mean and the term typical value to refer generically to measures of location.

2. median - the median is the value of the point which has half the data smaller than that point and half the data larger than that point. That is, if $X_1, X_2, \ldots, X_N$ is a random sample sorted from smallest value to largest value, then the median is defined as:

$$\tilde{Y} = Y_{(N+1)/2} \quad \text{if } N \text{ is odd}$$

$$\tilde{Y} = (Y_{N/2} + Y_{(N/2)+1})/2 \quad \text{if } N \text{ is even}$$

3. mode - the mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique. The mode is typically used in a qualitative fashion. For example, there may be a single dominant hump in the data perhaps two or more smaller humps in the data. This is usually evident from a histogram of the data.

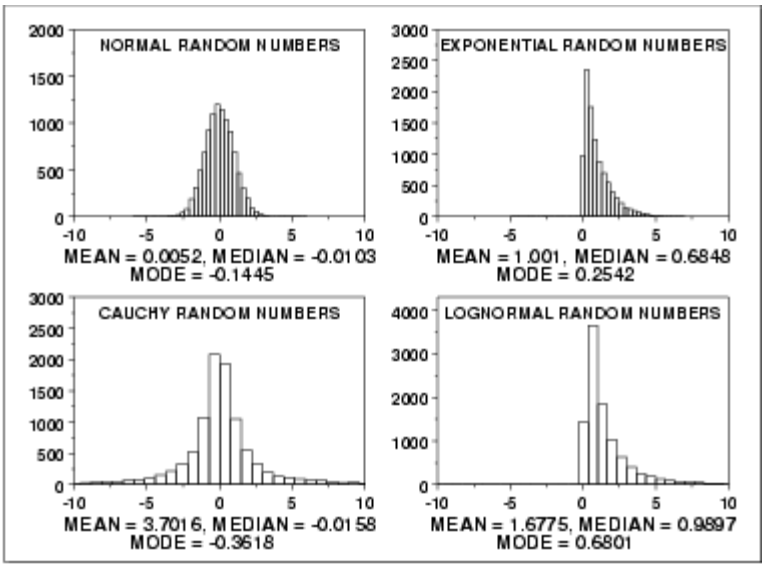
   When taking samples from continuous populations, we need to be somewhat careful in how we define the mode. That is, any specific value may not occur more than once if the data are continuous. What may be a more meaningful, if less exact measure, is the midpoint

of the class interval of the histogram with the highest peak.

Why Different Measures

A natural question is why we have more than one measure of the typical value. The following example helps to explain why these alternative definitions are useful and necessary.

This plot shows histograms for 10,000 random numbers generated from a normal, an exponential, a Cauchy, and a lognormal distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The mean is 0.005, the median is -0.010, and the mode is -0.144 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The normal distribution is a symmetric distribution with well-behaved tails and a single peak at the center of the distribution. By symmetric, we mean that the distribution can be folded about an axis so that the 2 sides coincide. That is, it behaves the same to the left and right of some center point. For a normal distribution, the mean, median, and mode are actually equivalent. The histogram above generates similar estimates for the mean, median, and mode. Therefore, if a histogram or normal probability plot indicates that your data are approximated well by a normal distribution, then it is reasonable to use the mean as the location estimator.

Exponential Distribution

The second histogram is a sample from an exponential distribution. The mean is 1.001, the median is 0.684, and the mode is 0.254 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The exponential distribution is a skewed, i. e., not symmetric, distribution. For skewed distributions, the mean and median are not the same. The mean will be pulled in the direction of

the skewness. That is, if the right tail is heavier than the left tail, the mean will be greater than the median. Likewise, if the left tail is heavier than the right tail, the mean will be less than the median.

For skewed distributions, it is not at all obvious whether the mean, the median, or the mode is the more meaningful measure of the typical value. In this case, all three measures are useful.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution. The mean is 3.70, the median is -0.016, and the mode is -0.362 (the mode is computed as the midpoint of the histogram interval with the highest peak).

For better visual comparison with the other data sets, we restricted the histogram of the Cauchy distribution to values between -10 and 10. The full Cauchy data set in fact has a minimum of approximately -29,000 and a maximum of approximately 89,000.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. The Cauchy distribution has the interesting property that collecting more data does not provide a more accurate estimate of the mean. That is, the sampling distribution of the mean is equivalent to the sampling distribution of the original data. This means that for the Cauchy distribution the mean is useless as a measure of the typical value. For this histogram, the mean of 3.7 is well above the vast majority of the data. This is caused by a few very extreme values in the tail. However, the median does provide a useful measure for the typical value.

Although the Cauchy distribution is an extreme case, it does illustrate the importance of heavy tails in measuring the mean. Extreme values in the tails distort the mean. However, these extreme values do not distort the median since the median is based on ranks. In general, for data with extreme values in the tails, the median provides a better estimate of location than does the mean.

Lognormal Distribution

The fourth histogram is a sample from a lognormal distribution. The mean is 1.677, the median is 0.989, and the mode is 0.680 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The lognormal is also a skewed distribution. Therefore the mean and median do not provide similar estimates for the location. As with the exponential distribution, there is no obvious answer to the question of which is the more meaningful measure of location.

Robustness

There are various alternatives to the mean and median for

measuring location. These alternatives were developed to address non-normal data since the mean is an optimal estimator if in fact your data are normal.

Tukey and Mosteller defined two types of robustness where robustness is a lack of susceptibility to the effects of nonnormality.

1. Robustness of validity means that the confidence intervals for the population location have a 95% chance of covering the population location regardless of what the underlying distribution is.

2. Robustness of efficiency refers to high effectiveness in the face of non-normal tails. That is, confidence intervals for the population location tend to be almost as narrow as the best that could be done if we knew the true shape of the distributuion.

The mean is an example of an estimator that is the best we can do if the underlying distribution is normal. However, it lacks robustness of validity. That is, confidence intervals based on the mean tend not to be precise if the underlying distribution is in fact not normal.

The median is an example of a an estimator that tends to have robustness of validity but not robustness of efficiency.

The alternative measures of location try to balance these two concepts of robustness. That is, the confidence intervals for the case when the data are normal should be almost as narrow as the confidence intervals based on the mean. However, they should maintain their validity even if the underlying data are not normal. In particular, these alternatives address the problem of heavy-tailed distributions.

Alternative Measures of Location

A few of the more common alternative location measures are:

1. Mid-Mean - computes a mean using the data between the 25th and 75th percentiles.

2. Trimmed Mean - similar to the mid-mean except different percentile values are used. A common choice is to trim 5% of the points in both the lower and upper tails, i.e., calculate the mean for data between the 5th and 95th percentiles.

3. Winsorized Mean - similar to the trimmed mean. However, instead of trimming the points, they are set to the lowest (or highest) value. For example, all data below the 5th percentile are set equal to the value of the 5th percentile and all data greater than the 95th percentile are set equal to the 95th percentile.

4. Mid-range = (smallest + largest)/2.

The first three alternative location estimators defined above have the advantage of the median in the sense that they are not unduly affected by extremes in the tails. However, they generate estimates that are closer to the mean for data that are normal (or nearly so).

The mid-range, since it is based on the two most extreme points, is not robust. Its use is typically restricted to situations in which the behavior at the extreme points is relevant.

Case Study The uniform random numbers case study compares the performance of several different location estimators for a particular non-normal distribution.

Software Most general purpose statistical software programs can compute at least some of the measures of location discussed above.

1.3.5.2. Confidence Limits for the Mean

Purpose: Interval Estimate for Mean

Confidence limits for the mean (Snedecor and Cochran, 1989) are an interval estimate for the mean. Interval estimates are often desirable because the estimate of the mean varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a lower and upper limit for the mean. The interval estimate gives an indication of how much uncertainty there is in our estimate of the true mean. The narrower the interval, the more precise is our estimate.

Confidence limits are expressed in terms of a confidence coefficient. Although the choice of confidence coefficient is somewhat arbitrary, in practice 90 %, 95 %, and 99 % intervals are often used, with 95 % being the most commonly used.

As a technical note, a 95 % confidence interval does **not** mean that there is a 95 % probability that the interval contains the true mean. The interval computed from a given sample either contains the true mean or it does not. Instead, the level of confidence is associated with the method of calculating the interval. The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean. That is, for a 95 % confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95 % of these intervals would contain the true mean.

Definition: Confidence Interval

Confidence limits are defined as:

$$\bar{Y} \pm t_{1-\alpha/2,\,N-1}\,\frac{s}{\sqrt{N}}$$

where $\bar{Y}$ is the sample mean, s is the sample standard deviation, N is the sample size, α is the desired significance level, and $t_{1-\alpha/2,\,N-1}$ is the 100(1-α/2) percentile of the t distribution with N - 1 degrees of freedom. Note that the confidence coefficient is 1 - α.

From the formula, it is clear that the width of the interval is controlled by two factors:

1. As N increases, the interval gets narrower from the $\sqrt{N}$ term.

   That is, one way to obtain more precise estimates for the mean is to increase the sample size.

2. The larger the sample standard deviation, the larger the confidence interval. This simply means that noisy data, i.e., data with a large standard deviation, are going to generate wider intervals than data with a smaller standard deviation.

Definition: Hypothesis Test

To test whether the population mean has a specific value, μ_0, against the two-sided alternative that it does not have a value μ_0, the confidence interval is converted to hypothesis-test form. The test is a one-sample *t*-test, and it is defined as:

H_0: $\mu = \mu_0$

H_a: $\mu \neq \mu_0$

Test Statistic: $T = (\bar{Y} - \mu_0)/(s/\sqrt{N})$

where $\bar{Y}$, N, and s are defined as above.

Significance Level: α. The most commonly used value for α is 0.05.

Critical Region: Reject the null hypothesis that the mean is a specified value, μ_0, if

$$T < t_{\alpha/2, N-1}$$

or

$$T > t_{1-\alpha/2, N-1}$$

Confidence Interval Example

We generated a 95 %, two-sided confidence interval for the ZARR13.DAT data set based on the following information.

```
N                              = 195
MEAN                           =   9.261460
STANDARD DEVIATION             =   0.022789
t_{1-0.025,N-1}                =   1.9723

LOWER LIMIT = 9.261460 - 1.9723*0.022789/√195
UPPER LIMIT = 9.261460 + 1.9723*0.022789/√195
```

Thus, a 95 % confidence interval for the mean is (9.258242, 9.264679).

t-Test Example

We performed a two-sided, one-sample *t*-test using the ZARR13.DAT data set to test the null hypothesis that the population mean is equal to 5.

```
H0:  μ = 5
Ha:  μ ≠ 5

Test statistic:  T = 2611.284
Degrees of freedom:  ν = 194
Significance level:  α = 0.05
Critical value:  t_{1-α/2,ν} = 1.9723
Critical region:  Reject H0 if |T| > 1.9723
```

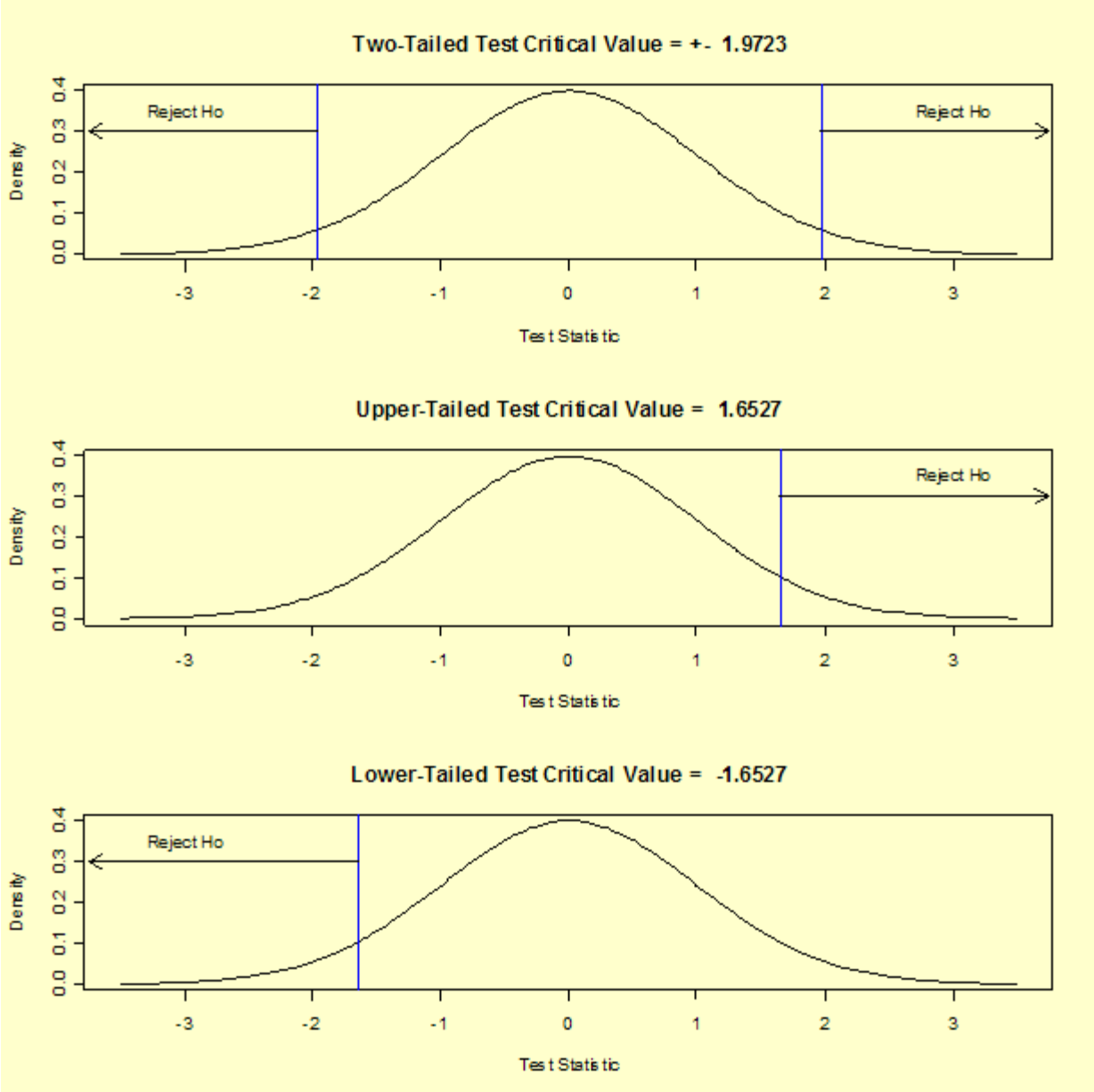
We reject the null hypotheses for our two-tailed *t*-test because the absolute value of the test statistic is greater than the critical value. If we were to perform an upper, one-tailed test, the critical value would be $t_{1-\alpha,\nu} = 1.6527$, and we would still reject the null hypothesis.

The confidence interval provides an alternative to the hypothesis test. If the confidence interval contains 5, then H_0 cannot be rejected. In our example, the confidence interval (9.258242, 9.264679) does not contain 5, indicating that the population mean does not equal 5 at the 0.05 level of significance.

In general, there are three possible alternative hypotheses and rejection regions for the one-sample *t*-test:

Alternative Hypothesis	Rejection Region
H_a: $\mu \neq \mu_0$	$\lvert T \rvert > t_{1-\alpha/2,\nu}$
H_a: $\mu > \mu_0$	$T > t_{1-\alpha,\nu}$
H_a: $\mu < \mu_0$	$T < t_{\alpha,\nu}$

The rejection regions for three posssible alternative hypotheses using our example data are shown in the following graphs.



Questions Confidence limits for the mean can be used to answer the following questions:

1. What is a reasonable estimate for the mean?
2. How much variability is there in the estimate of the mean?
3. Does a given target value fall within the confidence limits?

Related Techniques

Two-Sample *t*-Test

Confidence intervals for other location estimators such as the median or mid-mean tend to be mathematically difficult or intractable. For these cases, confidence intervals can be obtained using the bootstrap.

Case Study

Heat flow meter data.

Software

Confidence limits for the mean and one-sample *t*-tests are available in just about all general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.3. Two-Sample *t*-Test for Equal Means

Purpose: Test if two population means are equal

The two-sample *t*-test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application is to test if a new process or treatment is superior to a current process or treatment.

There are several variations on this test.

1. The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ are the two samples, then X_i corresponds to Y_i. For paired samples, the difference $X_i - Y_i$ is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.

2. The variances of the two samples may be assumed to be equal or unequal. Equal variances yields somewhat simpler formulas, although with computers this is no longer a significant issue.

3. In some applications, you may want to adopt a new process or treatment only if it exceeds the current treatment by some threshold. In this case, we can state the null hypothesis in the form that the difference between the two populations means is equal to some constant ($\mu_1 - \mu_2 = d_0$) where the constant is the desired threshold.

Definition

The two-sample *t*-test for unpaired data is defined as:

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

Test Statistic:
$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

where N_1 and N_2 are the sample sizes, $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances.

If equal variances are assumed, then the formula reduces to:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Significance Level: α.

Critical Region: Reject the null hypothesis that the two means are equal if

$$|T| > t_{1-\alpha/2,\nu}$$

where $t_{1-\alpha/2,\nu}$ is the critical value of the *t* distribution with ν degrees of freedom where

$$\upsilon = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

If equal variances are assumed, then

$$\upsilon = N_1 + N_2 - 2$$

Two-Sample t-Test Example

The following two-sample *t*-test was generated for the AUTO83B.DAT data set. The data set contains miles per gallon for U.S. cars (sample 1) and for Japanese cars (sample 2); the summary statistics for each sample are shown below.

```
SAMPLE 1:
    NUMBER OF OBSERVATIONS      = 249
    MEAN                         =  20.14458
    STANDARD DEVIATION           =   6.41470
    STANDARD ERROR OF THE MEAN   =   0.40652

SAMPLE 2:
    NUMBER OF OBSERVATIONS      = 79
    MEAN                         = 30.48101
    STANDARD DEVIATION           =  6.10771
    STANDARD ERROR OF THE MEAN   =  0.68717
```

We are testing the hypothesis that the population means are equal for the two samples. We assume that the variances for the two samples are equal.

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Test statistic: $T = -12.62059$
Pooled standard deviation: $s_p = 6.34260$
Degrees of freedom: $\nu = 326$
Significance level: $\alpha = 0.05$
Critical value (upper tail): $t_{1-\alpha/2,\nu} = 1.9673$
Critical region: Reject H_0 if $|T| > 1.9673$

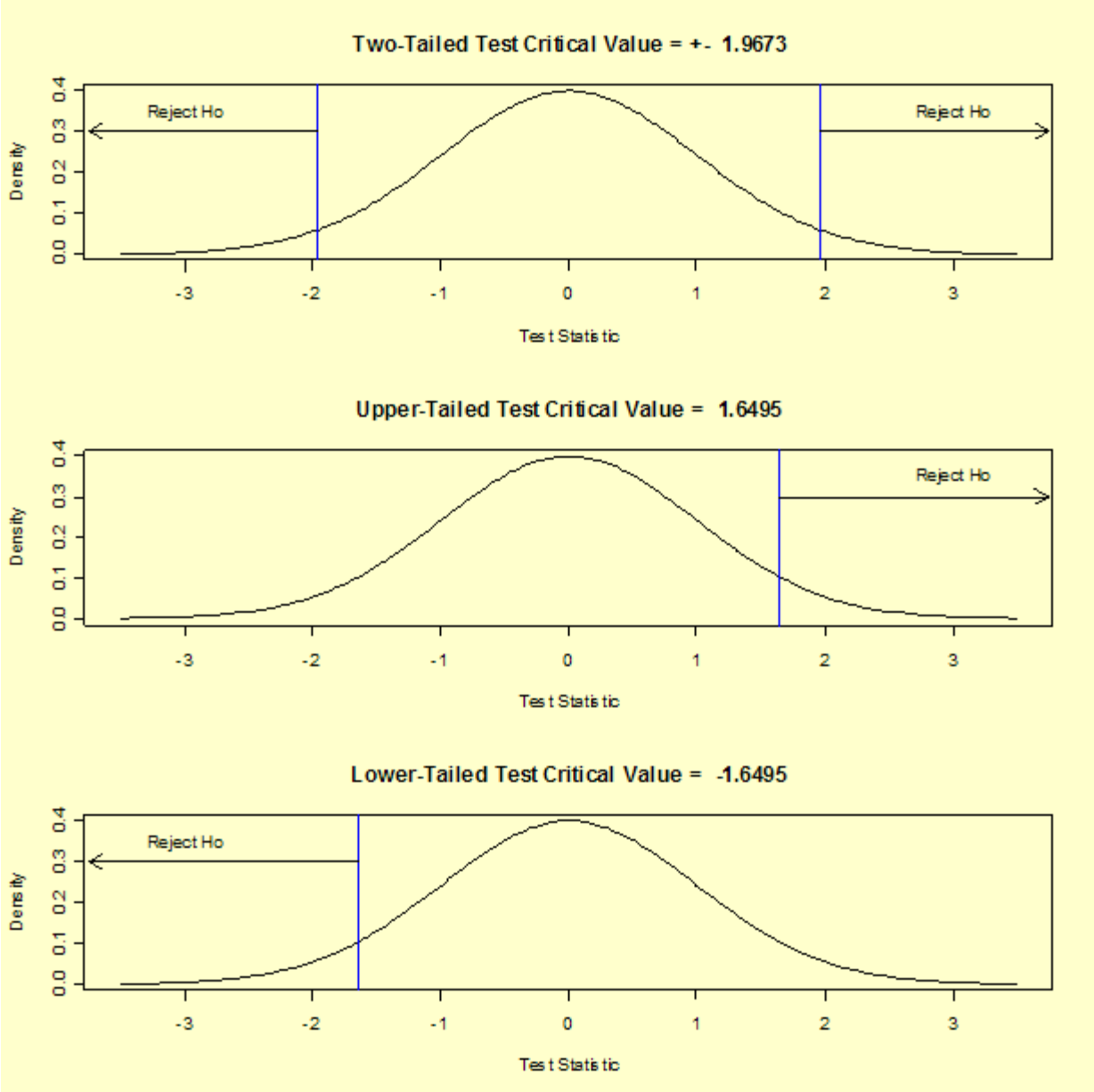
The absolute value of the test statistic for our example, 12.62059, is greater than the critical value of 1.9673, so we reject the null hypothesis and conclude that the two population means are different at the 0.05 significance level.

In general, there are three possible alternative hypotheses and rejection regions for the one-sample *t*-test:

Alternative Hypothesis	Rejection Region

H_a: $\mu_1 \neq \mu_2$	$\lvert T \rvert > t_{1-\alpha/2,\nu}$
H_a: $\mu_1 > \mu_2$	$T > t_{1-\alpha,\nu}$
H_a: $\mu_1 < \mu_2$	$T < t_{\alpha,\nu}$

For our two-tailed *t*-test, the critical value is $t_{1-\alpha/2,\nu} = 1.9673$, where $\alpha = 0.05$ and $\nu = 326$. If we were to perform an upper, one-tailed test, the critical value would be $t_{1-\alpha,\nu} = 1.6495$. The rejection regions for three posssible alternative hypotheses using our example data are shown below.



Questions Two-sample *t*-tests can be used to answer the following questions:

1. Is process 1 equivalent to process 2?
2. Is the new process better than the current process?
3. Is the new process better than the current process by at least some pre-determined threshold amount?

Related Techniques	Confidence Limits for the Mean Analysis of Variance
Case Study	Ceramic strength data.
Software	Two-sample *t*-tests are available in just about all general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.3.1. Data Used for Two-Sample *t*-Test

Data Used for Two-Sample t-Test Example

The following is the data used for the two-sample *t*-test example. The first column is miles per gallon for U.S. cars and the second column is miles per gallon for Japanese cars. For the *t*-test example, rows with the second column equal to -999 were deleted.

18	24
15	27
18	27
16	25
17	31
15	35
14	24
14	19
14	28
15	23
15	27
14	20
15	22
14	18
22	20
18	31
21	32
21	31
10	32
10	24
11	26
9	29
28	24
25	24
19	33
16	33
17	32
19	28
18	19
14	32
14	34
14	26
14	30
12	22
13	22
13	33
18	39
22	36
19	28
18	27
23	21
26	24
25	30
20	34
21	32
13	38
14	37
15	30
14	31
17	37
11	32
13	47
12	41

13	45
15	34
13	33
13	24
14	32
22	39
28	35
13	32
14	37
13	38
14	34
15	34
12	32
13	33
13	32
14	25
13	24
12	37
13	31
18	36
16	36
18	34
18	38
23	32
11	38
12	32
13	-999
12	-999
18	-999
21	-999
19	-999
21	-999
15	-999
16	-999
15	-999
11	-999
20	-999
21	-999
19	-999
15	-999
26	-999
25	-999
16	-999
16	-999
18	-999
16	-999
13	-999
14	-999
14	-999
14	-999
28	-999
19	-999
18	-999
15	-999
15	-999
16	-999
15	-999
16	-999
14	-999
17	-999
16	-999
15	-999
18	-999
21	-999
20	-999
13	-999
23	-999
20	-999
23	-999
18	-999
19	-999
25	-999
26	-999
18	-999
16	-999
16	-999
15	-999
22	-999
22	-999
24	-999
23	-999

29	-999
25	-999
20	-999
18	-999
19	-999
18	-999
27	-999
13	-999
17	-999
13	-999
13	-999
13	-999
30	-999
26	-999
18	-999
17	-999
16	-999
15	-999
18	-999
21	-999
19	-999
19	-999
16	-999
16	-999
16	-999
16	-999
25	-999
26	-999
31	-999
34	-999
36	-999
20	-999
19	-999
20	-999
19	-999
21	-999
20	-999
25	-999
21	-999
19	-999
21	-999
21	-999
19	-999
18	-999
19	-999
18	-999
18	-999
18	-999
30	-999
31	-999
23	-999
24	-999
22	-999
20	-999
22	-999
20	-999
21	-999
17	-999
18	-999
17	-999
18	-999
17	-999
16	-999
19	-999
19	-999
36	-999
27	-999
23	-999
24	-999
34	-999
35	-999
28	-999
29	-999
27	-999
34	-999
32	-999
28	-999
26	-999
24	-999
19	-999
28	-999

24	-999
27	-999
27	-999
26	-999
24	-999
30	-999
39	-999
35	-999
34	-999
30	-999
22	-999
27	-999
20	-999
18	-999
28	-999
27	-999
34	-999
31	-999
29	-999
27	-999
24	-999
23	-999
38	-999
36	-999
25	-999
38	-999
26	-999
22	-999
36	-999
27	-999
27	-999
32	-999
28	-999
31	-999

1.3.5.4. One-Factor ANOVA

Purpose: Test for Equal Means Across Groups

One factor analysis of variance (Snedecor and Cochran, 1989) is a special case of analysis of variance (ANOVA), for one factor of interest, and a generalization of the two-sample *t*-test. The two-sample *t*-test is used to decide whether two groups (levels) of a factor have the same mean. One-way analysis of variance generalizes this to levels where *k*, the number of levels, is greater than or equal to 2.

For example, data collected on, say, five instruments have one factor (instruments) at five levels. The ANOVA tests whether instruments have a significant effect on the results.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of one-factor ANOVA, including the details for the mathematical computations of one-way analysis of variance.

The model for the analysis of variance can be stated in two mathematically equivalent ways. In the following discussion, each level of each factor is called a cell. For the one-way case, a cell and a level are equivalent since there is only one factor. In the following, the subscript ***i*** refers to the level and the subscript ***j*** refers to the observation within a level. For example, Y_{23} refers to the third observation in the second level.

The first model is

$$Y_{ij} = \mu_i + E_{ij}$$

This model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These estimated cell means are the predicted values of the model and the differences between the response variable and the estimated cell means are the residuals. That is

$$\hat{Y}_{ij} = \hat{\mu}_i$$

$$R_{ij} = Y_{ij} - \hat{\mu}_i$$

The second model is

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

This model decomposes the response into an overall (grand) mean, the effect of the *i*th factor level, and an error term. The analysis of variance provides estimates of the grand mean and the effect of the *i*th factor level. The predicted values and the residuals of the model are

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

$$R_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i$$

The distinction between these models is that the second model divides the cell mean into an overall mean and the effect of the *i*th factor level. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, $\boldsymbol{E_{ij}}$, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

One-Way ANOVA Example

A one-way analysis of variance was generated for the GEAR.DAT data set. The data set contains 10 measurements of gear diameter for ten different batches for a total of 100 measurements.

```
                        DEGREES OF      SUM OF       MEAN
SOURCE                   FREEDOM       SQUARES      SQUARE
F STATISTIC
----------------    ----------    --------    ------
--   -----------
BATCH                         9        0.000729
0.000081       2.2969
RESIDUAL                     90        0.003174
0.000035
TOTAL (CORRECTED)            99        0.003903
0.000039

RESIDUAL STANDARD DEVIATION = 0.00594


BATCH      N       MEAN     SD(MEAN)
---------------------------------
  1       10     0.99800     0.00178
  2       10     0.99910     0.00178
  3       10     0.99540     0.00178
  4       10     0.99820     0.00178
  5       10     0.99190     0.00178
  6       10     0.99880     0.00178
  7       10     1.00150     0.00178
  8       10     1.00040     0.00178
  9       10     0.99830     0.00178
 10       10     0.99480     0.00178
```

The ANOVA table decomposes the variance into the following component sum of squares:

- Total sum of squares. The degrees of freedom for this

entry is the number of observations minus one.
- Sum of squares for the factor. The degrees of freedom for this entry is the number of levels minus one. The mean square is the sum of squares divided by the number of degrees of freedom.
- Residual sum of squares. The degrees of freedom is the total degrees of freedom minus the factor degrees of freedom. The mean square is the sum of squares divided by the number of degrees of freedom.

The sums of squares summarize how much of the variance in the data (total sum of squares) is accounted for by the factor effect (batch sum of squares) and how much is random error (residual sum of squares). Ideally, we would like most of the variance to be explained by the factor effect.

The ANOVA table provides a formal *F* test for the factor effect. For our example, we are testing the following hypothesis.

H_0: All individual batch means are equal.

H_a: At least one batch mean is not equal to the others.

The *F* statistic is the batch mean square divided by the residual mean square. This statistic follows an *F* distribution with (k-1) and (N-k) degrees of freedom. For our example, the critical *F* value (upper tail) for $\alpha = 0.05$, (k-1) = 10, and (N-k) = 90 is 1.9376. Since the *F* statistic, 2.2969, is greater than the critical value, we conclude that there is a significant batch effect at the 0.05 level of significance.

Once we have determined that there is a significant batch effect, we might be interested in comparing individual batch means. The batch means and the standard errors of the batch means provide some information about the individual batches, however, we may want to employ multiple comparison methods for a more formal analysis. (See Box, Hunter, and Hunter for more information.)

In addition to the quantitative ANOVA output, it is recommended that any analysis of variance be complemented with model validation. At a minimum, this should include:

1. a run sequence plot of the residuals,
2. a normal probability plot of the residuals, and
3. a scatter plot of the predicted values against the residuals.

Question The analysis of variance can be used to answer the following question

- Are means the same across groups in the data?

Importance	The analysis of uncertainty depends on whether the factor significantly affects the outcome.
Related Techniques	Two-sample *t*-test Multi-factor analysis of variance Regression Box plot
Software	Most general purpose statistical software programs can generate an analysis of variance. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.5. Multi-factor Analysis of Variance

Purpose: Detect significant factors

The analysis of variance (ANOVA) (Neter, Wasserman, and Kunter, 1990) is used to detect significant factors in a multi-factor model. In the multi-factor model, there is a response (dependent) variable and one or more factor (independent) variables. This is a common model in designed experiments where the experimenter sets the values for each of the factor variables and then measures the response variable.

Each factor can take on a certain number of values. These are referred to as the levels of a factor. The number of levels can vary betweeen factors. For designed experiments, the number of levels for a given factor tends to be small. Each factor and level combination is a cell. Balanced designs are those in which the cells have an equal number of observations and unbalanced designs are those in which the number of observations varies among cells. It is customary to use balanced designs in designed experiments.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of two-factor ANOVA, including the details for the mathematical computations.

The model for the analysis of variance can be stated in two mathematically equivalent ways. We explain the model for a two-way ANOVA (the concepts are the same for additional factors). In the following discussion, each combination of factors and levels is called a cell. In the following, the subscript i refers to the level of factor 1, j refers to the level of factor 2, and the subscript k refers to the kth observation within the (i,j)th cell. For example, Y_{235} refers to the fifth observation in the second level of factor 1 and the third level of factor 2.

The first model is

$$Y_{ijk} = \mu_{ij} + E_{ijk}$$

This model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These cell means are the predicted values of the model and the differences between the response

variable and the estimated cell means are the residuals. That is

$$\hat{Y}_{ijk} = \hat{\mu}_{ij}$$

$$R_{ijk} = Y_{ijk} - \hat{\mu}_{ij}$$

The second model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$$

This model decomposes the response into an overall (grand) mean, factor effects ($\hat{\alpha}_i$ and $\hat{\beta}_j$ represent the effects of the *i*th level of the first factor and the *j*th level of the second factor, respectively), and an error term. The analysis of variance provides estimates of the grand mean and the factor effects. The predicted values and the residuals of the model are

$$\hat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$$

$$R_{ijk} = Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

The distinction between these models is that the second model divides the cell mean into an overall mean and factor effects. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, $\boldsymbol{E_{ijk}}$, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

Multi-Factor ANOVA Example

An analysis of variance was performed for the JAHANMI2.DAT data set. The data contains four, two-level factors: table speed, down feed rate, wheel grit size, and batch. There are 30 measurements of ceramic strength for each factor combination for a total of 480 measurements.

```
 SOURCE               DF SUM OF SQUARES     MEAN
SQUARE    F STATISTIC
 -------------------------------------------------
-----------------
 TABLE SPEED           1   26672.726562
26672.726562         6.7080
 DOWN FEED RATE        1   11524.053711
11524.053711         2.8982
 WHEEL GRIT SIZE       1   14380.633789
14380.633789         3.6166
 BATCH                 1  727143.125000
727143.125000       182.8703
 RESIDUAL            475 1888731.500000
3976.276855
 TOTAL (CORRECTED)   479 2668446.000000
5570.868652

 RESIDUAL STANDARD DEVIATION = 63.05772781

 FACTOR           LEVEL    N       MEAN      SD(MEAN)
 ------------------------------------------------------
 TABLE SPEED       -1     240   657.53168     2.87818
                    1     240   642.62286     2.87818
 DOWN FEED RATE    -1     240   645.17755     2.87818
                    1     240   654.97723     2.87818
```

```
WHEEL GRIT SIZE  -1    240  655.55084    2.87818
                  1    240  644.60376    2.87818
BATCH             1    240  688.99890    2.87818
                  2    240  611.15594    2.87818
```

The ANOVA decomposes the variance into the following component sum of squares:

- Total sum of squares. The degrees of freedom for this entry is the number of observations minus one.
- Sum of squares for each of the factors. The degrees of freedom for these entries are the number of levels for the factor minus one. The mean square is the sum of squares divided by the number of degrees of freedom.
- Residual sum of squares. The degrees of freedom is the total degrees of freedom minus the sum of the factor degrees of freedom. The mean square is the sum of squares divided by the number of degrees of freedom.

The analysis of variance summarizes how much of the variance in the data (total sum of squares) is accounted for by the factor effects (factor sum of squares) and how much is due to random error (residual sum of squares). Ideally, we would like most of the variance to be explained by the factor effects. The ANOVA table provides a formal *F* test for the factor effects. To test the overall batch effect in our example we use the following hypotheses.

H_0: All individual batch means are equal.

H_a: At least one batch mean is not equal to the others.

The *F* statistic is the mean square for the factor divided by the residual mean square. This statistic follows an *F* distribution with (k-1) and (N-k) degrees of freedom where k is the number of levels for the given factor. Here, we see that the size of the "direction" effect dominates the size of the other effects. For our example, the critical *F* value (upper tail) for $\alpha = 0.05$, (k-1) = 1, and (N-k) = 475 is 3.86111. Thus, "table speed" and "batch" are significant at the 5 % level while "down feed rate" and "wheel grit size" are not significant at the 5 % level.

In addition to the quantitative ANOVA output, it is recommended that any analysis of variance be complemented with model validation. At a minimum, this should include

1. A run sequence plot of the residuals.
2. A normal probability plot of the residuals.
3. A scatter plot of the predicted values against the residuals.

Questions The analysis of variance can be used to answer the following questions:

1. Do any of the factors have a significant effect?

2. Which is the most important factor?
3. Can we account for most of the variability in the data?

Related Techniques
- One-factor analysis of variance
- Two-sample *t*-test
- Box plot
- Block plot
- DOE mean plot

Case Study
The quantitative ANOVA approach can be contrasted with the more graphical EDA approach in the ceramic strength case study.

Software
Most general purpose statistical software programs can perform multi-factor analysis of variance. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.6. Measures of Scale

Scale, Variability, or Spread

A fundamental task in many statistical analyses is to characterize the *spread*, or variability, of a data set. Measures of scale are simply attempts to estimate this variability.

When assessing the variability of a data set, there are two key components:

1. How spread out are the data values near the center?
2. How spread out are the tails?

Different numerical summaries will give different weight to these two elements. The choice of scale estimator is often driven by which of these components you want to emphasize.

The histogram is an effective graphical technique for showing both of these components of the spread.

Definitions of Variability

For univariate data, there are several common numerical measures of the spread:

1. variance - the variance is defined as
$$s^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2/(N-1)$$
where $\bar{Y}$ is the mean of the data.

   The variance is roughly the arithmetic average of the squared distance from the mean. Squaring the distance from the mean has the effect of giving greater weight to values that are further from the mean. For example, a point 2 units from the mean adds 4 to the above sum while a point 10 units from the mean adds 100 to the sum. Although the variance is intended to be an overall measure of spread, it can be greatly affected by the tail behavior.

2. standard deviation - the standard deviation is the square root of the variance. That is,
$$s = \sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2/(N-1)}$$

The standard deviation restores the units of the spread to the original data units (the variance squares the units).

3. range - the range is the largest value minus the smallest value in a data set. Note that this measure is based only on the lowest and highest extreme values in the sample. The spread near the center of the data is not captured at all.

4. average absolute deviation - the average absolute deviation (AAD) is defined as

$$AAD = \sum_{i=1}^{N}(|Y_i - \bar{Y}|)/N$$

where $\bar{Y}$ is the mean of the data and $|Y|$ is the absolute value of Y. This measure does not square the distance from the mean, so it is less affected by extreme observations than are the variance and standard deviation.

5. median absolute deviation - the median absolute deviation (MAD) is defined as

$$MAD = median(|Y_i - \tilde{Y}|)$$

where $\tilde{Y}$ is the median of the data and $|Y|$ is the absolute value of Y. This is a variation of the average absolute deviation that is even less affected by extremes in the tail because the data in the tails have less influence on the calculation of the median than they do on the mean.

6. interquartile range - this is the value of the 75th percentile minus the value of the 25th percentile. This measure of scale attempts to measure the variability of points near the center.

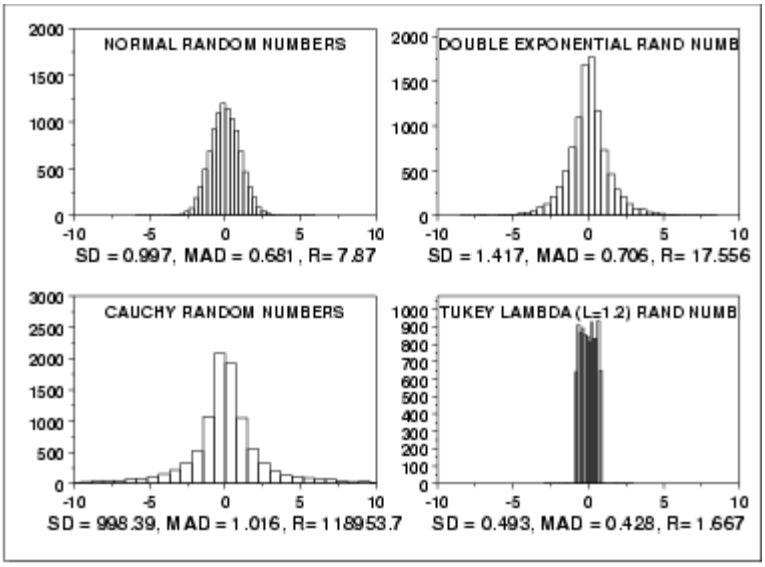
In summary, the variance, standard deviation, average absolute deviation, and median absolute deviation measure both aspects of the variability; that is, the variability near the center and the variability in the tails. They differ in that the average absolute deviation and median absolute deviation do not give undue weight to the tail behavior. On the other hand, the range only uses the two most extreme points and the interquartile range only uses the middle portion of the data.

Why Different

The following example helps to clarify why these alternative defintions of spread are useful and necessary.

Measures?

This plot shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Tukey-Lambda distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The standard deviation is 0.997, the median absolute deviation is 0.681, and the range is 7.87.

The normal distribution is a symmetric distribution with well-behaved tails and a single peak at the center of the distribution. By symmetric, we mean that the distribution can be folded about an axis so that the two sides coincide. That is, it behaves the same to the left and right of some center point. In this case, the median absolute deviation is a bit less than the standard deviation due to the downweighting of the tails. The range of a little less than 8 indicates the extreme values fall within about 4 standard deviations of the mean. If a histogram or normal probability plot indicates that your data are approximated well by a normal distribution, then it is reasonable to use the standard deviation as the spread estimator.

Double Exponential Distribution

The second histogram is a sample from a double exponential distribution. The standard deviation is 1.417, the median absolute deviation is 0.706, and the range is 17.556.

Comparing the double exponential and the normal histograms shows that the double exponential has a stronger peak at the center, decays more rapidly near the center, and has much longer tails. Due to the longer tails, the standard deviation tends to be inflated compared to the normal. On the other hand, the median absolute deviation is only slightly larger than it is for the normal data. The longer tails are clearly reflected in the value of the range, which shows that the extremes fall about 6 standard deviations from the mean compared to about 4 for the normal data.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution. The standard deviation is 998.389, the median absolute deviation is 1.16, and the range is 118,953.6.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. The Cauchy distribution has the interesting property that collecting more data does not provide a more accurate estimate for the mean or standard deviation. That is, the sampling distribution of the means and standard deviation are equivalent to the sampling distribution of the original data. That means that for the Cauchy distribution the standard deviation is useless as a measure of the spread. From the histogram, it is clear that just about all the data are between about -5 and 5. However, a few very extreme values cause both the standard deviation and range to be extremely large. However, the median absolute deviation is only slightly larger than it is for the normal distribution. In this case, the median absolute deviation is clearly the better measure of spread.

Although the Cauchy distribution is an extreme case, it does illustrate the importance of heavy tails in measuring the spread. Extreme values in the tails can distort the standard deviation. However, these extreme values do not distort the median absolute deviation since the median absolute deviation is based on ranks. In general, for data with extreme values in the tails, the median absolute deviation or interquartile range can provide a more stable estimate of spread than the standard deviation.

Tukey-Lambda Distribution

The fourth histogram is a sample from a Tukey lambda distribution with shape parameter α = 1.2. The standard deviation is 0.49, the median absolute deviation is 0.427, and the range is 1.666.

The Tukey lambda distribution has a range limited to $(-1/\lambda, 1/\lambda)$. That is, it has truncated tails. In this case the standard deviation and median absolute deviation have closer values than for the other three examples which have significant tails.

Robustness

Tukey and Mosteller defined two types of robustness where robustness is a lack of susceptibility to the effects of nonnormality.

1. Robustness of validity means that the confidence intervals for a measure of the population spread (e.g., the standard deviation) have a 95 % chance of covering the true value (i.e., the population value) of that measure of spread regardless of the underlying distribution.

2. Robustness of efficiency refers to high effectiveness in the face of non-normal tails. That is, confidence intervals for the measure of spread tend to be almost as narrow as the best that could be done if we knew the true shape of the distribution.

The standard deviation is an example of an estimator that is the best we can do if the underlying distribution is normal. However, it lacks robustness of validity. That is, confidence intervals based on the standard deviation tend to lack precision if the underlying distribution is in fact not normal.

The median absolute deviation and the interquartile range are estimates of scale that have robustness of validity. However, they are not particularly strong for robustness of efficiency.

If histograms and probability plots indicate that your data are in fact reasonably approximated by a normal distribution, then it makes sense to use the standard deviation as the estimate of scale. However, if your data are not normal, and in particular if there are long tails, then using an alternative measure such as the median absolute deviation, average absolute deviation, or interquartile range makes sense. The range is used in some applications, such as quality control, for its simplicity. In addition, comparing the range to the standard deviation gives an indication of the spread of the data in the tails.

Since the range is determined by the two most extreme points in the data set, we should be cautious about its use for large values of N.

Tukey and Mosteller give a scale estimator that has both robustness of validity and robustness of efficiency. However, it is more complicated and we do not give the formula here.

Software

Most general purpose statistical software programs can generate at least some of the measures of scale discusssed above.

1.3.5.7. Bartlett's Test

Purpose: Test for Homogeneity of Variances

Bartlett's test (Snedecor and Cochran, 1983) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality.

Definition

The Bartlett test is defined as:

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).

Test Statistic: The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln s_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i - 1)) - 1/(N-k))}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{k}(N_i - 1)s_i^2/(N-k)$$

Significance Level: α

Critical Region: The variances are judged to be unequal if,

$$T > \chi^2_{1-\alpha,\,k-1}$$

where $\chi^2_{1-\alpha, k-1}$ is the critical value of the chi-square distribution with k - 1 degrees of freedom and a significance level of α.

An alternate definition (Dixon and Massey, 1969) is based on an approximation to the F distribution. This definition is given in the Product and Process Comparisons chapter (chapter 7).

Example

Bartlett's test was performed for the GEAR.DAT data set. The data set contains 10 measurements of gear diameter for ten different batches for a total of 100 measurements.

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{10}^2$
H_a: At least one σ_i^2 is not equal to the others.

Test statistic: $T = 20.78580$
Degrees of freedom: $k - 1 = 9$
Significance level: $\alpha = 0.05$
Critical value: $X^2_{1-\alpha, k-1} = 16.919$
Critical region: Reject H_0 if $T > 16.919$

We are testing the null hypothesis that the batch variances are all equal. Because the test statistic is larger than the critical value, we reject the null hypotheses at the 0.05 significance level and conclude that at least one batch variance is different from the others.

Question

Bartlett's test can be used to answer the following question:

- Is the assumption of equal variances valid?

Importance

Bartlett's test is useful whenever the assumption of equal variances is made. In particular, this assumption is made for the frequently used one-way analysis of variance. In this case, Bartlett's or Levene's test should be applied to verify the assumption.

Related Techniques

Standard Deviation Plot
Box Plot
Levene Test
Chi-Square Test
Analysis of Variance

Case Study

Heat flow meter data

Software

The Bartlett test is available in many general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.8. Chi-Square Test for the Variance

Purpose: Test if the variance is equal to a specified value

A chi-square test (Snedecor and Cochran, 1983) can be used to test if the variance of a population is equal to a specified value. This test can be either a two-sided test or a one-sided test. The two-sided version tests against the alternative that the true variance is either less than or greater than the specified value. The one-sided version only tests in one direction. The choice of a two-sided or one-sided test is determined by the problem. For example, if we are testing a new process, we may only be concerned if its variability is greater than the variability of the current process.

Definition

The chi-square hypothesis test is defined as:

H_0: $\sigma^2 = \sigma_0^2$

H_a: $\sigma^2 < \sigma_0^2$ for a lower one-tailed test

$\sigma^2 > \sigma_0^2$ for an upper one-tailed test

$\sigma^2 \neq \sigma_0^2$ for a two-tailed test

Test Statistic: $T = (N-1)/(s/\sigma_0)^2$

where N is the sample size and s is the sample standard deviation. The key element of this formula is the ratio s/σ_0 which compares the ratio of the sample standard deviation to the target standard deviation. The more this ratio deviates from 1, the more likely we are to reject the null hypothesis.

Significance Level: α.

Critical Region: Reject the null hypothesis that the variance is a specified value, σ_0^2, if

$T > \chi^2_{1-\alpha, N-1}$ for an upper one-tailed alternative

$T < \chi^2_{\alpha, N-1}$ for a lower one-tailed alternative

$T < \chi^2_{\alpha/2, N-1}$ or $T > \chi^2_{1-\alpha/2, N-1}$ for a two-tailed test

where $\chi^2_{\alpha, N-1}$ is the critical value of the chi-square distribution with N - 1 degrees of freedom.

The formula for the hypothesis test can easily be converted to form an interval estimate for the variance:

$$\frac{(N-1)s^2}{\chi^2_{1-\alpha/2, N-1}} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_{\alpha/2, N-1}}$$

A confidence interval for the standard deviation is computed by taking the square root of the upper and lower limits of the confidence interval for the variance.

Chi-Square Test Example

A chi-square test was performed for the GEAR.DAT data set. The observed variance for the 100 measurements of gear diameter is 0.00003969 (the standard deviation is 0.0063). We will test the null hypothesis that the true variance is equal to 0.01.

H_0: $\sigma^2 = 0.01$
H_a: $\sigma^2 \neq 0.01$

Test statistic: $T = 0.3903$
Degrees of freedom: $N - 1 = 99$
Significance level: $\alpha = 0.05$
Critical values: $\chi^2_{\alpha/2, N-1} = 73.361$
$\chi^2_{1-\alpha/2, N-1} = 128.422$
Critical region: Reject H_0 if $T < 73.361$ or $T > 128.422$

The test statistic value of 0.3903 is much smaller than the lower critical value, so we reject the null hypothesis and conclude that the variance is not equal to 0.01.

Questions

The chi-square test can be used to answer the following questions:

1. Is the variance equal to some pre-determined threshold value?
2. Is the variance greater than some pre-determined threshold value?
3. Is the variance less than some pre-determined threshold value?

Related Techniques

F Test
Bartlett Test
Levene Test

Software

The chi-square test for the variance is available in many general purpose statistical software programs. Both Dataplot code and R code can be used to generate the

analyses in this section.

1.3.5.8.1. Data Used for Chi-Square Test for the Variance

Data Used for Chi-Square Test for the Variance Example

The following are the data used for the chi-square test for the variance example. The first column is gear diameter and the second column is batch number. Only the first column is used for this example.

1.006	1.000
0.996	1.000
0.998	1.000
1.000	1.000
0.992	1.000
0.993	1.000
1.002	1.000
0.999	1.000
0.994	1.000
1.000	1.000
0.998	2.000
1.006	2.000
1.000	2.000
1.002	2.000
0.997	2.000
0.998	2.000
0.996	2.000
1.000	2.000
1.006	2.000
0.988	2.000
0.991	3.000
0.987	3.000
0.997	3.000
0.999	3.000
0.995	3.000
0.994	3.000
1.000	3.000
0.999	3.000
0.996	3.000
0.996	3.000
1.005	4.000
1.002	4.000
0.994	4.000
1.000	4.000
0.995	4.000
0.994	4.000
0.998	4.000
0.996	4.000
1.002	4.000
0.996	4.000
0.998	5.000
0.998	5.000
0.982	5.000
0.990	5.000
1.002	5.000
0.984	5.000
0.996	5.000
0.993	5.000
0.980	5.000
0.996	5.000
1.009	6.000
1.013	6.000

1.009	6.000
0.997	6.000
0.988	6.000
1.002	6.000
0.995	6.000
0.998	6.000
0.981	6.000
0.996	6.000
0.990	7.000
1.004	7.000
0.996	7.000
1.001	7.000
0.998	7.000
1.000	7.000
1.018	7.000
1.010	7.000
0.996	7.000
1.002	7.000
0.998	8.000
1.000	8.000
1.006	8.000
1.000	8.000
1.002	8.000
0.996	8.000
0.998	8.000
0.996	8.000
1.002	8.000
1.006	8.000
1.002	9.000
0.998	9.000
0.996	9.000
0.995	9.000
0.996	9.000
1.004	9.000
1.004	9.000
0.998	9.000
0.999	9.000
0.991	9.000
0.991	10.000
0.995	10.000
0.984	10.000
0.994	10.000
0.997	10.000
0.997	10.000
0.991	10.000
0.998	10.000
1.004	10.000
0.997	10.000

1.3.5.9. *F*-Test for Equality of Two Variances

Purpose: Test if variances from two populations are equal

An *F*-test (Snedecor and Cochran, 1983) is used to test if the variances of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the variances are not equal. The one-tailed version only tests in one direction, that is the variance from the first population is either greater than or less than (but not both) the second population variance. The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new process is less variable than the old process.

Definition

The *F* hypothesis test is defined as:

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 < \sigma_2^2$ for a lower one-tailed test

$\sigma_1^2 > \sigma_2^2$ for an upper one-tailed test

$\sigma_1^2 \neq \sigma_2^2$ for a two-tailed test

Test Statistic: $F = s_1^2 / s_2^2$

where s_1^2 and s_2^2 are the sample variances. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

Significance Level: α

Critical Region: The hypothesis that the two variances are equal is rejected if

$F > F_{\alpha, N_1-1, N_2-1}$ for an upper one-tailed test

$F < F_{1-\alpha, N_1-1, N_2-1}$ for a lower one-tailed test

$F < F_{1-\alpha/2, N_1-1, N_2-1}$ for a two-tailed test

or

$$F > F_{\alpha/2,\, N_1\text{-}1,\, N_2\text{-}1}$$

where $F_{\alpha,\, N_1\text{-}1,\, N_2\text{-}1}$ is the critical value of the *F* distribution with N_1-1 and N_2-1 degrees of freedom and a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that F_α is the upper critical value from the *F* distribution and $F_{1\text{-}\alpha}$ is the lower critical value from the *F* distribution. Note that this is the opposite of the designation used by some texts and software programs.

F Test Example

The following *F*-test was generated for the AUTO83B.DAT data set. The data set contains 480 ceramic strength measurements for two batches of material. The summary statistics for each batch are shown below.

```
BATCH 1:
   NUMBER OF OBSERVATIONS      =      240
   MEAN                        =    688.9987
   STANDARD DEVIATION          =    65.54909

BATCH 2:
   NUMBER OF OBSERVATIONS      =      240
   MEAN                        =    611.1559
   STANDARD DEVIATION          =    61.85425
```

We are testing the null hypothesis that the variances for the two batches are equal.

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

Test statistic: $F = 1.123037$
Numerator degrees of freedom: $N_1 - 1 = 239$
Denominator degrees of freedom: $N_2 - 1 = 239$
Significance level: $\alpha = 0.05$
Critical values: $F(1\text{-}\alpha/2, N_1\text{-}1, N_2\text{-}1) = 0.7756$
$F(\alpha/2, N_1\text{-}1, N_2\text{-}1) = 1.2894$
Rejection region: Reject H_0 if $F < 0.7756$ or $F > 1.2894$

The *F* test indicates that there is not enough evidence to reject the null hypothesis that the two batch variancess are equal at the 0.05 significance level.

Questions

The *F*-test can be used to answer the following questions:

1. Do two samples come from populations with equal variancess?
2. Does a new process, treatment, or test reduce the variability of the current process?

Related Techniques	Quantile-Quantile Plot Bihistogram Chi-Square Test Bartlett's Test Levene Test
Case Study	Ceramic strength data.
Software	The *F*-test for equality of two variances is available in many general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.10. Levene Test for Equality of Variances

Purpose: Test for Homogeneity of Variances

Levene's test (Levene 1960) is used to test if k samples have equal variances. Equal variances across samples is called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene test can be used to verify that assumption.

Levene's test is an alternative to the Bartlett test. The Levene test is less sensitive than the Bartlett test to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal, distribution, then Bartlett's test has better performance.

Definition

The Levene test is defined as:

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).

Test Statistic: Given a variable Y with sample of size N divided into k subgroups, where N_i is the sample size of the ith subgroup, the Levene test statistic is defined as:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i.})^2}$$

where $\boldsymbol{Z_{ij}}$ can have one of the following three definitions:

1. $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$

   where $\bar{Y}_{i.}$ is the mean of the ith subgroup.

2. $Z_{ij} = |Y_{ij} - \tilde{Y}_{i.}|$

   where $\tilde{Y}_{i.}$ is the median of the ith subgroup.

3. $Z_{ij} = |Y_{ij} - \bar{Y}'_{i.}|$

where $\bar{Y}_{i.}'$ is the 10% trimmed mean of the ith subgroup.

$\bar{Z}_{i.}$ are the group means of the Z_{ij} and $\bar{Z}_{..}$ is the overall mean of the Z_{ij}.

The three choices for defining Z_{ij} determine the robustness and power of Levene's test. By robustness, we mean the ability of the test to not falsely detect unequal variances when the underlying data are not normally distributed and the variables are in fact equal. By power, we mean the ability of the test to detect unequal variances when the variances are in fact unequal.

Levene's original paper only proposed using the mean. Brown and Forsythe (1974)) extended Levene's test to use either the median or the trimmed mean in addition to the mean. They performed Monte Carlo studies that indicated that using the trimmed mean performed best when the underlying data followed a Cauchy distribution (i.e., heavy-tailed) and the median performed best when the underlying data followed a χ^2_4 (i.e., skewed) distribution. Using the mean provided the best power for symmetric, moderate-tailed, distributions.

Although the optimal choice depends on the underlying distribution, the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. If you have knowledge of the underlying distribution of the data, this may indicate using one of the other choices.

Significance Level: α

Critical Region: The Levene test rejects the hypothesis that the variances are equal if

$$W > F_{\alpha,\, k\text{-}1,\, N\text{-}k}$$

where $F_{\alpha,\, k\text{-}1,\, N\text{-}k}$ is the upper critical value of the *F* distribution with k-1 and N-k degrees of freedom at a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that F_α is the upper critical value from the *F* distribution

and $F_{1-\alpha}$ is the lower critical value. Note that this is the opposite of some texts and software programs.

Levene's Test Example

Levene's test, based on the median, was performed for the GEAR.DAT data set. The data set includes ten measurements of gear diameter for each of ten batches for a total of 100 measurements.

$H_0: \quad \sigma_1^2 = \ldots = \sigma_{10}^2$
$H_a: \quad \sigma_1^2 \neq \ldots \neq \sigma_{10}^2$

Test statistic: $W = 1.705910$
Degrees of freedom: $k-1 = 10-1 = 9$
$N-k = 100-10 = 90$
Significance level: $\alpha = 0.05$
Critical value (upper tail): $F_{\alpha, k-1, N-k} = 1.9855$
Critical region: Reject H_0 if $F > 1.9855$

We are testing the hypothesis that the group variances are equal. We fail to reject the null hypothesis at the 0.05 significance level since the value of the Levene test statistic is less than the critical value. We conclude that there is insufficient evidence to claim that the variances are not equal.

Question

Levene's test can be used to answer the following question:

- Is the assumption of equal variances valid?

Related Techniques

Standard Deviation Plot
Box Plot
Bartlett Test
Chi-Square Test
Analysis of Variance

Software

The Levene test is available in some general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.11. Measures of Skewness and Kurtosis

Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

Definition of Skewness

For univariate data $Y_1, Y_2, ..., Y_N$, the formula for skewness is:

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

Definition of Kurtosis

For univariate data $Y_1, Y_2, ..., Y_N$, the formula for kurtosis is:

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points.

Alternative Definition of Kurtosis

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

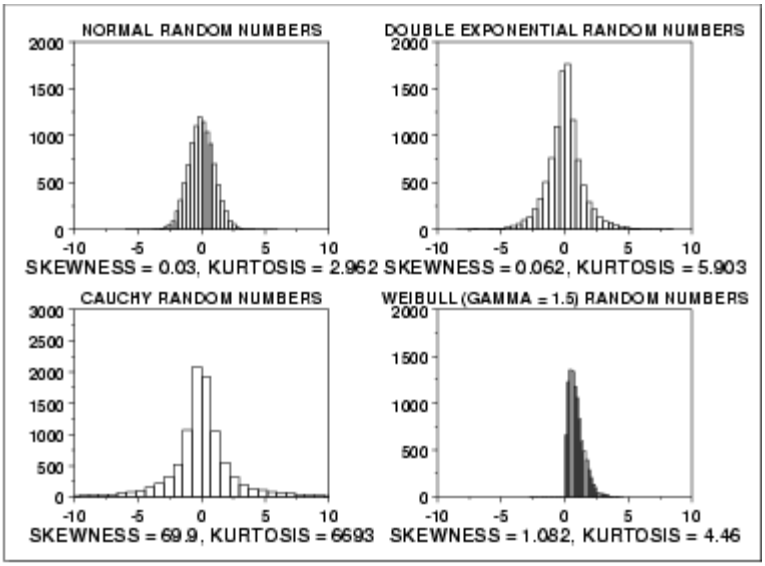
$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4} - 3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

Which definition of kurtosis is used is a matter of convention (this handbook uses the original definition). When using software to compute the sample kurtosis, you need to be aware of which convention is being followed. Many sources use the term kurtosis when they are actually computing "excess kurtosis", so it may not always be clear.

Examples

The following example shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Weibull distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The normal distribution is a symmetric distribution with well-behaved tails. This is indicated by the skewness of 0.03. The kurtosis of 2.96 is near the expected value of 3. The histogram verifies the symmetry.

Double Exponential Distribution

The second histogram is a sample from a double exponential distribution. The double exponential is a symmetric distribution. Compared to the normal, it has a stronger peak, more rapid decay, and heavier tails. That is, we would expect a skewness near zero and a kurtosis higher than 3. The skewness is 0.06 and the kurtosis is 5.9.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution.

For better visual comparison with the other data sets, we restricted the histogram of the Cauchy distribution to values between -10 and 10. The full data set for the Cauchy data in fact has a minimum of approximately -29,000 and a maximum of approximately 89,000.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. Since it is symmetric, we would expect a skewness near zero. Due to the heavier tails, we might expect the kurtosis to be larger than for a normal distribution. In fact the skewness is 69.99 and the kurtosis is 6,693. These extremely high values can be explained by the heavy tails. Just as the mean and standard deviation can be distorted by extreme values in the tails, so too can the skewness and kurtosis measures.

Weibull Distribution

The fourth histogram is a sample from a Weibull distribution with shape parameter 1.5. The Weibull distribution is a skewed distribution with the amount of skewness depending on the value of the shape parameter. The degree of decay as we move away from the center also depends on the value of the shape parameter. For this data set, the skewness is 1.08 and the kurtosis is 4.46, which indicates moderate skewness and kurtosis.

Dealing with Skewness and Kurtosis

Many classical statistical tests and intervals depend on normality assumptions. Significant skewness and kurtosis clearly indicate that data are not normal. If a data set exhibits significant skewness or kurtosis (as indicated by a histogram or the numerical measures), what can we do about it?

One approach is to apply some type of transformation to try to make the data normal, or more nearly normal. The Box-Cox transformation is a useful technique for trying to normalize a data set. In particular, taking the log or square root of a data set is often useful for data that exhibit moderate right skewness.

Another approach is to use techniques based on distributions other than the normal. For example, in reliability studies, the exponential, Weibull, and lognormal distributions are typically used as a basis for modeling rather than using the normal distribution. The probability plot correlation coefficient plot and the probability plot are useful tools for determining a good distributional model for the data.

Software The skewness and kurtosis coefficients are available in most general purpose statistical software programs.

1.3.5.12. Autocorrelation

Purpose: Detect Non-Randomness, Time Series Modeling

The autocorrelation (Box and Jenkins, 1976) function can be used for the following two purposes:

1. To detect non-randomness in data.
2. To identify an appropriate time series model if the data are not random.

Definition

Given measurements, Y_1, Y_2, ..., Y_N at time X_1, X_2, ..., X_N, the lag k autocorrelation function is defined as

$$r_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Although the time variable, X, is not used in the formula for autocorrelation, the assumption is that the observations are equi-spaced.

Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times $\boldsymbol{X_i}$ and $\boldsymbol{X_{i+k}}$.

When the autocorrelation is used to detect non-randomness, it is usually only the first (lag 1) autocorrelation that is of interest. When the autocorrelation is used to identify an appropriate time series model, the autocorrelations are usually plotted for many lags.

Autocorrelation Example

Lag-one autocorrelations were computed for the the LEW.DAT data set.

lag	autocorrelation
0.	1.00
1.	-0.31
2.	-0.74
3.	0.77
4.	0.21
5.	-0.90
6.	0.38
7.	0.63
8.	-0.77
9.	-0.12
10.	0.82
11.	-0.40
12.	-0.55
13.	0.73
14.	0.07
15.	-0.76

```
16.      0.40
17.      0.48
18.     -0.70
19.     -0.03
20.      0.70
21.     -0.41
22.     -0.43
23.      0.67
24.      0.00
25.     -0.66
26.      0.42
27.      0.39
28.     -0.65
29.      0.03
30.      0.63
31.     -0.42
32.     -0.36
33.      0.64
34.     -0.05
35.     -0.60
36.      0.43
37.      0.32
38.     -0.64
39.      0.08
40.      0.58
41.     -0.45
42.     -0.28
43.      0.62
44.     -0.10
45.     -0.55
46.      0.45
47.      0.25
48.     -0.61
49.      0.14
```

Questions

The autocorrelation function can be used to answer the following questions.

1. Was this sample data set generated from a random process?
2. Would a non-linear or time series model be a more appropriate model for these data than a simple constant plus error model?

Importance

Randomness is one of the key assumptions in determining if a univariate statistical process is in control. If the assumptions of constant location and scale, randomness, and fixed distribution are reasonable, then the univariate process can be modeled as:

$$Y_i = A_0 + E_i$$

where E_i is an error term.

If the randomness assumption is not valid, then a different model needs to be used. This will typically be either a time series model or a non-linear model (with time as the independent variable).

Related Techniques

Autocorrelation Plot
Run Sequence Plot
Lag Plot
Runs Test

Case Study

The heat flow meter data demonstrate the use of autocorrelation in determining if the data are from a random process.

Software

The autocorrelation capability is available in most general

purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.13. Runs Test for Detecting Non-randomness

Purpose: Detect Non-Randomness

The runs test (Bradley, 1968) can be used to decide if a data set is from a random process.

A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run. In a random data set, the probability that the (I+1)th value is larger or smaller than the Ith value follows a binomial distribution, which forms the basis of the runs test.

Typical Analysis and Test Statistics

The first step in the runs test is to count the number of runs in the data sequence. There are several ways to define runs in the literature, however, in all cases the formulation must produce a dichotomous sequence of values. For example, a series of 20 coin tosses might produce the following sequence of heads (H) and tails (T).

H H T T H T H H H H T H H T T T T T H H

The number of runs for this series is nine. There are 11 heads and 9 tails in the sequence.

Definition

We will code values above the median as positive and values below the median as negative. A run is defined as a series of consecutive positive (or negative) values. The runs test is defined as:

H_0:	the sequence was produced in a random manner
H_a:	the sequence was not produced in a random manner
Test Statistic:	The test statistic is

$$Z = \frac{R - \bar{R}}{s_R}$$

where R is the observed number of runs, $\overline{R}$, is the expected number of runs, and s_R is the standard deviation of the number of runs. The values of $\overline{R}$

and s_R are computed as follows:

$$\bar{R} = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$s_R^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

where n_1 and n_2 are the number of positive and negative values in the series.

Significance Level: α

Critical Region: The runs test rejects the null hypothesis if

$$|Z| > Z_{1-\alpha/2}$$

For a large-sample runs test (where $n_1 > 10$ and $n_2 > 10$), the test statistic is compared to a standard normal table. That is, at the 5 % significance level, a test statistic with an absolute value greater than 1.96 indicates non-randomness. For a small-sample runs test, there are tables to determine critical values that depend on values of n_1 and n_2 (Mendenhall, 1982).

Runs Test Example

A runs test was performed for 200 measurements of beam deflection contained in the LEW.DAT data set.

```
H0:  the sequence was produced in a random manner
Ha:  the sequence was not produced in a random
manner

Test statistic:  Z = 2.6938
Significance level:  α = 0.05
Critical value (upper tail):  Z1-α/2 = 1.96
Critical region: Reject H0 if |Z| > 1.96
```

Since the test statistic is greater than the critical value, we conclude that the data are not random at the 0.05 significance level.

Question

The runs test can be used to answer the following question:

- Were these sample data generated from a random process?

Importance

Randomness is one of the key assumptions in determining if a univariate statistical process is in control. If the assumptions of constant location and scale, randomness, and fixed distribution

are reasonable, then the univariate process can be modeled as:

$$Y_i = A_0 + E_i$$

where E_i is an error term.

If the randomness assumption is not valid, then a different model needs to be used. This will typically be either a times series model or a non-linear model (with time as the independent variable).

Related Techniques
Autocorrelation
Run Sequence Plot
Lag Plot

Case Study
Heat flow meter data

Software
Most general purpose statistical software programs support a runs test. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.14. Anderson-Darling Test

Purpose: Test for Distributional Adequacy

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions. We do not provide the tables of critical values in this Handbook (see Stephens 1974, 1976, 1977, and 1979) since this test is usually applied with a statistical software program that will print the relevant critical values.

The Anderson-Darling test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests.

Definition

The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution.

H_a: The data do not follow the specified distribution

Test Statistic: The Anderson-Darling test statistic is defined as

$$A^2 = -N - S$$

where

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\ln F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right]$$

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the *ordered* data.

Significance Level: α

Critical Region: The critical values for the Anderson-Darling test are dependent on the specific distribution that is being

tested. Tabulated values and formulas have been published (Stephens, 1974, 1976, 1977, 1979) for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.

Note that for a given distribution, the Anderson-Darling statistic may be multiplied by a constant (which usually depends on the sample size, *n*). These constants are given in the various papers by Stephens. In the sample output below, the test statistic values are adjusted. Also, be aware that different constants (and therefore critical values) have been published. You just need to be aware of what constant was used for a given set of critical values (the needed constant is typically given with the critical values).

Sample Output

We generated 1,000 random numbers for normal, double exponential, Cauchy, and lognormal distributions. In all four cases, the Anderson-Darling test was applied to test for a normal distribution.

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the Cauchy random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

```
      Distribution                 Mean       Standard
Deviation
      ------------               --------     ---------
---------
      Normal (Y1)                0.004360
1.001816
      Double Exponential (Y2)    0.020349
1.321627
      Cauchy (Y3)                1.503854
35.130590
      Lognormal (Y4)             1.518372
1.719969

      H0:  the data are normally distributed
      Ha:  the data are not normally distributed

      Y1 adjusted test statistic:  A2 =   0.2576
      Y2 adjusted test statistic:  A2 =   5.8492
      Y3 adjusted test statistic:  A2 = 288.7863
      Y4 adjusted test statistic:  A2 =  83.3935

      Significance level:  α = 0.05
      Critical value:  0.752
      Critical region:  Reject H0 if A2 > 0.752
```

When the data were generated using a normal distribution, the test statistic was small and the hypothesis of normality was not rejected. When the data were generated using the double exponential, Cauchy, and lognormal distributions, the test statistics were large, and the hypothesis of an underlying normal distribution was

rejected at the 0.05 significance level.

Questions

The Anderson-Darling test can be used to answer the following questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?

Importance

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the data follow a Weibull distribution.

There are many non-parametric and robust techniques that do not make strong distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than non-parametric and robust techniques. Therefore, if the distributional assumptions can be validated, they are generally preferred.

Related Techniques

Chi-Square goodness-of-fit Test
Kolmogorov-Smirnov Test
Shapiro-Wilk Normality Test
Probability Plot
Probability Plot Correlation Coefficient Plot

Case Study

Josephson junction cryothermometry case study.

Software

The Anderson-Darling goodness-of-fit test is available in some general purpose statistical software programs. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.15. Chi-Square Goodness-of-Fit Test

Purpose: Test for distributional adequacy

The chi-square test (Snedecor and Cochran, 1989) is used to test if a sample of data came from a population with a specific distribution.

An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the chi-square test. However, the value of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

Additional discussion of the chi-square goodness-of-fit test is contained in the product and process comparisons chapter (chapter 7).

Definition

The chi-square test is defined for the hypothesis:

H_0:	The data follow a specified distribution.
H_a:	The data do not follow the specified distribution.
Test Statistic:	For the chi-square goodness-of-fit computation, the data are divided into *k* bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

where O_i is the observed frequency for bin *i* and E_i is the expected frequency for bin *i*. The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l))$$

where F is the cumulative Distribution function for the distribution being tested, Y_u is the upper limit for class i, Y_l is the lower limit for class i, and *N* is the sample size.

This test is sensitive to the choice of bins. There is no optimal choice for the bin width (since the optimal bin width depends on the distribution). Most reasonable choices should produce similar, but not identical, results. For the chi-square approximation to be valid, the expected frequency should be at least 5. This test is not valid for small samples, and if some of the counts are less than five, you may need to combine some bins in the tails.

Significance Level: α.

Critical Region: The test statistic follows, approximately, a chi-square distribution with (k - c) degrees of freedom where k is the number of non-empty cells and c = the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution + 1. For example, for a 3-parameter Weibull distribution, c = 4.

Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if

$$\chi^2 > \chi^2_{1-\alpha,\, k-c}$$

where $\chi^2_{1-\alpha,\, k-c}$ is the chi-square critical value with k - c degrees of freedom and significance level α.

Chi-Square Test Example

We generated 1,000 random numbers for normal, double exponential, *t* with 3 degrees of freedom, and lognormal distributions. In all cases, a chi-square test with k = 32 bins was applied to test for normally distributed data. Because the normal distribution has two parameters, c = 2 + 1 = 3

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the *t* random numbers were stored in the

variable Y3, and the lognormal random numbers were stored in the variable Y4.

H_0: the data are normally distributed
H_a: the data are not normally distributed

Y1 Test statistic: X^2 = 32.256
Y2 Test statistic: X^2 = 91.776
Y3 Test statistic: X^2 = 101.488
Y4 Test statistic: X^2 = 1085.104

Significance level: $\alpha = 0.05$
Degrees of freedom: $k - c = 32 - 3 = 29$
Critical value: $X^2_{1-\alpha,\ k-c} = 42.557$
Critical region: Reject H_0 if $X^2 > 42.557$

As we would hope, the chi-square test fails to reject the null hypothesis for the normally distributed data set and rejects the null hypothesis for the three non-normal data sets.

Questions

The chi-square test can be used to answer the following types of questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?
- Are the data from a binomial distribution?

Importance

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the distribution of the data follows a Weibull distribution.

There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique, such as the sign test, that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumption can be confirmed, the parametric techniques are generally preferred.

If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, a non-parametric or robust technique may be required.

Related Techniques	Anderson-Darling Goodness-of-Fit Test Kolmogorov-Smirnov Test Shapiro-Wilk Normality Test Probability Plots Probability Plot Correlation Coefficient Plot
Software	Some general purpose statistical software programs provide a chi-square goodness-of-fit test for at least some of the common distributions. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test

Purpose: Test for Distributional Adequacy

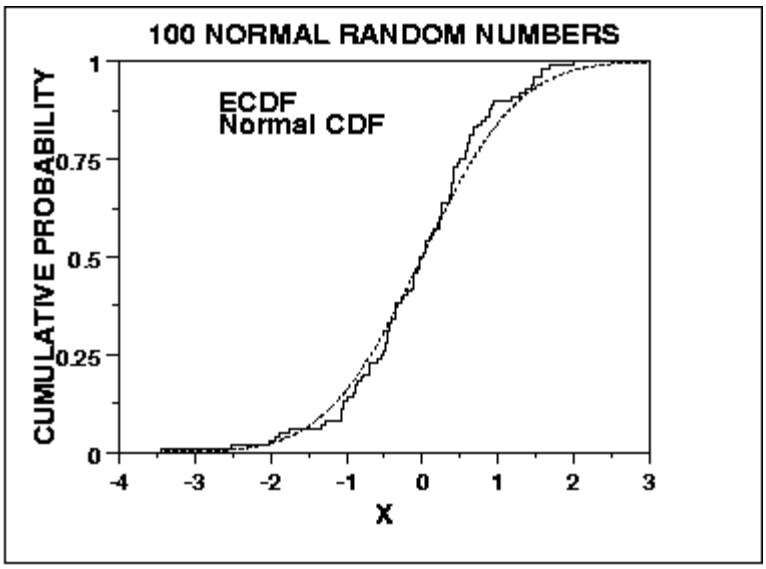
The Kolmogorov-Smirnov test (Chakravart, Laha, and Roy, 1967) is used to decide if a sample comes from a population with a specific distribution.

The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given *N ordered* data points $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, ..., $\boldsymbol{Y_N}$, the ECDF is defined as

$$E_N = n(i)/N$$

where ***n(i)*** is the number of points less than $\boldsymbol{Y_i}$ and the $\boldsymbol{Y_i}$ are ordered from smallest to largest value. This is a step function that increases by 1/*N* at the value of each ordered data point.

The graph below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.



Characteristics and Limitations of

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is

the K-S Test

an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important limitations:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test. However, the Anderson-Darling test is only available for a few specific distributions.

Definition

The Kolmogorov-Smirnov test is defined by:

H_0:	The data follow a specified distribution
H_a:	The data do not follow the specified distribution
Test Statistic:	The Kolmogorov-Smirnov test statistic is defined as $$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$ where ***F*** is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).
Significance Level:	α.
Critical Values:	The hypothesis regarding the distributional form is rejected if the test statistic, ***D***, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated. We do not provide the K-S tables in the Handbook since software programs that perform a K-S test will provide the relevant critical values.

Technical Note

Previous editions of e-Handbook gave the following formula for the computation of the Kolmogorov-Smirnov goodness of fit

statistic:

$$D = \max_{1 \le i \le N} |F(Y_i) - \frac{i}{N}|$$

This formula is in fact not correct. Note that this formula can be rewritten as:

$$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i}{N}, \frac{i}{N} - F(Y_i) \right)$$

This form makes it clear that an upper bound on the difference between these two formulas is i/N. For actual data, the difference is likely to be less than the upper bound.

For example, for $N = 20$, the upper bound on the difference between these two formulas is 0.05 (for comparison, the 5% critical value is 0.294). For $N = 100$, the upper bound is 0.001. In practice, if you have moderate to large sample sizes (say $N \ge 50$), these formulas are essentially equivalent.

Kolmogorov-Smirnov Test Example

We generated 1,000 random numbers for normal, double exponential, *t* with 3 degrees of freedom, and lognormal distributions. In all cases, the Kolmogorov-Smirnov test was applied to test for a normal distribution.

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the *t* random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

```
H0:  the data are normally distributed
Ha:  the data are not normally distributed

Y1 test statistic:  D = 0.0241492
Y2 test statistic:  D = 0.0514086
Y3 test statistic:  D = 0.0611935
Y4 test statistic:  D = 0.5354889

Significance level:  α = 0.05
Critical value:  0.04301
Critical region:  Reject H0 if D > 0.04301
```

As expected, the null hypothesis is not rejected for the normally distributed data, but is rejected for the remaining three data sets that are not normally distributed.

Questions

The Kolmogorov-Smirnov test can be used to answer the following types of questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?

Importance

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the data follow a Weibull distribution.

There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique, such as the sign test, that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumptions can be confirmed, the parametric techniques are generally preferred.

If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, using a non-parametric or robust technique may be required.

Related Techniques

Anderson-Darling goodness-of-fit Test
Chi-Square goodness-of-fit Test
Shapiro-Wilk Normality Test
Probability Plots
Probability Plot Correlation Coefficient Plot

Software

Some general purpose statistical software programs support the Kolmogorov-Smirnov goodness-of-fit test, at least for the more common distributions. Both Dataplot code and R code can be used to generate the analyses in this section.

1.3.5.17. Detection of Outliers

Introduction

An outlier is an observation that appears to deviate markedly from other observations in the sample.

Identification of potential outliers is important for the following reasons.

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques.

Labeling, Accomodation, Identification

Iglewicz and Hoaglin distinguish the three following issues with regards to outliers.

1. outlier labeling - flag potential outliers for further investigation (i.e., are the potential outliers erroneous data, indicative of an inappropriate distributional model, and so on).

2. outlier accomodation - use robust statistical techniques that will not be unduly affected by outliers. That is, if we cannot determine that potential outliers are erroneous observations, do we need modify our statistical analysis to more appropriately account for these observations?

3. outlier identification - formally test whether observations are outliers.

This section focuses on the labeling and identification

issues.

Normality Assumption

Identifying an observation as an outlier depends on the underlying distribution of the data. In this section, we limit the discussion to univariate data sets that are assumed to follow an approximately normal distribution. If the normality assumption for the data being tested is not valid, then a determination that there is an outlier may in fact be due to the non-normality of the data rather than the prescence of an outlier.

For this reason, it is recommended that you generate a normal probability plot of the data before applying an outlier test. Although you can also perform formal tests for normality, the prescence of one or more outliers may cause the tests to reject normality when it is in fact a reasonable assumption for applying the outlier test.

In addition to checking the normality assumption, the lower and upper tails of the normal probability plot can be a useful graphical technique for identifying potential outliers. In particular, the plot can help determine whether we need to check for a single outlier or whether we need to check for multiple outliers.

The box plot and the histogram can also be useful graphical tools in checking the normality assumption and in identifying potential outliers.

Single Versus Multiple Outliers

Some outlier tests are designed to detect the prescence of a single outlier while other tests are designed to detect the prescence of multiple outliers. It is not appropriate to apply a test for a single outlier sequentially in order to detect multiple outliers.

In addition, some tests that detect multiple outliers may require that you specify the number of suspected outliers exactly.

Masking and Swamping

Masking can occur when we specify too few outliers in the test. For example, if we are testing for a single outlier when there are in fact two (or more) outliers, these additional outliers may influence the value of the test statistic enough so that no points are declared as outliers.

On the other hand, swamping can occur when we specify too many outliers in the test. For example, if we are testing for two or more outliers when there is in fact only a single outlier, both points may be declared outliers (many tests will declare either all or none of the tested points as outliers).

Due to the possibility of masking and swamping, it is useful to complement formal outlier tests with graphical

methods. Graphics can often help identify cases where masking or swamping may be an issue. Swamping and masking are also the reason that many tests require that the exact number of outliers being tested must be specified.

Also, masking is one reason that trying to apply a single outlier test sequentially can fail. For example, if there are multiple outliers, masking may cause the outlier test for the first outlier to return a conclusion of no outliers (and so the testing for any additional outliers is not performed).

Z-Scores and Modified Z-Scores

The Z-score of an observation is defined as

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

with $\bar{Y}$ and s denoting the sample mean and sample standard deviation, respectively. In other words, data is given in units of how many standard deviations it is from the mean.

Although it is common practice to use Z-scores to identify possible outliers, this can be misleading (partiucarly for small sample sizes) due to the fact that the maximum Z-score is at most $(n-1)/\sqrt{n}$.

Iglewicz and Hoaglin recommend using the modified Z-score

$$M_i = \frac{0.6745(x_i - \tilde{x})}{\text{MAD}}$$

with MAD denoting the median absolute deviation and $\tilde{x}$ denoting the median.

These authors recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.

Formal Outlier Tests

A number of formal outlier tests have proposed in the literature. These can be grouped by the following characteristics:

- What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution.
- Is the test designed for a single outlier or is it designed for multiple outliers?
- If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers?

The following are a few of the more commonly used outlier tests for normally distributed data. This list is not exhaustive (a large number of outlier tests have been proposed in the literature). The tests given here are essentially based on the criterion of "distance from the mean". This is not the only criterion that could be used. For example, the Dixon test, which is not discussed here, is based a value being too large (or small) compared to its nearest neighbor.

1. Grubbs' Test - this is the recommended test when testing for a single outlier.

2. Tietjen-Moore Test - this is a generalization of the Grubbs' test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.

3. Generalized Extreme Studentized Deviate (ESD) Test - this test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

Lognormal Distribution

The tests discussed here are specifically based on the assumption that the data follow an approximately normal disribution. If your data follow an approximately lognormal distribution, you can transform the data to normality by taking the logarithms of the data and then applying the outlier tests discussed here.

Further Information

Iglewicz and Hoaglin provide an extensive discussion of the outlier tests given above (as well as some not given above) and also give a good tutorial on the subject of outliers. Barnett and Lewis provide a book length treatment of the subject.

In addition to discussing additional tests for data that follow an approximately normal distribution, these sources also discuss the case where the data are not normally distributed.

1.3.5.18. Yates Algorithm

Purpose: Estimate Factor Effects in a 2-Level Factorial Design

Full factorial and fractional factorial designs are common in designed experiments for engineering and scientific applications.

In these designs, each factor is assigned two levels. These are typically called the low and high levels. For computational purposes, the factors are scaled so that the low level is assigned a value of -1 and the high level is assigned a value of +1. These are also commonly referred to as "-" and "+".

A full factorial design contains all possible combinations of low/high levels for all the factors. A fractional factorial design contains a carefully chosen subset of these combinations. The criterion for choosing the subsets is discussed in detail in the process improvement chapter.

The Yates algorithm exploits the special structure of these designs to generate least squares estimates for factor effects for all factors and all relevant interactions.

The mathematical details of the Yates algorithm are given in chapter 10 of Box, Hunter, and Hunter (1978). Natrella (1963) also provides a procedure for testing the significance of effect estimates.

The effect estimates are typically complemented by a number of graphical techniques such as the DOE mean plot and the DOE contour plot ("DOE" represents "design of experiments"). These are demonstrated in the eddy current case study.

Yates Order

Before performing the Yates algorithm, the data should be arranged in "Yates order". That is, given k factors, the kth column consists of 2^{k-1} minus signs (i.e., the low level of the factor) followed by 2^{k-1} plus signs (i.e., the high level of the factor). For example, for a full factorial design with three factors, the design matrix is

```
- - -
+ - -
- + -
```

```
+ + -
- - +
+ - +
- + +
+ + +
```

Determining the Yates order for fractional factorial designs requires knowledge of the confounding structure of the fractional factorial design.

Yates Algorithm

The Yates algorithm is demonstrated for the eddy current data set. The data set contains eight measurements from a two-level, full factorial design with three factors. The purpose of the experiment is to identify factors that have the most effect on eddy current measurements.

In the "Effect" column, we list the main effects and interactions from our factorial experiment in standard order. In the "Response" column, we list the measurement results from our experiment in Yates order.

Effect	Response	Col 1	Col 2	Col 3	Estimate
Mean	1.70	6.27	10.21	21.27	2.65875
X1	4.57	3.94	11.06	12.41	1.55125
X2	0.55	6.10	5.71	-3.47	-0.43375
X1*X2	3.39	4.96	6.70	0.51	0.06375
X3	1.51	2.87	-2.33	0.85	0.10625
X1*X3	4.59	2.84	-1.14	0.99	0.12375
X2*X3	0.67	3.08	-0.03	1.19	0.14875
X1*X2*X3	4.29	3.62	0.54	0.57	0.07125

Sum of responses:	21.27
Sum-of-squared responses:	77.7707
Sum-of-squared Col 3:	622.1656

The first four values in Col 1 are obtained by adding adjacent pairs of responses, for example 4.57 + 1.70 = 6.27, and 3.39 + 0.55 = 3.94. The second four values in Col 1 are obtained by subtracting the same adjacent pairs of responses, for example, 4.57 - 1.70 = 2.87, and 3.39 - 0.55 = 2.84. The values in Col 2 are calculated in the same way, except that we are adding and subtracting adjacent values from Col 1. Col 3 is computed using adjacent values from Col 2. Finally, we obtain the "Estimate" column by dividing the values in Col 3 by the total number of responses, 8.

We can check our calculations by making sure that the first value in Col 3 (21.27) is the sum of all the responses. In addition, the sum-of-squared responses (77.7707) should equal the sum-of-squared Col 3 values divided by 8 (622.1656/8 = 77.7707).

Practical Considerations

The Yates algorithm provides a convenient method for computing effect estimates; however, the same information is easily obtained from statistical software using either an analysis of variance or regression procedure. The methods for analyzing data from a designed experiment are discussed more fully in the chapter on Process Improvement.

Graphical Presentation

The following plots may be useful to complement the quantitative information from the Yates algorithm.

1. Ordered data plot
2. Ordered absolute effects plot
3. Cumulative residual standard deviation plot

Questions

The Yates algorithm can be used to answer the following question.

1. What is the estimated effect of a factor on the response?

Related Techniques

Multi-factor analysis of variance
DOE mean plot
Block plot
DOE contour plot

Case Study

The analysis of a full factorial design is demonstrated in the eddy current case study.

Software

All statistical software packages are capable of estimating effects using an analysis of variance or least squares regression procedure.

1.3.5.18.1. Defining Models and Prediction Equations

For Orthogonal Designs, Parameter Estimates Don't Change as Additional Terms Are Added

In most cases of least-squares fitting, the model coefficients for previously added terms change depending on what was successively added. For example, the X1 coefficient might change depending on whether or not an X2 term was included in the model. This is **not** the case when the design is orthogonal, as is a 2^3 full factorial design. For orthogonal designs, the estimates for the previously included terms do not change as additional terms are added. This means the ranked list of parameter estimates are the least-squares coefficient estimates for progressively more complicated models.

Example Prediction Equation

We use the parameter estimates derived from a least-squares analysis for the eddy current data set to create an example prediction equation.

Parameter	Estimate
Mean	2.65875
X1	1.55125
X2	-0.43375
X1*X2	0.06375
X3	0.10625
X1*X3	0.12375
X2*X3	0.14875
X1*X2*X3	0.07125

A prediction equation predicts a value of the reponse variable for given values of the factors. The equation we select can include all the factors shown above, or it can include a subset of the factors. For example, one possible prediction equation using only two factors, X1 and X2, is:

$$\hat{Y} = 2.65875 + 1.55125 \cdot X_1 - 0.43375 \cdot X_2$$

The least-squares parameter estimates in the prediction equation reflect the change in response for a one-unit change in the factor value. To obtain "full" effect estimates (as computed using the Yates algorithm) for the change in factor levels from -1 to +1, the effect estimates (except for the intercept) would be multiplied by two.

Remember that the Yates algorithm is just a convenient

method for computing effects, any statistical software package with least-squares regression capabilities will produce the same effects as well as many other useful analyses.

Model Selection

We want to select the most appropriate model for our data while balancing the following two goals.

1. We want the model to include all important factors.
2. We want the model to be parsimonious. That is, the model should be as simple as possible.

Note that the residual standard deviation alone is insufficient for determining the most appropriate model as it will always be decreased by adding additional factors. The next section describes a number of approaches for determining which factors (and interactions) to include in the model.

1.3.5.18.2. Important Factors

Identify Important Factors

We want to select the most appropriate model to represent our data. This requires balancing the following two goals.

1. We want the model to include all important factors.
2. We want the model to be parsimonious. That is, the model should be as simple as possible.

In short, we want our model to include all the important factors and interactions and to omit the unimportant factors and interactions.

Seven criteria are utilized to define important factors. These seven criteria are not all equally important, nor will they yield identical subsets, in which case a consensus subset or a weighted consensus subset must be extracted. In practice, some of these criteria may not apply in all situations.

These criteria will be examined in the context of the eddy current data set. The parameter estimates computed using least-squares analysis are shown below.

Parameter	Estimate
Mean	2.65875
X1	1.55125
X2	-0.43375
X1*X2	0.06375
X3	0.10625
X1*X3	0.12375
X2*X3	0.14875
X1*X2*X3	0.07125

In practice, not all of these criteria will be used with every analysis (and some analysts may have additional criteria). These critierion are given as useful guidelines. Most analysts will focus on those criteria that they find most useful.

Criteria for Including Terms in the Model

The seven criteria that we can use in determining whether to keep a factor in the model can be summarized as follows.

1. Parameters: Engineering Significance
2. Parameters: Order of Magnitude
3. Parameters: Statistical Significance
4. Parameters: Probability Plots
5. Effects: Youden Plot
6. Residual Standard Deviation: Engineering Significance
7. Residual Standard Deviation: Statistical Significance

The first four criteria focus on parameter estimates with three numeric criteria and one

graphical criteria. The fifth criteria focuses on effects, which are twice the parameter estimates. The last two criteria focus on the residual standard deviation of the model. We discuss each of these seven criteria in detail in the sections that following.

Parameters: Engineering Significance

The minimum engineering significant difference is defined as

$$|\hat{\beta}_i| > \Delta$$

where $|\hat{\beta}_i|$ is the absolute value of the parameter estimate and Δ is the minimum engineering significant difference.

That is, declare a factor as "important" if the parameter estimate is greater than some a priori declared engineering difference. This implies that the engineering staff have in fact stated what a minimum difference will be. Oftentimes this is not the case. In the absence of an a priori difference, a good rough rule for the minimum engineering significant Δ is to keep only those factors whose parameter estimate is greater than, say, 10% of the current production average. In this case, let's say that the average detector has a sensitivity of 2.5 ohms. This would suggest that we would declare all factors whose parameter is greater than 10 % of 2.5 ohms = 0.25 ohm to be significant (from an engineering point of view).

Based on this minimum engineering significant difference criterion, we conclude that we should keep two terms: X1 and X2.

Parameters: Order of Magnitude

The order of magnitude criterion is defined as

$$|\hat{\beta}_i| < 0.10 * max|\hat{\beta}_i|$$

That is, exclude any factor that is less than 10 % of the maximum parameter size. We may or may not keep the other factors. This criterion is neither engineering nor statistical, but it does offer some additional numerical insight. For the current example, the largest parameter is from X1 (1.55125 ohms), and so 10 % of that is 0.155 ohms, which suggests keeping all factors whose parameters exceed 0.155 ohms.

Based on the order-of-magnitude criterion, we thus conclude that we should keep two terms: X1 and X2. A third term, X2*X3 (0.14875), is just slightly under the cutoff level, so we may consider keeping it based on the other criterion.

Parameters: Statistical Significance

Statistical significance is defined as

$$|\hat{\beta}_i| > 2 \text{ s.e.}(\hat{\beta}_i)$$

That is, declare a factor as important if its parameter is more than 2 standard deviations away from 0 (0, by definition, meaning "no effect").

The "2" comes from normal theory (more specifically, a value of 1.96 yields a 95 % confidence interval). More precise values would come from *t*-distribution theory.

The difficulty with this is that in order to invoke this criterion we need the standard deviation, σ, of an observation. This is problematic because

1. the engineer may not know σ;
2. the experiment might not have replication, and so a model-free estimate of σ is not obtainable;

3. obtaining an estimate of σ by assuming the sometimes- employed assumption of ignoring 3-term interactions and higher may be incorrect from an engineering point of view.

For the eddy current example:

1. the engineer did **not** know σ;
2. the design (a 2^3 full factorial) did **not** have replication;
3. ignoring 3-term interactions and higher interactions leads to an estimate of σ based on omitting only a single term: the X1*X2*X3 interaction.

For the eddy current example, if one assumes that the 3-term interaction is nil and hence represents a single drawing from a population centered at zero, then an estimate of the standard deviation of a parameter is simply the estimate of the 3-factor interaction (0.07125). Two standard deviations is thus 0.1425. For this example, the rule is thus to keep all $|\hat{\beta}_i| >$ 0.1425.

This results in keeping three terms: X1 (1.55125), X2 (-0.43375), and X1*X2 (0.14875).

Parameters: Probability Plots

Probability plots can be used in the following manner.

1. Normal Probability Plot: Keep a factor as "important" if it is well off the line through zero on a normal probability plot of the parameter estimates.

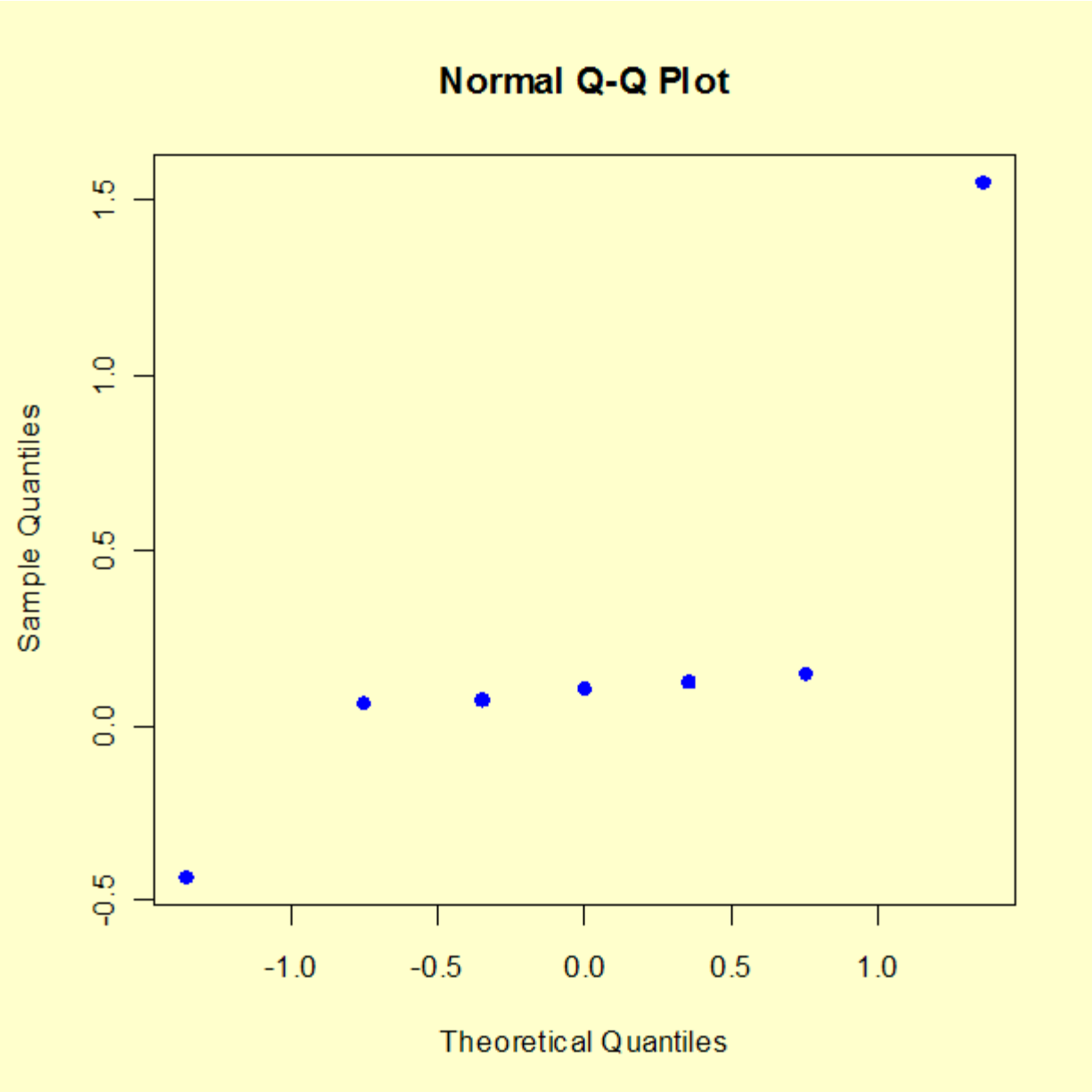
2. Half-Normal Probability Plot: Keep a factor as "important" if it is well off the line near zero on a half-normal probability plot of the absolute value of parameter estimates.

Both of these methods are based on the fact that the least-squares estimates of parameters for these two-level orthogonal designs are simply half the difference of averages and so the central limit theorem, loosely applied, suggests that (if no factor were important) the parameter estimates should have approximately a normal distribution with mean zero and the absolute value of the estimates should have a half-normal distribution.

Since the half-normal probability plot is only concerned with parmeter magnitudes as opposed to signed parameters (which are subject to the vagaries of how the initial factor codings +1 and -1 were assigned), the half-normal probability plot is preferred by some over the normal probability plot.

Normal Probablity Plot of Parameters

The following normal probability plot shows the parameter estimates for the eddy current data.



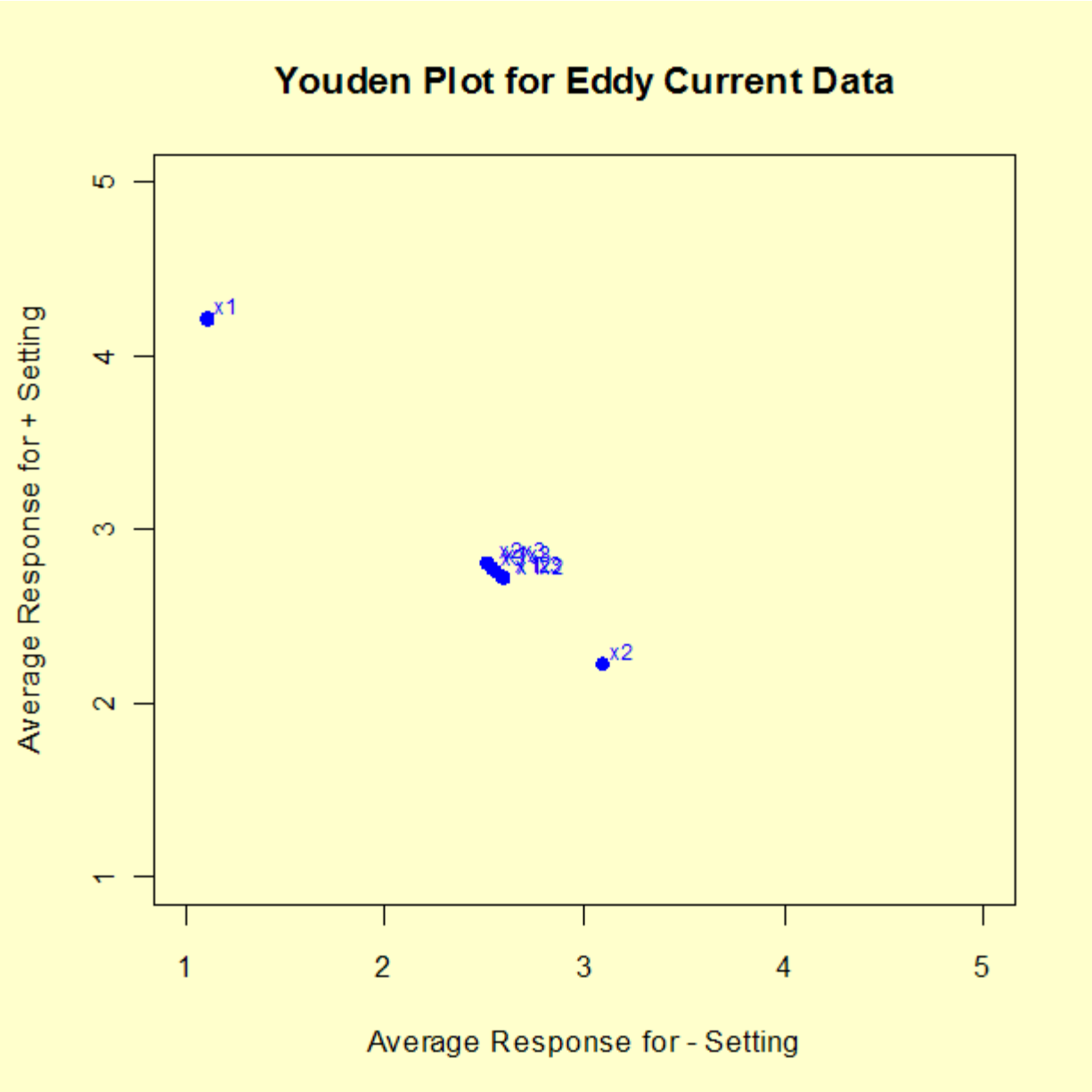
For the example at hand, the probability plot clearly shows two factors (X1 and X2) displaced off the line. All of the remaining five parameters are behaving like random drawings from a normal distribution centered at zero, and so are deemed to be statistically non-significant. In conclusion, this rule keeps two factors: X1 (1.55125) and X2 (-0.43375).

Averages: Youden Plot

A Youden plot can be used in the following way. Keep a factor as "important" if it is displaced away from the central-tendancy "bunch" in a Youden plot of high and low averages. By definition, a factor is important when its average response for the low (-1) setting is significantly different from its average response for the high (+1) setting. (Note that effects are twice the parameter estimates.) Conversely, if the low and high averages are about the same, then what difference does it make which setting to use and so why would such a factor be considered important? This fact in combination with the intrinsic benefits of the Youden plot for comparing pairs of items leads to the technique of generating a Youden plot of the low and high averages.

Youden Plot of Effect Estimates

The following is the Youden plot of the effect estimatess for the eddy current data.



For the example at hand, the Youden plot clearly shows a cluster of points near the grand average (2.65875) with two displaced points above (factor 1) and below (factor 2). Based on the Youden plot, we conclude to keep two factors: X1 (1.55125) and X2 (-0.43375).

Residual Standard Deviation: Engineering Significance

This criterion is defined as

Residual Standard Deviation > Cutoff

That is, declare a factor as "important" if the cumulative model that includes the factor (and all larger factors) has a residual standard deviation smaller than an a priori engineering-specified minimum residual standard deviation.

This criterion is different from the others in that it is model focused. In practice, this criterion states that starting with the largest parameter, we cumulatively keep adding terms to the model and monitor how the residual standard deviation for each progressively more complicated model becomes smaller. At some point, the cumulative model will become complicated enough and comprehensive enough that the resulting residual standard deviation will drop below the pre-specified engineering cutoff for the residual standard deviation. At that point, we stop adding terms and declare all of the model-included terms to be "important" and

everything not in the model to be "unimportant".

This approach implies that the engineer has considered what a minimum residual standard deviation should be. In effect, this relates to what the engineer can tolerate for the magnitude of the typical residual (the difference between the raw data and the predicted value from the model). In other words, how good does the engineer want the prediction equation to be. Unfortunately, this engineering specification has not always been formulated and so this criterion can become moot.

In the absence of a prior specified cutoff, a good rough rule for the minimum engineering residual standard deviation is to keep adding terms until the residual standard deviation just dips below, say, 5 % of the current production average. For the eddy current data, let's say that the average detector has a sensitivity of 2.5 ohms. Then this would suggest that we would keep adding terms to the model until the residual standard deviation falls below 5 % of 2.5 ohms = 0.125 ohms.

Model	Residual Std. Dev.
Mean + X1	0.57272
Mean + X1 + X2	0.30429
Mean + X1 + X2 + X2*X3	0.26737
Mean + X1 + X2 + X2*X3 + X1*X3	0.23341
Mean + X1 + X2 + X2*X3 + X1*X3 + X3	0.19121
Mean + X1 + X2 + X2*X3 + X1*X3 + X3 + X1*X2*X3	0.18031
Mean + X1 + X2 + X2*X3 + X1*X3 + X3 + X1*X2*X3 + X1*X2	NA

Based on the minimum residual standard deviation criteria, and we would include **all** terms in order to drive the residual standard deviation below 0.125. Again, the 5 % rule is a rough-and-ready rule that has no basis in engineering or statistics, but is simply a "numerics". Ideally, the engineer has a better cutoff for the residual standard deviation that is based on how well he/she wants the equation to peform in practice. If such a number were available, then for this criterion and data set we would select something less than the entire collection of terms.

Residual Standard Deviation: Statistical Significance

This criterion is defined as

Residual Standard Deviation > σ

where σ is the standard deviation of an observation under replicated conditions.

That is, declare a term as "important" until the cumulative model that includes the term has a residual standard deviation smaller than σ. In essence, we are allowing that we cannot demand a model fit any better than what we would obtain if we had replicated data; that is, we cannot demand that the residual standard deviation from any fitted model be any smaller than the (theoretical or actual) replication standard deviation. We can drive the fitted standard deviation down (by adding terms) until it achieves a value close to σ, but to attempt to drive it down further means that we are, in effect, trying to fit noise.

In practice, this criterion may be difficult to apply because

1. the engineer may not know σ;
2. the experiment might not have replication, and so a model-free estimate of σ is not obtainable.

For the current case study:

1. the engineer did **not** know σ;

2. the design (a 2^3 full factorial) did **not** have replication. The most common way of having replication in such designs is to have replicated center points at the center of the cube ((X1,X2,X3) = (0,0,0)).

Thus for this current case, this criteria could **not** be used to yield a subset of "important" factors.

Conclusions In summary, the seven criteria for specifying "important" factors yielded the following for the eddy current data:

1. Parameters, Engineering Significance: X1, X2
2. Parameters, Numerically Significant: X1, X2
3. Parameters, Statistically Significant: X1, X2, X2*X3
4. Parameters, Probability Plots: X1, X2
5. Effects, Youden Plot: X1, X2
6. Residual SD, Engineering Significance: all 7 terms
7. Residual SD, Statistical Significance: not applicable

Such conflicting results are common. Arguably, the three most important criteria (listed in order of most important) are:

4. Parameters, Probability Plots: X1, X2
1. Parameters, Engineering Significance: X1, X2
3. Residual SD, Engineering Significance: all 7 terms

Scanning all of the above, we thus declare the following consensus for the eddy current data:

1. Important Factors: X1 and X2
2. Parsimonious Prediction Equation:

$$\hat{Y} = 2.65875 + 1.55125 \cdot X_1 - 0.43375 \cdot X_2$$

(with a residual standard deviation of 0.30429 ohms)

Note that this is the initial model selection. We still need to perform model validation with a residual analysis.

1.3.6. Probability Distributions

Probability Distributions

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level.

Some practical uses of probability distributions are:

- To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
- For univariate data, it is often useful to determine a reasonable distributional model for the data.
- Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies with random numbers generated from using a specific probability distribution are often needed.

Table of Contents

1. What is a probability distribution?
2. Related probability functions
3. Families of distributions
4. Location and scale parameters
5. Estimating the parameters of a distribution
6. A gallery of common distributions
7. Tables for probability distributions

1.3.6.1. What is a Probability Distribution

Discrete Distributions

The mathematical definition of a discrete probability function, p(x), is a function that satisfies the following properties.

1. The probability that x can take a specific value is p(x). That is

   $$P[X = x] = p(x) = p_x$$

2. p(x) is non-negative for all real x.

3. The sum of p(x) over all possible values of x is 1, that is

   $$\sum_j p_j = 1$$

   where j represents all possible values that x can have and p_j is the probability at x_j.

   One consequence of properties 2 and 3 is that 0 <= p(x) <= 1.

What does this actually mean? A discrete probability function is a function that can take a discrete number of values (not necessarily finite). This is most often the non-negative integers or some subset of the non-negative integers. There is no mathematical restriction that discrete probability functions only be defined at integers, but in practice this is usually what makes sense. For example, if you toss a coin 6 times, you can get 2 heads or 3 heads but not 2 1/2 heads. Each of the discrete values has a certain probability of occurrence that is between zero and one. That is, a discrete function that allows negative values or values greater than one is not a probability function. The condition that the probabilities sum to one means that at least one of the values has to occur.

Continuous Distributions

The mathematical definition of a continuous probability function, f(x), is a function that satisfies the following properties.

1. The probability that x is between two points a and b is

$$p[a \leq x \leq b] = \int_a^b f(x)dx$$

2. It is non-negative for all real x.

3. The integral of the probability function is one, that is

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

What does this actually mean? Since continuous probability functions are defined for an infinite number of points over a continuous interval, the probability at a single point is always zero. Probabilities are measured over intervals, not single points. That is, the area under the curve between two distinct points defines the probability for that interval. This means that the height of the probability function can in fact be greater than one. The property that the integral must equal one is equivalent to the property for discrete distributions that the sum of all the probabilities must equal one.

Probability Mass Functions Versus Probability Density Functions

Discrete probability functions are referred to as probability mass functions and continuous probability functions are referred to as probability density functions. The term probability functions covers both discrete and continuous distributions. When we are referring to probability functions in generic terms, we may use the term probability density functions to mean both discrete and continuous probability functions.

1.3.6.2. Related Distributions

Probability distributions are typically defined in terms of the probability density function. However, there are a number of probability functions used in applications.

Probability Density Function

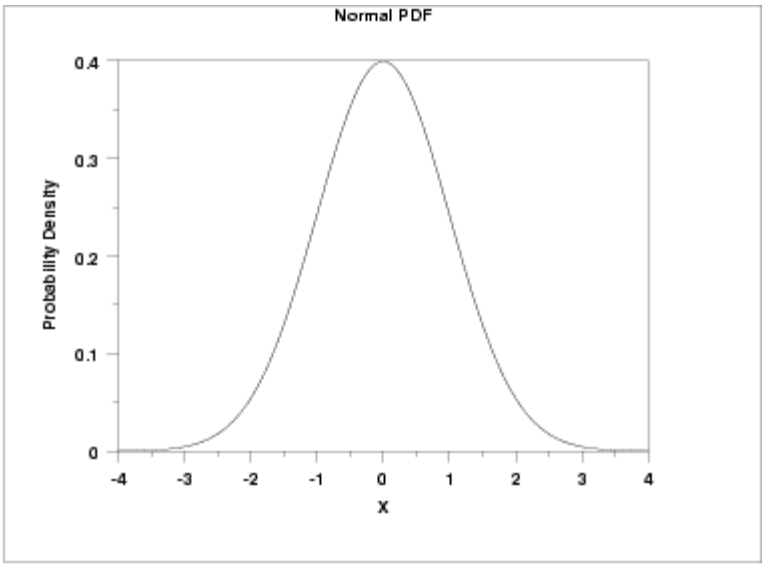
For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

$$\int_a^b f(x)dx = Pr[a \leq X \leq b]$$

For a discrete distribution, the pdf is the probability that the variate takes the value x.

$$f(x) = Pr[X = x]$$

The following is the plot of the normal probability density function.



Cumulative Distribution Function

The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x. That is

$$F(x) = Pr[X \leq x] = \alpha$$

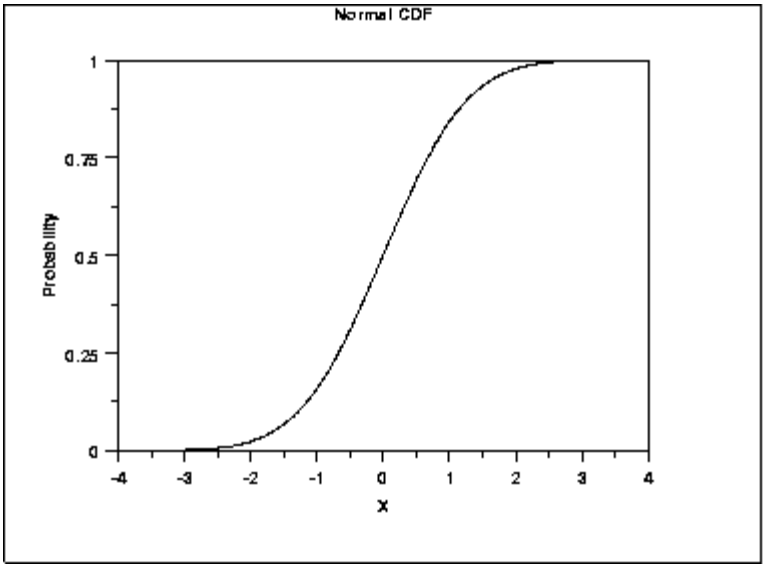
For a continuous distribution, this can be expressed mathematically as

$$F(x) = \int_{-\infty}^{x} f(\mu) d\mu$$

For a discrete distribution, the cdf can be expressed as

$$F(x) = \sum_{i=0}^{x} f(i)$$

The following is the plot of the normal cumulative distribution function.



The horizontal axis is the allowable domain for the given probability function. Since the vertical axis is a probability, it must fall between zero and one. It increases from zero to one as we go from left to right on the horizontal axis.

Percent Point Function

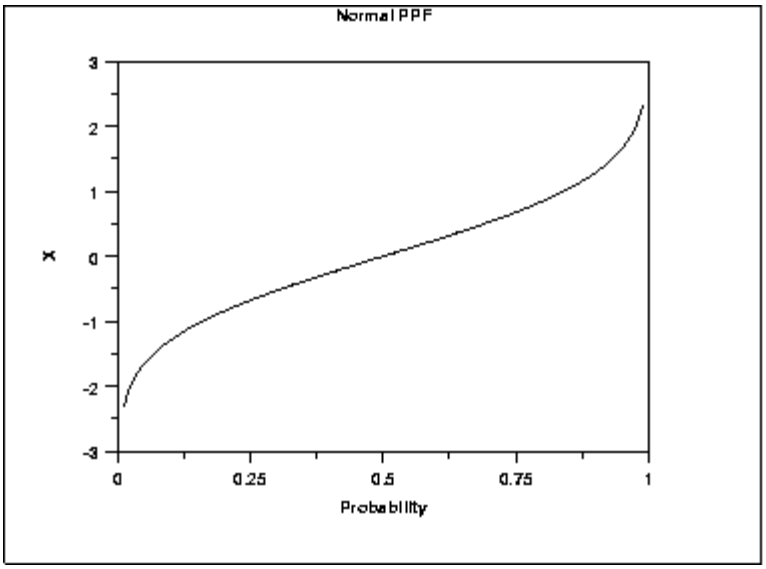
The percent point function (ppf) is the inverse of the cumulative distribution function. For this reason, the percent point function is also commonly referred to as the inverse distribution function. That is, for a distribution function we calculate the probability that the variable is less than or equal to x for a given x. For the percent point function, we start with the probability and compute the corresponding x for the cumulative distribution. Mathematically, this can be expressed as

$$Pr[X \leq G(\alpha)] = \alpha$$

or alternatively

$$x = G(\alpha) = G(F(x))$$

The following is the plot of the normal percent point function.



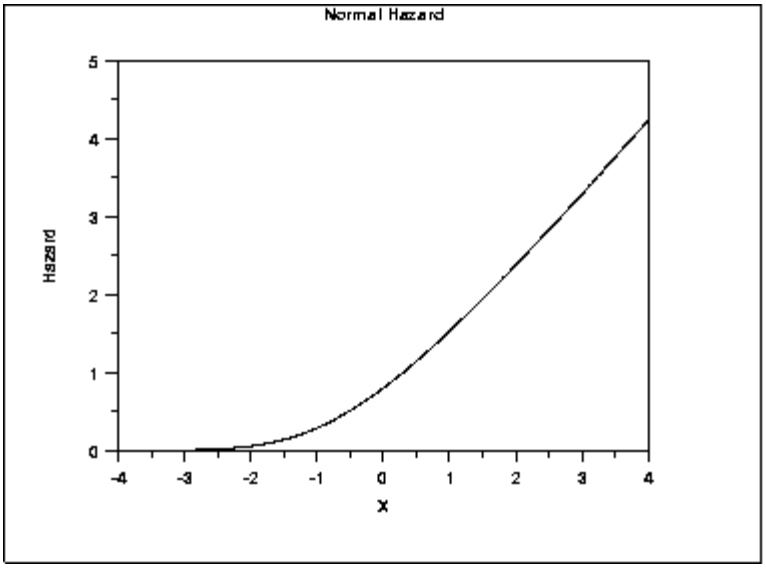
Since the horizontal axis is a probability, it goes from zero to one. The vertical axis goes from the smallest to the largest value of the cumulative distribution function.

Hazard Function

The hazard function is the ratio of the probability density function to the survival function, $S(x)$.

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

The following is the plot of the normal distribution hazard function.



Hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the conditional failure density function rather than the hazard function.

Cumulative Hazard

The cumulative hazard function is the integral of the hazard function.

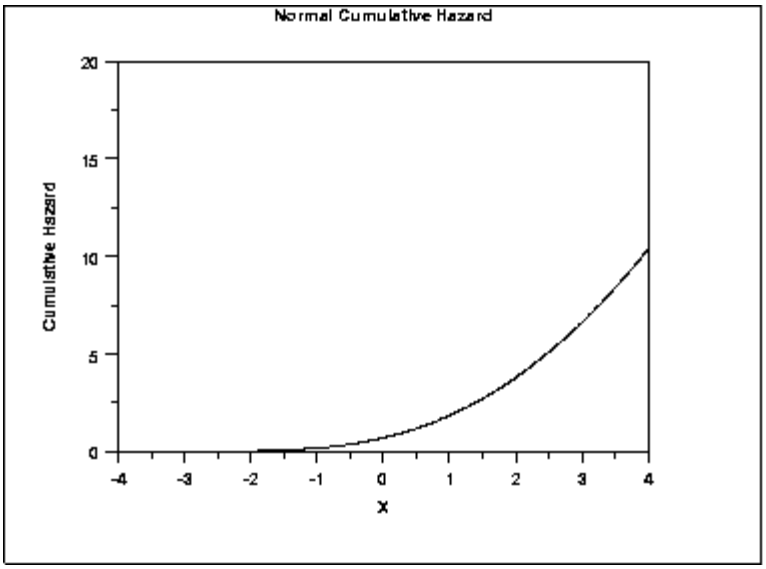
Function

$$H(x) = \int_{-\infty}^{x} h(\mu)d\mu$$

This can alternatively be expressed as

$$H(x) = -\ln(1 - F(x))$$

The following is the plot of the normal cumulative hazard function.



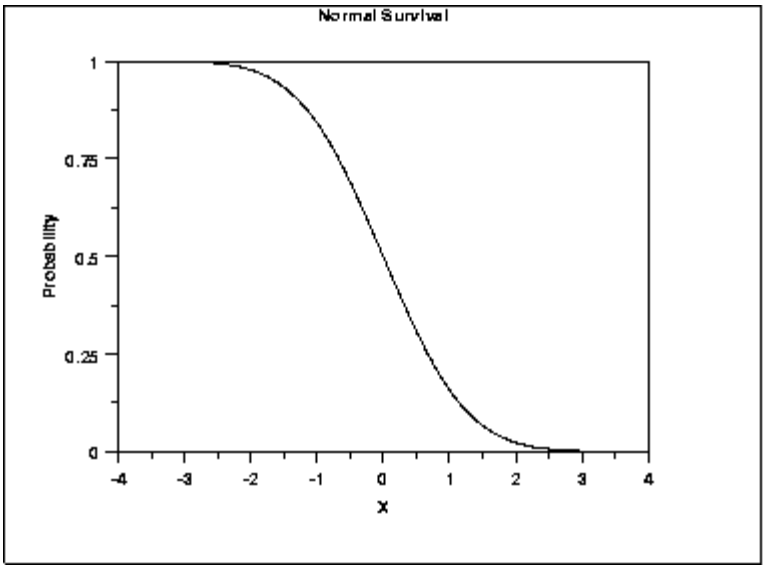
Cumulative hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the hazard function rather than the cumulative hazard function.

Survival Function

Survival functions are most often used in reliability and related fields. The survival function is the probability that the variate takes a value greater than x.

$$S(x) = Pr[X > x] = 1 - F(x)$$

The following is the plot of the normal distribution survival function.



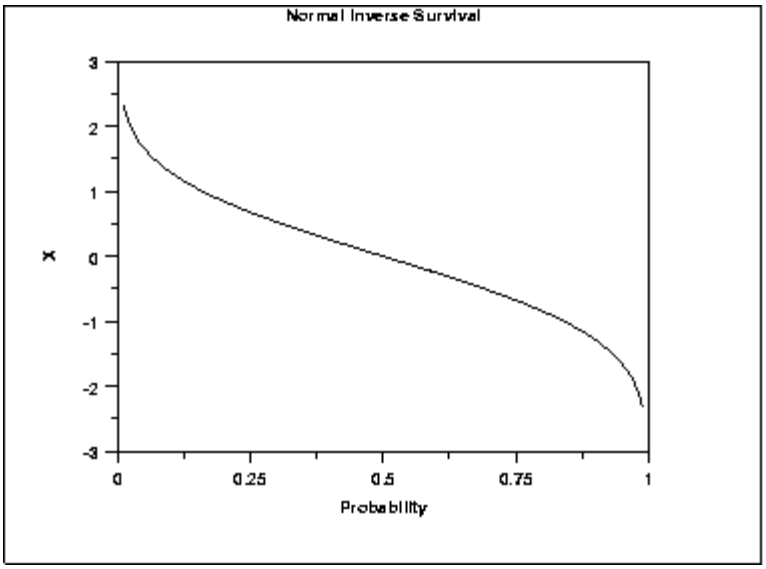
For a survival function, the y value on the graph starts at 1 and monotonically decreases to zero. The survival function should be compared to the cumulative distribution function.

Inverse Survival Function

Just as the percent point function is the inverse of the cumulative distribution function, the survival function also has an inverse function. The inverse survival function can be defined in terms of the percent point function.

$$Z(\alpha) = G(1 - \alpha)$$

The following is the plot of the normal distribution inverse survival function.



As with the percent point function, the horizontal axis is a probability. Therefore the horizontal axis goes from 0 to 1 regardless of the particular distribution. The appearance is similar to the percent point function. However, instead of going from the smallest to the largest value on the vertical axis, it goes from the largest to the smallest value.

1.3.6.3. Families of Distributions

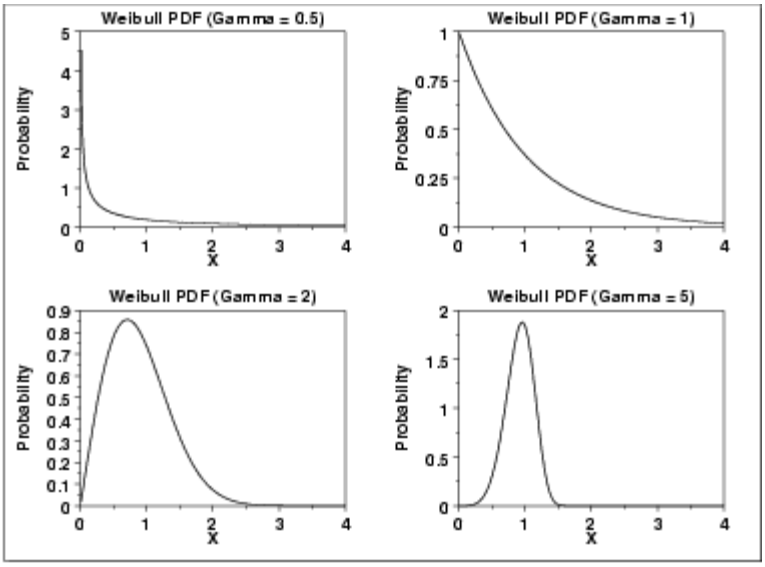
Shape Parameters

Many probability distributions are not a single distribution, but are in fact a family of distributions. This is due to the distribution having one or more shape parameters.

Shape parameters allow a distribution to take on a variety of shapes, depending on the value of the shape parameter. These distributions are particularly useful in modeling applications since they are flexible enough to model a variety of data sets.

Example: Weibull Distribution

The Weibull distribution is an example of a distribution that has a shape parameter. The following graph plots the Weibull pdf with the following values for the shape parameter: 0.5, 1.0, 2.0, and 5.0.



The shapes above include an exponential distribution, a right-skewed distribution, and a relatively symmetric distribution.

The Weibull distribution has a relatively simple distributional form. However, the shape parameter allows the Weibull to assume a wide variety of shapes. This combination of simplicity and flexibility in the shape of the Weibull distribution has made it an effective distributional model in reliability applications. This ability to model a wide variety of distributional shapes using a relatively simple distributional form is possible with many other distributional

families as well.

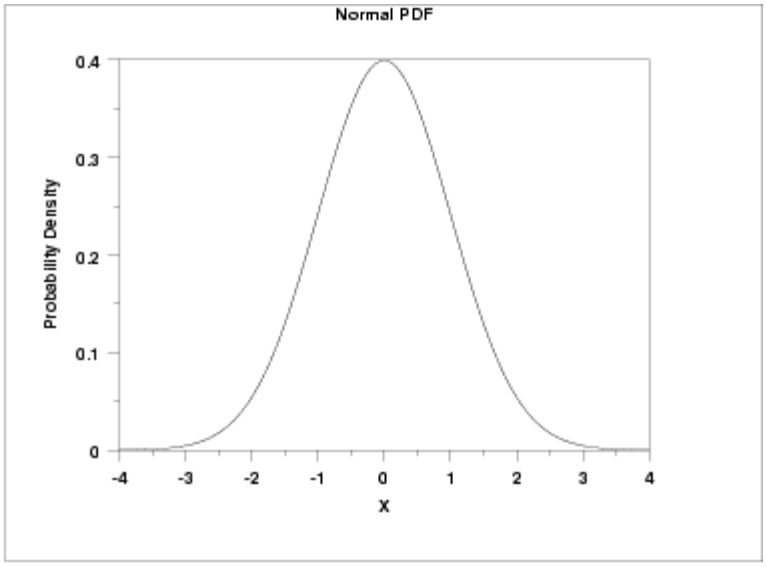
PPCC Plots The PPCC plot is an effective graphical tool for selecting the member of a distributional family with a single shape parameter that best fits a given set of data.

1.3.6.4. Location and Scale Parameters

Normal PDF

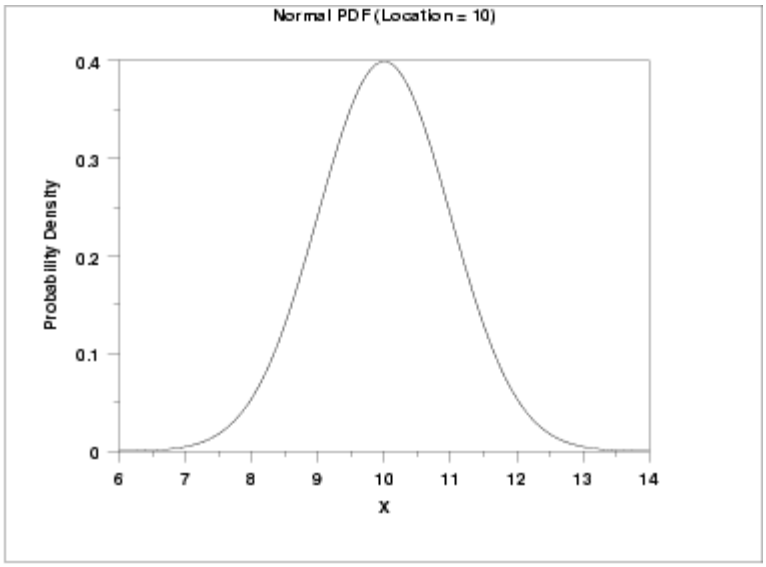
A probability distribution is characterized by location and scale parameters. Location and scale parameters are typically used in modeling applications.

For example, the following graph is the probability density function for the standard normal distribution, which has the location parameter equal to zero and scale parameter equal to one.



Location Parameter

The next plot shows the probability density function for a normal distribution with a location parameter of 10 and a scale parameter of 1.

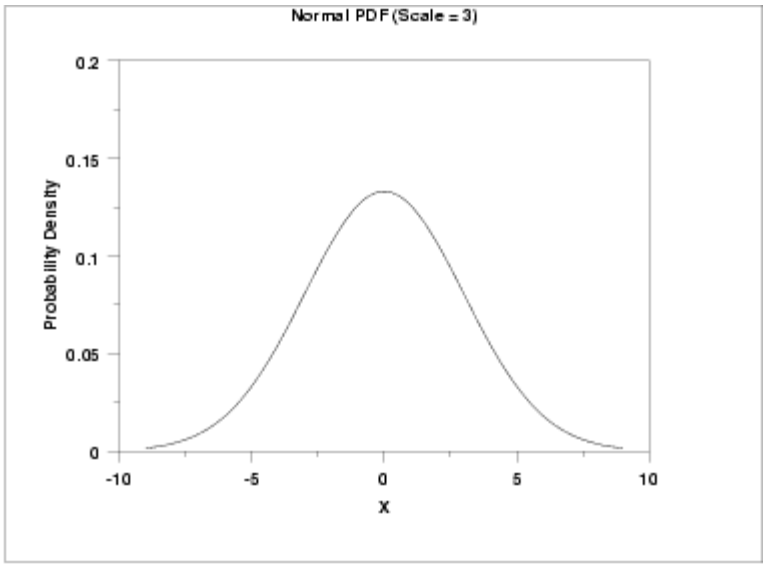


The effect of the location parameter is to translate the graph, relative to the standard normal distribution, 10 units to the right on the horizontal axis. A location parameter of -10 would have shifted the graph 10 units to the left on the horizontal axis.

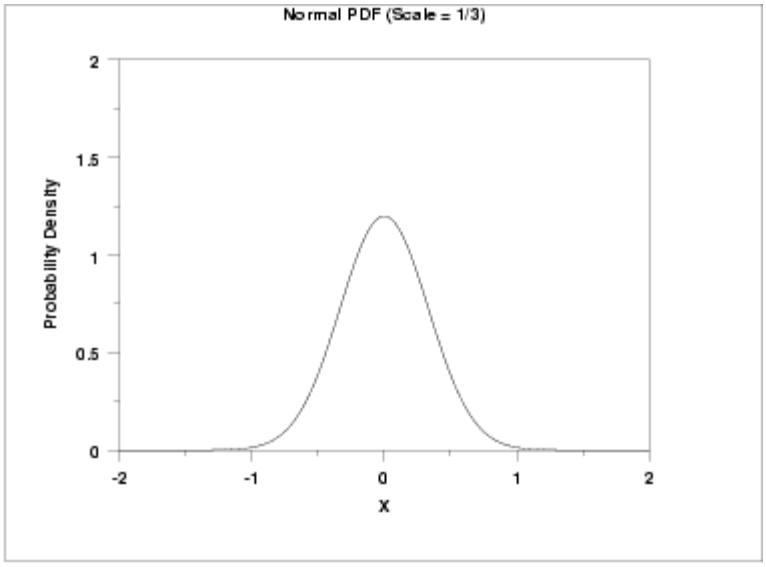
That is, a location parameter simply shifts the graph left or right on the horizontal axis.

Scale Parameter

The next plot has a scale parameter of 3 (and a location parameter of zero). The effect of the scale parameter is to stretch out the graph. The maximum y value is approximately 0.13 as opposed 0.4 in the previous graphs. The y value, i.e., the vertical axis value, approaches zero at about (+/-) 9 as opposed to (+/-) 3 with the first graph.



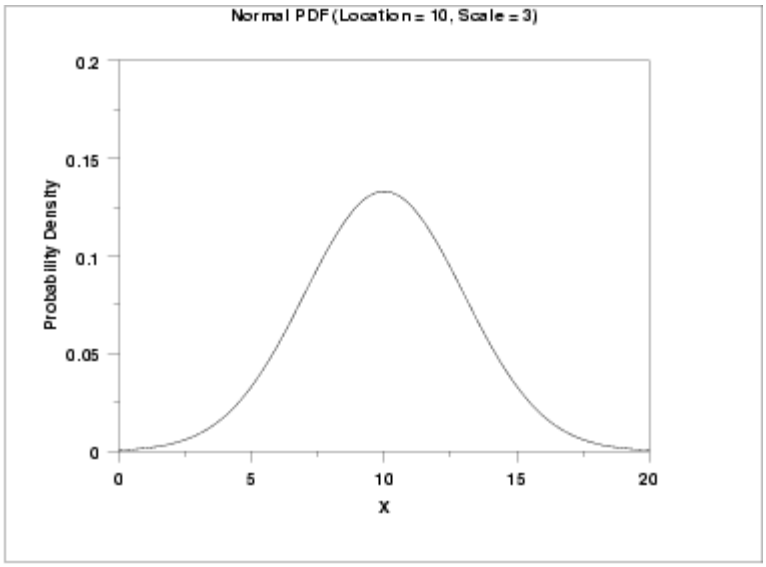
In contrast, the next graph has a scale parameter of 1/3 (=0.333). The effect of this scale parameter is to squeeze the pdf. That is, the maximum y value is approximately 1.2 as opposed to 0.4 and the y value is near zero at (+/-) 1 as opposed to (+/-) 3.



The effect of a scale parameter greater than one is to stretch the pdf. The greater the magnitude, the greater the stretching. The effect of a scale parameter less than one is to compress the pdf. The compressing approaches a spike as the scale parameter goes to zero. A scale parameter of 1 leaves the pdf unchanged (if the scale parameter is 1 to begin with) and non-positive scale parameters are not allowed.

Location and Scale Together

The following graph shows the effect of both a location and a scale parameter. The plot has been shifted right 10 units and stretched by a factor of 3.



Standard Form

The standard form of any distribution is the form that has location parameter zero and scale parameter one.

It is common in statistical software packages to only compute the standard form of the distribution. There are formulas for converting from the standard form to the form with other location and scale parameters. These formulas are independent of the particular probability distribution.

Formulas for Location and Scale Based on the Standard Form

The following are the formulas for computing various probability functions based on the standard form of the distribution. The parameter *a* refers to the location parameter and the parameter *b* refers to the scale parameter. Shape parameters are not included.

Cumulative Distribution Function	$F(x;a,b) = F((x-a)/b;0,1)$
Probability Density Function	$f(x;a,b) = (1/b)f((x-a)/b;0,1)$
Percent Point Function	$G(\alpha;a,b) = a + bG(\alpha;0,1)$
Hazard Function	$h(x;a,b) = (1/b)h((x-a)/b;0,1)$
Cumulative Hazard Function	$H(x;a,b) = H((x-a)/b;0,1)$
Survival Function	$S(x;a,b) = S((x-a)/b;0,1)$
Inverse Survival Function	$Z(\alpha;a,b) = a + bZ(\alpha;0,1)$
Random Numbers	$Y(a,b) = a + bY(0,1)$

Relationship to Mean and Standard Deviation

For the normal distribution, the location and scale parameters correspond to the mean and standard deviation, respectively. However, this is not necessarily true for other distributions. In fact, it is not true for most distributions.

1.3.6.5. Estimating the Parameters of a Distribution

Model a univariate data set with a probability distribution

One common application of probability distributions is modeling univariate data with a specific probability distribution. This involves the following two steps:

1. Determination of the "best-fitting" distribution.
2. Estimation of the parameters (shape, location, and scale parameters) for that distribution.

Various Methods

There are various methods, both numerical and graphical, for estimating the parameters of a probability distribution.

1. Method of moments
2. Maximum likelihood
3. Least squares
4. PPCC and probability plots

1.3.6.5.1. Method of Moments

Method of Moments

The method of moments equates sample moments to parameter estimates. When moment methods are available, they have the advantage of simplicity. The disadvantage is that they are often not available and they do not have the desirable optimality properties of maximum likelihood and least squares estimators.

The primary use of moment estimates is as starting values for the more precise maximum likelihood and least squares estimates.

Software

Most general purpose statistical software does not include explicit method of moments parameter estimation commands. However, when utilized, the method of moment formulas tend to be straightforward and can be easily implemented in most statistical software programs.

1.3.6.5.2. Maximum Likelihood

Maximum Likelihood

Maximum likelihood estimation begins with the mathematical expression known as a likelihood function of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data given the chosen probability model. This expression contains the unknown parameters. Those values of the parameter that maximize the sample likelihood are known as the maximum likelihood estimates.

The reliability chapter contains some examples of the likelihood functions for a few of the commonly used distributions in reliability analysis.

Advantages

The advantages of this method are:

- Maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations. For example, they can be applied in reliability analysis to censored data under various censoring models.

- Maximum likelihood methods have desirable mathematical and optimality properties. Specifically,
  1. They become minimum variance unbiased estimators as the sample size increases. By unbiased, we mean that if we take (a very large number of) random samples with replacement from a population, the average value of the parameter estimates will be theoretically exactly equal to the population value. By minimum variance, we mean that the estimator has the smallest variance, and thus the narrowest confidence interval, of all estimators of that type.
  2. They have approximate normal distributions and approximate sample variances that can be used to generate confidence bounds and hypothesis tests for the parameters.

- Several popular statistical software packages provide

excellent algorithms for maximum likelihood estimates for many of the commonly used distributions. This helps mitigate the computational complexity of maximum likelihood estimation.

Disadvantages The disadvantages of this method are:

- The likelihood equations need to be specifically worked out for a given distribution and estimation problem. The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired.
- The numerical estimation is usually non-trivial. Except for a few cases where the maximum likelihood formulas are in fact simple, it is generally best to rely on high quality statistical software to obtain maximum likelihood estimates. Fortunately, high quality maximum likelihood software is becoming increasingly common.
- Maximum likelihood estimates can be heavily biased for small samples. The optimality properties may not apply for small samples.
- Maximum likelihood can be sensitive to the choice of starting values.

Software Most general purpose statistical software programs support maximum likelihood estimation (MLE) in some form. MLE estimation can be supported in two ways.

1. A software program may provide a generic function minimization (or equivalently, maximization) capability. This is also referred to as function optimization. Maximum likelihood estimation is essentially a function optimization problem.

   This type of capability is particularly common in mathematical software programs.

2. A software program may provide MLE computations for a specific problem. For example, it may generate ML estimates for the parameters of a Weibull distribution.

   Statistical software programs will often provide ML estimates for many specific problems even when they do not support general function optimization.

The advantage of function minimization software is that it can be applied to many different MLE problems. The drawback is that you have to specify the maximum

likelihood equations to the software. As the functions can be non-trivial, there is potential for error in entering the equations.

The advantage of the specific MLE procedures is that greater efficiency and better numerical stability can often be obtained by taking advantage of the properties of the specific estimation problem. The specific methods often return explicit confidence intervals. In addition, you do not have to know or specify the likelihood equations to the software. The disadvantage is that each MLE problem must be specifically coded.

1.3.6.5.3. Least Squares

Least Squares Non-linear least squares provides an alternative to maximum likelihood.

Advantages The advantages of this method are:

- Non-linear least squares software may be available in many statistical software packages that do not support maximum likelihood estimates.
- It can be applied more generally than maximum likelihood. That is, if your software provides non-linear fitting and it has the ability to specify the probability function you are interested in, then you can generate least squares estimates for that distribution. This will allow you to obtain reasonable estimates for distributions even if the software does not provide maximum likelihood estimates.

Disadvantages The disadvantages of this method are:

- It is not readily applicable to censored data.
- It is generally considered to have less desirable optimality properties than maximum likelihood.
- It can be quite sensitive to the choice of starting values.

Software Non-linear least squares fitting is available in many general purpose statistical software programs.

1.3.6.5.4. PPCC and Probability Plots

PPCC and Probability Plots

The PPCC plot can be used to estimate the shape parameter of a distribution with a single shape parameter. After finding the best value of the shape parameter, the probability plot can be used to estimate the location and scale parameters of a probability distribution.

Advantages

The advantages of this method are:

- It is based on two well-understood concepts.
  1. The linearity (i.e., straightness) of the probability plot is a good measure of the adequacy of the distributional fit.
  2. The correlation coefficient between the points on the probability plot is a good measure of the linearity of the probability plot.
- It is an easy technique to implement for a wide variety of distributions with a single shape parameter. The basic requirement is to be able to compute the percent point function, which is needed in the computation of both the probability plot and the PPCC plot.
- The PPCC plot provides insight into the sensitivity of the shape parameter. That is, if the PPCC plot is relatively flat in the neighborhood of the optimal value of the shape parameter, this is a strong indication that the fitted model will not be sensitive to small deviations, or even large deviations in some cases, in the value of the shape parameter.
- The maximum correlation value provides a method for comparing across distributions as well as identifying the best value of the shape parameter for a given distribution. For example, we could use the PPCC and probability fits for the Weibull, lognormal, and possibly several other distributions. Comparing the maximum correlation coefficient achieved for each distribution can help in selecting which is the best distribution to use.

Disadvantages The disadvantages of this method are:

- It is limited to distributions with a single shape parameter.
- PPCC plots are not widely available in statistical software packages other than Dataplot (Dataplot provides PPCC plots for 40+ distributions). Probability plots are generally available. However, many statistical software packages only provide them for a limited number of distributions.
- Significance levels for the correlation coefficient (i.e., if the maximum correlation value is above a given value, then the distribution provides an adequate fit for the data with a given confidence level) have only been worked out for a limited number of distributions.

Other Graphical Methods For reliability applications, the hazard plot and the Weibull plot are alternative graphical methods that are commonly used to estimate parameters.

1.3.6.6. Gallery of Distributions

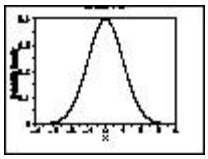
Gallery of Common Distributions

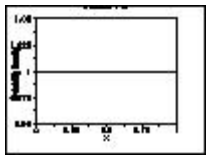
Detailed information on a few of the most common distributions is available below. There are a large number of distributions used in statistical applications. It is beyond the scope of this Handbook to discuss more than a few of these. Two excellent sources for additional detailed information on a large array of distributions are Johnson, Kotz, and Balakrishnan and Evans, Hastings, and Peacock. Equations for the probability functions are given for the standard form of the distribution. Formulas exist for defining the functions with location and scale parameters in terms of the standard form of the distribution.

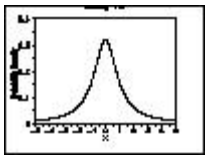
The sections on parameter estimation are restricted to the method of moments and maximum likelihood. This is because the least squares and PPCC and probability plot estimation procedures are generic. The maximum likelihood equations are not listed if they involve solving simultaneous equations. This is because these methods require sophisticated computer software to solve. Except where the maximum likelihood estimates are trivial, you should depend on a statistical software program to compute them. References are given for those who are interested.

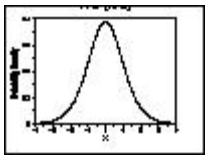
Be aware that different sources may give formulas that are different from those shown here. In some cases, these are simply mathematically equivalent formulations. In other cases, a different parameterization may be used.

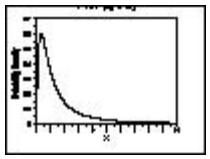
Continuous Distributions

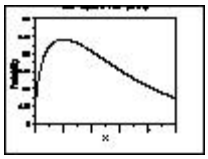

Normal Distribution


Uniform Distribution

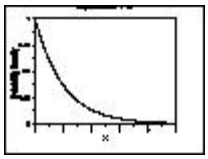

Cauchy Distribution

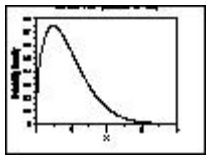


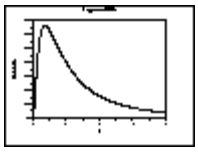


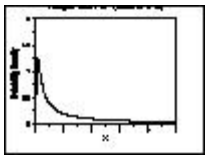


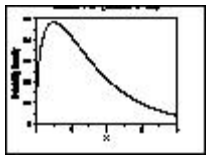
[t Distribution](#) [F Distribution](#) [Chi-Square Distribution](#)

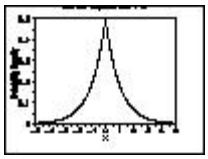

[Exponential Distribution](#)

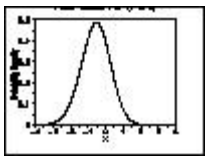

[Weibull Distribution](#)

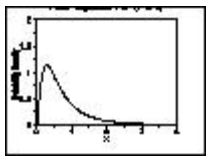

[Lognormal Distribution](#)

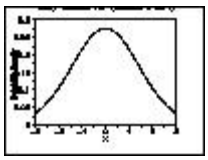

[Birnbaum-Saunders (Fatigue Life) Distribution](#)

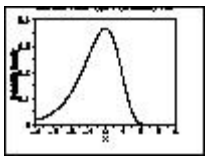

[Gamma Distribution](#)

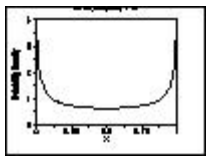

[Double Exponential Distribution](#)


[Power Normal Distribution](#)

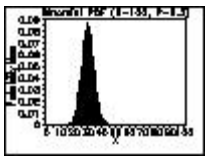

[Power Lognormal Distribution](#)

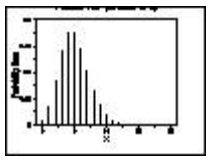

[Tukey-Lambda Distribution](#)


[Extreme Value Type I Distribution](#)


[Beta Distribution](#)

Discrete Distributions


[Binomial Distribution](#)


[Poisson Distribution](#)

1.3.6.6.1. Normal Distribution

Probability Density Function

The general formula for the probability density function of the normal distribution is

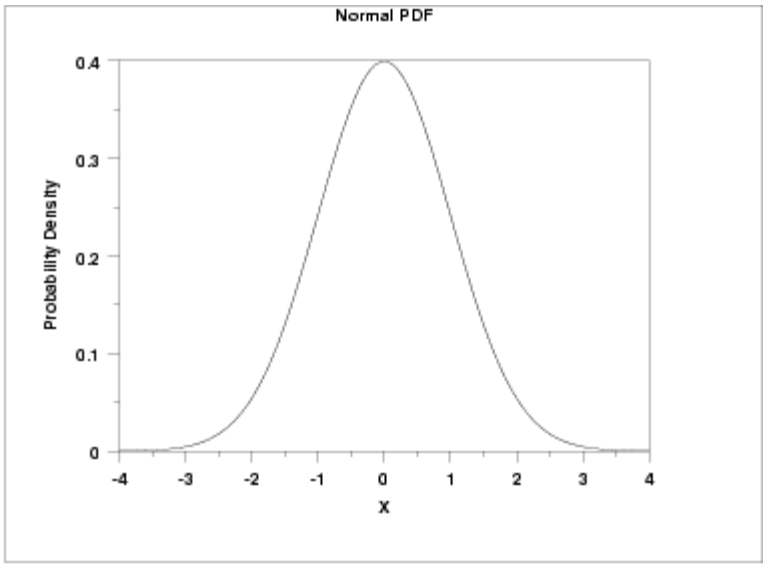
$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where μ is the location parameter and σ is the scale parameter. The case where $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**. The equation for the standard normal distribution is

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the standard normal probability density function.



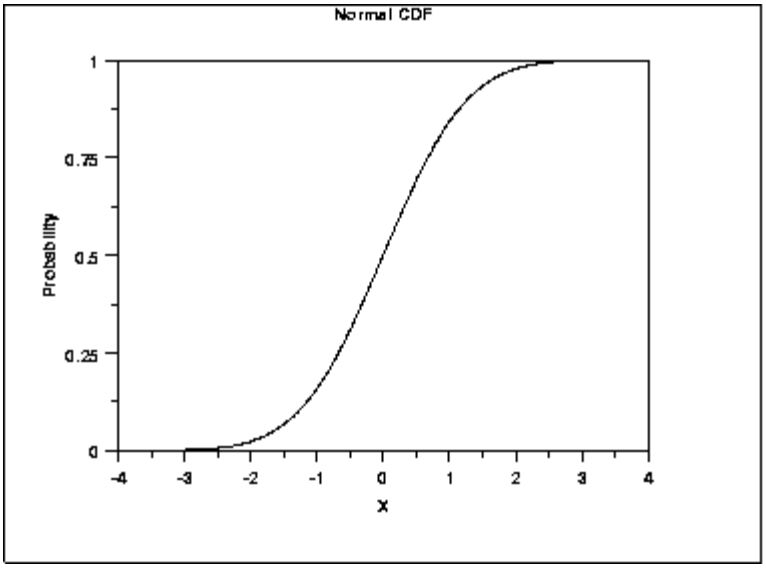
Cumulative

The formula for the cumulative distribution function of the

Distribution Function

normal distribution does not exist in a simple closed formula. It is computed numerically.

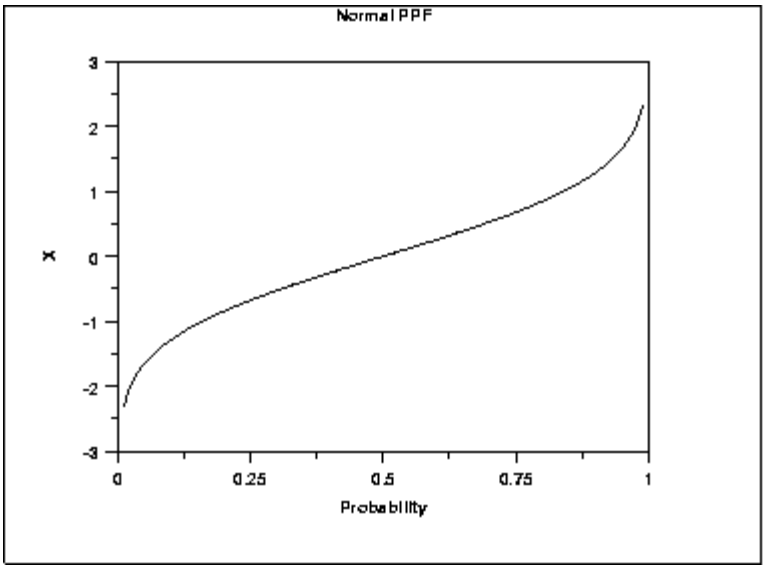
The following is the plot of the normal cumulative distribution function.



Percent Point Function

The formula for the percent point function of the normal distribution does not exist in a simple closed formula. It is computed numerically.

The following is the plot of the normal percent point function.



Hazard Function

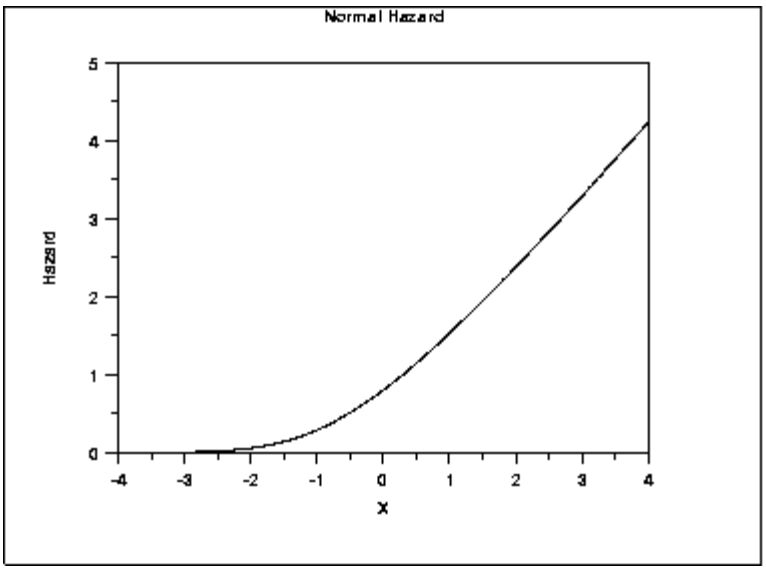
The formula for the hazard function of the normal distribution is

$$h(x) = \frac{\phi(x)}{\Phi(-x)}$$

where Φ is the cumulative distribution function of the ϕ

standard normal distribution and is the probability density function of the standard normal distribution.

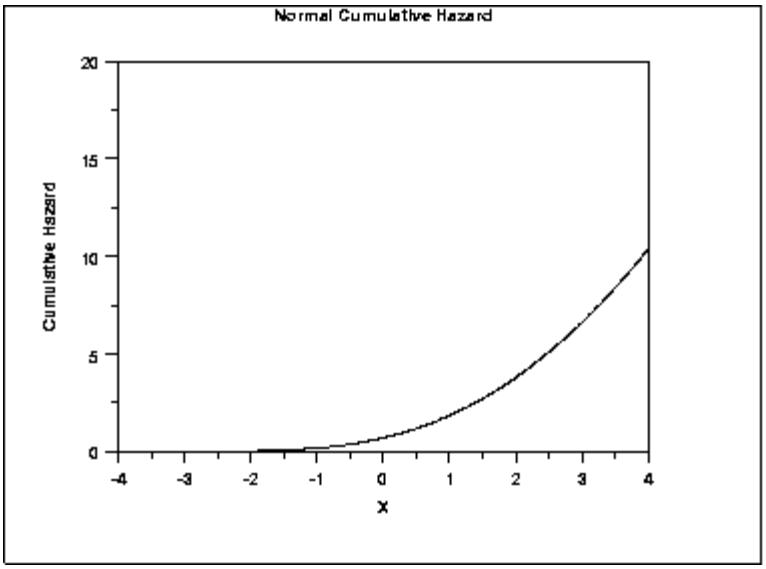
The following is the plot of the normal hazard function.



Cumulative Hazard Function

The normal cumulative hazard function can be computed from the normal cumulative distribution function.

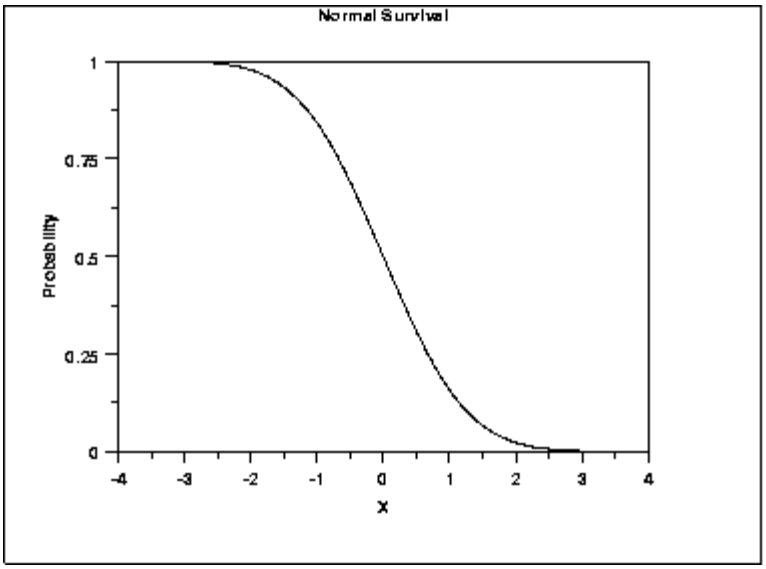
The following is the plot of the normal cumulative hazard function.



Survival Function

The normal survival function can be computed from the normal cumulative distribution function.

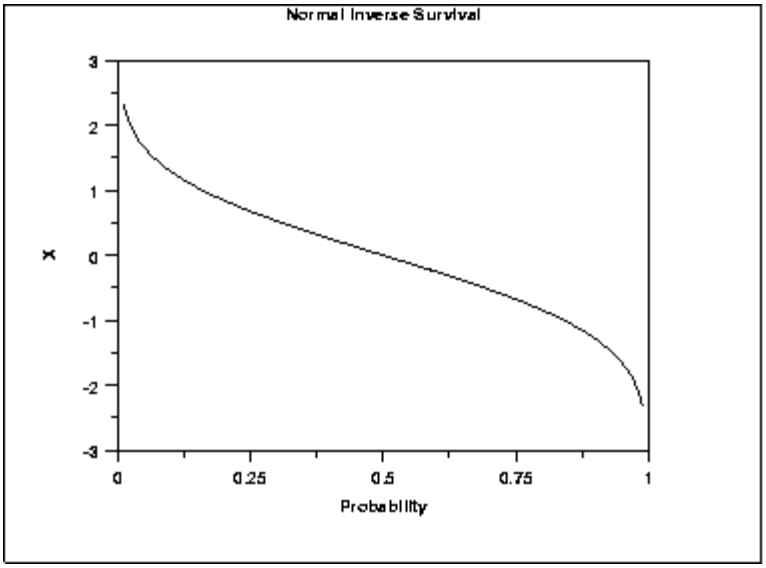
The following is the plot of the normal survival function.



Inverse Survival Function

The normal [inverse survival function](#) can be computed from the normal percent point function.

The following is the plot of the normal inverse survival function.



Common Statistics

Mean	The location parameter μ.
Median	The location parameter μ.
Mode	The location parameter μ.
Range	Infinity in both directions.
Standard Deviation	The scale parameter σ.
Coefficient of Variation	σ/μ
Skewness	0
Kurtosis	3

Parameter

The location and scale parameters of the normal distribution

Estimation can be estimated with the sample mean and sample standard deviation, respectively.

Comments For both theoretical and practical reasons, the normal distribution is probably the most important distribution in statistics. For example,

- Many classical statistical tests are based on the assumption that the data follow a normal distribution. This assumption should be tested before applying these tests.
- In modeling applications, such as linear and non-linear regression, the error term is often assumed to follow a normal distribution with fixed location and scale.
- The normal distribution is used to find significance levels in many hypothesis tests and confidence intervals.

Theroretical Justification - Central Limit Theorem The normal distribution is widely used. Part of the appeal is that it is well behaved and mathematically tractable. However, the central limit theorem provides a theoretical basis for why it has wide applicability.

The central limit theorem basically states that as the sample size (N) becomes large, the following occur:

1. The sampling distribution of the mean becomes approximately normal regardless of the distribution of the original variable.
2. The sampling distribution of the mean is centered at the population mean, μ, of the original variable. In addition, the standard deviation of the sampling distribution of the mean approaches $\sigma/\sqrt{N}$.

Software Most general purpose statistical software programs support at least some of the probability functions for the normal distribution.

1.3.6.6.2. Uniform Distribution

Probability Density Function

The general formula for the probability density function of the uniform distribution is

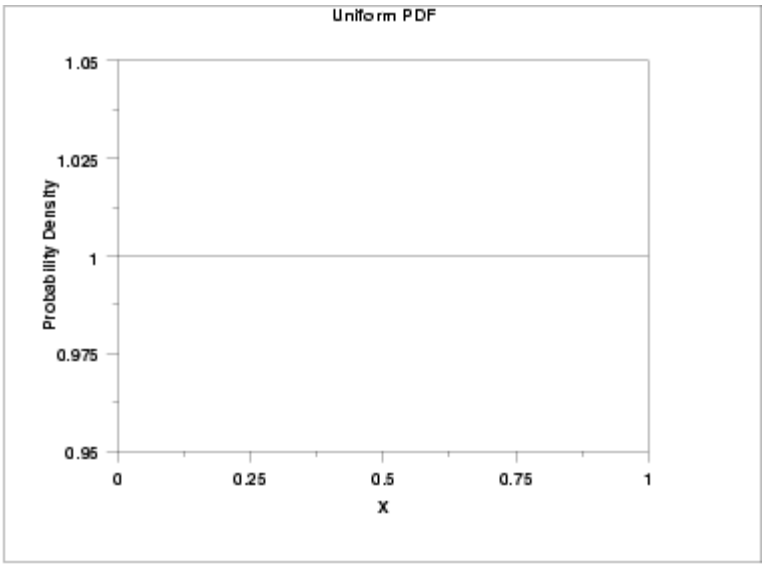
$$f(x) = \frac{1}{B - A} \qquad \text{for } A \le x \le B$$

where A is the location parameter and (B - A) is the scale parameter. The case where A = 0 and B = 1 is called the **standard uniform distribution**. The equation for the standard uniform distribution is

$$f(x) = 1 \qquad \text{for } 0 \le x \le 1$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the uniform probability density function.

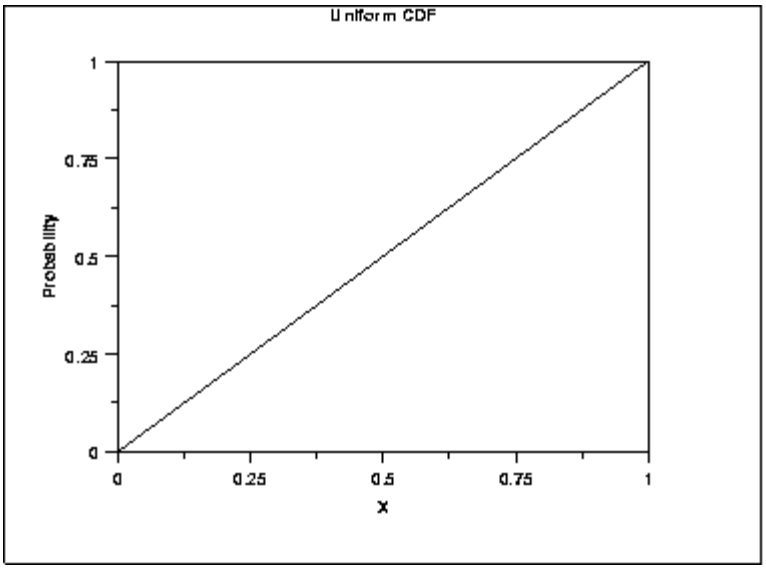


Cumulative Distribution Function

The formula for the cumulative distribution function of the uniform distribution is

$$F(x) = x \qquad \text{for } 0 \le x \le 1$$

The following is the plot of the uniform cumulative distribution function.

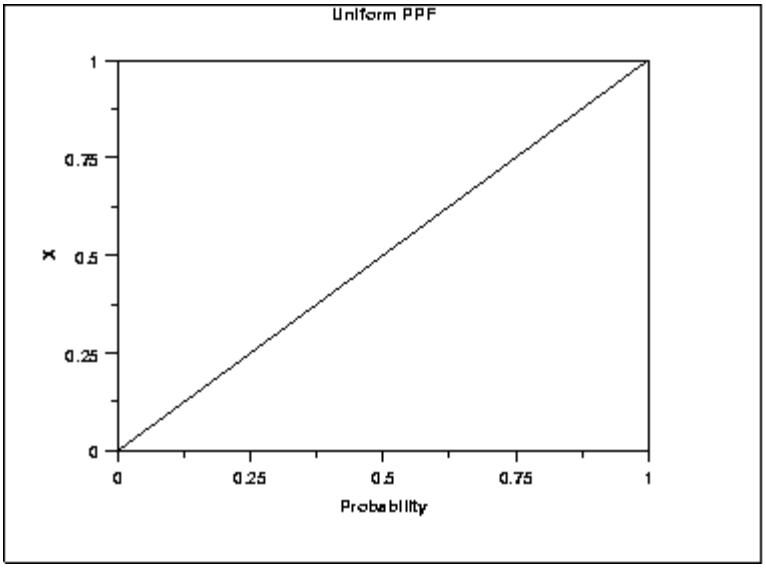


Percent Point Function

The formula for the percent point function of the uniform distribution is

$$G(p) = p \qquad \text{for } 0 \leq p \leq 1$$

The following is the plot of the uniform percent point function.

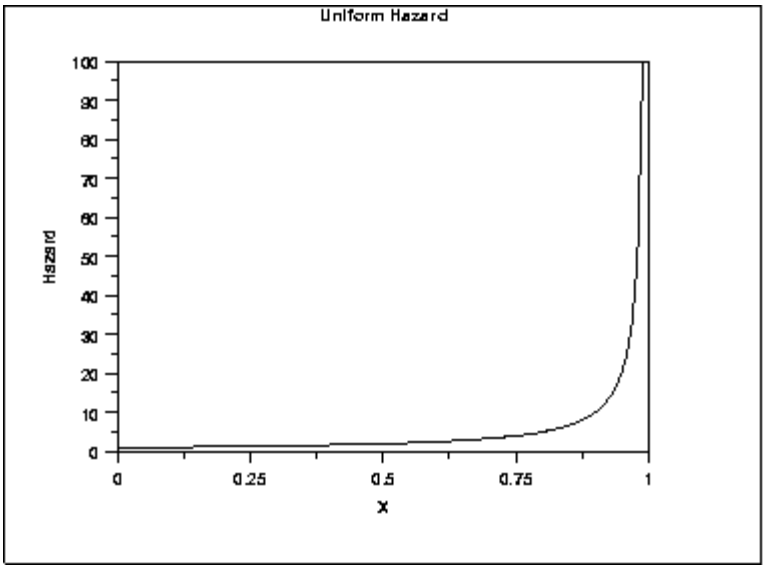


Hazard Function

The formula for the hazard function of the uniform distribution is

$$h(x) = \frac{1}{1-x} \qquad \text{for } 0 \leq x < 1$$

The following is the plot of the uniform hazard function.

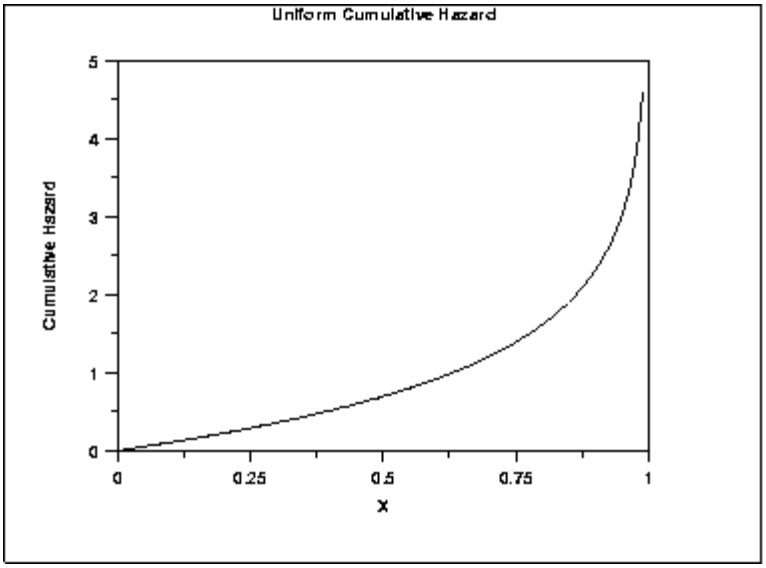


Cumulative Hazard Function

The formula for the cumulative hazard function of the uniform distribution is

$$H(x) = -ln(1-x) \qquad \text{for } 0 \le x < 1$$

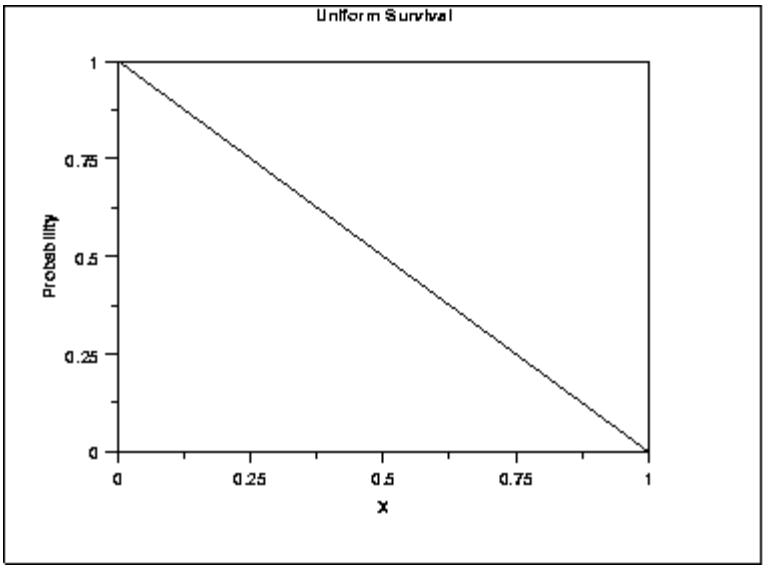
The following is the plot of the uniform cumulative hazard function.



Survival Function

The uniform survival function can be computed from the uniform cumulative distribution function.

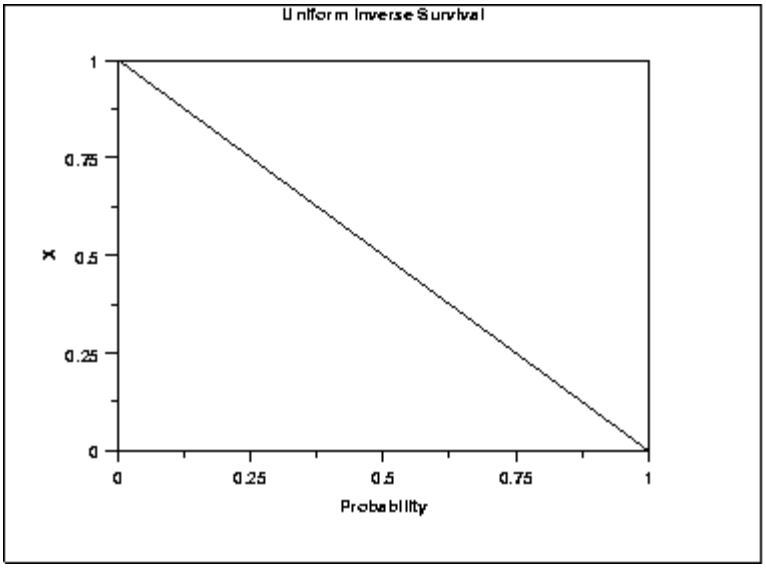
The following is the plot of the uniform survival function.



Inverse Survival Function

The uniform inverse survival function can be computed from the uniform percent point function.

The following is the plot of the uniform inverse survival function.



Common Statistics

Mean	(A + B)/2
Median	(A + B)/2
Range	B - A
Standard Deviation	$\sqrt{\frac{(B-A)^2}{12}}$
Coefficient of Variation	$\frac{(B-A)}{\sqrt{3}(B+A)}$
Skewness	0
Kurtosis	9/5

Parameter Estimation

The method of moments estimators for A and B are

$$\hat{A} = \bar{x} - \sqrt{3}s$$

$$\hat{B} = \bar{x} + \sqrt{3}s$$

The maximum likelihood estimators are usually given in terms of the parameters *a* and *h* where

A = *a* - *h*
B = *a* + *h*

The maximum likelihood estimators for *a* and *h* are

$$\hat{a} = \text{midrange}(Y_1, Y_2, ..., Y_n)$$
$$\hat{h} = 0.5[\text{range}(Y_1, Y_2, ..., Y_n)]$$

This gives the following maximum likelihood estimators for A and B

$$\hat{A} = \text{midrange}(Y_1, Y_2, ..., Y_n) - 0.5[\text{range}(Y_1, Y_2, ..., Y_n)] = Y_1$$

$$\hat{B} = \text{midrange}(Y_1, Y_2, ..., Y_n) + 0.5[\text{range}(Y_1, Y_2, ..., Y_n)] = Y_n$$

Comments

The uniform distribution defines equal probability over a given range for a continuous distribution. For this reason, it is important as a reference distribution.

One of the most important applications of the uniform distribution is in the generation of random numbers. That is, almost all random number generators generate random numbers on the (0,1) interval. For other distributions, some transformation is applied to the uniform random numbers.

Software

Most general purpose statistical software programs support at least some of the probability functions for the uniform distribution.

1.3.6.6.3. Cauchy Distribution

Probability Density Function

The general formula for the probability density function of the Cauchy distribution is

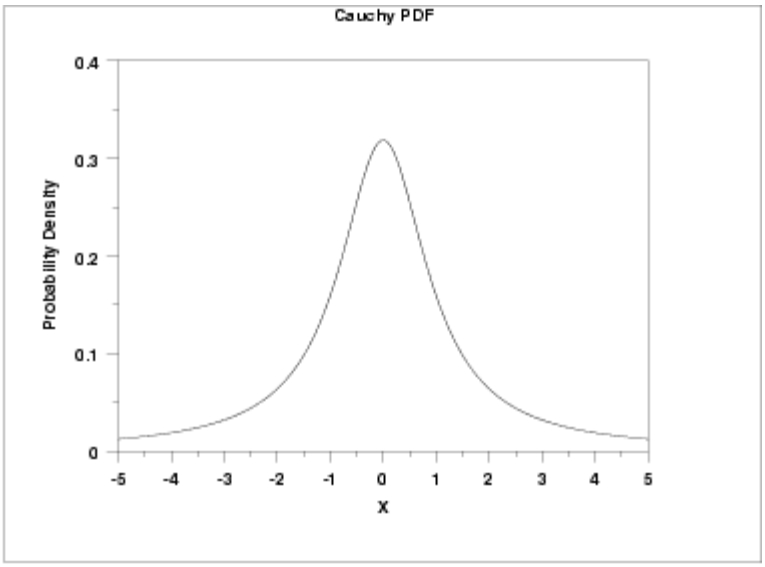
$$f(x) = \frac{1}{s\pi(1 + ((x-t)/s)^2)}$$

where ***t*** is the location parameter and ***s*** is the scale parameter. The case where ***t*** = 0 and ***s*** = 1 is called the **standard Cauchy distribution**. The equation for the standard Cauchy distribution reduces to

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the standard Cauchy probability density function.



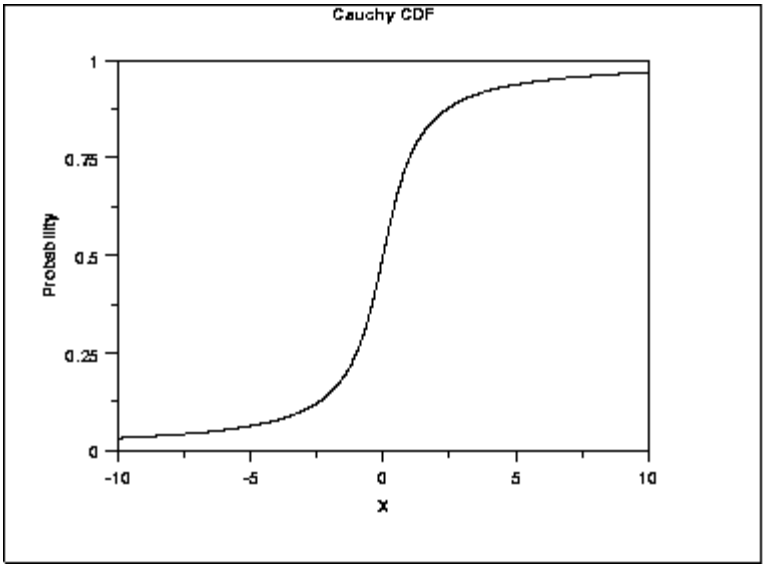
Cumulative Distribution

The formula for the cumulative distribution function for the Cauchy distribution is

Function

$$F(x) = 0.5 + \frac{\arctan(x)}{\pi}$$

The following is the plot of the Cauchy cumulative distribution function.

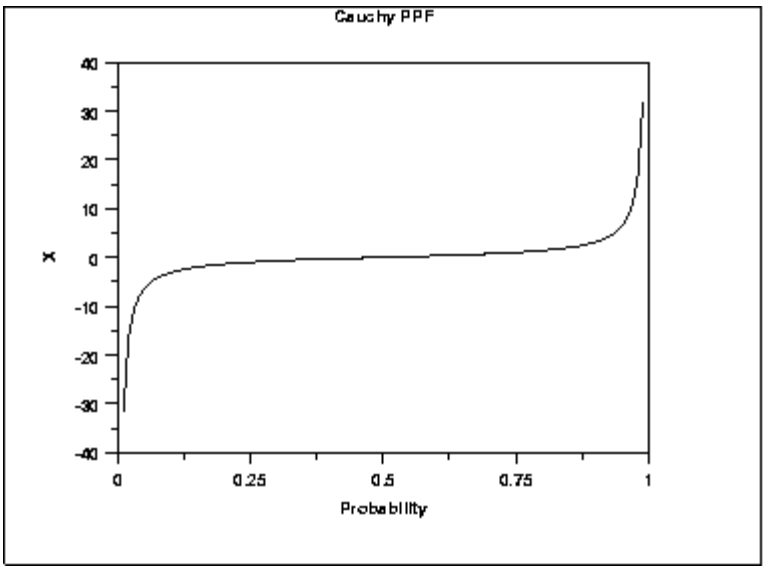


Percent Point Function

The formula for the percent point function of the Cauchy distribution is

$$G(p) = -\cot(\pi p)$$

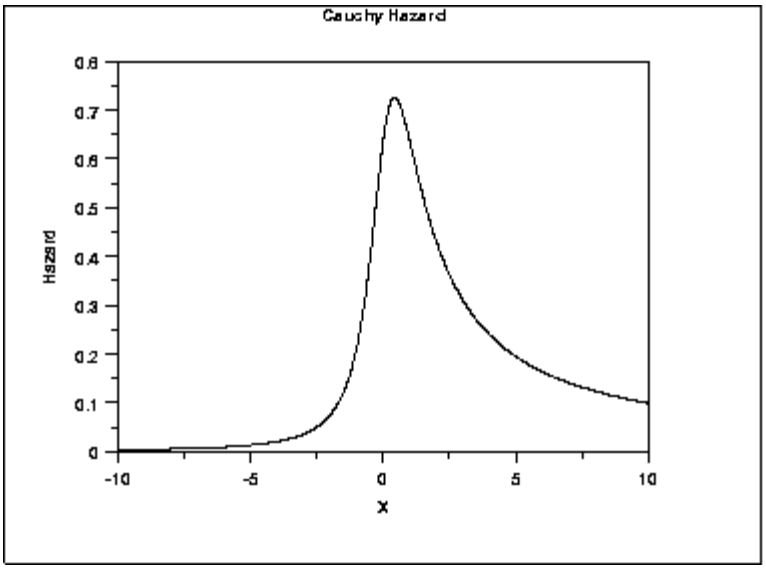
The following is the plot of the Cauchy percent point function.



Hazard Function

The Cauchy hazard function can be computed from the Cauchy probability density and cumulative distribution functions.

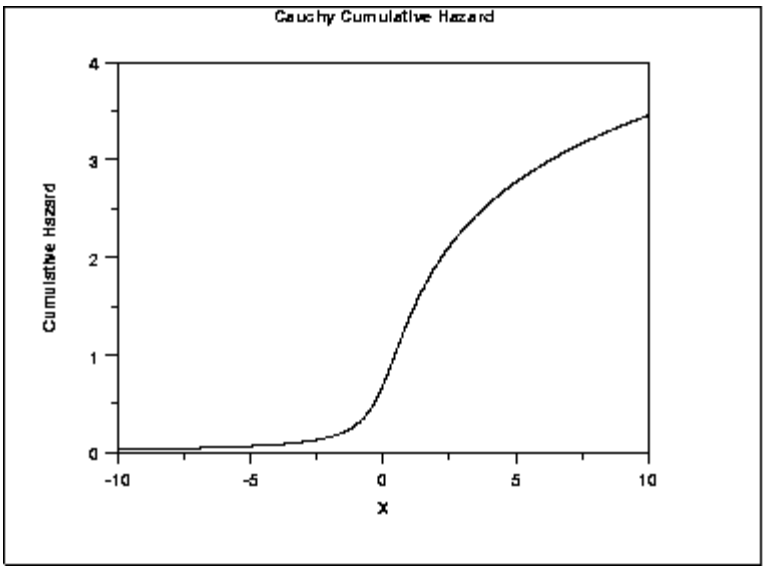
The following is the plot of the Cauchy hazard function.



Cumulative Hazard Function

The Cauchy cumulative hazard function can be computed from the Cauchy cumulative distribution function.

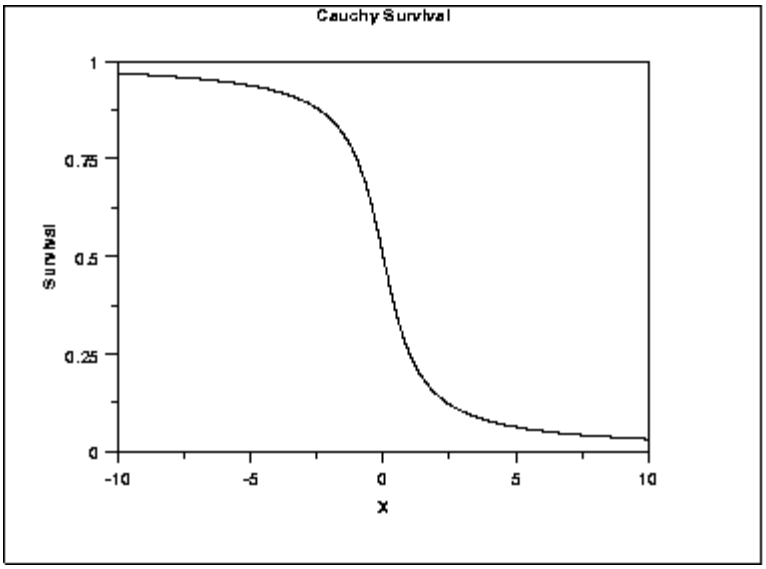
The following is the plot of the Cauchy cumulative hazard function.



Survival Function

The Cauchy survival function can be computed from the Cauchy cumulative distribution function.

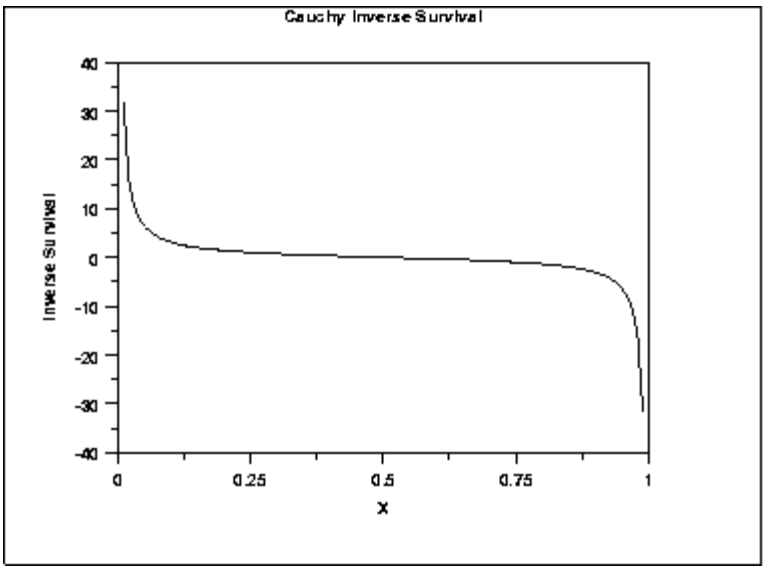
The following is the plot of the Cauchy survival function.



Inverse Survival Function

The Cauchy inverse survival function can be computed from the Cauchy percent point function.

The following is the plot of the Cauchy inverse survival function.



Common Statistics

Mean	The mean is undefined.
Median	The location parameter *t*.
Mode	The location parameter *t*.
Range	Infinity in both directions.
Standard Deviation	The standard deviation is undefined.
Coefficient of Variation	The coefficient of variation is undefined.
Skewness	The skewness is undefined.
Kurtosis	The kurtosis is undefined.

Parameter

The likelihood functions for the Cauchy maximum likelihood

Estimation estimates are given in chapter 16 of Johnson, Kotz, and Balakrishnan. These equations typically must be solved numerically on a computer.

Comments The Cauchy distribution is important as an example of a pathological case. Cauchy distributions look similar to a normal distribution. However, they have much heavier tails. When studying hypothesis tests that assume normality, seeing how the tests perform on data from a Cauchy distribution is a good indicator of how sensitive the tests are to heavy-tail departures from normality. Likewise, it is a good check for robust techniques that are designed to work well under a wide variety of distributional assumptions.

The mean and standard deviation of the Cauchy distribution are undefined. The practical meaning of this is that collecting 1,000 data points gives no more accurate an estimate of the mean and standard deviation than does a single point.

Software Many general purpose statistical software programs support at least some of the probability functions for the Cauchy distribution.

1.3.6.6.4. t Distribution

Probability Density Function

The formula for the probability density function of the *t* distribution is

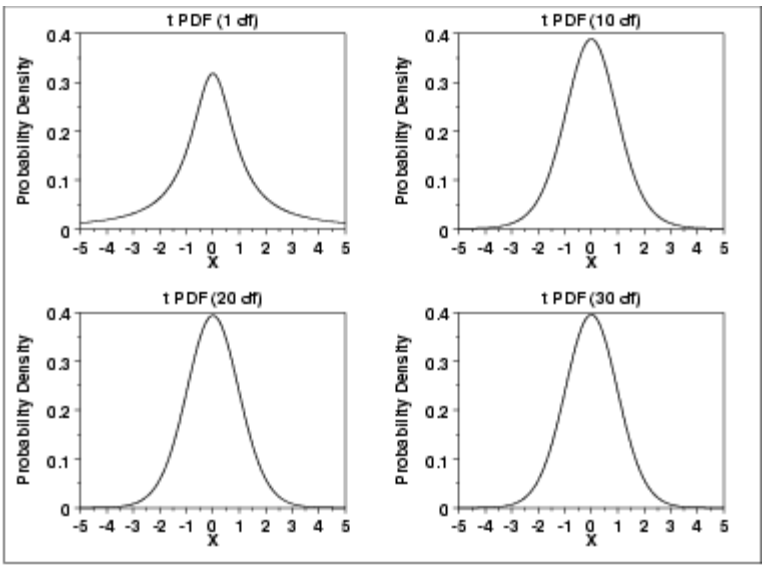
$$f(x) = \frac{\left(1 + \frac{x^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}}{B(0.5, 0.5\nu)\sqrt{\nu}}$$

where B is the beta function and ν is a positive integer shape parameter. The formula for the beta function is

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

In a testing context, the *t* distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the *t* distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the *t* probability density function for 4 different values of the shape parameter.



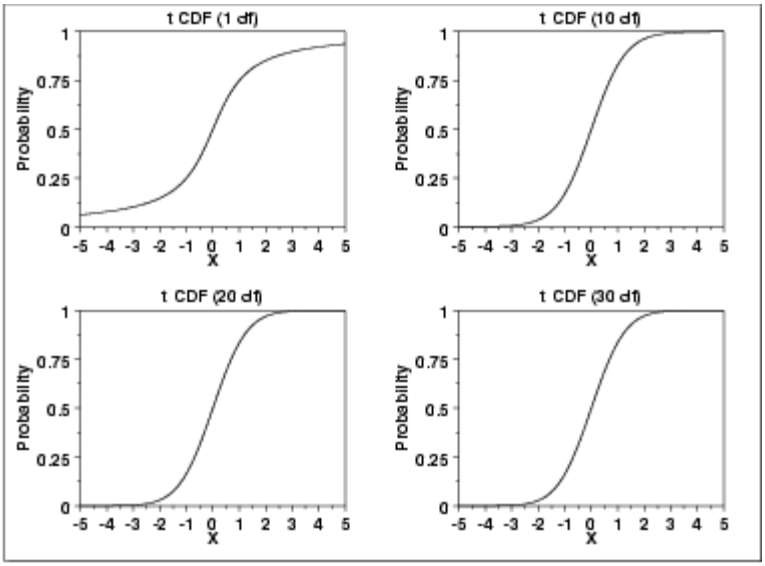
These plots all have a similar shape. The difference is in the heaviness of the tails. In fact, the *t* distribution with ν equal

to 1 is a Cauchy distribution. The *t* distribution approaches a normal distribution as ν becomes large. The approximation is quite good for values of $\nu > 30$.

Cumulative Distribution Function

The formula for the cumulative distribution function of the *t* distribution is complicated and is not included here. It is given in the Evans, Hastings, and Peacock book.

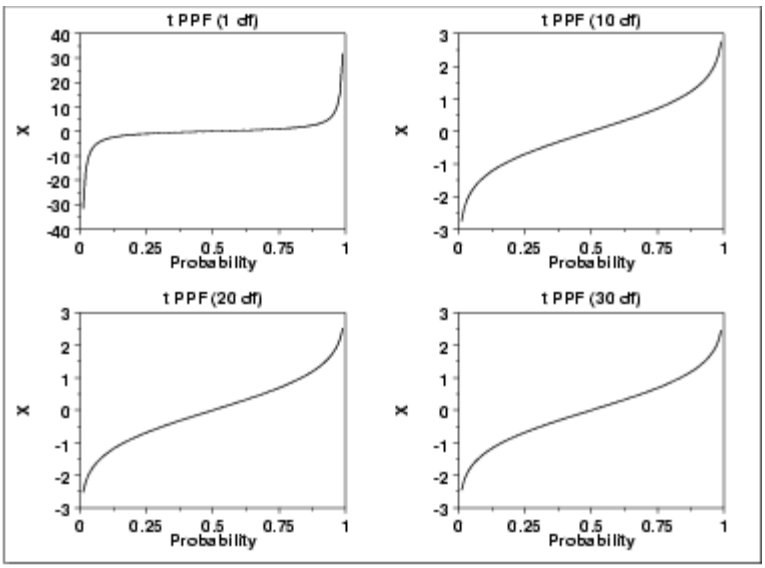
The following are the plots of the *t* cumulative distribution function with the same values of ν as the pdf plots above.



Percent Point Function

The formula for the percent point function of the *t* distribution does not exist in a simple closed form. It is computed numerically.

The following are the plots of the *t* percent point function with the same values of ν as the pdf plots above.



Other Probability Functions

Since the *t* distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard,

cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

Mean	0 (It is undefined for ν equal to 1.)
Median	0
Mode	0
Range	Infinity in both directions.
Standard Deviation	$\sqrt{\frac{\nu}{(\nu-2)}}$ It is undefined for ν equal to 1 or 2.
Coefficient of Variation	Undefined
Skewness	0. It is undefined for ν less than or equal to 3. However, the t distribution is symmetric in all cases.
Kurtosis	$\frac{3(\nu-2)}{(\nu-4)}$ It is undefined for ν less than or equal to 4.

Parameter Estimation

Since the *t* distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.

Comments

The *t* distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. The most common example is testing if data are consistent with the assumed process mean.

Software

Most general purpose statistical software programs support at least some of the probability functions for the *t* distribution.

1.3.6.6.5. F Distribution

Probability Density Function

The F distribution is the ratio of two chi-square distributions with degrees of freedom ν_1 and ν_2, respectively, where each chi-square has first been divided by its degrees of freedom. The formula for the probability density function of the F distribution is

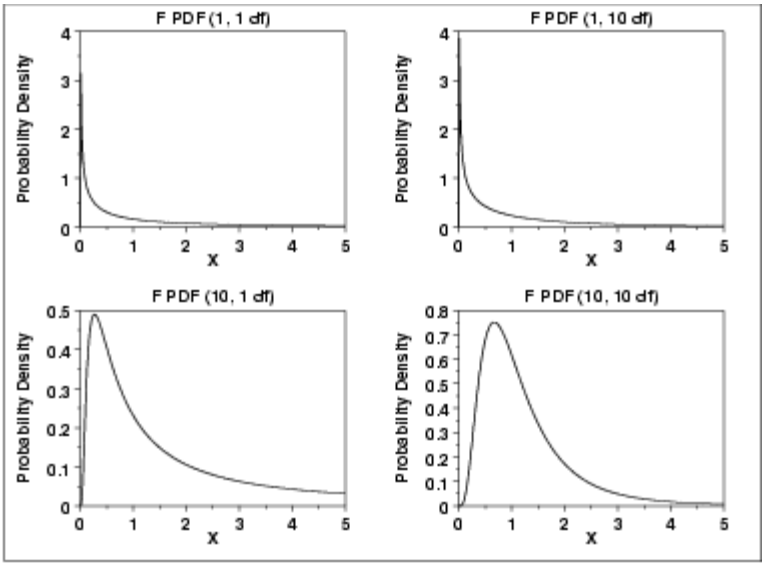
$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}$$

where ν_1 and ν_2 are the shape parameters and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

In a testing context, the F distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the F distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the F probability density function for 4 different values of the shape parameters.



Cumulative Distribution Function

The formula for the Cumulative distribution function of the F distribution is

$$F(x) = 1 - I_k(\frac{\nu_2}{2}, \frac{\nu_1}{2})$$

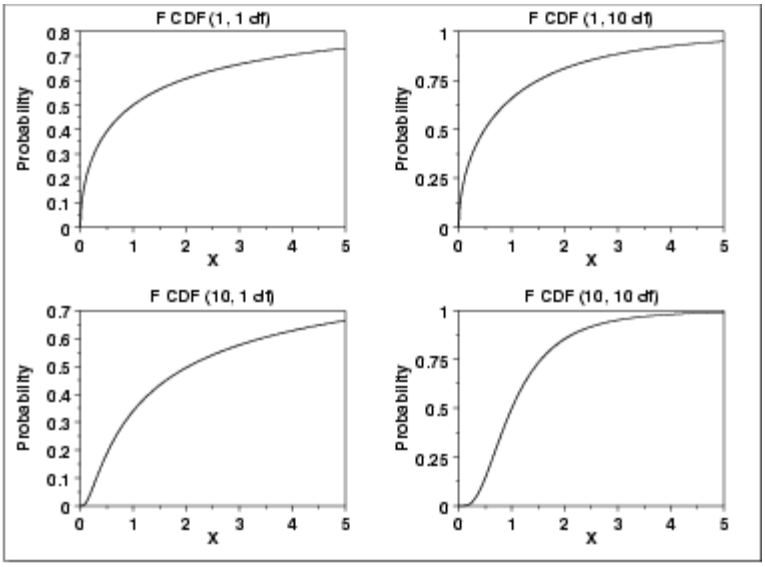
where k = ν_2/ (ν_2 + ν_1*x) and I_k is the incomplete beta function. The formula for the incomplete beta function is

$$I_k(x, \alpha, \beta) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt}{B(\alpha, \beta)}$$

where B is the beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

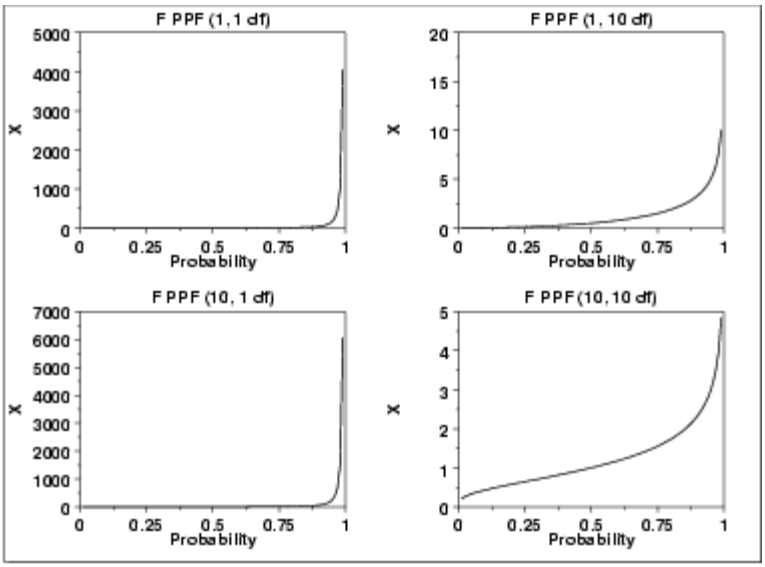
The following is the plot of the F cumulative distribution function with the same values of ν_1 and ν_2 as the pdf plots above.



Percent Point Function

The formula for the percent point function of the F distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the F percent point function with the same values of ν_1 and ν_2 as the pdf plots above.



Other Probability Functions

Since the F distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

The formulas below are for the case where the location parameter is zero and the scale parameter is one.

Mean	$\frac{\nu_2}{(\nu_2 - 2)} \quad \nu_2 > 2$
Mode	$\frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)} \quad \nu_1 > 2$
Range	0 to positive infinity
Standard Deviation	$\sqrt{\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}} \quad \nu_2 > 4$
Coefficient of Variation	$\sqrt{\frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}} \quad \nu_2 > 4$
Skewness	$\frac{(2\nu_1 + \nu_2 - 2)\sqrt{8(\nu_2 - 4)}}{\sqrt{\nu_1}(\nu_2 - 6)\sqrt{(\nu_1 + \nu_2 - 2)}} \quad \nu_2 > 6$

Parameter Estimation

Since the F distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.

Comments

The F distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. Two common examples are the analysis of variance and the F test to determine if the variances of two populations are equal.

Software Most general purpose statistical software programs support at least some of the probability functions for the F distribution.

1.3.6.6.6. Chi-Square Distribution

Probability Density Function

The chi-square distribution results when ν independent variables with standard normal distributions are squared and summed. The formula for the probability density function of the chi-square distribution is

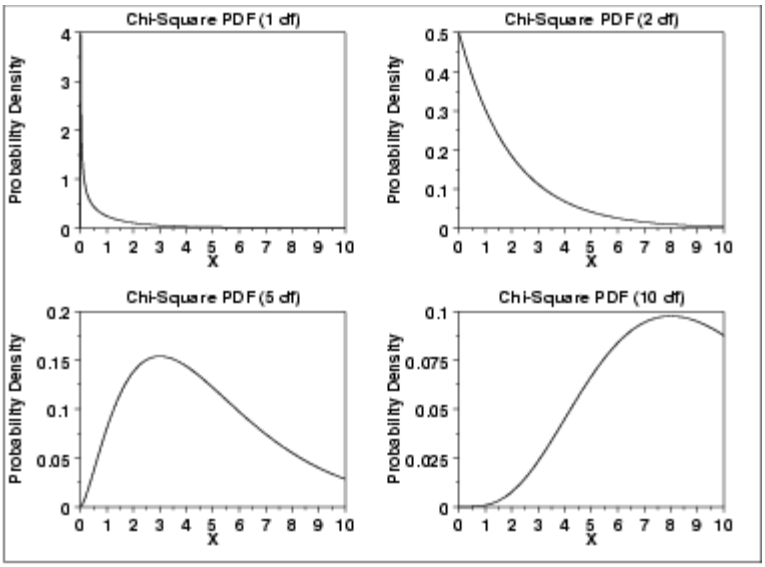
$$f(x) = \frac{e^{\frac{-x}{2}} x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} \qquad \text{for } x \geq 0$$

where ν is the shape parameter and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

In a testing context, the chi-square distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the chi-square distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the chi-square probability density function for 4 different values of the shape parameter.



Cumulative Distribution Function

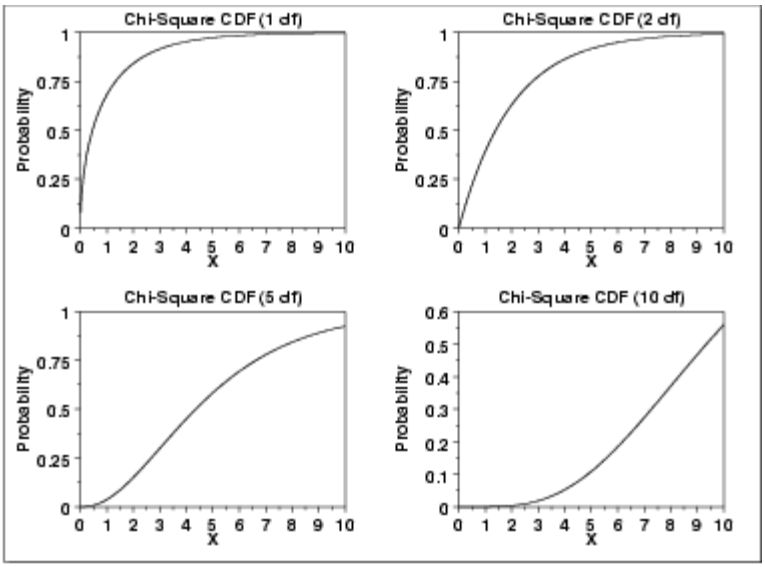
The formula for the cumulative distribution function of the chi-square distribution is

$$F(x) = \frac{\gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})} \qquad \text{for } x \geq 0$$

where Γ is the gamma function defined above and γ is the incomplete gamma function. The formula for the incomplete gamma function is

$$\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} dt$$

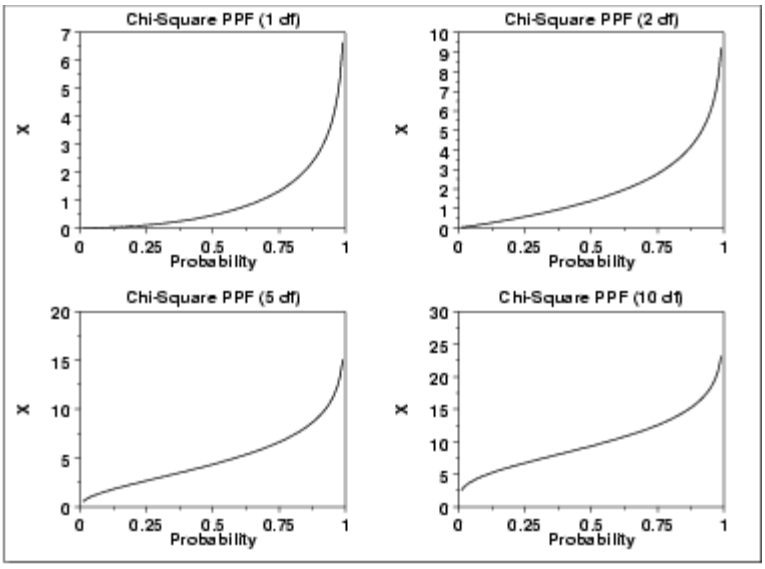
The following is the plot of the chi-square cumulative distribution function with the same values of ν as the pdf plots above.



Percent Point Function

The formula for the percent point function of the chi-square distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the chi-square percent point function with the same values of ν as the pdf plots above.



Other Probability Functions

Since the chi-square distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

Mean	ν
Median	approximately ν - 2/3 for large ν
Mode	$\nu - 2$ for $\nu > 2$
Range	0 to positive infinity
Standard Deviation	$\sqrt{2\nu}$
Coefficient of Variation	$\sqrt{\frac{2}{\nu}}$
Skewness	$\frac{2^{1.5}}{\sqrt{\nu}}$
Kurtosis	$3 + \frac{12}{\nu}$

Parameter Estimation

Since the chi-square distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.

Comments

The chi-square distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. Two common examples are the chi-square test for independence in an ***RxC*** contingency table and the chi-square test to determine if the standard deviation of a population is equal to a pre-specified value.

Software

Most general purpose statistical software programs support at least some of the probability functions for the chi-square distribution.

1.3.6.6.7. Exponential Distribution

Probability Density Function

The general formula for the probability density function of the exponential distribution is

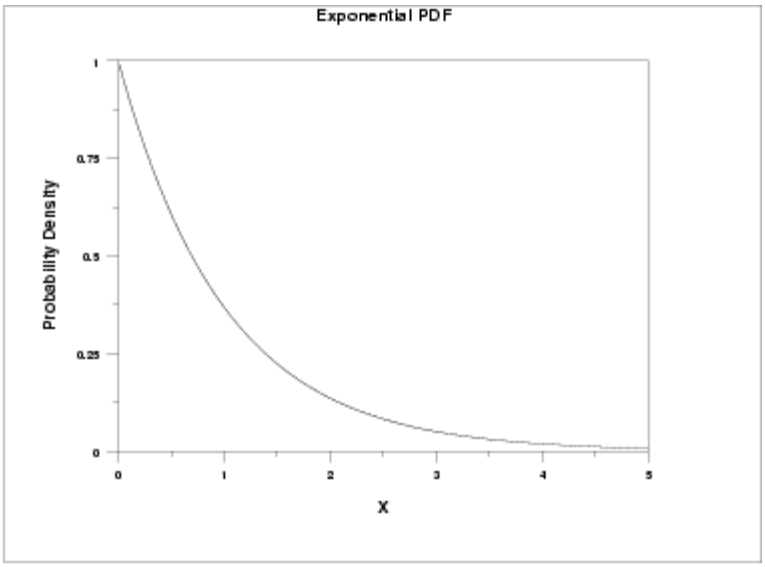
$$f(x) = \frac{1}{\beta} e^{-(x-\mu)/\beta} \qquad x \geq \mu; \beta > 0$$

where μ is the location parameter and β is the scale parameter (the scale parameter is often referred to as λ which equals $1/\beta$). The case where $\mu = 0$ and $\beta = 1$ is called the **standard exponential distribution**. The equation for the standard exponential distribution is

$$f(x) = e^{-x} \qquad \text{for } x \geq 0$$

The general form of probability functions can be expressed in terms of the standard distribution. Subsequent formulas in this section are given for the 1-parameter (i.e., with scale parameter) form of the function.

The following is the plot of the exponential probability density function.



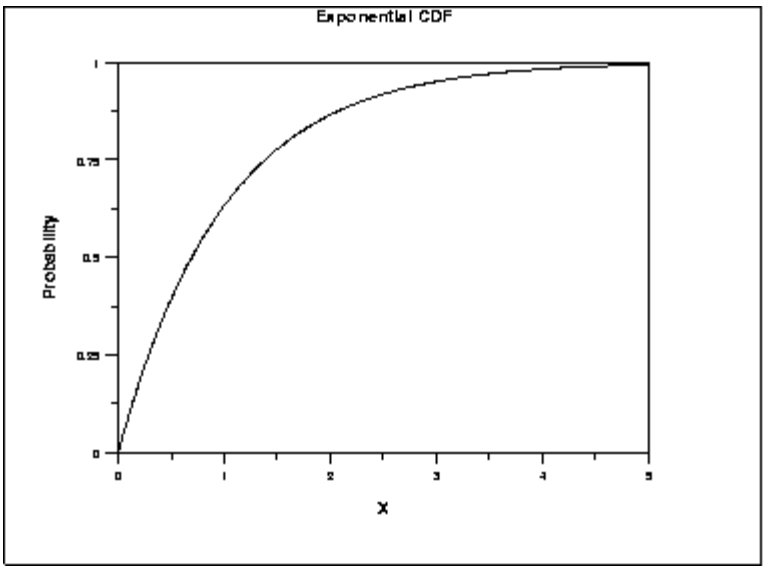
Cumulative Distribution

The formula for the cumulative distribution function of the exponential distribution is

Function

$$F(x) = 1 - e^{-x/\beta} \qquad x \geq 0; \beta > 0$$

The following is the plot of the exponential cumulative distribution function.

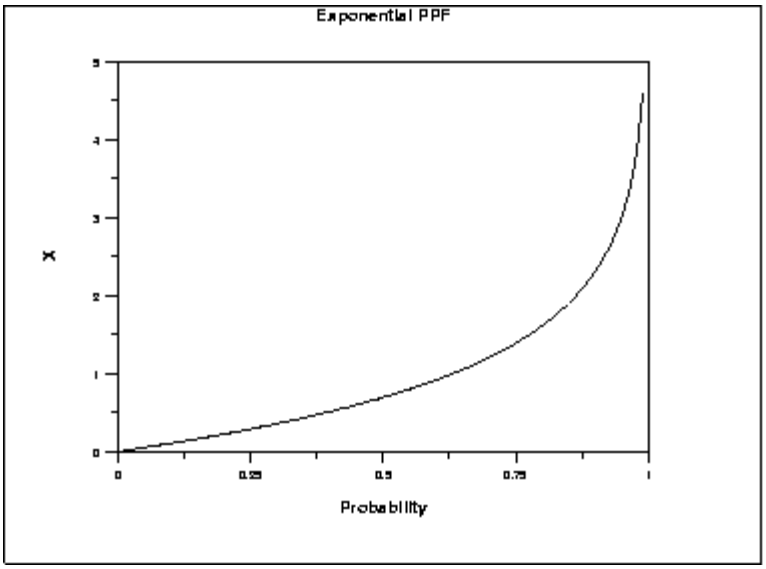


Percent Point Function

The formula for the percent point function of the exponential distribution is

$$G(p) = -\beta \ln(1 - p) \qquad 0 \leq p < 1; \beta > 0$$

The following is the plot of the exponential percent point function.

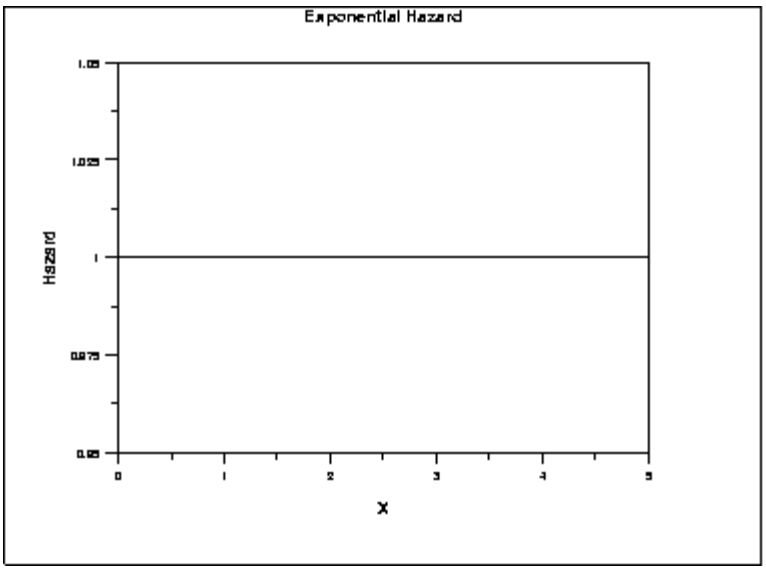


Hazard Function

The formula for the hazard function of the exponential distribution is

$$h(x) = \frac{1}{\beta} \qquad x \geq 0; \beta > 0$$

The following is the plot of the exponential hazard function.

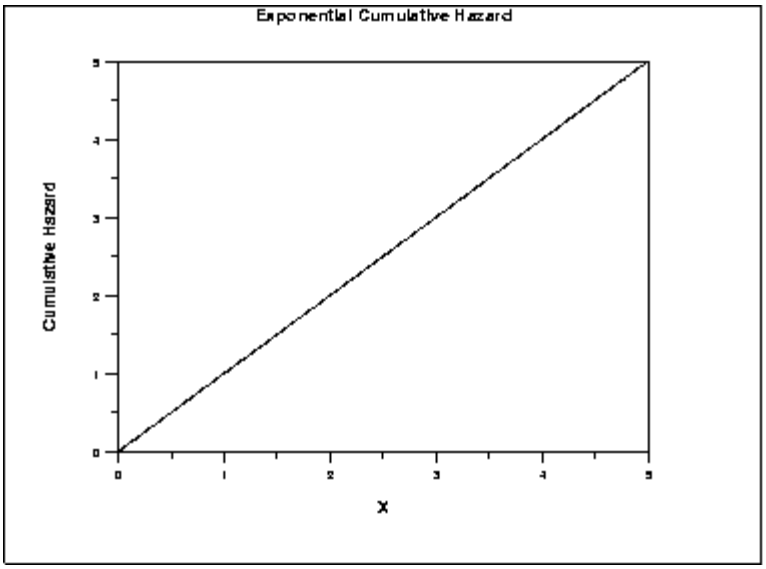


Cumulative Hazard Function

The formula for the cumulative hazard function of the exponential distribution is

$$H(x) = \frac{x}{\beta} \qquad x \geq 0; \beta > 0$$

The following is the plot of the exponential cumulative hazard function.

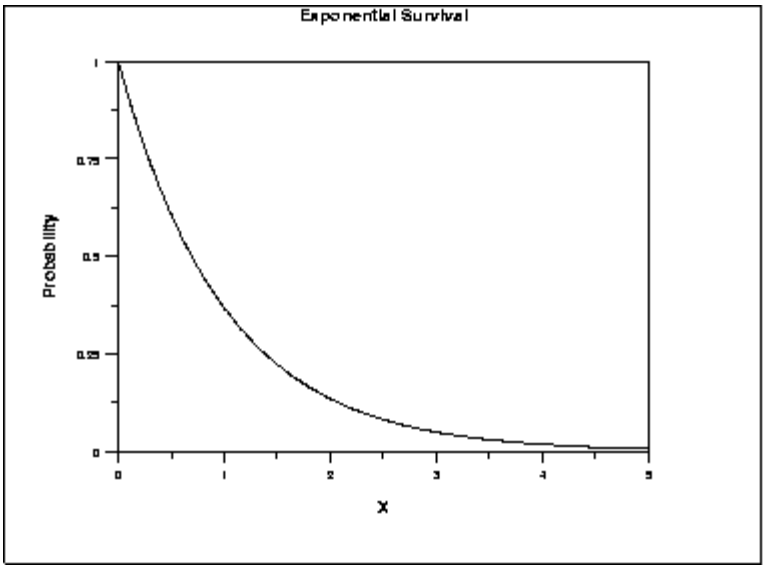


Survival Function

The formula for the survival function of the exponential distribution is

$$S(x) = e^{-x/\beta} \qquad x \geq 0; \beta > 0$$

The following is the plot of the exponential survival function.

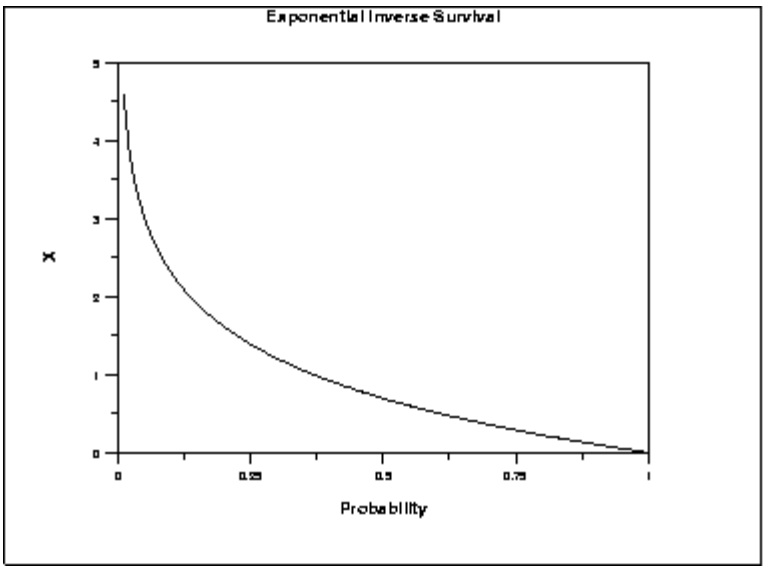


Inverse Survival Function

The formula for the inverse survival function of the exponential distribution is

$$Z(p) = -\beta \ln(p) \qquad 0 \leq p < 1; \beta > 0$$

The following is the plot of the exponential inverse survival function.



Common Statistics

Mean	β
Median	$\beta \ln 2$
Mode	Zero
Range	Zero to plus infinity
Standard Deviation	β
Coefficient of Variation	1
Skewness	2
Kurtosis	9

Parameter Estimation	For the full sample case, the maximum likelihood estimator of the scale parameter is the sample mean. Maximum likelihood estimation for the exponential distribution is discussed in the chapter on reliability (Chapter 8). It is also discussed in chapter 19 of Johnson, Kotz, and Balakrishnan.
Comments	The exponential distribution is primarily used in reliability applications. The exponential distribution is used to model data with a constant failure rate (indicated by the hazard plot which is simply equal to a constant).
Software	Most general purpose statistical software programs support at least some of the probability functions for the exponential distribution.

1.3.6.6.8. Weibull Distribution

Probability Density Function

The formula for the probability density function of the general Weibull distribution is

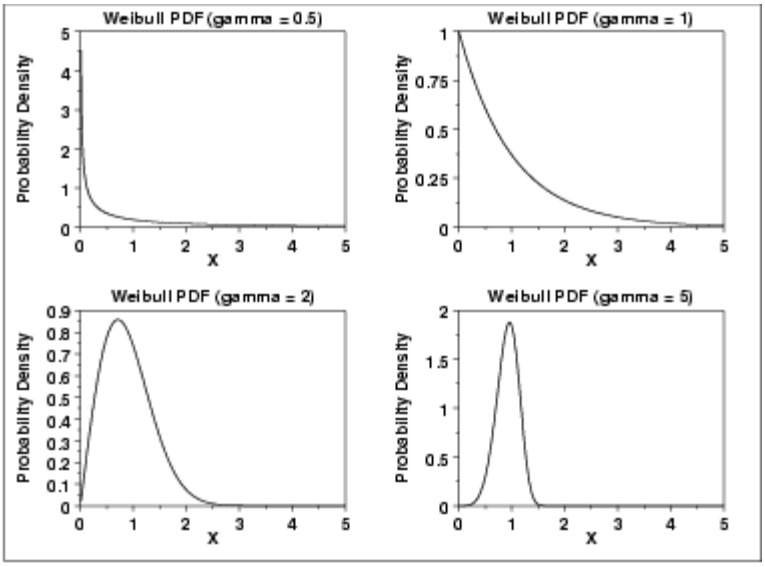
$$f(x) = \frac{\gamma}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{(\gamma-1)} \exp\left(-((x-\mu)/\alpha)^{\gamma}\right) \qquad x \geq \mu; \gamma, \alpha > 0$$

where γ is the shape parameter, μ is the location parameter and α is the scale parameter. The case where $\mu = 0$ and $\alpha = 1$ is called the **standard Weibull distribution**. The case where $\mu = 0$ is called the 2-parameter Weibull distribution. The equation for the standard Weibull distribution reduces to

$$f(x) = \gamma x^{(\gamma-1)} \exp(-(x^{\gamma})) \qquad x \geq 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the Weibull probability density function.

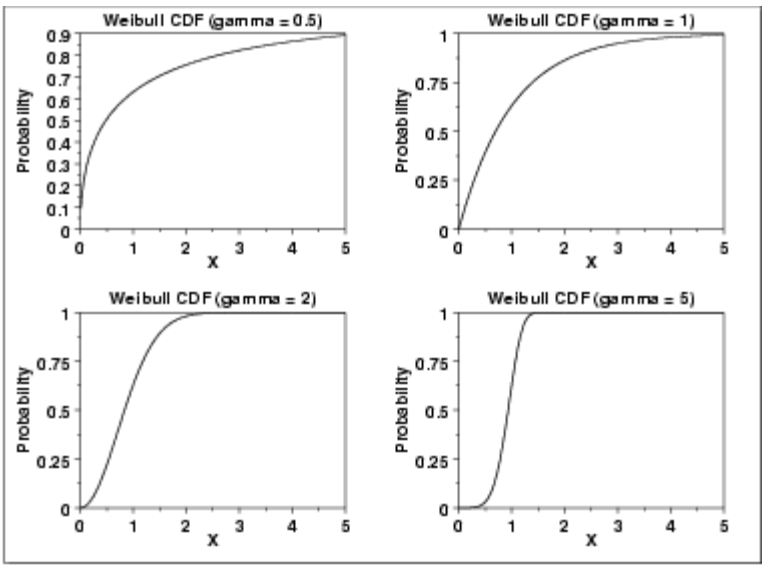


Cumulative Distribution Function

The formula for the cumulative distribution function of the Weibull distribution is

$$F(x) = 1 - e^{-(x^{\gamma})} \qquad x \geq 0; \gamma > 0$$

The following is the plot of the Weibull cumulative distribution function with the same values of γ as the pdf plots above.

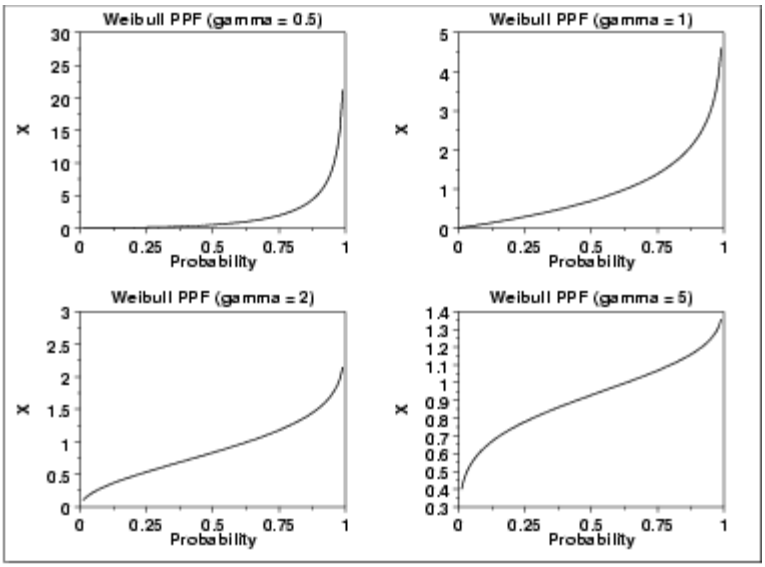


Percent Point Function

The formula for the percent point function of the Weibull distribution is

$$G(p) = (-\ln(1-p))^{1/\gamma} \qquad 0 \le p < 1; \gamma > 0$$

The following is the plot of the Weibull percent point function with the same values of γ as the pdf plots above.

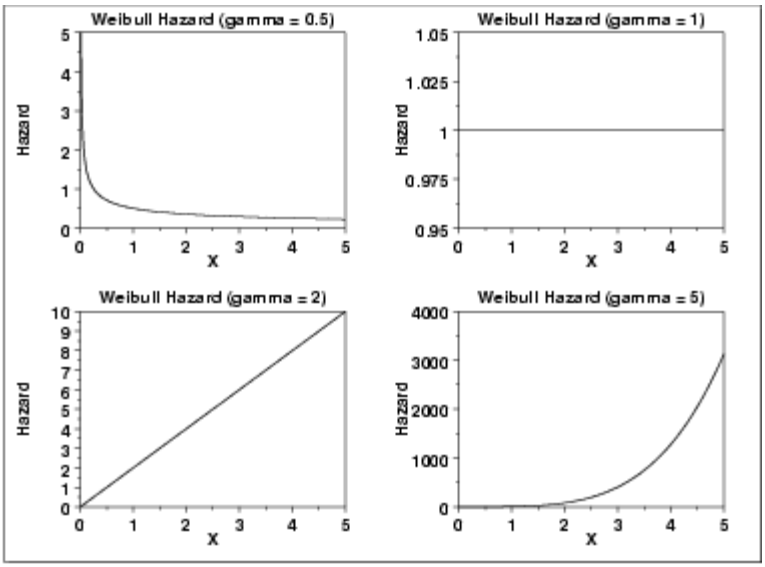


Hazard Function

The formula for the hazard function of the Weibull distribution is

$$h(x) = \gamma x^{(\gamma-1)} \qquad x \ge 0; \gamma > 0$$

The following is the plot of the Weibull hazard function with the same values of γ as the pdf plots above.

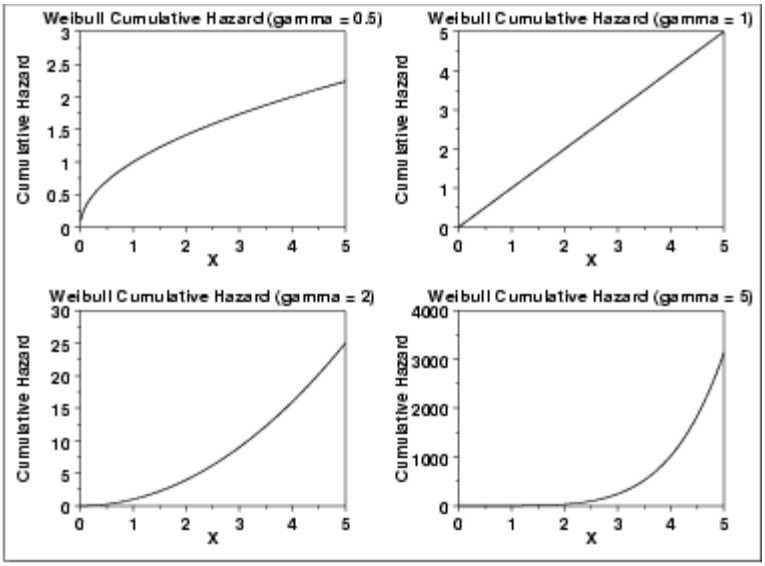


Cumulative Hazard Function

The formula for the cumulative hazard function of the Weibull distribution is

$$H(x) = x^{\gamma} \qquad x \geq 0; \gamma > 0$$

The following is the plot of the Weibull cumulative hazard function with the same values of γ as the pdf plots above.

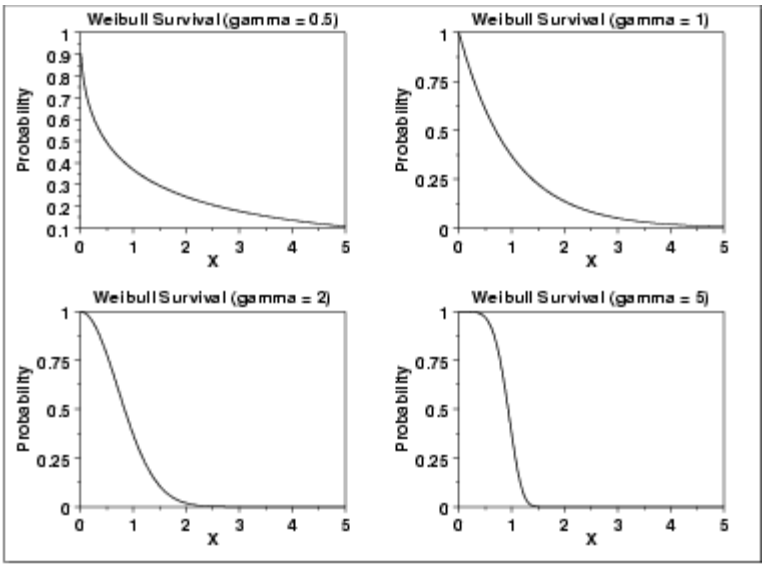


Survival Function

The formula for the survival function of the Weibull distribution is

$$S(x) = \exp -(x^{\gamma}) \qquad x \geq 0; \gamma > 0$$

The following is the plot of the Weibull survival function with the same values of γ as the pdf plots above.

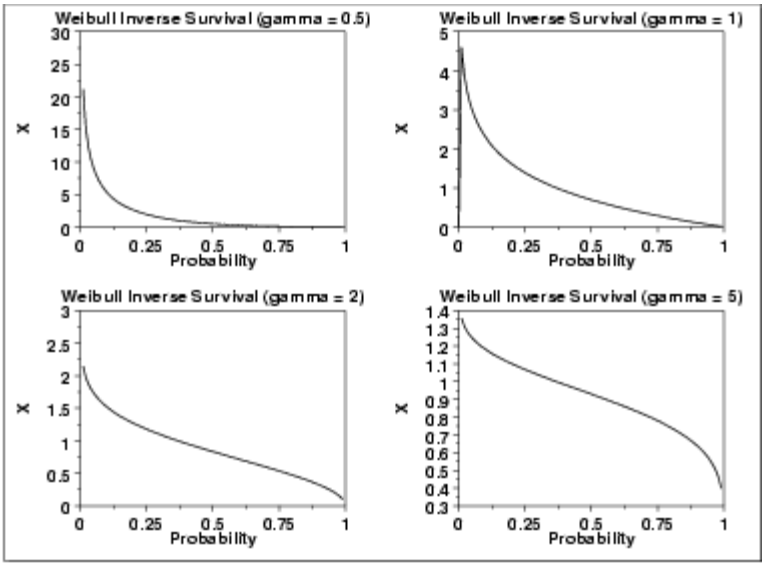


Inverse Survival Function

The formula for the [inverse survival function](#) of the Weibull distribution is

$$Z(p) = (-\ln(p))^{1/\gamma} \qquad 0 \leq p < 1; \gamma > 0$$

The following is the plot of the Weibull inverse survival function with the same values of γ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean

$$\Gamma(\frac{\gamma+1}{\gamma})$$

where Γ is the gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$$

Median

$$\ln(2)^{1/\gamma}$$

Mode

	$(1-\frac{1}{\gamma})^{1/\gamma} \quad \gamma > 1$ $0 \quad \gamma \leq 1$
Range	Zero to positive infinity.
Standard Deviation	$\sqrt{\Gamma(\frac{\gamma+2}{\gamma}) - (\Gamma(\frac{\gamma+1}{\gamma}))^2}$
Coefficient of Variation	$\sqrt{\frac{\Gamma(\frac{\gamma+2}{\gamma})}{(\Gamma(\frac{\gamma+1}{\gamma}))^2} - 1}$

Parameter Estimation

Maximum likelihood estimation for the Weibull distribution is discussed in the Reliability chapter (Chapter 8). It is also discussed in Chapter 21 of Johnson, Kotz, and Balakrishnan.

Comments

The Weibull distribution is used extensively in reliability applications to model failure times.

Software

Most general purpose statistical software programs support at least some of the probability functions for the Weibull distribution.

1.3.6.6.9. Lognormal Distribution

Probability Density Function

A variable X is lognormally distributed if Y = LN(X) is normally distributed with "LN" denoting the natural logarithm. The general formula for the probability density function of the lognormal distribution is

$$f(x) = \frac{e^{-((\ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \qquad x \geq \theta; m, \sigma > 0$$

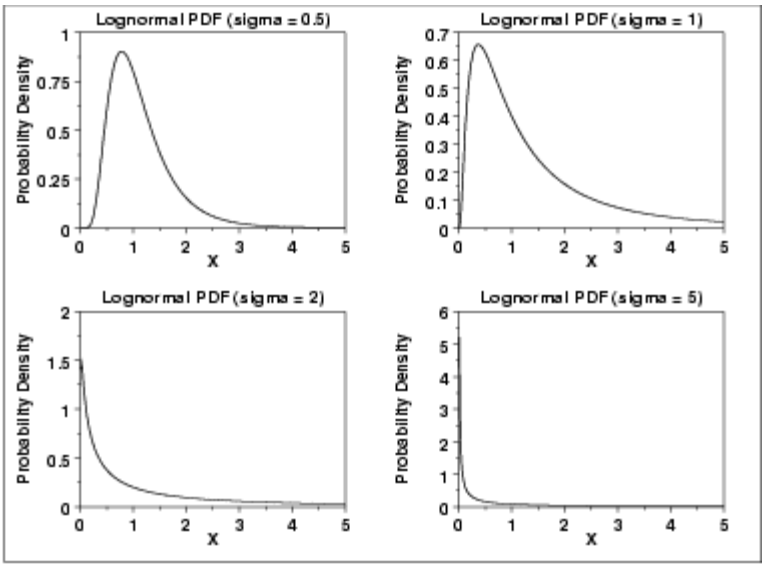
where σ is the shape parameter, θ is the location parameter and m is the scale parameter. The case where $\theta = 0$ and $m = 1$ is called the **standard lognormal distribution**. The case where θ equals zero is called the 2-parameter lognormal distribution.

The equation for the standard lognormal distribution is

$$f(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}} \qquad x \geq 0; \sigma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the lognormal probability density function for four values of σ.



There are several common parameterizations of the lognormal distribution. The form given here is from Evans, Hastings, and Peacock.

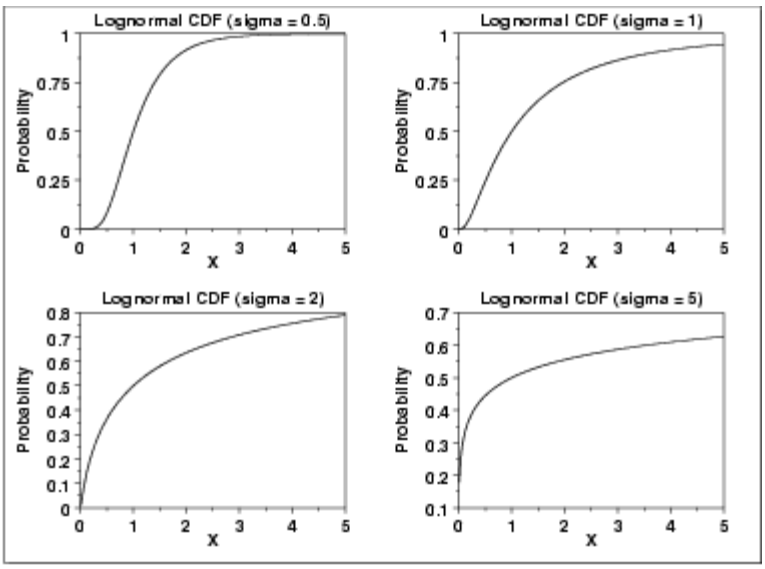
Cumulative Distribution Function

The formula for the cumulative distribution function of the lognormal distribution is

$$F(x) = \Phi\left(\frac{\ln(x)}{\sigma}\right) \qquad x \geq 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal cumulative distribution function with the same values of σ as the pdf plots above.



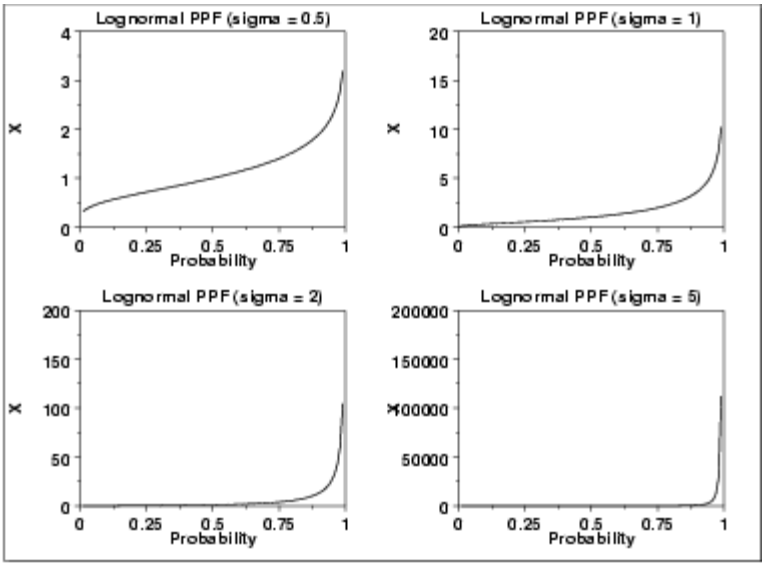
Percent Point Function

The formula for the percent point function of the lognormal distribution is

$$G(p) = \exp(\sigma\Phi^{-1}(p)) \qquad 0 \leq p < 1; \sigma > 0$$

where Φ^{-1} is the percent point function of the normal distribution.

The following is the plot of the lognormal percent point function with the same values of σ as the pdf plots above.



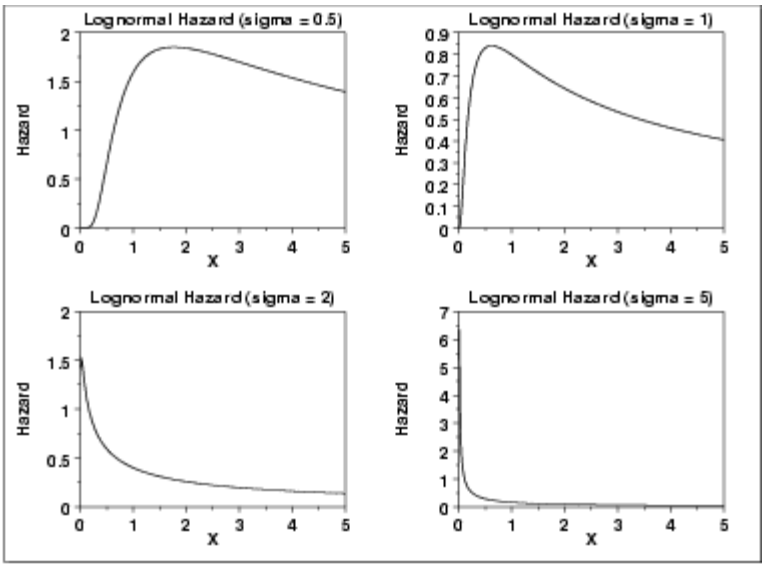
Hazard Function

The formula for the hazard function of the lognormal distribution is

$$h(x,\sigma) = \frac{\left(\frac{1}{x\sigma}\right)\phi\left(\frac{\ln x}{\sigma}\right)}{\Phi\left(\frac{-\ln x}{\sigma}\right)} \qquad x > 0; \sigma > 0$$

where ϕ is the probability density function of the normal distribution and Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal hazard function with the same values of σ as the pdf plots above.



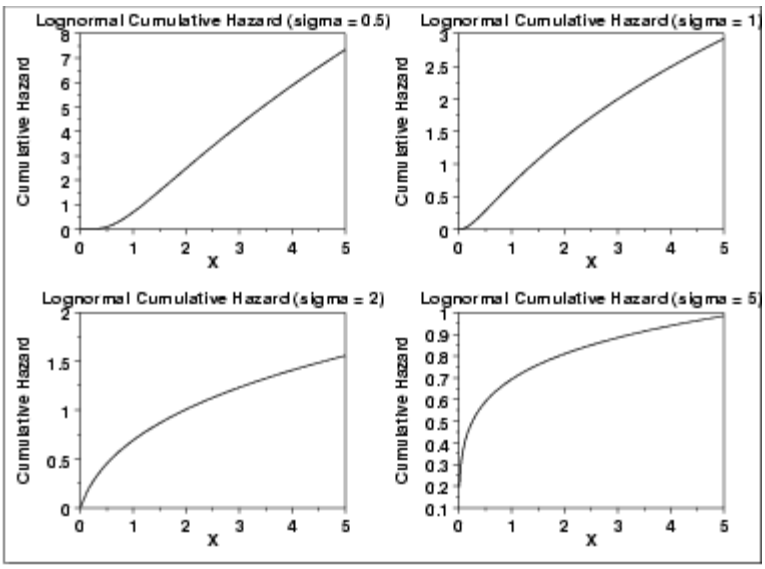
Cumulative Hazard Function

The formula for the cumulative hazard function of the lognormal distribution is

$$H(x) = -\ln(1 - \Phi(\frac{\ln(x)}{\sigma})) \qquad x \geq 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal cumulative hazard function with the same values of σ as the pdf plots above.



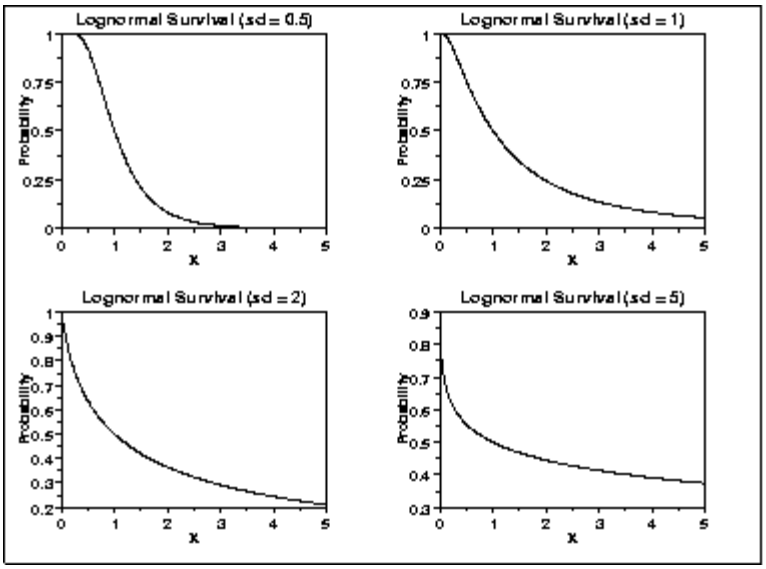
Survival Function

The formula for the survival function of the lognormal distribution is

$$S(x) = 1 - \Phi(\frac{\ln(x)}{\sigma}) \qquad x \geq 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal survival function with the same values of σ as the pdf plots above.



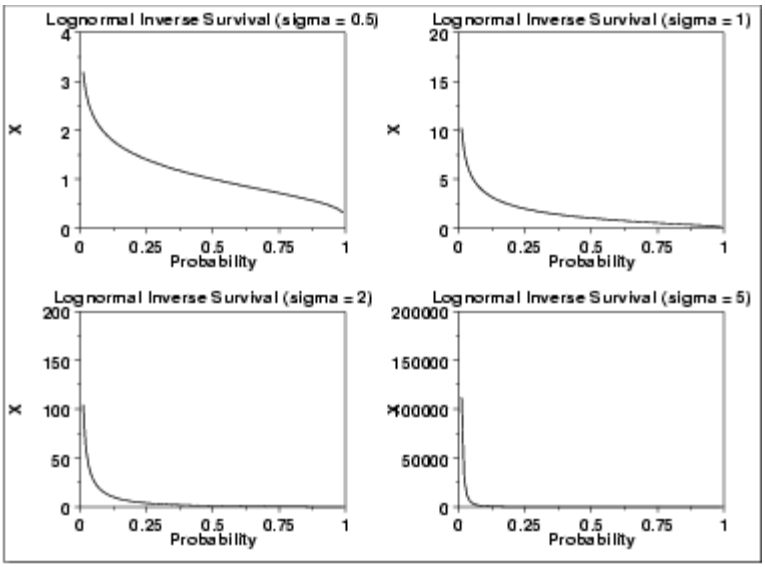
Inverse Survival Function

The formula for the inverse survival function of the lognormal distribution is

$$Z(p) = \exp(\sigma\Phi^{-1}(1-p)) \qquad 0 \le p < 1; \sigma > 0$$

where Φ^{-1} is the percent point function of the normal distribution.

The following is the plot of the lognormal inverse survival function with the same values of σ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	$e^{0.5\sigma^2}$
Median	Scale parameter ***m*** (= 1 if scale parameter not specified).
Mode	$\frac{1}{e^{\sigma^2}}$
Range	Zero to positive infinity

Standard Deviation	$\sqrt{e^{\sigma^2}(e^{\sigma^2}-1)}$
Skewness	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$
Kurtosis	$(e^{\sigma^2})^4+2(e^{\sigma^2})^3+3(e^{\sigma^2})^2-3$
Coefficient of Variation	$\sqrt{e^{\sigma^2}-1}$

Parameter Estimation

The maximum likelihood estimates for the scale parameter, ***m***, and the shape parameter, σ, are

$$\hat{m} = \exp \hat{\mu}$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} (\ln (X_i) - \hat{\mu})^2}{N}}$$

where

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \ln X_i}{N}$$

If the location parameter is known, it can be subtracted from the original data points before computing the maximum likelihood estimates of the shape and scale parameters.

Comments

The lognormal distribution is used extensively in reliability applications to model failure times. The lognormal and Weibull distributions are probably the most commonly used distributions in reliability applications.

Software

Most general purpose statistical software programs support at least some of the probability functions for the lognormal distribution.

1.3.6.6.10. Birnbaum-Saunders (Fatigue Life) Distribution

Probability Density Function

The Birnbaum-Saunders distribution is also commonly known as the fatigue life distribution. There are several alternative formulations of the Birnbaum-Saunders distribution in the literature.

The general formula for the probability density function of the Birnbaum-Saunders distribution is

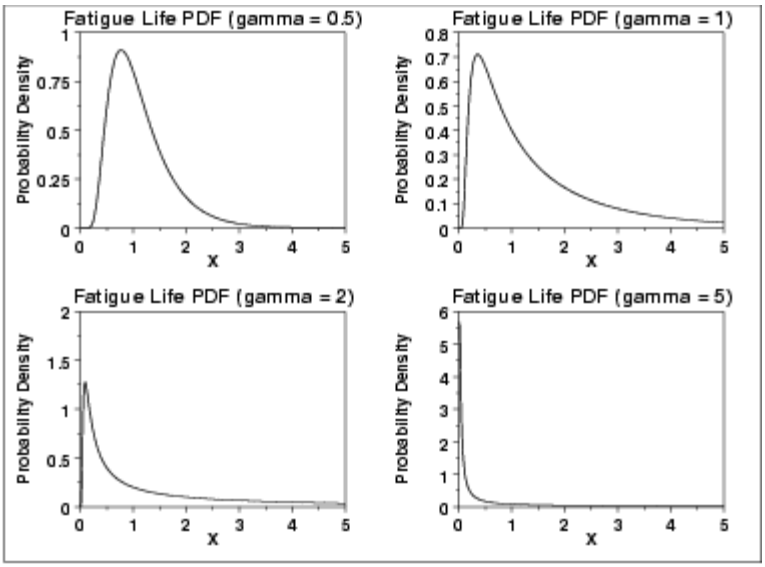
$$f(x) = \left(\frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)}\right)\phi\left(\frac{\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}}{\gamma}\right) \qquad x > \mu; \gamma, \beta > 0$$

where γ is the shape parameter, μ is the location parameter, β is the scale parameter, ϕ is the probability density function of the standard normal distribution, and Φ is the cumulative distribution function of the standard normal distribution. The case where $\mu = 0$ and $\beta = 1$ is called the **standard Birnbaum-Saunders distribution**. The equation for the standard Birnbaum-Saunders distribution reduces to

$$f(x) = \left(\frac{\sqrt{x} + \sqrt{\frac{1}{x}}}{2\gamma x}\right)\phi\left(\frac{\sqrt{x} - \sqrt{\frac{1}{x}}}{\gamma}\right) \qquad x > 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the Birnbaum-Saunders probability density function.

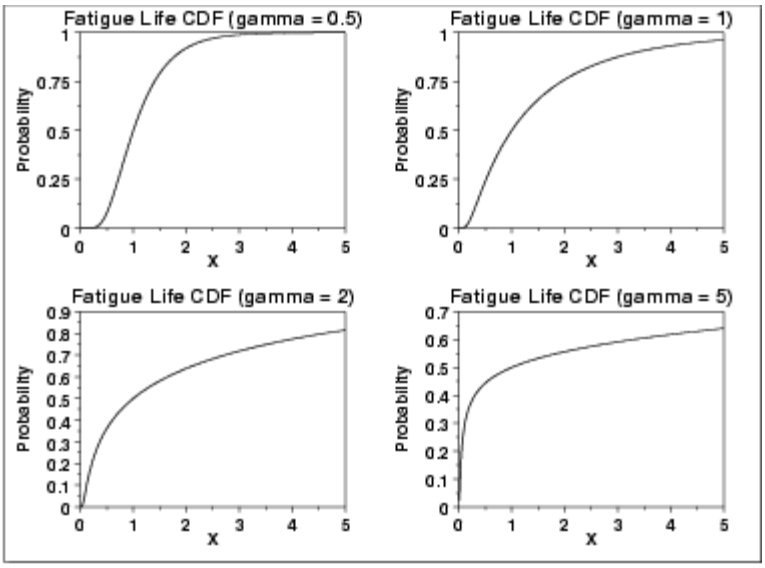


Cumulative Distribution Function

The formula for the cumulative distribution function of the Birnbaum-Saunders distribution is

$$F(x) = \Phi\left(\frac{\sqrt{x} - \sqrt{\frac{1}{x}}}{\gamma}\right) \qquad x > 0; \gamma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution. The following is the plot of the Birnbaum-Saunders cumulative distribution function with the same values of γ as the pdf plots above.



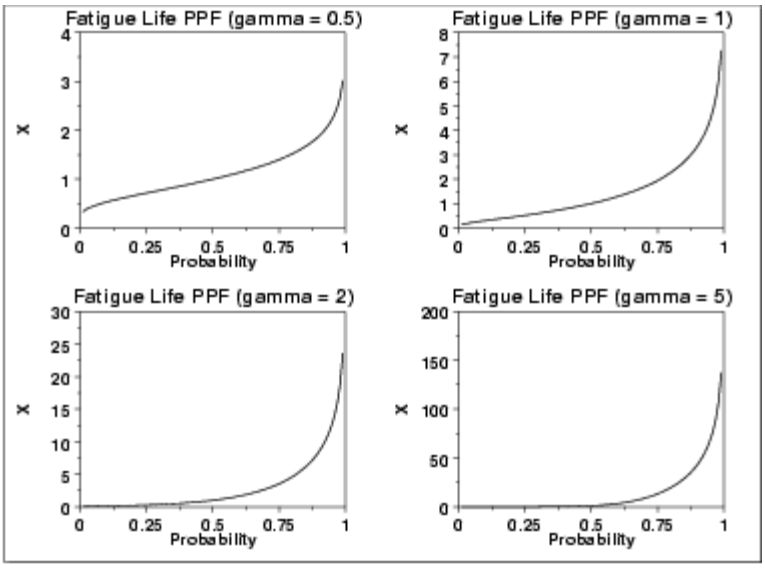
Percent Point Function

The formula for the percent point function of the Birnbaum-Saunders distribution is

$$G(p) = \frac{1}{4}\left[\gamma\Phi^{-1}(p) + \sqrt{4 + (\gamma\Phi^{-1}(p))^2}\right]^2$$

where Φ^{-1} is the percent point function of the standard normal distribution. The following is the plot of the Birnbaum-Saunders percent point function with the same values of γ as the pdf plots

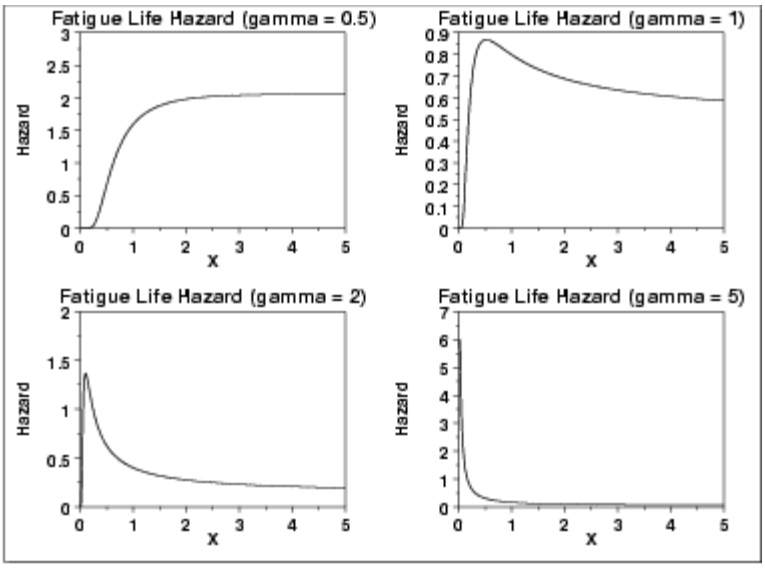
above.



Hazard Function

The Birnbaum-Saunders hazard function can be computed from the Birnbaum-Saunders probability density and cumulative distribution functions.

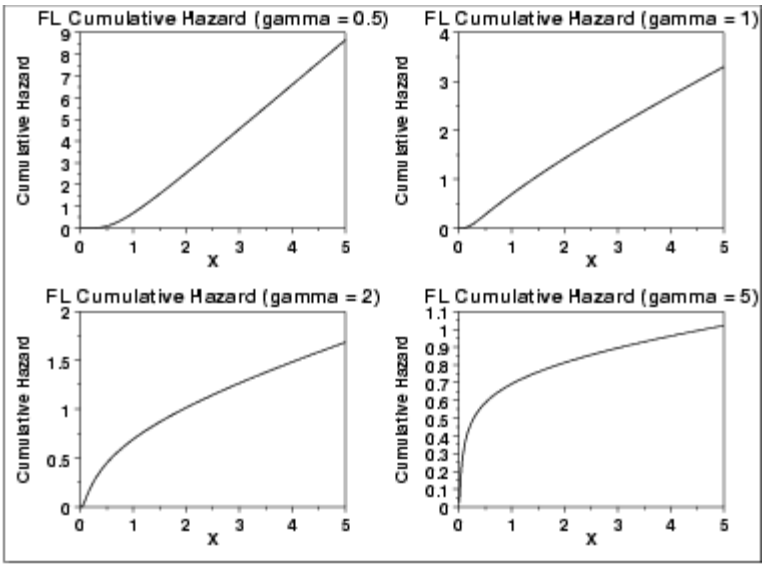
The following is the plot of the Birnbaum-Saunders hazard function with the same values of γ as the pdf plots above.



Cumulative Hazard Function

The Birnbaum-Saunders cumulative hazard function can be computed from the Birnbaum-Saunders cumulative distribution function.

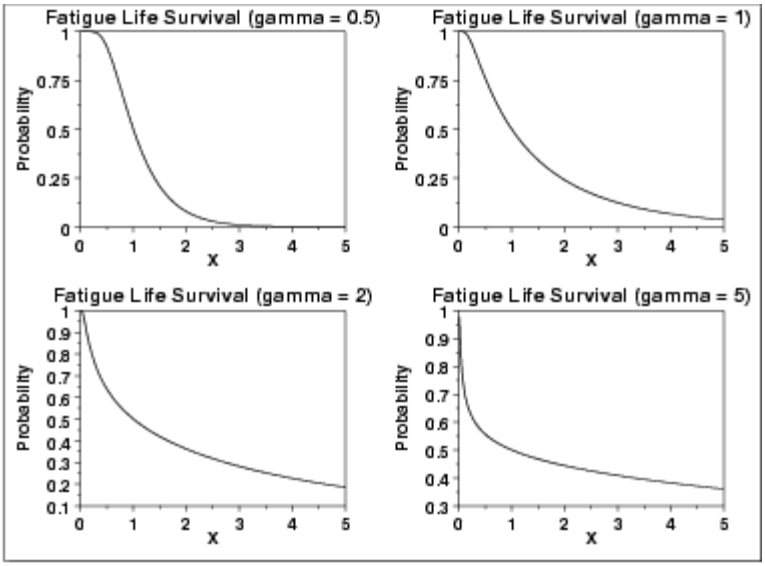
The following is the plot of the Birnbaum-Saunders cumulative hazard function with the same values of γ as the pdf plots above.



Survival Function

The Birnbaum-Saunders survival function can be computed from the Birnbaum-Saunders cumulative distribution function.

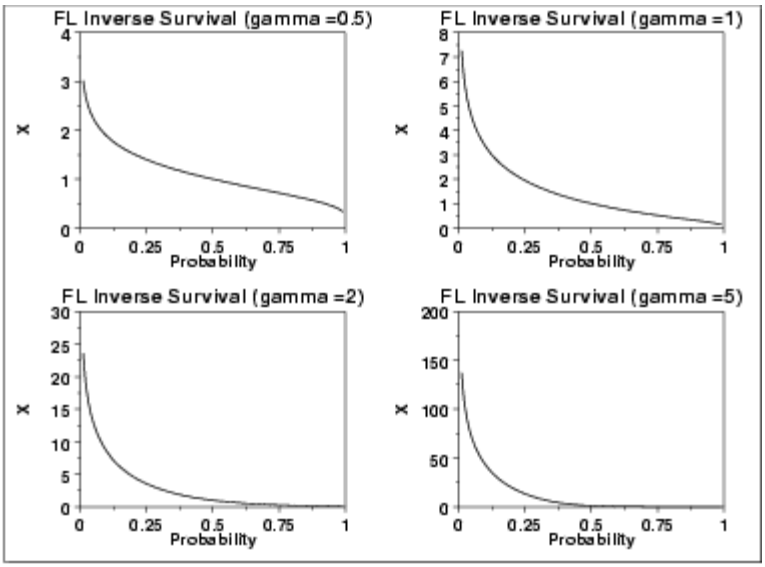
The following is the plot of the Birnbaum-Saunders survival function with the same values of γ as the pdf plots above.



Inverse Survival Function

The Birnbaum-Saunders inverse survival function can be computed from the Birnbaum-Saunders percent point function.

The following is the plot of the gamma inverse survival function with the same values of γ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	$1 + \frac{\gamma^2}{2}$
Range	Zero to positive infinity.
Standard Deviation	$\gamma\sqrt{1 + \frac{5\gamma^2}{4}}$
Coefficient of Variation	$\frac{2 + \gamma^2}{\gamma\sqrt{1 + 5\gamma^2}}$

Parameter Estimation

Maximum likelihood estimation for the Birnbaum-Saunders distribution is discussed in the Reliability chapter.

Comments

The Birnbaum-Saunders distribution is used extensively in reliability applications to model failure times.

Software

Some general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the Birnbaum-Saunders distribution. Support for this distribution is likely to be available for statistical programs that emphasize reliability applications.

The "bs" package implements support for the Birnbaum-Saunders distribution for the R package. See

> Leiva, V., Hernandez, H., and Riquelme, M. (2006). A New Package for the Birnbaum-Saunders Distribution. *Rnews*, 6/4, 35-40. (http://www.r-project.org)

1.3.6.6.11. Gamma Distribution

Probability Density Function

The general formula for the probability density function of the gamma distribution is

$$f(x) = \frac{\left(\frac{x-\mu}{\beta}\right)^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta\Gamma(\gamma)} \qquad x \geq \mu; \gamma, \beta > 0$$

where γ is the shape parameter, μ is the location parameter, β is the scale parameter, and Γ is the gamma function which has the formula

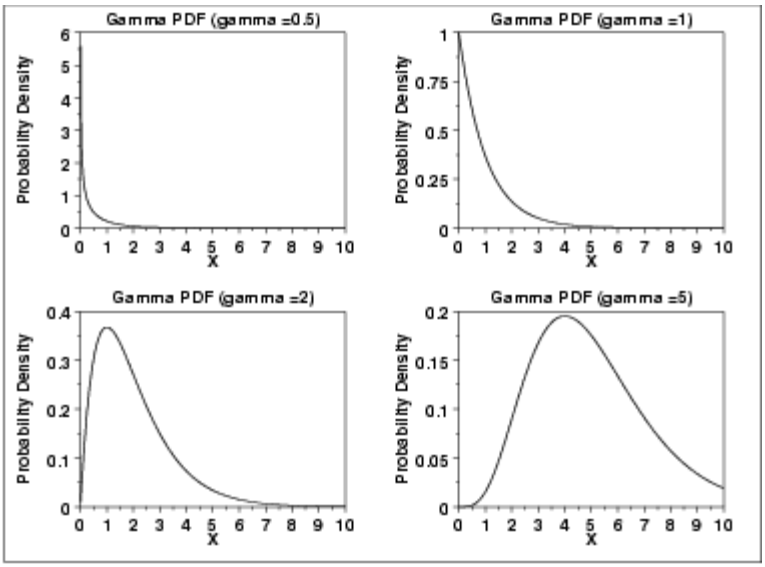
$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

The case where $\mu = 0$ and $\beta = 1$ is called the **standard gamma distribution**. The equation for the standard gamma distribution reduces to

$$f(x) = \frac{x^{\gamma-1} e^{-x}}{\Gamma(\gamma)} \qquad x \geq 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the gamma probability density function.



Cumulative Distribution Function

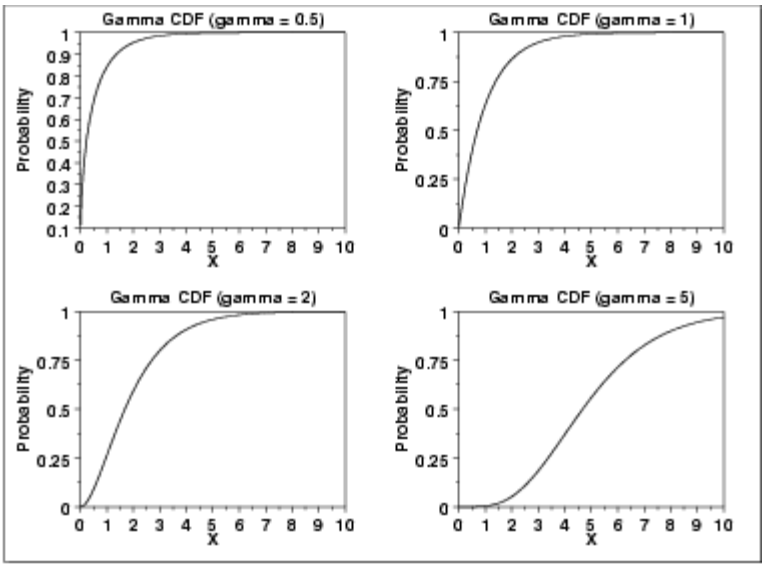
The formula for the cumulative distribution function of the gamma distribution is

$$F(x) = \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)} \qquad x \geq 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function. The incomplete gamma function has the formula

$$\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} dt$$

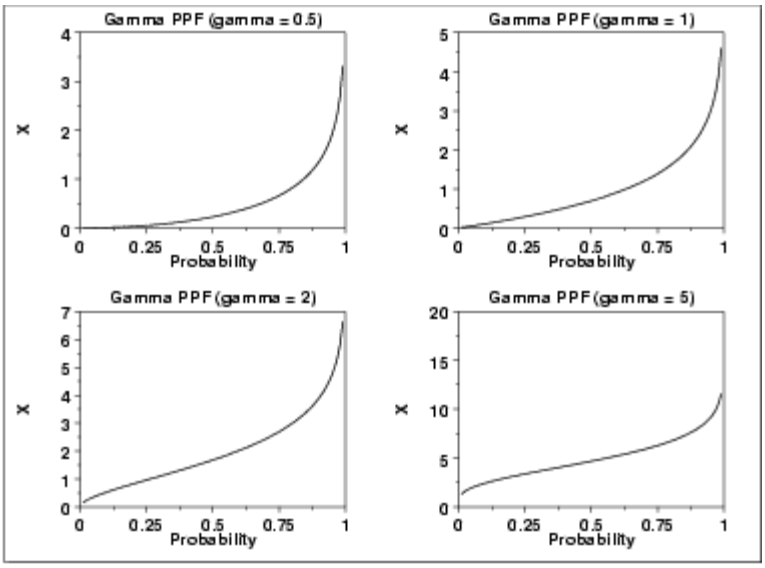
The following is the plot of the gamma cumulative distribution function with the same values of γ as the pdf plots above.



Percent Point Function

The formula for the percent point function of the gamma distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the gamma percent point function with the same values of γ as the pdf plots above.

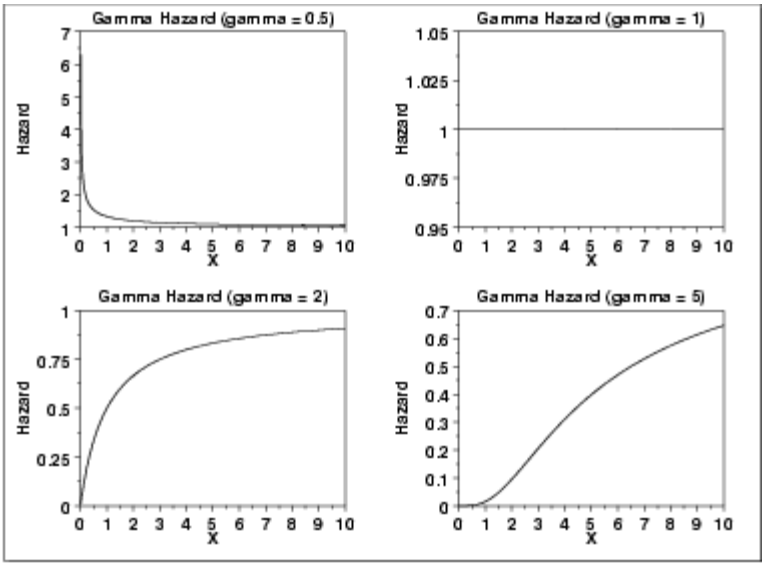


Hazard Function

The formula for the hazard function of the gamma distribution is

$$h(x) = \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma) - \Gamma_x(\gamma)} \qquad x \geq 0; \gamma > 0$$

The following is the plot of the gamma hazard function with the same values of γ as the pdf plots above.



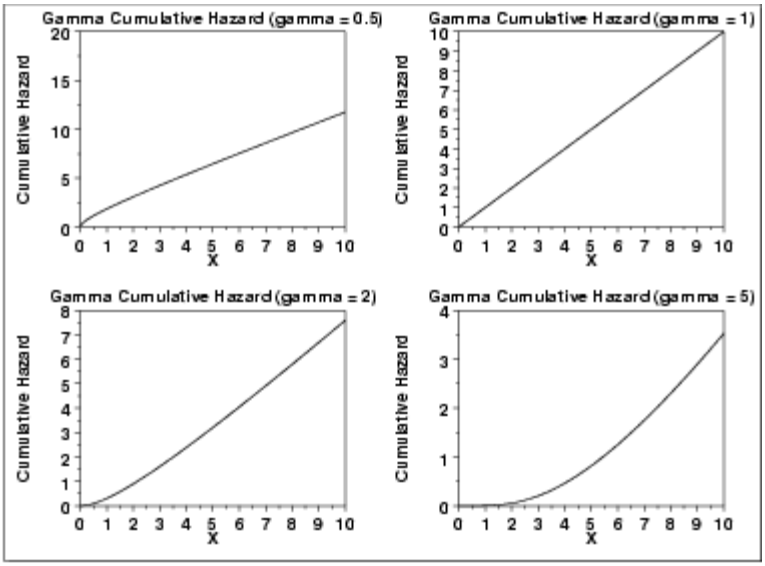
Cumulative Hazard Function

The formula for the cumulative hazard function of the gamma distribution is

$$H(x) = -\log\left(1 - \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)}\right) \qquad x \geq 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function defined above.

The following is the plot of the gamma cumulative hazard function with the same values of γ as the pdf plots above.



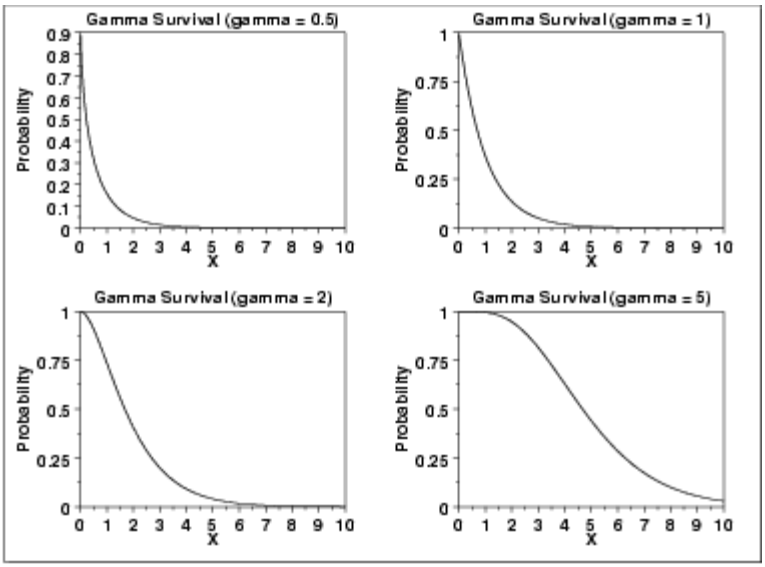
Survival Function

The formula for the survival function of the gamma distribution is

$$S(x) = 1 - \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)} \qquad x \geq 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function defined above.

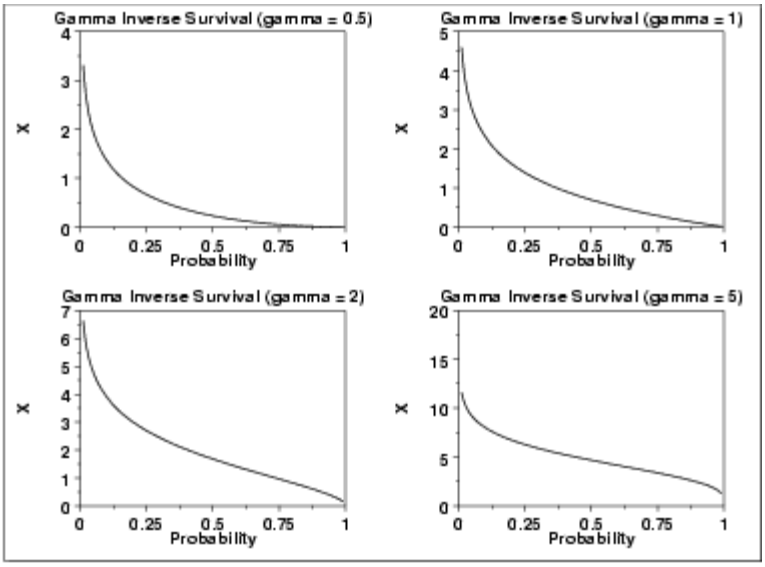
The following is the plot of the gamma survival function with the same values of γ as the pdf plots above.



Inverse Survival Function

The gamma inverse survival function does not exist in simple closed form. It is computed numberically.

The following is the plot of the gamma inverse survival function with the same values of γ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	γ
Mode	$\gamma - 1 \qquad \gamma \geq 1$
Range	Zero to positive infinity.
Standard Deviation	$\sqrt{\gamma}$
Skewness	$\frac{2}{\sqrt{\gamma}}$
Kurtosis	$3 + \frac{6}{\gamma}$
Coefficient of Variation	$\frac{1}{\sqrt{\gamma}}$

Parameter Estimation

The method of moments estimators of the gamma distribution are

$$\hat{\gamma} = (\frac{\bar{x}}{s})^2$$

$$\hat{\beta} = \frac{s^2}{\bar{x}}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

The equations for the maximum likelihood estimation of the shape and scale parameters are given in Chapter 18 of Evans, Hastings, and Peacock and Chapter 17 of Johnson, Kotz, and Balakrishnan. These equations need to be solved numerically; this is typically accomplished by using statistical software packages.

Software Some general purpose statistical software programs support at least some of the probability functions for the gamma distribution.

1.3.6.6.12. Double Exponential Distribution

Probability Density Function

The general formula for the probability density function of the double exponential distribution is

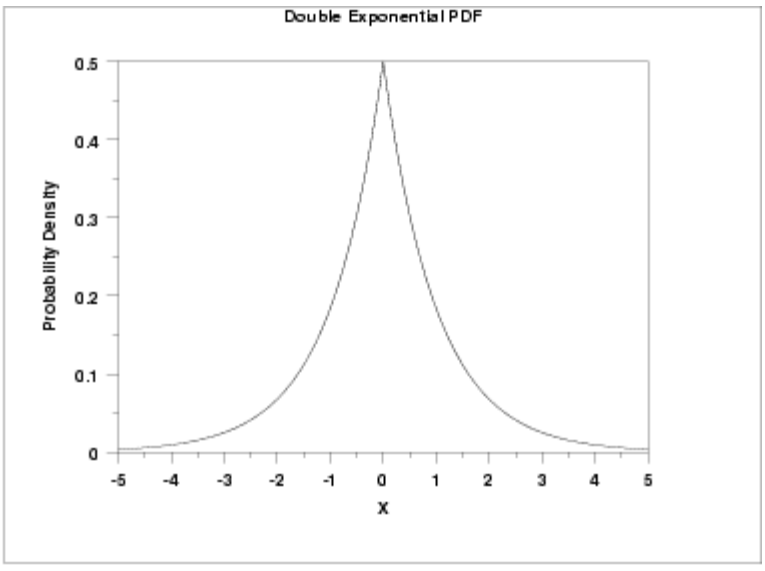
$$f(x) = \frac{e^{-|\frac{x-\mu}{\beta}|}}{2\beta}$$

where μ is the location parameter and β is the scale parameter. The case where $\mu = 0$ and $\beta = 1$ is called the **standard double exponential distribution**. The equation for the standard double exponential distribution is

$$f(x) = \frac{e^{-|x|}}{2}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the double exponential probability density function.



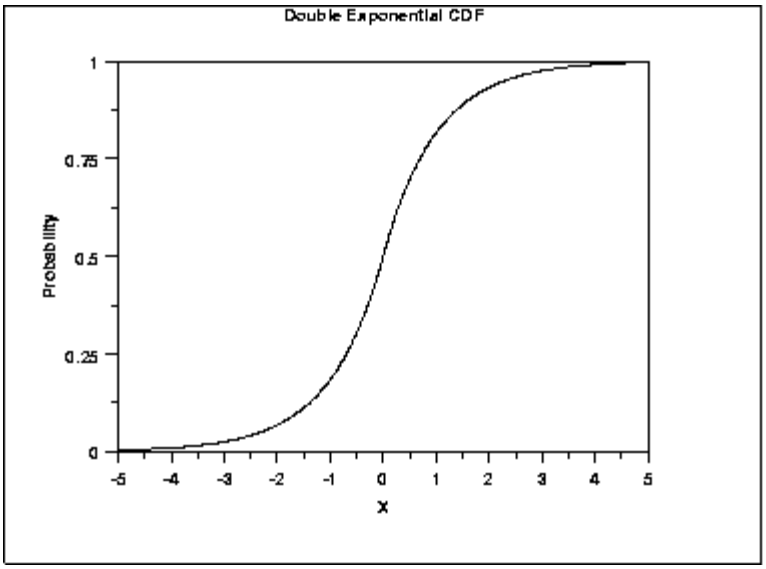
Cumulative Distribution

The formula for the cumulative distribution function of the double exponential distribution is

Function

$$F(x) = \begin{array}{ll} \frac{e^x}{2} & \text{for } x < 0 \\ 1 - \frac{e^{-x}}{2} & \text{for } x \geq 0 \end{array}$$

The following is the plot of the double exponential cumulative distribution function.

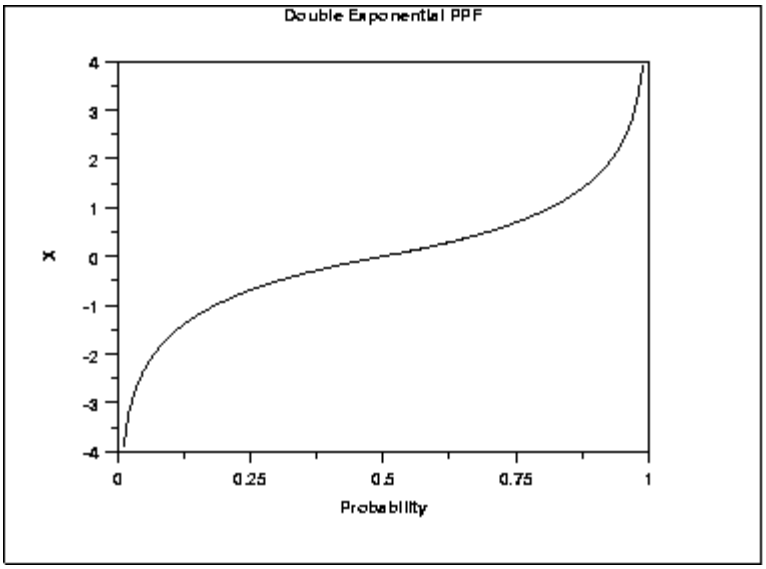


Percent Point Function

The formula for the percent point function of the double exponential distribution is

$$G(P) = \begin{array}{ll} \log(2p) & \text{for } p \leq 0.5 \\ -\log(2(1-p)) & \text{for } p > 0.5 \end{array}$$

The following is the plot of the double exponential percent point function.

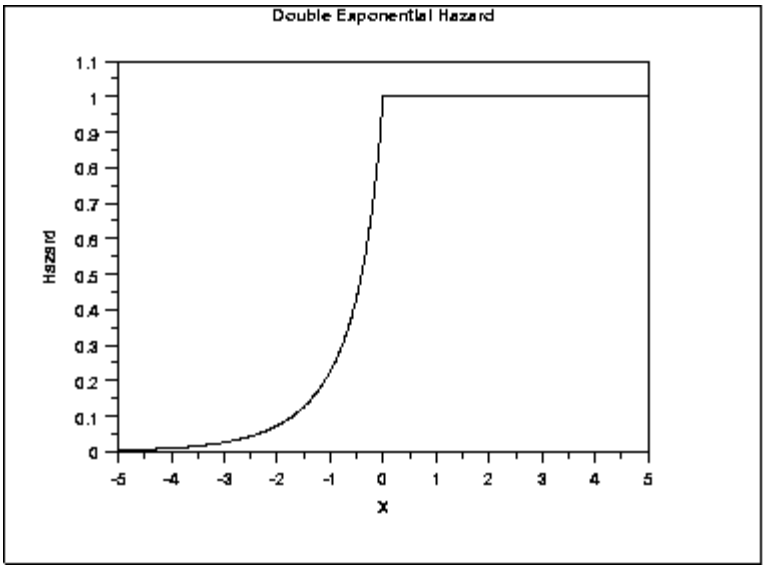


Hazard Function

The formula for the hazard function of the double exponential distribution is

$$h(x) = \begin{cases} \frac{e^x}{2-e^x} & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

The following is the plot of the double exponential hazard function.

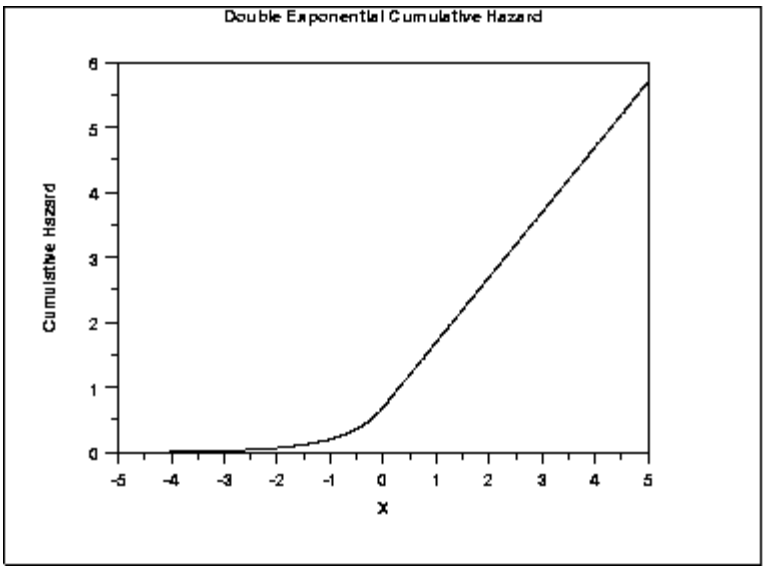


Cumulative Hazard Function

The formula for the cumulative hazard function of the double exponential distribution is

$$H(x) = \begin{cases} -log(1 - \frac{e^x}{2}) & \text{for } x < 0 \\ x + \log(2) & \text{for } x \geq 0 \end{cases}$$

The following is the plot of the double exponential cumulative hazard function.

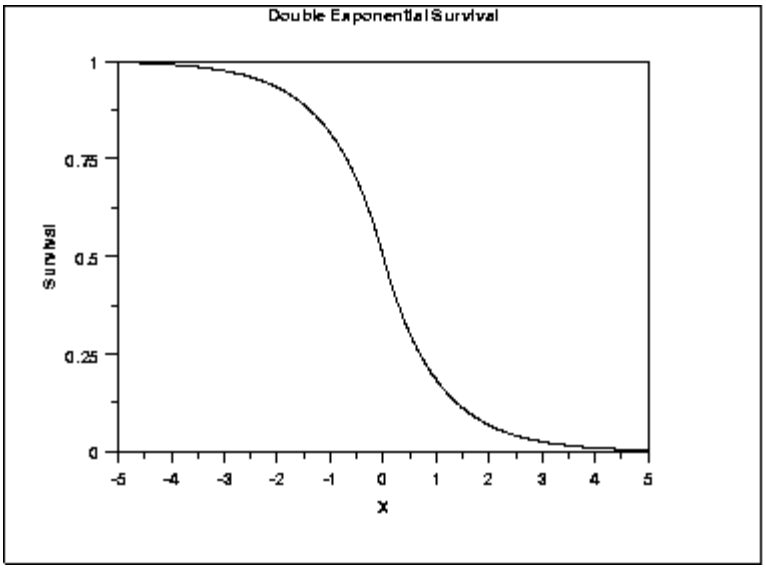


Survival Function

The double exponential survival function can be computed from the cumulative distribution function of the double exponential distribution.

The following is the plot of the double exponential survival

function.

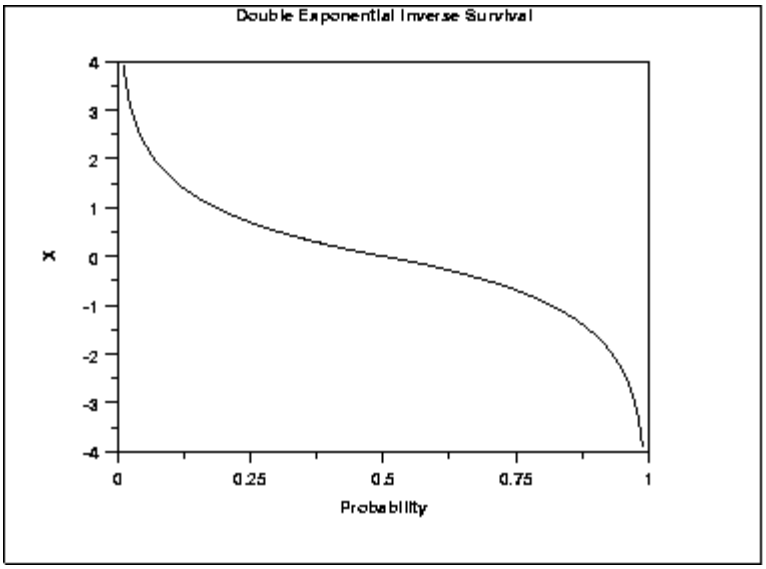


Inverse Survival Function

The formula for the inverse survival function of the double exponential distribution is

$$Z(P) = \begin{array}{ll} \log(2(1-p)) & \text{for } p \leq 0.5 \\ -\log(2p) & \text{for } p > 0.5 \end{array}$$

The following is the plot of the double exponential inverse survival function.



Common Statistics

Mean	μ
Median	μ
Mode	μ
Range	Negative infinity to positive infinity
Standard Deviation	$\sqrt{2}\beta$
Skewness	0
Kurtosis	6

Coefficient of Variation	$\sqrt{2}(\frac{\beta}{\mu})$

Parameter Estimation

The maximum likelihood estimators of the location and scale parameters of the double exponential distribution are

$$\hat{\mu} = \tilde{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} |X_i - \tilde{X}|}{N}$$

where $\tilde{X}$ is the sample median.

Software

Some general purpose statistical software programs support at least some of the probability functions for the double exponential distribution.

1.3.6.6.13. Power Normal Distribution

Probability Density Function

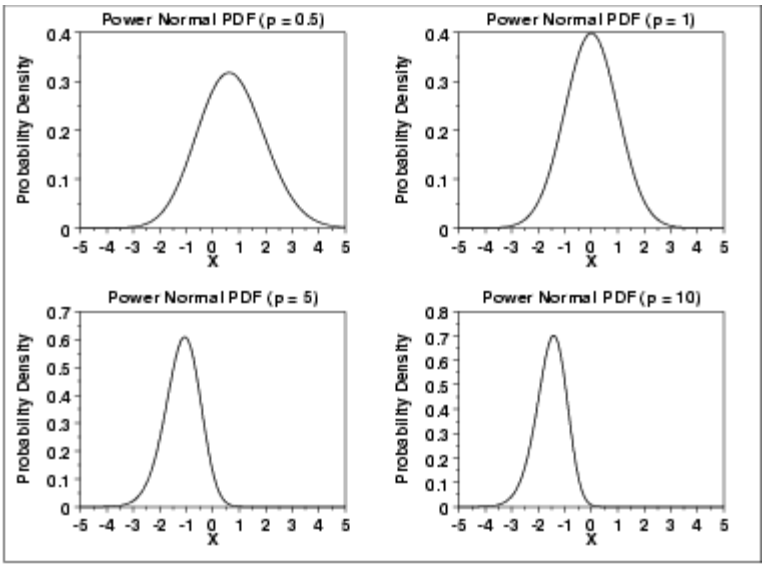
The formula for the probability density function of the standard form of the power normal distribution is

$$f(x,p) = p\phi(x)(\Phi(-x))^{p-1} \qquad x, p > 0$$

where ***p*** is the shape parameter (also referred to as the power parameter), Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

As with other probability distributions, the power normal distribution can be transformed with a location parameter, μ, and a scale parameter, σ. We omit the equation for the general form of the power normal distribution. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the power normal probability density function with four values of ***p***.



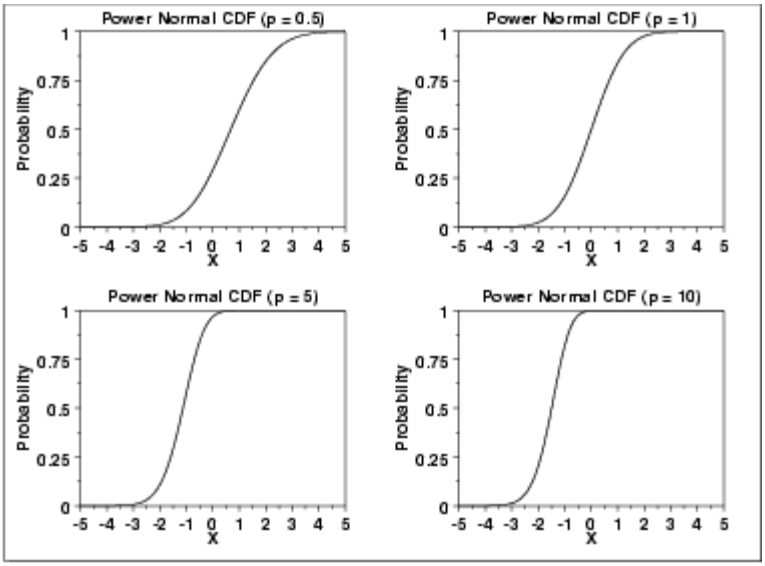
Cumulative Distribution Function

The formula for the cumulative distribution function of the power normal distribution is

$$F(x,p) = 1 - (\Phi(-x))^p \qquad x, p > 0$$

where Φ is the cumulative distribution function of the standard normal distribution.

The following is the plot of the power normal cumulative distribution function with the same values of ***p*** as the pdf plots above.



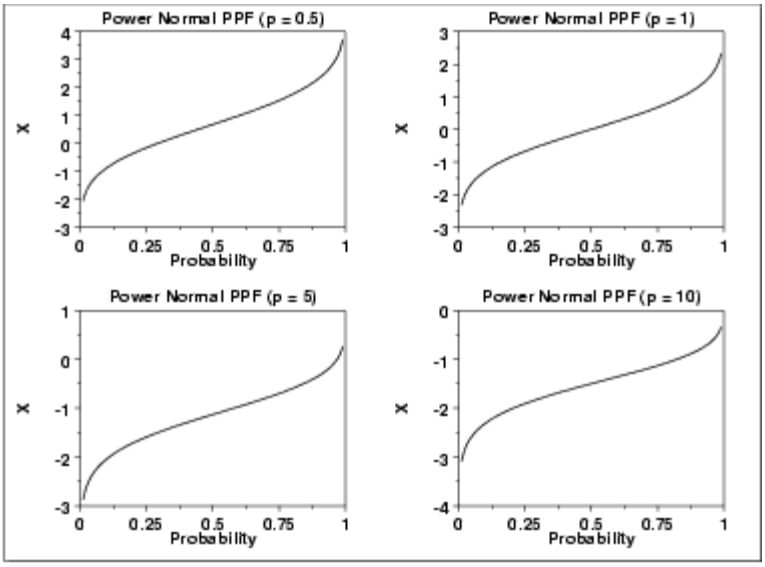
Percent Point Function

The formula for the percent point function of the power normal distribution is

$$G(f) = \Phi^{-1}(1 - (1-f)^{1/p}) \qquad 0 < f < 1; p > 0$$

where Φ^{-1} is the percent point function of the standard normal distribution.

The following is the plot of the power normal percent point function with the same values of ***p*** as the pdf plots above.



Hazard

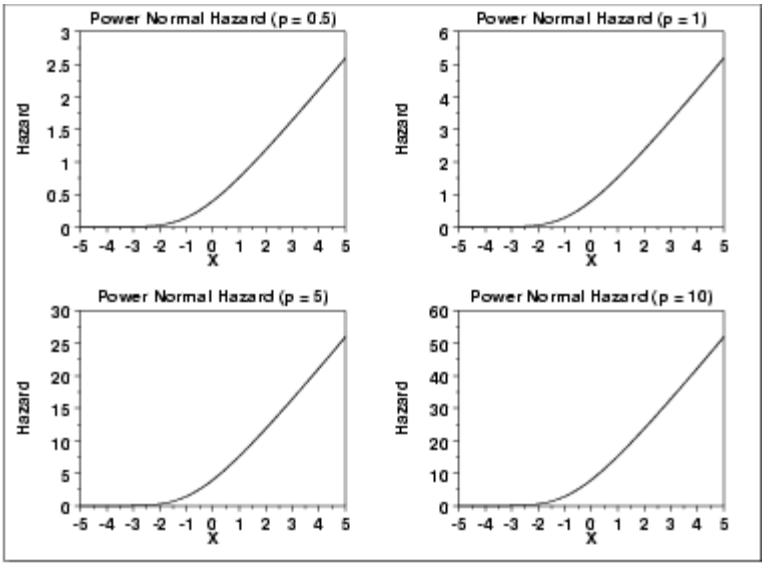
The formula for the hazard function of the power normal

Function

distribution is

$$h(x,p) = \frac{p\phi(x)}{\Phi(-x)} \qquad x, p > 0$$

The following is the plot of the power normal hazard function with the same values of ***p*** as the pdf plots above.

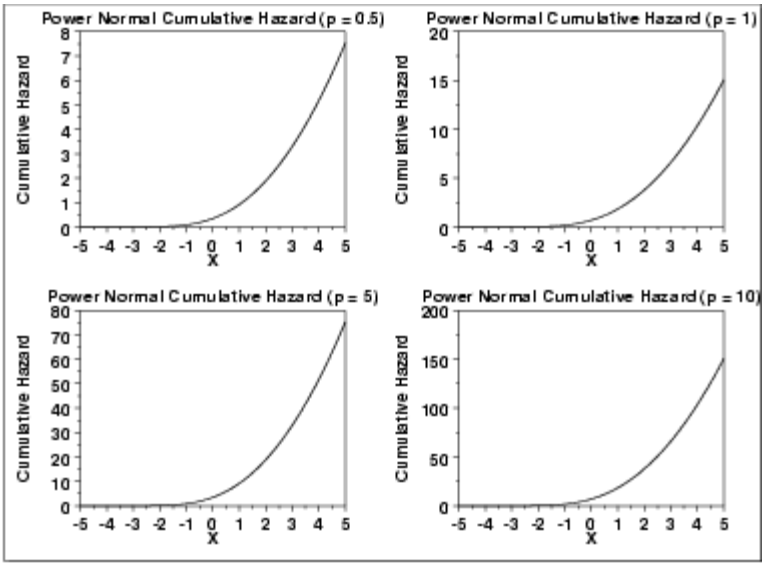


Cumulative Hazard Function

The formula for the cumulative hazard function of the power normal distribution is

$$H(x,p) = -\log\left((\Phi(-x))^p\right) \qquad x, p > 0$$

The following is the plot of the power normal cumulative hazard function with the same values of ***p*** as the pdf plots above.

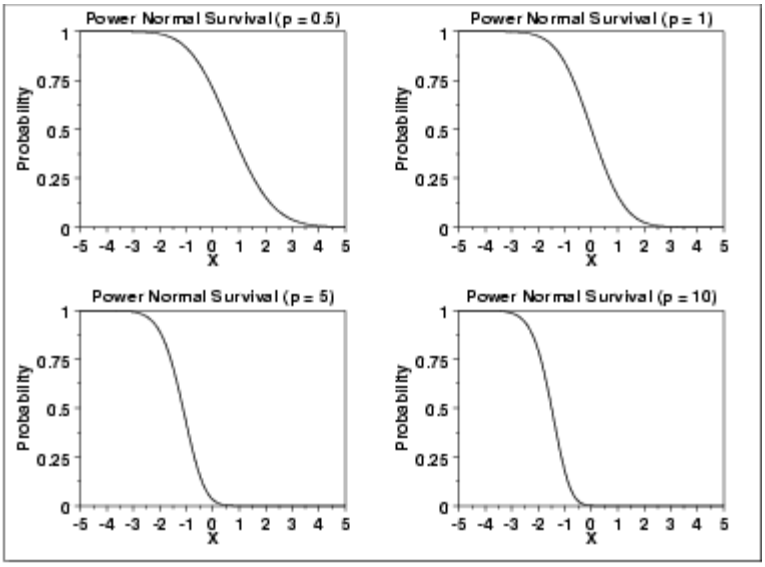


Survival Function

The formula for the survival function of the power normal distribution is

$$S(x,p) = (\Phi(-x))^p \qquad x, p > 0$$

The following is the plot of the power normal survival function with the same values of ***p*** as the pdf plots above.

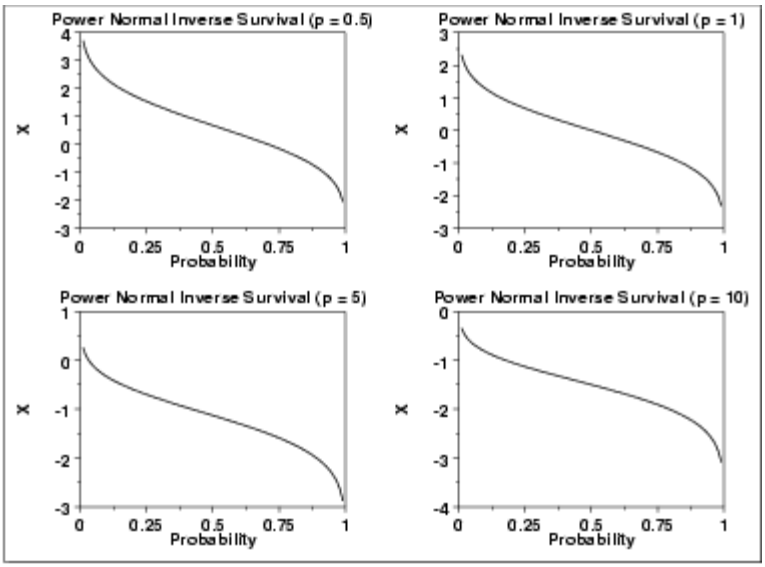


Inverse Survival Function

The formula for the inverse survival function of the power normal distribution is

$$Z(f) = \Phi^{-1}(1 - f^{1/p}) \qquad 0 < f < 1; p > 0$$

The following is the plot of the power normal inverse survival function with the same values of ***p*** as the pdf plots above.



Common Statistics

The statistics for the power normal distribution are complicated and require tables. Nelson discusses the mean, median, mode, and standard deviation of the power normal distribution and provides references to the appropriate tables.

Software

Most general purpose statistical software programs do not support the probability functions for the power normal distribution.

1.3.6.6.14. Power Lognormal Distribution

Probability Density Function

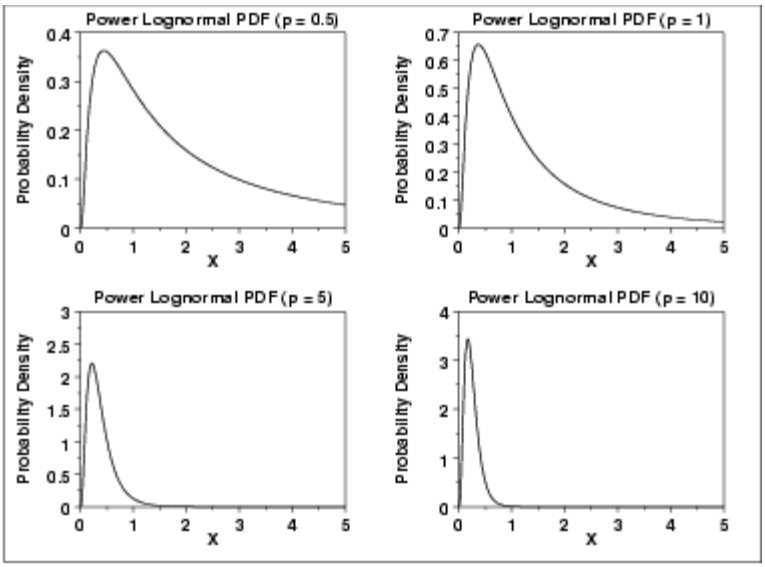
The formula for the probability density function of the standard form of the power lognormal distribution is

$$f(x,p,\sigma) = \left(\frac{p}{x\sigma}\right)\phi\left(\frac{\log x}{\sigma}\right)\left(\Phi\left(\frac{-\log x}{\sigma}\right)\right)^{p-1} \qquad x,p,\sigma > 0$$

where p (also referred to as the power parameter) and σ are the shape parameters, Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

As with other probability distributions, the power lognormal distribution can be transformed with a location parameter, μ, and a scale parameter, B. We omit the equation for the general form of the power lognormal distribution. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the power lognormal probability density function with four values of p and σ set to 1.



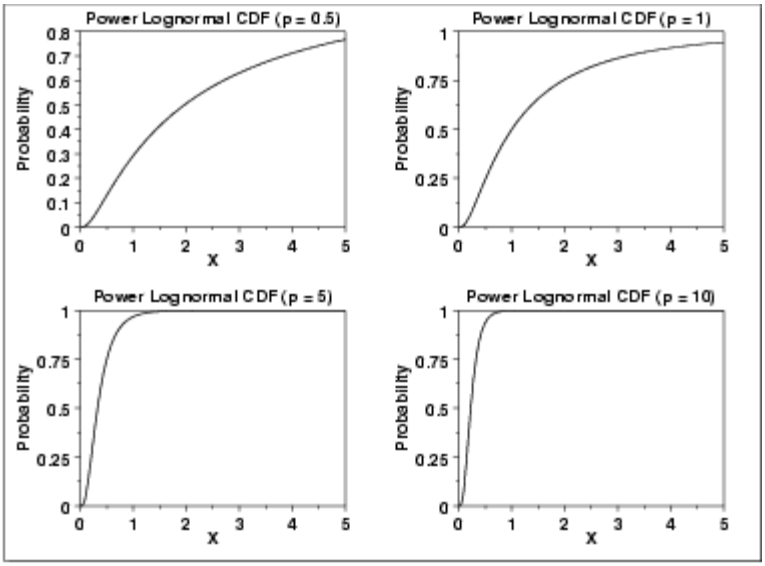
Cumulative Distribution Function

The formula for the cumulative distribution function of the power lognormal distribution is

$$F(x,p,\sigma) = 1 - (\Phi(\frac{-\log x}{\sigma}))^p \qquad x,p,\sigma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution.

The following is the plot of the power lognormal cumulative distribution function with the same values of ***p*** as the pdf plots above.



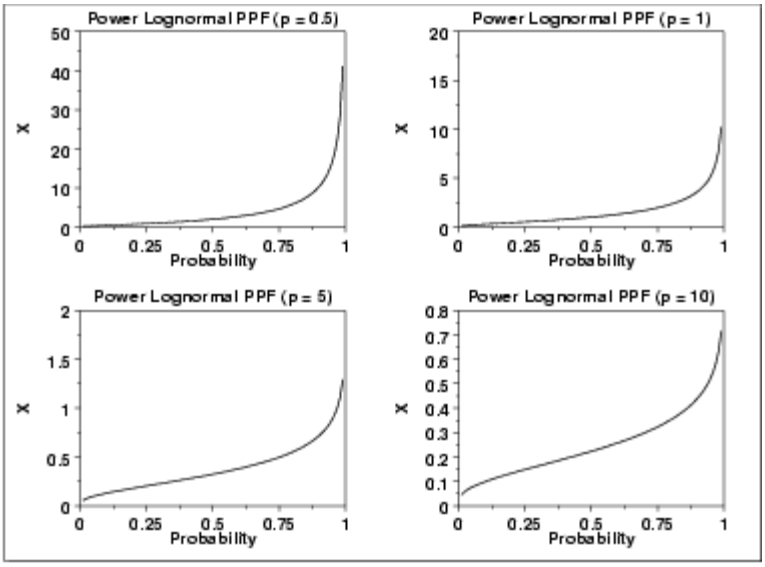
Percent Point Function

The formula for the percent point function of the power lognormal distribution is

$$G(f,p,\sigma) = \exp\left(\Phi^{-1}(1-(1-f)^{1/p})\sigma\right) \qquad 0 < p < 1; p,\sigma > 0$$

where Φ^{-1} is the percent point function of the standard normal distribution.

The following is the plot of the power lognormal percent point function with the same values of ***p*** as the pdf plots above.



Hazard

The formula for the hazard function of the power lognormal distribution

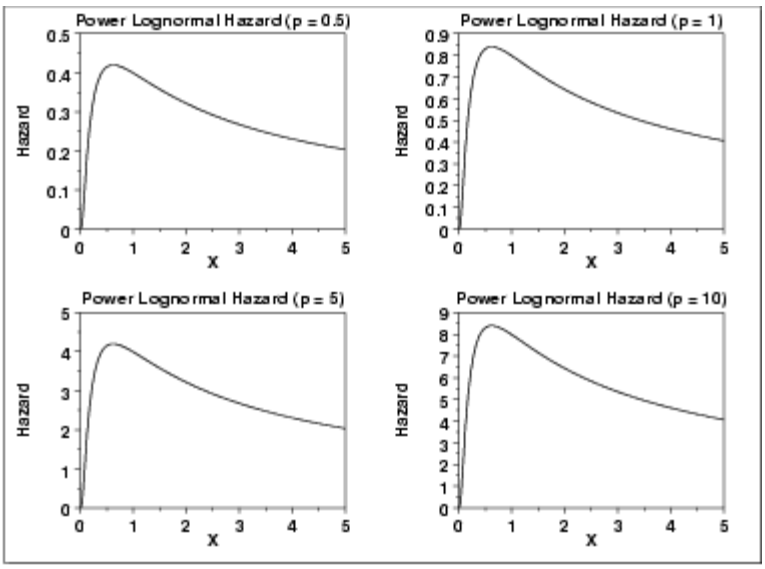
Function is

$$h(x,p,\sigma) = \frac{p(\frac{1}{x\sigma})\phi(\frac{\log x}{\sigma})}{\Phi(\frac{-\log x}{\sigma})} \qquad x,p,\sigma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

Note that this is simply a multiple (p) of the lognormal hazard function.

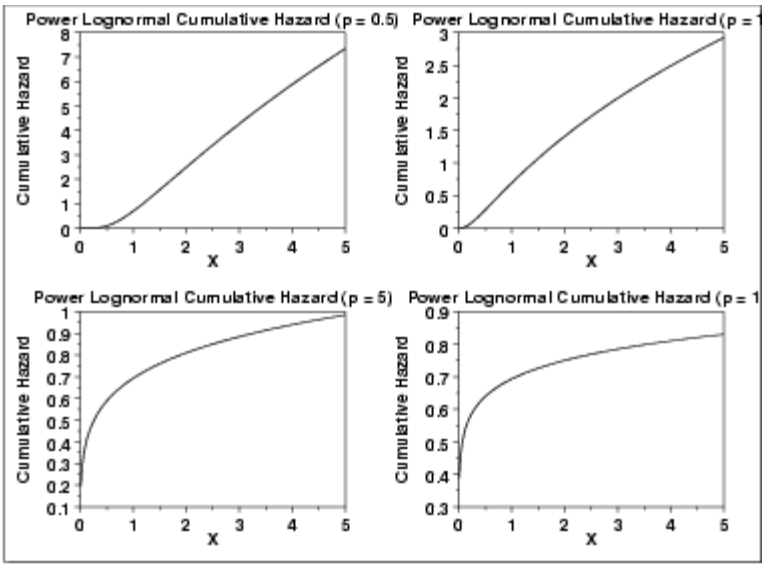
The following is the plot of the power lognormal hazard function with the same values of ***p*** as the pdf plots above.



Cumulative Hazard Function The formula for the cumulative hazard function of the power lognormal distribution is

$$H(x,p,\sigma) = -\log\left((\Phi(\frac{-\log x}{\sigma}))^p\right) \qquad x,p,\sigma > 0$$

The following is the plot of the power lognormal cumulative hazard function with the same values of ***p*** as the pdf plots above.

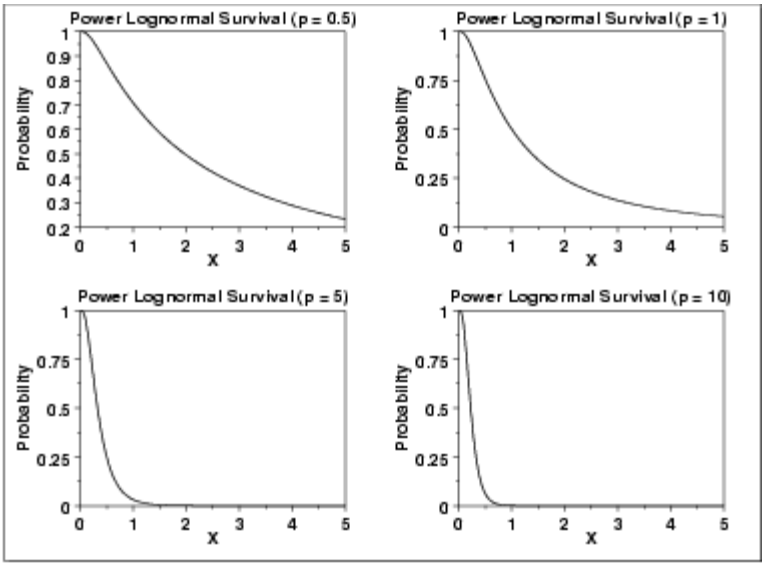


Survival Function

The formula for the survival function of the power lognormal distribution is

$$S(x,p,\sigma) = (\Phi(\frac{-\log x}{\sigma}))^p \qquad x,p,\sigma > 0$$

The following is the plot of the power lognormal survival function with the same values of ***p*** as the pdf plots above.

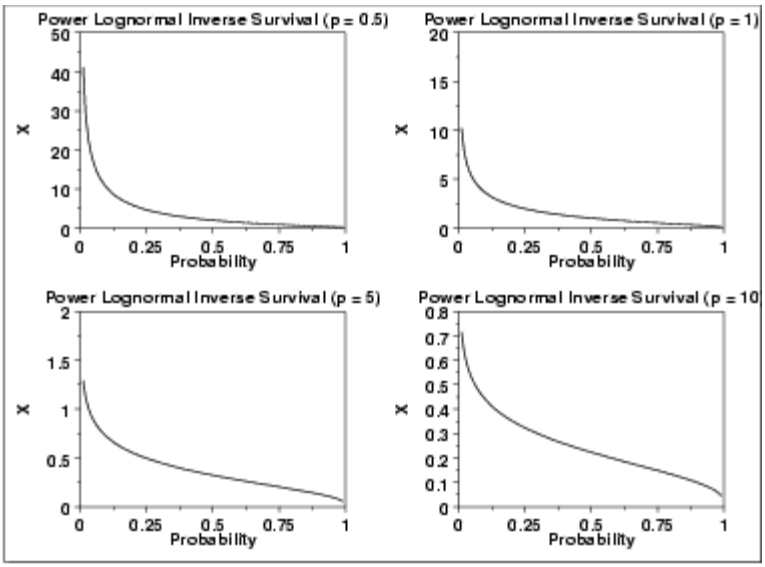


Inverse Survival Function

The formula for the inverse survival function of the power lognormal distribution is

$$Z(f,p,\sigma) = \exp\left(\Phi^{-1}(1-f^{1/p})\sigma\right) \qquad 0 < p < 1; p,\sigma > 0$$

The following is the plot of the power lognormal inverse survival function with the same values of ***p*** as the pdf plots above.



Common Statistics

The statistics for the power lognormal distribution are complicated and require tables. Nelson discusses the mean, median, mode, and standard deviation of the power lognormal distribution and provides references to the appropriate tables.

Parameter Estimation

Nelson discusses maximum likelihood estimation for the power lognormal distribution. These estimates need to be performed with computer software. Software for maximum likelihood estimation of the parameters of the power lognormal distribution is not as readily available as for other reliability distributions such as the exponential, Weibull, and lognormal.

Software

Most general purpose statistical software programs do not support the probability functions for the power lognormal distribution.

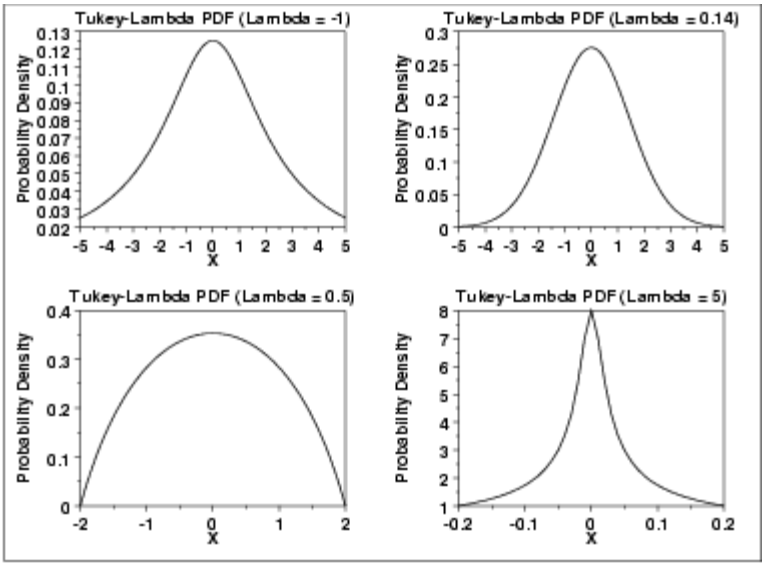
1.3.6.6.15. Tukey-Lambda Distribution

Probability Density Function

The Tukey-Lambda density function does not have a simple, closed form. It is computed numerically.

The Tukey-Lambda distribution has the shape parameter λ. As with other probability distributions, the Tukey-Lambda distribution can be transformed with a location parameter, μ, and a scale parameter, σ. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

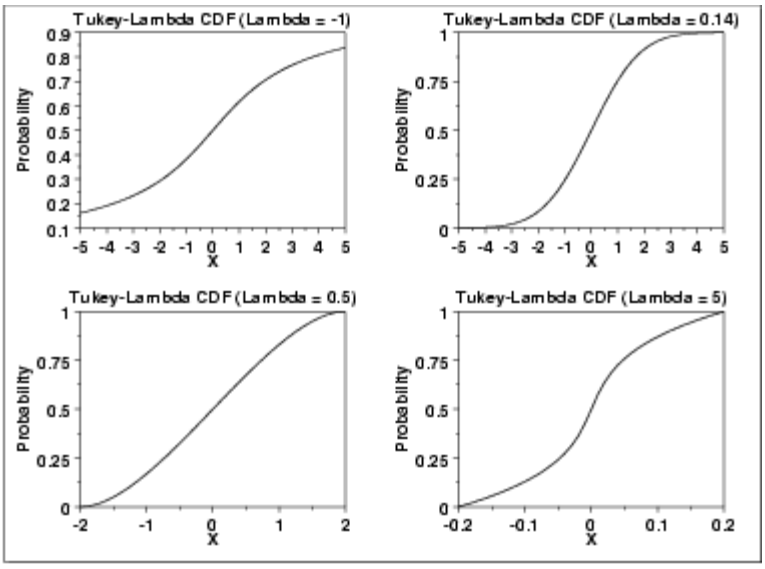
The following is the plot of the Tukey-Lambda probability density function for four values of λ.



Cumulative Distribution Function

The Tukey-Lambda distribution does not have a simple, closed form. It is computed numerically.

The following is the plot of the Tukey-Lambda cumulative distribution function with the same values of λ as the pdf plots above.

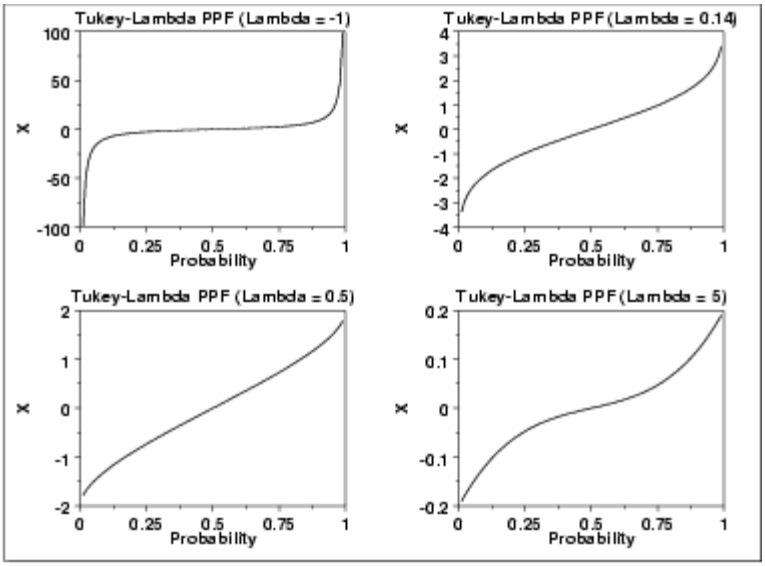


Percent Point Function

The formula for the percent point function of the standard form of the Tukey-Lambda distribution is

$$G(p) = \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda}$$

The following is the plot of the Tukey-Lambda percent point function with the same values of λ as the pdf plots above.



Other Probability Functions

The Tukey-Lambda distribution is typically used to identify an appropriate distribution (see the comments below) and not used in statistical models directly. For this reason, we omit the formulas, and plots for the hazard, cumulative hazard, survival, and inverse survival functions. We also omit the common statistics and parameter estimation sections.

Comments

The Tukey-Lambda distribution is actually a family of distributions that can approximate a number of common distributions. For example,

λ = -1 approximately Cauchy

$\lambda = 0$ exactly logistic
$\lambda = 0.14$ approximately normal
$\lambda = 0.5$ U-shaped
$\lambda = 1$ exactly uniform (from -1 to +1)

The most common use of this distribution is to generate a Tukey-Lambda PPCC plot of a data set. Based on the ppcc plot, an appropriate model for the data is suggested. For example, if the maximum correlation occurs for a value of λ at or near 0.14, then the data can be modeled with a normal distribution. Values of λ less than this imply a heavy-tailed distribution (with -1 approximating a Cauchy). That is, as the optimal value of λ goes from 0.14 to -1, increasingly heavy tails are implied. Similarly, as the optimal value of λ becomes greater than 0.14, shorter tails are implied.

As the Tukey-Lambda distribution is a symmetric distribution, the use of the Tukey-Lambda PPCC plot to determine a reasonable distribution to model the data only applies to symmetric distributuins. A histogram of the data should provide evidence as to whether the data can be reasonably modeled with a symmetric distribution.

Software Most general purpose statistical software programs do not support the probability functions for the Tukey-Lambda distribution.

1.3.6.6.16. Extreme Value Type I Distribution

Probability Density Function

The extreme value type I distribution has two forms. One is based on the smallest extreme and the other is based on the largest extreme. We call these the minimum and maximum cases, respectively. Formulas and plots for both cases are given. The extreme value type I distribution is also referred to as the Gumbel distribution.

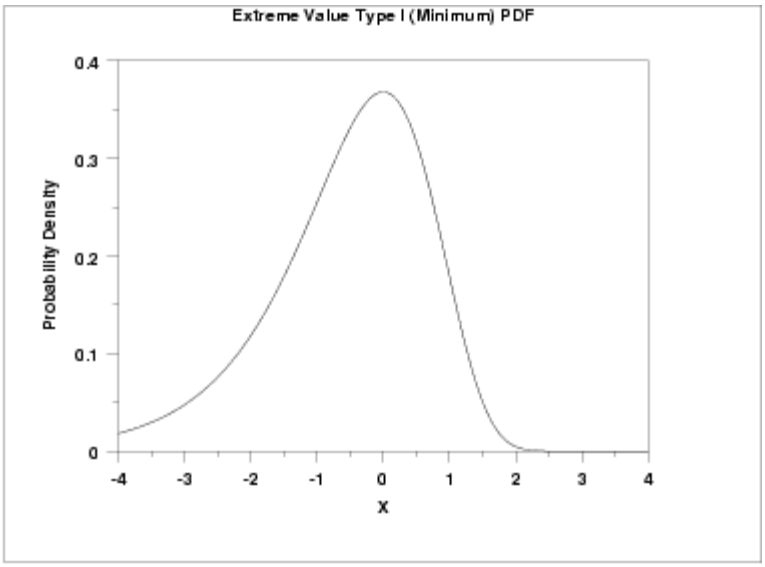
The general formula for the probability density function of the Gumbel (minimum) distribution is

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}}$$

where μ is the location parameter and β is the scale parameter. The case where $\mu = 0$ and $\beta = 1$ is called the **standard Gumbel distribution**. The equation for the standard Gumbel distribution (minimum) reduces to

$$f(x) = e^{x} e^{-e^{x}}$$

The following is the plot of the Gumbel probability density function for the minimum case.



The general formula for the probability density function of

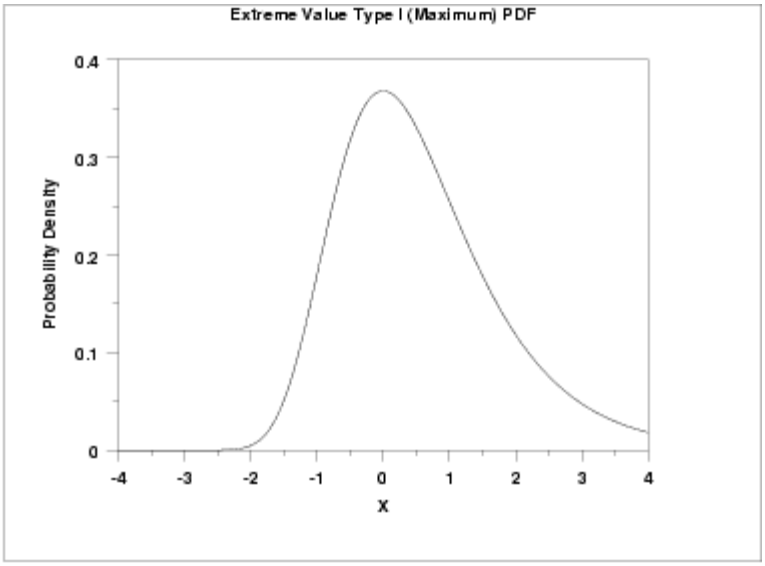
the Gumbel (maximum) distribution is

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}}$$

where μ is the location parameter and β is the scale parameter. The case where $\mu = 0$ and $\beta = 1$ is called the **standard Gumbel distribution**. The equation for the standard Gumbel distribution (maximum) reduces to

$$f(x) = e^{-x} e^{-e^{-x}}$$

The following is the plot of the Gumbel probability density function for the maximum case.



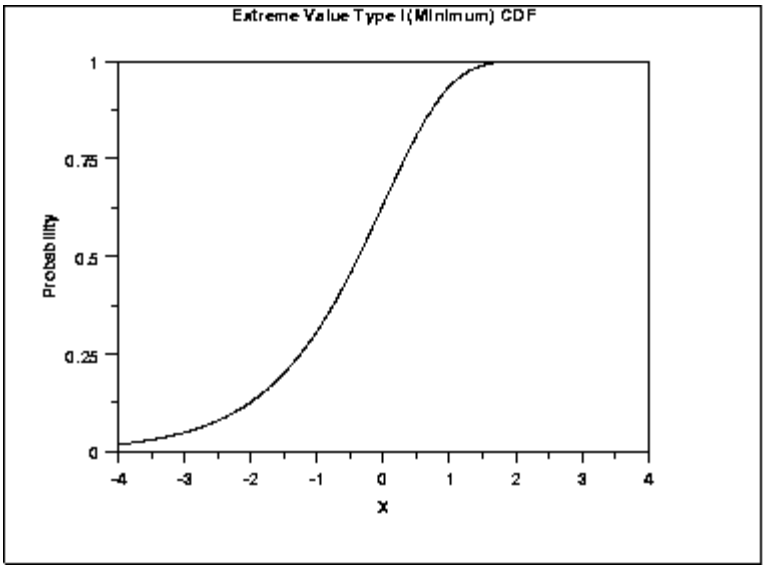
Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

Cumulative Distribution Function

The formula for the cumulative distribution function of the Gumbel distribution (minimum) is

$$F(x) = 1 - e^{-e^{x}}$$

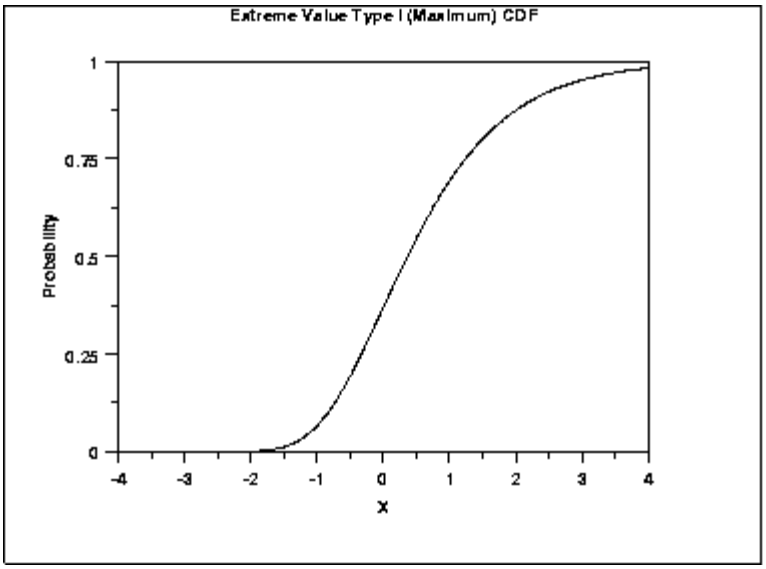
The following is the plot of the Gumbel cumulative distribution function for the minimum case.



The formula for the cumulative distribution function of the Gumbel distribution (maximum) is

$$F(x) = e^{-e^{-x}}$$

The following is the plot of the Gumbel cumulative distribution function for the maximum case.

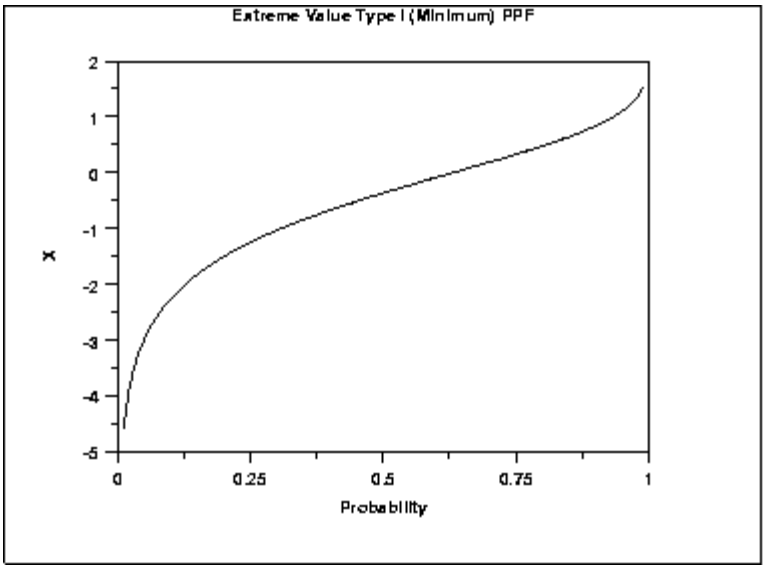


Percent Point Function

The formula for the percent point function of the Gumbel distribution (minimum) is

$$G(p) = \ln(\ln(\frac{1}{1-p}))$$

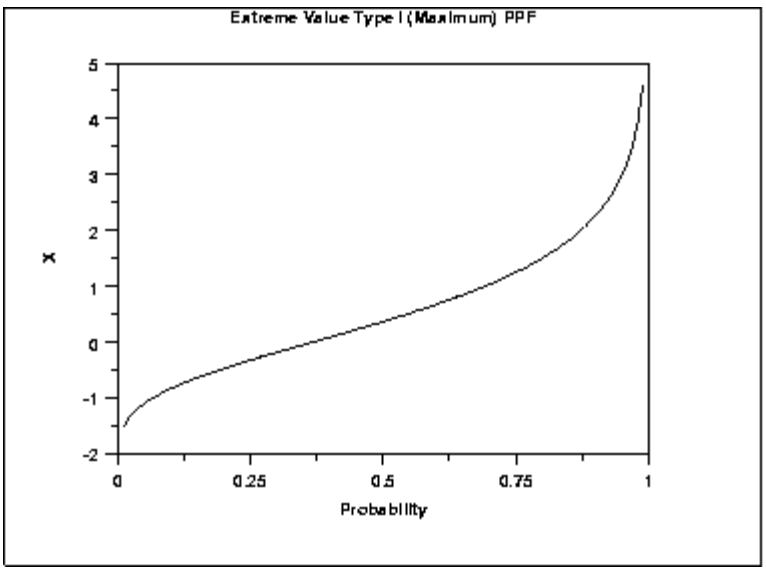
The following is the plot of the Gumbel percent point function for the minimum case.



The formula for the percent point function of the Gumbel distribution (maximum) is

$$G(p) = -\ln(\ln(\frac{1}{p}))$$

The following is the plot of the Gumbel percent point function for the maximum case.

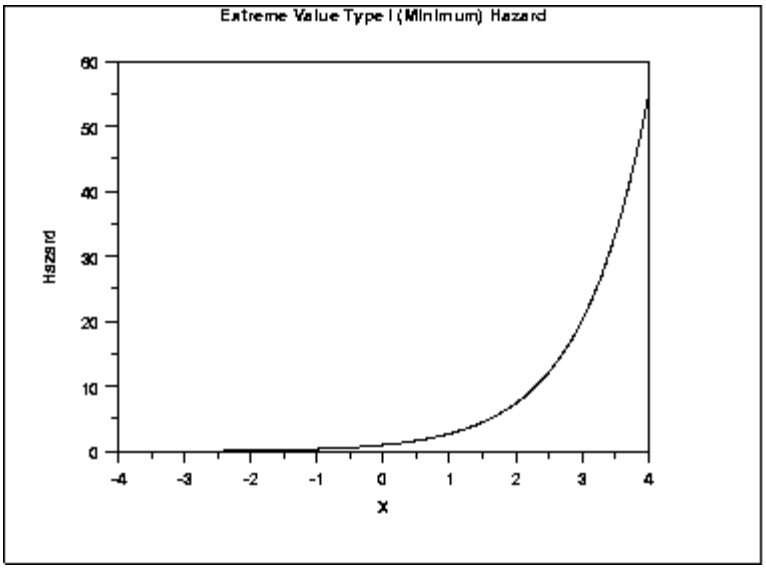


Hazard Function

The formula for the hazard function of the Gumbel distribution (minimum) is

$$h(x) = e^x$$

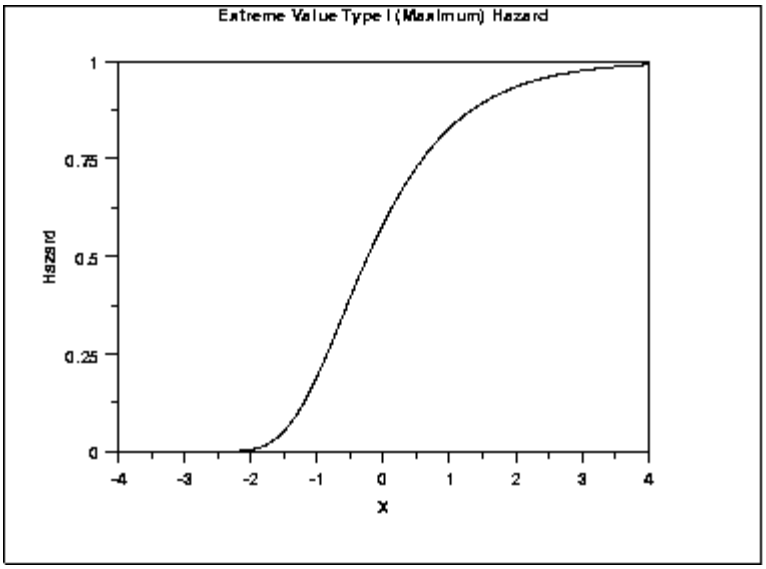
The following is the plot of the Gumbel hazard function for the minimum case.



The formula for the hazard function of the Gumbel distribution (maximum) is

$$h(x) = \frac{e^{-x}}{e^{e^{-x}} - 1}$$

The following is the plot of the Gumbel hazard function for the maximum case.

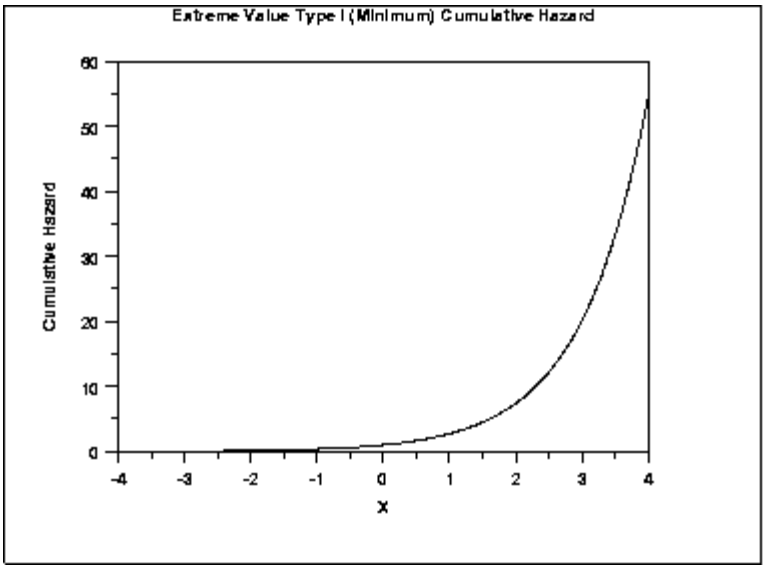


Cumulative Hazard Function

The formula for the cumulative hazard function of the Gumbel distribution (minimum) is

$$H(x) = e^{x}$$

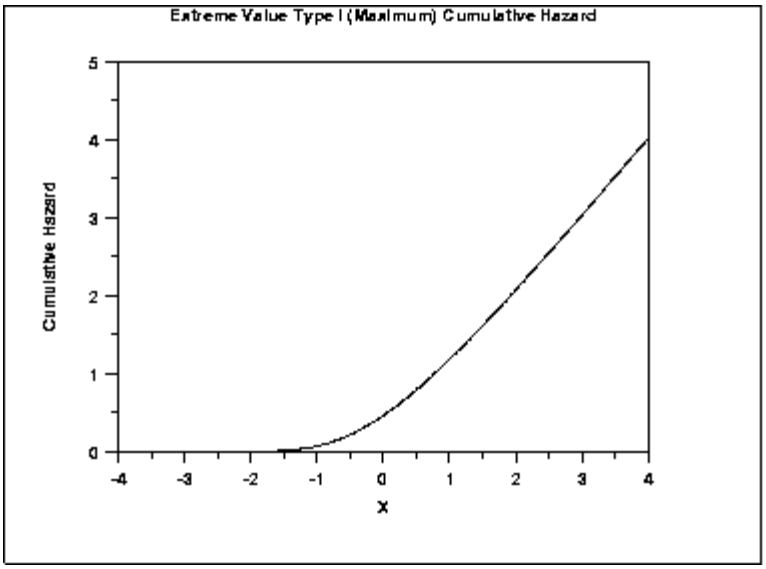
The following is the plot of the Gumbel cumulative hazard function for the minimum case.



The formula for the cumulative hazard function of the Gumbel distribution (maximum) is

$$H(x) = -\ln(1 - e^{-e^{-x}})$$

The following is the plot of the Gumbel cumulative hazard function for the maximum case.

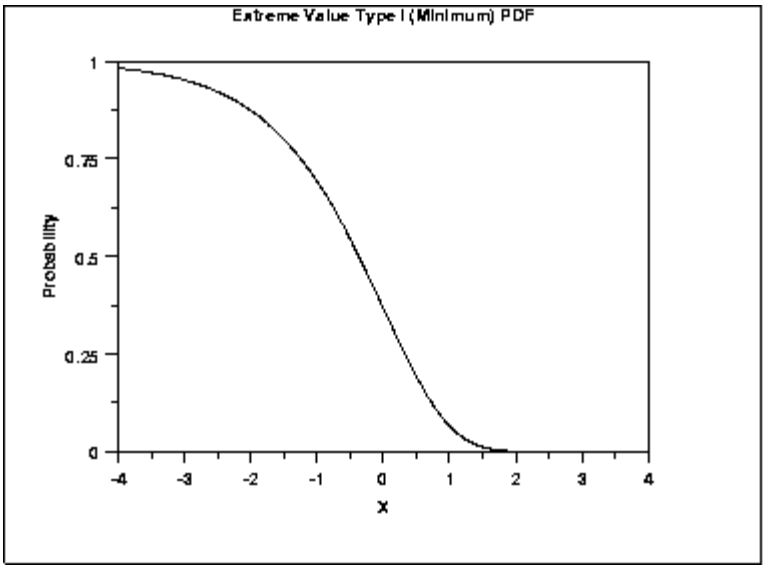


Survival Function

The formula for the survival function of the Gumbel distribution (minimum) is

$$S(x) = e^{-e^{x}}$$

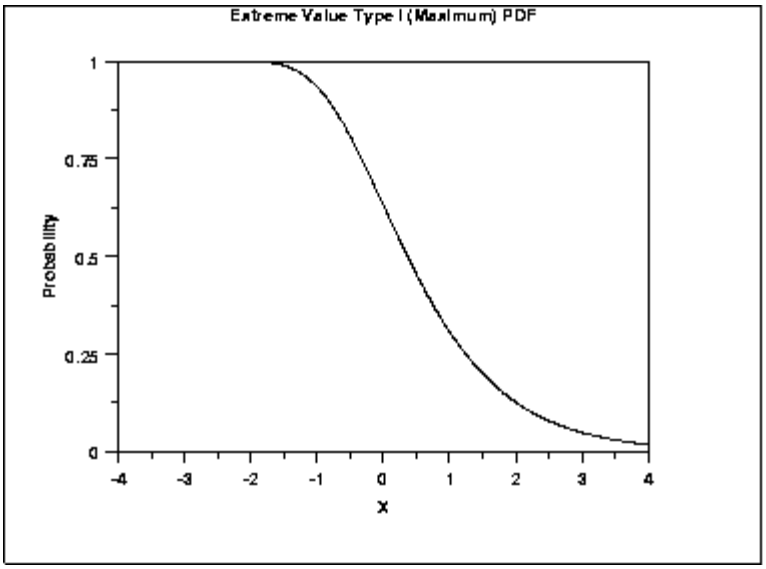
The following is the plot of the Gumbel survival function for the minimum case.



The formula for the survival function of the Gumbel distribution (maximum) is

$$S(x) = 1 - e^{-e^{-x}}$$

The following is the plot of the Gumbel survival function for the maximum case.

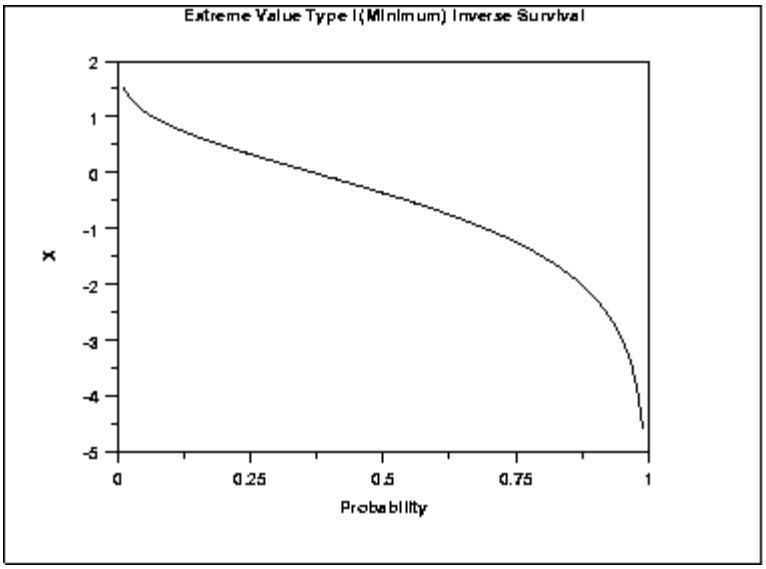


Inverse Survival Function

The formula for the inverse survival function of the Gumbel distribution (minimum) is

$$Z(p) = \ln(\ln(\frac{1}{p}))$$

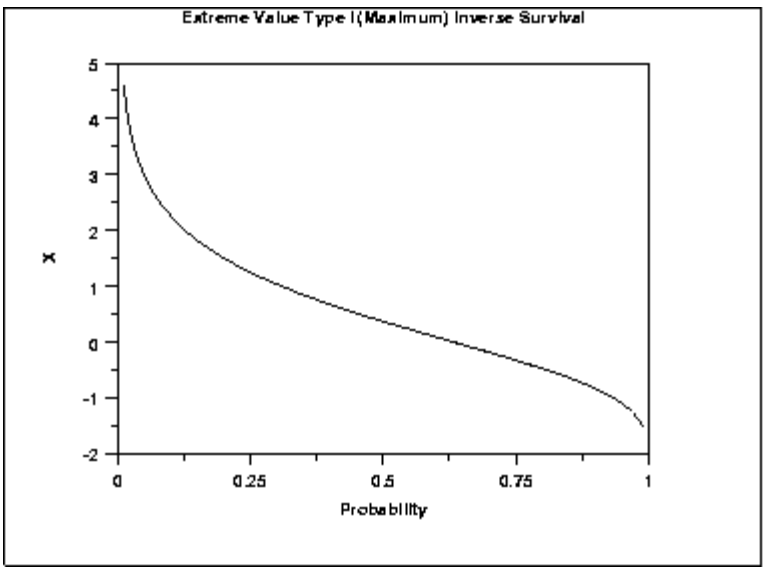
The following is the plot of the Gumbel inverse survival function for the minimum case.



The formula for the inverse survival function of the Gumbel distribution (maximum) is

$$Z(p) = -\ln(\ln(\frac{1}{1-p}))$$

The following is the plot of the Gumbel inverse survival function for the maximum case.



Common Statistics

The formulas below are for the maximum order statistic case.

Mean	$\mu + 0.5772\beta$ The constant 0.5772 is Euler's number.
Median	$\mu - \beta\ln(\ln(2))$
Mode	μ
Range	Negative infinity to positive infinity.
Standard Deviation	$\frac{\beta\pi}{\sqrt{6}}$

Skewness	1.13955
Kurtosis	5.4
Coefficient of Variation	$\frac{\beta\pi}{\sqrt{6}(\mu + 0.5772\beta)}$

Parameter Estimation

The method of moments estimators of the Gumbel (maximum) distribution are

$$\tilde{\beta} = \frac{s\sqrt{6}}{\pi}$$

$$\tilde{\mu} = \bar{X} - 0.5772\tilde{\beta}$$

where $\bar{X}$ and s are the sample mean and standard deviation, respectively.

The equations for the maximum likelihood estimation of the shape and scale parameters are discussed in Chapter 15 of Evans, Hastings, and Peacock and Chapter 22 of Johnson, Kotz, and Balakrishnan. These equations need to be solved numerically and this is typically accomplished by using statistical software packages.

Software

Some general purpose statistical software programs support at least some of the probability functions for the extreme value type I distribution.

1.3.6.6.17. Beta Distribution

Probability Density Function

The general formula for the probability density function of the beta distribution is

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \le x \le b; p, q > 0$$

where p and q are the shape parameters, a and b are the lower and upper bounds, respectively, of the distribution, and $B(p,q)$ is the beta function. The beta function has the formula

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

The case where $a = 0$ and $b = 1$ is called the **standard beta distribution**. The equation for the standard beta distribution is

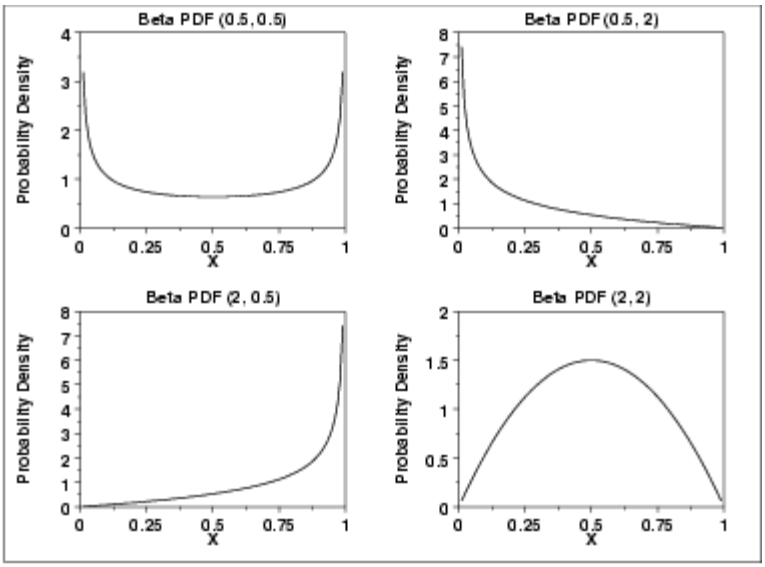
$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)} \qquad 0 \le x \le 1; p, q > 0$$

Typically we define the general form of a distribution in terms of location and scale parameters. The beta is different in that we define the general distribution in terms of the lower and upper bounds. However, the location and scale parameters can be defined in terms of the lower and upper limits as follows:

location = a
scale = b - a

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the beta probability density function for four different values of the shape parameters.



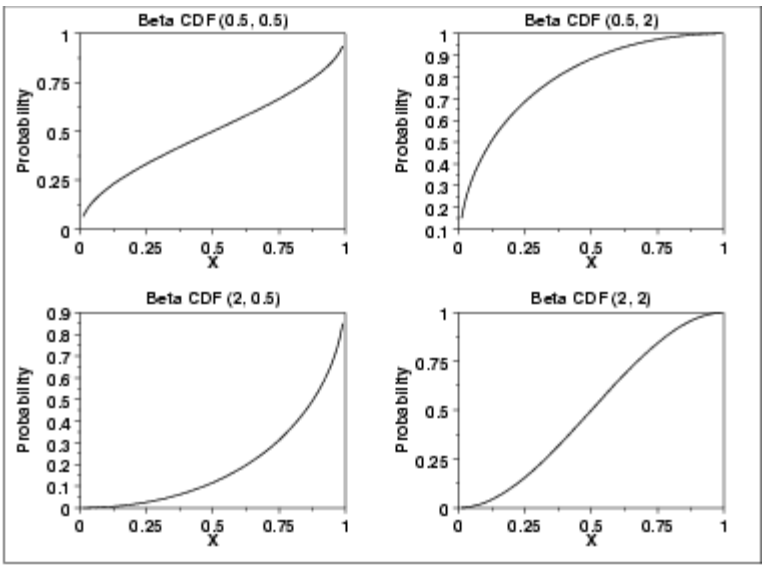
Cumulative Distribution Function

The formula for the cumulative distribution function of the beta distribution is also called the incomplete beta function ratio (commonly denoted by I_x) and is defined as

$$F(x) = I_x(p,q) = \frac{\int_0^x t^{p-1}(1-t)^{q-1}dt}{B(p,q)} \qquad 0 \le x \le 1; p,q > 0$$

where B is the beta function defined above.

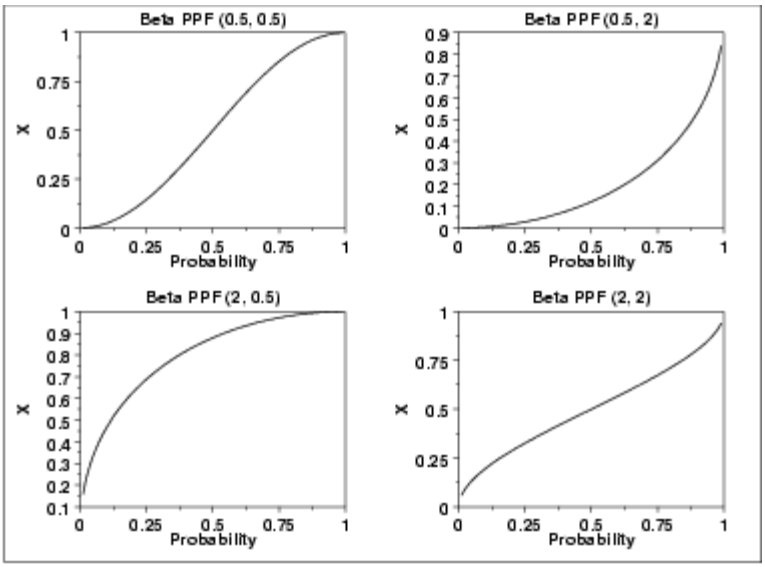
The following is the plot of the beta cumulative distribution function with the same values of the shape parameters as the pdf plots above.



Percent Point Function

The formula for the percent point function of the beta distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the beta percent point function with the same values of the shape parameters as the pdf plots above.



Other Probability Functions

Since the beta distribution is not typically used for reliability applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

The formulas below are for the case where the lower limit is zero and the upper limit is one.

Mean	$\frac{p}{p+q}$
Mode	$\frac{p-1}{p+q-2} \quad p,q>1$
Range	0 to 1
Standard Deviation	$\sqrt{\frac{pq}{(P+q)^2(p+q+1)}}$
Coefficient of Variation	$\sqrt{\frac{q}{p(p+q+1)}}$
Skewness	$\frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

Parameter Estimation

First consider the case where *a* and *b* are assumed to be known. For this case, the method of moments estimates are

$$p = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

$$q = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

where $\bar{x}$ is the sample mean and s^2 is the sample variance. If *a* and *b* are not 0 and 1, respectively, then replace $\bar{x}$ with $\frac{\bar{x}-a}{b-a}$ and s^2 with $\frac{s^2}{(b-a)^2}$ in the above equations.

For the case when *a* and *b* are known, the maximum likelihood estimates

can be obtained by solving the following set of equations

$$\psi(\hat{p}) - \psi(\hat{p}+\hat{q}) = \frac{1}{n}\sum_{i=1}^{n}\log(\frac{Y_i - a}{b-a})$$

$$\psi(\hat{q}) - \psi(\hat{p}+\hat{q}) = \frac{1}{n}\sum_{i=1}^{n}\log(\frac{b - Y_i}{b-a})$$

The maximum likelihood equations for the case when *a* and *b* are not known are given in pages 221-235 of Volume II of Johnson, Kotz, and Balakrishan.

Software Most general purpose statistical software programs support at least some of the probability functions for the beta distribution.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions
1.3.6.6. Gallery of Distributions

1.3.6.6.18. Binomial Distribution

Probability Mass Function

The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are appropriately labeled "success" and "failure". The binomial distribution is used to obtain the probability of observing *x* successes in *N* trials, with the probability of success on a single trial denoted by *p*. The binomial distribution assumes that *p* is fixed for all trials.

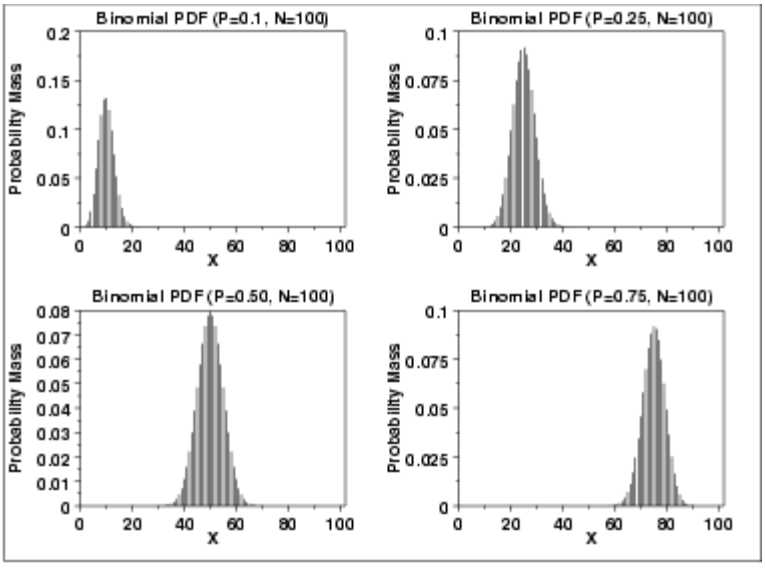
The formula for the binomial probability mass function is

$$P(x,p,n) = \left(\begin{array}{c} n \\ x \end{array} \right) (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

where

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}$$

The following is the plot of the binomial probability density function for four values of *p* and *n* = 100.

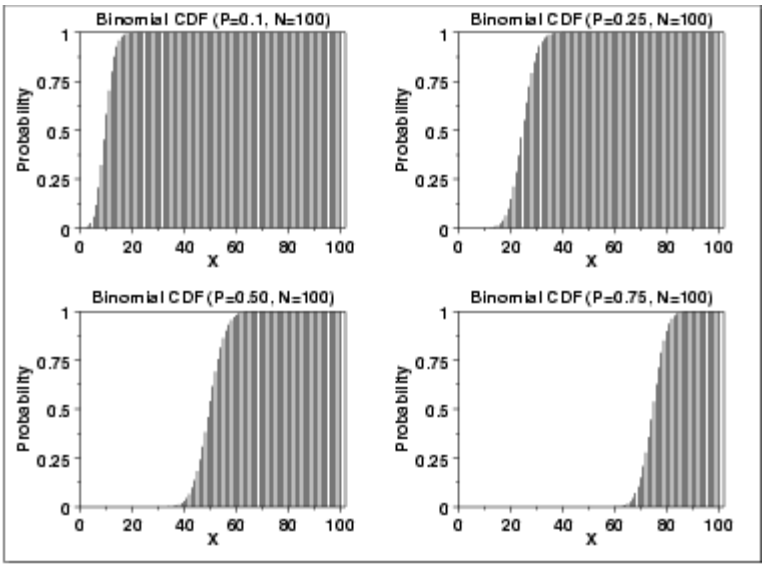


Cumulative Distribution Function

The formula for the binomial cumulative probability function is

$$F(x,p,n) = \sum_{i=0}^{x} \left(\begin{array}{c} n \\ i \end{array} \right) (p)^i (1-p)^{(n-i)}$$

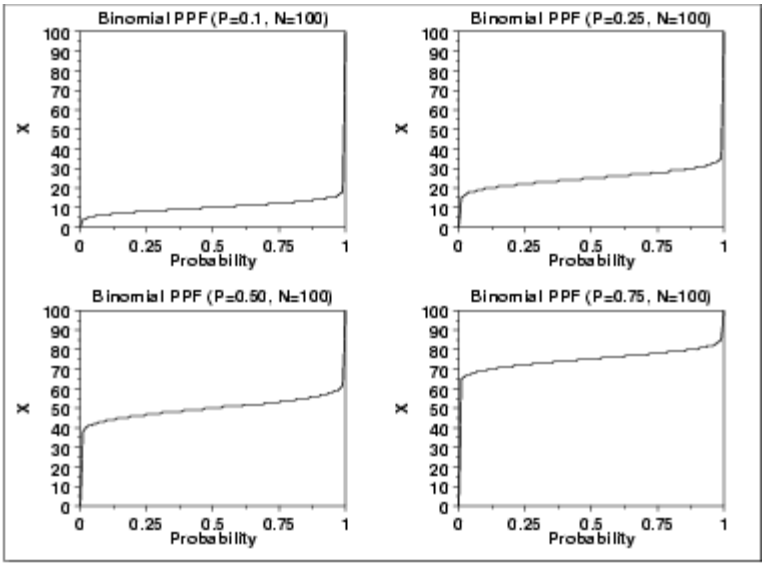
The following is the plot of the binomial cumulative distribution function with the same values of *p* as the pdf plots above.



Percent Point Function

The binomial percent point function does not exist in simple closed form. It is computed numerically. Note that because this is a discrete distribution that is only defined for integer values of *x*, the percent point function is not smooth in the way the percent point function typically is for a continuous distribution.

The following is the plot of the binomial percent point function with the same values of *p* as the pdf plots above.



Common Statistics

Mean	np
Mode	$p(n+1) - 1 \leq x \leq p(n+1)$
Range	0 to N
Standard Deviation	$\sqrt{np(1-p)}$

Coefficient of Variation	$\sqrt{\frac{(1-p)}{np}}$
Skewness	$\frac{(1-2p)}{\sqrt{np(1-p)}}$
Kurtosis	$3-\frac{6}{n}+\frac{1}{np(1-p)}$

Comments The binomial distribution is probably the most commonly used discrete distribution.

Parameter Estimation The maximum likelihood estimator of p (n is fixed) is

$$\tilde{p}=\frac{x}{n}$$

Software Most general purpose statistical software programs support at least some of the probability functions for the binomial distribution.

1.3.6.6.19. Poisson Distribution

Probability Mass Function

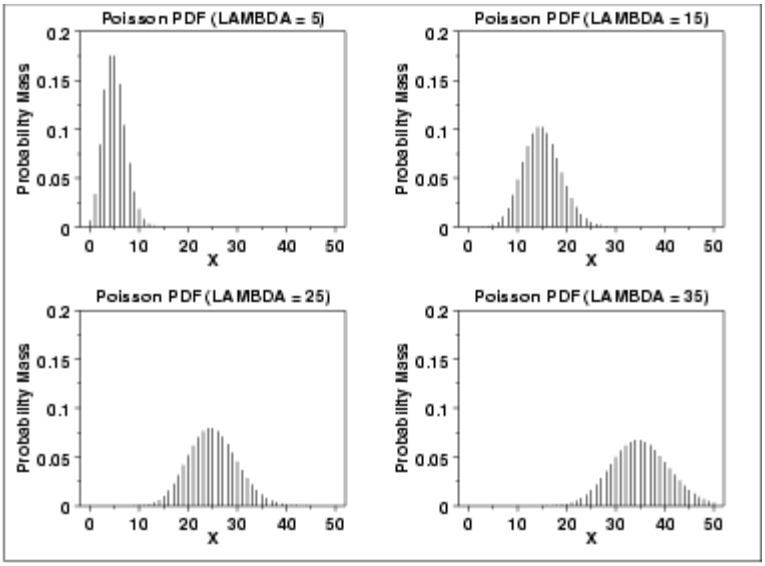
The Poisson distribution is used to model the number of events occurring within a given time interval.

The formula for the Poisson probability mass function is

$$p(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \cdots$$

λ is the shape parameter which indicates the average number of events in the given time interval.

The following is the plot of the Poisson probability density function for four values of λ.

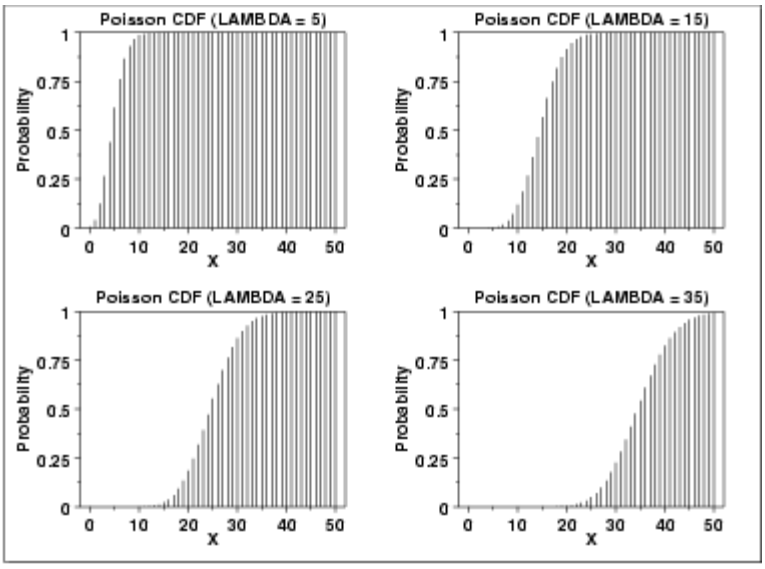


Cumulative Distribution Function

The formula for the Poisson cumulative probability function is

$$F(x, \lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

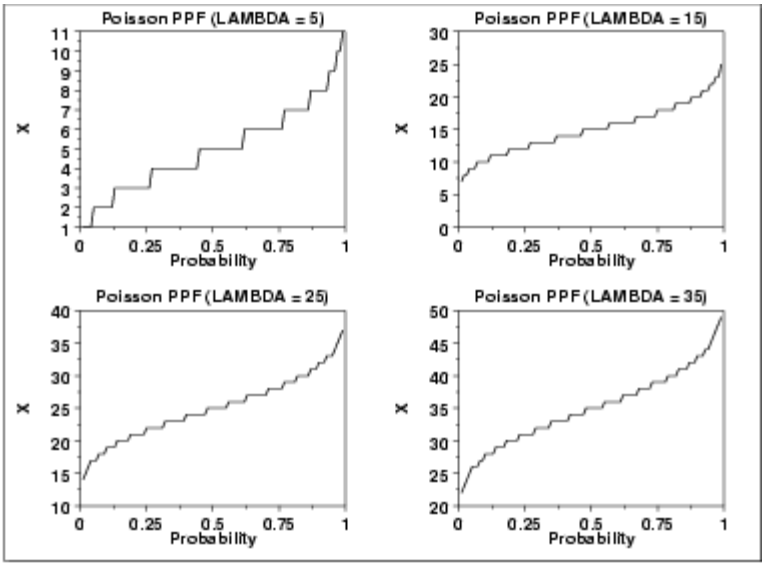
The following is the plot of the Poisson cumulative distribution function with the same values of λ as the pdf plots above.



Percent Point Function

The Poisson percent point function does not exist in simple closed form. It is computed numerically. Note that because this is a discrete distribution that is only defined for integer values of *x*, the percent point function is not smooth in the way the percent point function typically is for a continuous distribution.

The following is the plot of the Poisson percent point function with the same values of λ as the pdf plots above.



Common Statistics

Mean	λ
Mode	For non-integer λ, it is the largest integer less than λ. For integer λ, x = λ and x = λ - 1 are both the mode.
Range	0 to positive infinity
Standard Deviation	$\sqrt{\lambda}$
Coefficient of Variation	$\frac{1}{\sqrt{\lambda}}$

Skewness	$\frac{1}{\sqrt{\lambda}}$
Kurtosis	$3 + \frac{1}{\lambda}$

Parameter Estimation

The maximum likelihood estimator of λ is

$$\tilde{\lambda} = \bar{X}$$

where $\bar{X}$ is the sample mean.

Software

Most general purpose statistical software programs support at least some of the probability functions for the Poisson distribution.

1.3.6.7. Tables for Probability Distributions

Tables

Several commonly used tables for probability distributions can be referenced below.

The values from these tables can also be obtained from most general purpose statistical software programs. Most introductory statistics textbooks (e.g., Snedecor and Cochran) contain more extensive tables than are included here. These tables are included for convenience.

1. Cumulative distribution function for the standard normal distribution
2. Upper critical values of Student's t-distribution with ν degrees of freedom
3. Upper critical values of the F-distribution with ν_1 and ν_2 degrees of freedom
4. Upper critical values of the chi-square distribution with ν degrees of freedom
5. Critical values of t^* distribution for testing the output of a linear calibration line at 3 points
6. Upper critical values of the normal PPCC distribution

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions
1.3.6.7. Tables for Probability Distributions

1.3.6.7.1. Cumulative Distribution Function of the Standard Normal Distribution

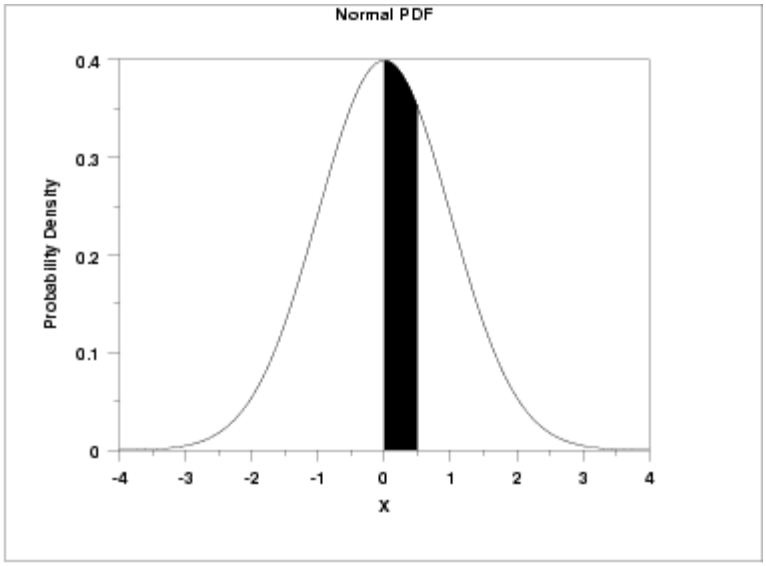
How to Use This Table

The table below contains the area under the standard normal curve from 0 to *z*. This can be used to compute the cumulative distribution function values for the standard normal distribution.

The table utilizes the symmetry of the normal distribution, so what in fact is given is

$$P[0 \le x \le |a|]$$

where *a* is the value of interest. This is demonstrated in the graph below for *a* = 0.5. The shaded area of the curve represents the probability that *x* is between 0 and *a*.



This can be clarified by a few simple examples.

1. What is the probability that *x* is less than or equal to 1.53? Look for 1.5 in the X column, go right to the 0.03 column to find the value 0.43699. Now add 0.5 (for the probability less than zero) to obtain the final result of 0.93699.

2. What is the probability that *x* is less than or equal to -

1.53? For negative values, use the relationship

$$P[x \leq a] = 1 - P[x \leq |a|] \quad \text{for } x < 0$$

From the first example, this gives 1 - 0.93699 = 0.06301.

3. What is the probability that *x* is between -1 and 0.5? Look up the values for 0.5 (0.5 + 0.19146 = 0.69146) and -1 (1 - (0.5 + 0.34134) = 0.15866). Then subtract the results (0.69146 - 0.15866) to obtain the result 0.5328.

To use this table with a non-standard normal distribution (either the location parameter is not 0 or the scale parameter is not 1), standardize your value by subtracting the mean and dividing the result by the standard deviation. Then look up the value for this standardized value.

A few particularly important numbers derived from the table below, specifically numbers that are commonly used in significance tests, are summarized in the following table:

p	0.001	0.005	0.010	0.025	0.050	0.100
Z_p	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282

p	0.999	0.995	0.990	0.975	0.950	0.900
Z_p	+3.090	+2.576	+2.326	+1.960	+1.645	+1.282

These are critical values for the normal distribution.

Area under the Normal Curve from 0 to X

X	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891

1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543
0.35769	0.35993	0.36214					
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698
0.37900	0.38100	0.38298					
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617
0.39796	0.39973	0.40147					
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308
0.41466	0.41621	0.41774					
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785
0.42922	0.43056	0.43189					
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062
0.44179	0.44295	0.44408					
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154
0.45254	0.45352	0.45449					
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080
0.46164	0.46246	0.46327					
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856
0.46926	0.46995	0.47062					
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500
0.47558	0.47615	0.47670					
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030
0.48077	0.48124	0.48169					
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461
0.48500	0.48537	0.48574					
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809
0.48840	0.48870	0.48899					
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086
0.49111	0.49134	0.49158					
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305
0.49324	0.49343	0.49361					
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477
0.49492	0.49506	0.49520					
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609
0.49621	0.49632	0.49643					
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711
0.49720	0.49728	0.49736					
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788
0.49795	0.49801	0.49807					
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846
0.49851	0.49856	0.49861					
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889
0.49893	0.49896	0.49900					
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921
0.49924	0.49926	0.49929					
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944
0.49946	0.49948	0.49950					
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961
0.49962	0.49964	0.49965					
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973
0.49974	0.49975	0.49976					
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981
0.49982	0.49983	0.49983					
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987
0.49988	0.49988	0.49989					
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992
0.49992	0.49992	0.49992					
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994
0.49995	0.49995	0.49995					
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996
0.49996	0.49997	0.49997					
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998
0.49998	0.49998	0.49998					

1.3.6.7.2. Critical Values of the Student's *t* Distribution

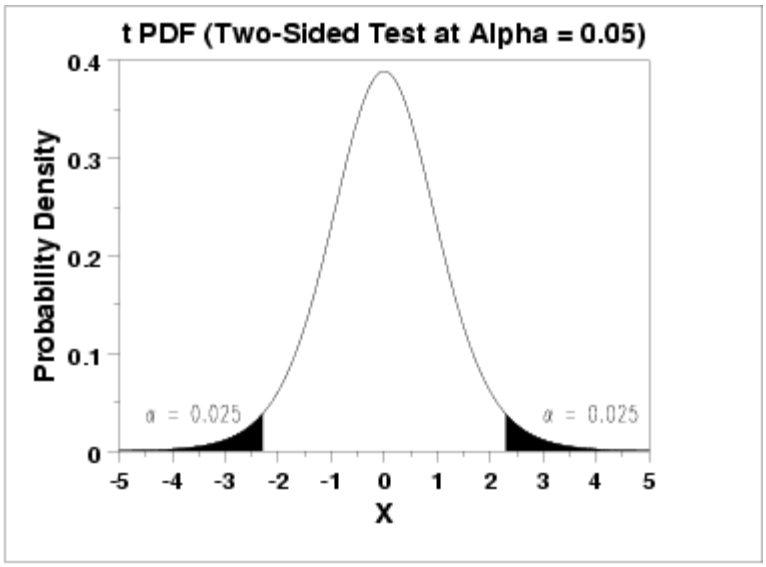
How to Use This Table

This table contains critical values of the Student's *t* distribution computed using the cumulative distribution function. The *t* distribution is symmetric so that

$$t_{1-\alpha,\nu} = -t_{\alpha,\nu}.$$

The *t* table can be used for both one-sided (lower and upper) and two-sided tests using the appropriate value of α.

The significance level, α, is demonstrated in the graph below, which displays a *t* distribution with 10 degrees of freedom. The most commonly used significance level is $\alpha = 0.05$. For a two-sided test, we compute $1 - \alpha/2$, or $1 - 0.05/2 = 0.975$ when $\alpha = 0.05$. If the absolute value of the test statistic is greater than the critical value (0.975), then we reject the null hypothesis. Due to the symmetry of the *t* distribution, we only tabulate the positive critical values in the table below.



Given a specified value for α :

1. For a two-sided test, find the column corresponding to $1-\alpha/2$ and reject the null hypothesis if the absolute value

of the test statistic is greater than the value of $t_{1-\alpha/2,\nu}$ in the table below.
2. For an upper, one-sided test, find the column corresponding to 1-α and reject the null hypothesis if the test statistic is greater than the table value.
3. For a lower, one-sided test, find the column corresponding to 1-α and reject the null hypothesis if the test statistic is less than the negative of the table value.

Critical values of Student's *t* distribution with *v* degrees of freedom

Probability less than the critical value ($t_{1-\alpha,\nu}$)

ν	0.90	0.95	0.975	0.99	0.995	0.999
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	

3.552					
21.	1.323	1.721	2.080	2.518	2.831
3.527					
22.	1.321	1.717	2.074	2.508	2.819
3.505					
23.	1.319	1.714	2.069	2.500	2.807
3.485					
24.	1.318	1.711	2.064	2.492	2.797
3.467					
25.	1.316	1.708	2.060	2.485	2.787
3.450					
26.	1.315	1.706	2.056	2.479	2.779
3.435					
27.	1.314	1.703	2.052	2.473	2.771
3.421					
28.	1.313	1.701	2.048	2.467	2.763
3.408					
29.	1.311	1.699	2.045	2.462	2.756
3.396					
30.	1.310	1.697	2.042	2.457	2.750
3.385					
31.	1.309	1.696	2.040	2.453	2.744
3.375					
32.	1.309	1.694	2.037	2.449	2.738
3.365					
33.	1.308	1.692	2.035	2.445	2.733
3.356					
34.	1.307	1.691	2.032	2.441	2.728
3.348					
35.	1.306	1.690	2.030	2.438	2.724
3.340					
36.	1.306	1.688	2.028	2.434	2.719
3.333					
37.	1.305	1.687	2.026	2.431	2.715
3.326					
38.	1.304	1.686	2.024	2.429	2.712
3.319					
39.	1.304	1.685	2.023	2.426	2.708
3.313					
40.	1.303	1.684	2.021	2.423	2.704
3.307					
41.	1.303	1.683	2.020	2.421	2.701
3.301					
42.	1.302	1.682	2.018	2.418	2.698
3.296					
43.	1.302	1.681	2.017	2.416	2.695
3.291					
44.	1.301	1.680	2.015	2.414	2.692
3.286					
45.	1.301	1.679	2.014	2.412	2.690
3.281					
46.	1.300	1.679	2.013	2.410	2.687
3.277					
47.	1.300	1.678	2.012	2.408	2.685
3.273					
48.	1.299	1.677	2.011	2.407	2.682
3.269					
49.	1.299	1.677	2.010	2.405	2.680
3.265					
50.	1.299	1.676	2.009	2.403	2.678
3.261					
51.	1.298	1.675	2.008	2.402	2.676
3.258					
52.	1.298	1.675	2.007	2.400	2.674

3.255					
53.	1.298	1.674	2.006	2.399	2.672
3.251					
54.	1.297	1.674	2.005	2.397	2.670
3.248					
55.	1.297	1.673	2.004	2.396	2.668
3.245					
56.	1.297	1.673	2.003	2.395	2.667
3.242					
57.	1.297	1.672	2.002	2.394	2.665
3.239					
58.	1.296	1.672	2.002	2.392	2.663
3.237					
59.	1.296	1.671	2.001	2.391	2.662
3.234					
60.	1.296	1.671	2.000	2.390	2.660
3.232					
61.	1.296	1.670	2.000	2.389	2.659
3.229					
62.	1.295	1.670	1.999	2.388	2.657
3.227					
63.	1.295	1.669	1.998	2.387	2.656
3.225					
64.	1.295	1.669	1.998	2.386	2.655
3.223					
65.	1.295	1.669	1.997	2.385	2.654
3.220					
66.	1.295	1.668	1.997	2.384	2.652
3.218					
67.	1.294	1.668	1.996	2.383	2.651
3.216					
68.	1.294	1.668	1.995	2.382	2.650
3.214					
69.	1.294	1.667	1.995	2.382	2.649
3.213					
70.	1.294	1.667	1.994	2.381	2.648
3.211					
71.	1.294	1.667	1.994	2.380	2.647
3.209					
72.	1.293	1.666	1.993	2.379	2.646
3.207					
73.	1.293	1.666	1.993	2.379	2.645
3.206					
74.	1.293	1.666	1.993	2.378	2.644
3.204					
75.	1.293	1.665	1.992	2.377	2.643
3.202					
76.	1.293	1.665	1.992	2.376	2.642
3.201					
77.	1.293	1.665	1.991	2.376	2.641
3.199					
78.	1.292	1.665	1.991	2.375	2.640
3.198					
79.	1.292	1.664	1.990	2.374	2.640
3.197					
80.	1.292	1.664	1.990	2.374	2.639
3.195					
81.	1.292	1.664	1.990	2.373	2.638
3.194					
82.	1.292	1.664	1.989	2.373	2.637
3.193					
83.	1.292	1.663	1.989	2.372	2.636
3.191					
84.	1.292	1.663	1.989	2.372	2.636

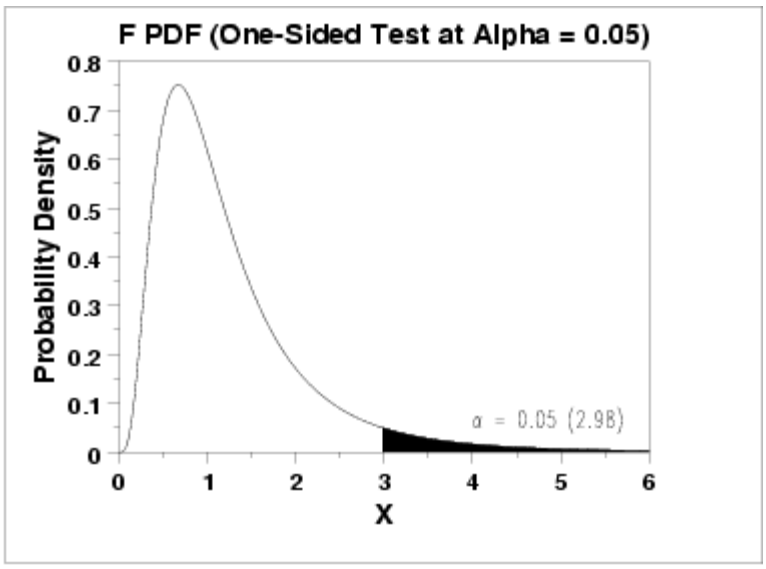
3.190					
85.	1.292	1.663	1.988	2.371	2.635
3.189					
86.	1.291	1.663	1.988	2.370	2.634
3.188					
87.	1.291	1.663	1.988	2.370	2.634
3.187					
88.	1.291	1.662	1.987	2.369	2.633
3.185					
89.	1.291	1.662	1.987	2.369	2.632
3.184					
90.	1.291	1.662	1.987	2.368	2.632
3.183					
91.	1.291	1.662	1.986	2.368	2.631
3.182					
92.	1.291	1.662	1.986	2.368	2.630
3.181					
93.	1.291	1.661	1.986	2.367	2.630
3.180					
94.	1.291	1.661	1.986	2.367	2.629
3.179					
95.	1.291	1.661	1.985	2.366	2.629
3.178					
96.	1.290	1.661	1.985	2.366	2.628
3.177					
97.	1.290	1.661	1.985	2.365	2.627
3.176					
98.	1.290	1.661	1.984	2.365	2.627
3.175					
99.	1.290	1.660	1.984	2.365	2.626
3.175					
100.	1.290	1.660	1.984	2.364	2.626
3.174					
∞	1.282	1.645	1.960	2.326	2.576
3.090					

1.3.6.7.3. Upper Critical Values of the F Distribution

How to Use This Table

This table contains the upper critical values of the F distribution. This table is used for one-sided F tests at the α = 0.05, 0.10, and 0.01 levels.

More specifically, a test statistic is computed with ν_1 and ν_2 degrees of freedom, and the result is compared to this table. For a one-sided test, the null hypothesis is rejected when the test statistic is greater than the tabled value. This is demonstrated with the graph of an F distribution with ν_1 = 10 and ν_2 = 10. The shaded area of the graph indicates the rejection region at the α significance level. Since this is a one-sided test, we have α probability in the upper tail of exceeding the critical value and zero in the lower tail. Because the F distribution is asymmetric, a two-sided test requires a set of of tables (not included here) that contain the rejection regions for both the lower and upper tails.



Contents

The following tables for ν_2 from 1 to 100 are included:

1. One sided, 5% significance level, ν_1 = 1 - 10
2. One sided, 5% significance level, ν_1 = 11 - 20
3. One sided, 10% significance level, ν_1 = 1 - 10
4. One sided, 10% significance level, ν_1 = 11 - 20

5. One sided, 1% significance level, ν_1 = 1 - 10
6. One sided, 1% significance level, ν_1 = 11 - 20

Upper critical values of the F distribution for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

5% significance level

$$F_{.05}(\nu_1, \nu_2)$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.882	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321

22	4.301	3.443	3.049	2.817	2.661
2.549	2.464	2.397	2.342	2.297	
23	4.279	3.422	3.028	2.796	2.640
2.528	2.442	2.375	2.320	2.275	
24	4.260	3.403	3.009	2.776	2.621
2.508	2.423	2.355	2.300	2.255	
25	4.242	3.385	2.991	2.759	2.603
2.490	2.405	2.337	2.282	2.236	
26	4.225	3.369	2.975	2.743	2.587
2.474	2.388	2.321	2.265	2.220	
27	4.210	3.354	2.960	2.728	2.572
2.459	2.373	2.305	2.250	2.204	
28	4.196	3.340	2.947	2.714	2.558
2.445	2.359	2.291	2.236	2.190	
29	4.183	3.328	2.934	2.701	2.545
2.432	2.346	2.278	2.223	2.177	
30	4.171	3.316	2.922	2.690	2.534
2.421	2.334	2.266	2.211	2.165	
31	4.160	3.305	2.911	2.679	2.523
2.409	2.323	2.255	2.199	2.153	
32	4.149	3.295	2.901	2.668	2.512
2.399	2.313	2.244	2.189	2.142	
33	4.139	3.285	2.892	2.659	2.503
2.389	2.303	2.235	2.179	2.133	
34	4.130	3.276	2.883	2.650	2.494
2.380	2.294	2.225	2.170	2.123	
35	4.121	3.267	2.874	2.641	2.485
2.372	2.285	2.217	2.161	2.114	
36	4.113	3.259	2.866	2.634	2.477
2.364	2.277	2.209	2.153	2.106	
37	4.105	3.252	2.859	2.626	2.470
2.356	2.270	2.201	2.145	2.098	
38	4.098	3.245	2.852	2.619	2.463
2.349	2.262	2.194	2.138	2.091	
39	4.091	3.238	2.845	2.612	2.456
2.342	2.255	2.187	2.131	2.084	
40	4.085	3.232	2.839	2.606	2.449
2.336	2.249	2.180	2.124	2.077	
41	4.079	3.226	2.833	2.600	2.443
2.330	2.243	2.174	2.118	2.071	
42	4.073	3.220	2.827	2.594	2.438
2.324	2.237	2.168	2.112	2.065	
43	4.067	3.214	2.822	2.589	2.432
2.318	2.232	2.163	2.106	2.059	
44	4.062	3.209	2.816	2.584	2.427
2.313	2.226	2.157	2.101	2.054	
45	4.057	3.204	2.812	2.579	2.422
2.308	2.221	2.152	2.096	2.049	
46	4.052	3.200	2.807	2.574	2.417
2.304	2.216	2.147	2.091	2.044	
47	4.047	3.195	2.802	2.570	2.413
2.299	2.212	2.143	2.086	2.039	
48	4.043	3.191	2.798	2.565	2.409
2.295	2.207	2.138	2.082	2.035	
49	4.038	3.187	2.794	2.561	2.404
2.290	2.203	2.134	2.077	2.030	
50	4.034	3.183	2.790	2.557	2.400
2.286	2.199	2.130	2.073	2.026	
51	4.030	3.179	2.786	2.553	2.397
2.283	2.195	2.126	2.069	2.022	
52	4.027	3.175	2.783	2.550	2.393
2.279	2.192	2.122	2.066	2.018	
53	4.023	3.172	2.779	2.546	2.389
2.275	2.188	2.119	2.062	2.015	

54	4.020	3.168	2.776	2.543	2.386
2.272	2.185	2.115	2.059	2.011	
55	4.016	3.165	2.773	2.540	2.383
2.269	2.181	2.112	2.055	2.008	
56	4.013	3.162	2.769	2.537	2.380
2.266	2.178	2.109	2.052	2.005	
57	4.010	3.159	2.766	2.534	2.377
2.263	2.175	2.106	2.049	2.001	
58	4.007	3.156	2.764	2.531	2.374
2.260	2.172	2.103	2.046	1.998	
59	4.004	3.153	2.761	2.528	2.371
2.257	2.169	2.100	2.043	1.995	
60	4.001	3.150	2.758	2.525	2.368
2.254	2.167	2.097	2.040	1.993	
61	3.998	3.148	2.755	2.523	2.366
2.251	2.164	2.094	2.037	1.990	
62	3.996	3.145	2.753	2.520	2.363
2.249	2.161	2.092	2.035	1.987	
63	3.993	3.143	2.751	2.518	2.361
2.246	2.159	2.089	2.032	1.985	
64	3.991	3.140	2.748	2.515	2.358
2.244	2.156	2.087	2.030	1.982	
65	3.989	3.138	2.746	2.513	2.356
2.242	2.154	2.084	2.027	1.980	
66	3.986	3.136	2.744	2.511	2.354
2.239	2.152	2.082	2.025	1.977	
67	3.984	3.134	2.742	2.509	2.352
2.237	2.150	2.080	2.023	1.975	
68	3.982	3.132	2.740	2.507	2.350
2.235	2.148	2.078	2.021	1.973	
69	3.980	3.130	2.737	2.505	2.348
2.233	2.145	2.076	2.019	1.971	
70	3.978	3.128	2.736	2.503	2.346
2.231	2.143	2.074	2.017	1.969	
71	3.976	3.126	2.734	2.501	2.344
2.229	2.142	2.072	2.015	1.967	
72	3.974	3.124	2.732	2.499	2.342
2.227	2.140	2.070	2.013	1.965	
73	3.972	3.122	2.730	2.497	2.340
2.226	2.138	2.068	2.011	1.963	
74	3.970	3.120	2.728	2.495	2.338
2.224	2.136	2.066	2.009	1.961	
75	3.968	3.119	2.727	2.494	2.337
2.222	2.134	2.064	2.007	1.959	
76	3.967	3.117	2.725	2.492	2.335
2.220	2.133	2.063	2.006	1.958	
77	3.965	3.115	2.723	2.490	2.333
2.219	2.131	2.061	2.004	1.956	
78	3.963	3.114	2.722	2.489	2.332
2.217	2.129	2.059	2.002	1.954	
79	3.962	3.112	2.720	2.487	2.330
2.216	2.128	2.058	2.001	1.953	
80	3.960	3.111	2.719	2.486	2.329
2.214	2.126	2.056	1.999	1.951	
81	3.959	3.109	2.717	2.484	2.327
2.213	2.125	2.055	1.998	1.950	
82	3.957	3.108	2.716	2.483	2.326
2.211	2.123	2.053	1.996	1.948	
83	3.956	3.107	2.715	2.482	2.324
2.210	2.122	2.052	1.995	1.947	
84	3.955	3.105	2.713	2.480	2.323
2.209	2.121	2.051	1.993	1.945	
85	3.953	3.104	2.712	2.479	2.322
2.207	2.119	2.049	1.992	1.944	

```
 86        3.952   3.103   2.711   2.478   2.321
2.206   2.118   2.048   1.991   1.943
 87        3.951   3.101   2.709   2.476   2.319
2.205   2.117   2.047   1.989   1.941
 88        3.949   3.100   2.708   2.475   2.318
2.203   2.115   2.045   1.988   1.940
 89        3.948   3.099   2.707   2.474   2.317
2.202   2.114   2.044   1.987   1.939
 90        3.947   3.098   2.706   2.473   2.316
2.201   2.113   2.043   1.986   1.938
 91        3.946   3.097   2.705   2.472   2.315
2.200   2.112   2.042   1.984   1.936
 92        3.945   3.095   2.704   2.471   2.313
2.199   2.111   2.041   1.983   1.935
 93        3.943   3.094   2.703   2.470   2.312
2.198   2.110   2.040   1.982   1.934
 94        3.942   3.093   2.701   2.469   2.311
2.197   2.109   2.038   1.981   1.933
 95        3.941   3.092   2.700   2.467   2.310
2.196   2.108   2.037   1.980   1.932
 96        3.940   3.091   2.699   2.466   2.309
2.195   2.106   2.036   1.979   1.931
 97        3.939   3.090   2.698   2.465   2.308
2.194   2.105   2.035   1.978   1.930
 98        3.938   3.089   2.697   2.465   2.307
2.193   2.104   2.034   1.977   1.929
 99        3.937   3.088   2.696   2.464   2.306
2.192   2.103   2.033   1.976   1.928
100        3.936   3.087   2.696   2.463   2.305
2.191   2.103   2.032   1.975   1.927
```

ν_2 \ ν_1 11 12 13 14 15 16 17 18 19 20

```
  1      242.983 243.906 244.690 245.364 245.950
246.464 246.918 247.323 247.686 248.013
  2       19.405  19.413  19.419  19.424  19.429
19.433  19.437  19.440  19.443  19.446
  3        8.763   8.745   8.729   8.715   8.703
8.692   8.683   8.675   8.667   8.660
  4        5.936   5.912   5.891   5.873   5.858
5.844   5.832   5.821   5.811   5.803
  5        4.704   4.678   4.655   4.636   4.619
4.604   4.590   4.579   4.568   4.558
  6        4.027   4.000   3.976   3.956   3.938
3.922   3.908   3.896   3.884   3.874
  7        3.603   3.575   3.550   3.529   3.511
3.494   3.480   3.467   3.455   3.445
  8        3.313   3.284   3.259   3.237   3.218
3.202   3.187   3.173   3.161   3.150
  9        3.102   3.073   3.048   3.025   3.006
2.989   2.974   2.960   2.948   2.936
 10        2.943   2.913   2.887   2.865   2.845
2.828   2.812   2.798   2.785   2.774
 11        2.818   2.788   2.761   2.739   2.719
2.701   2.685   2.671   2.658   2.646
 12        2.717   2.687   2.660   2.637   2.617
2.599   2.583   2.568   2.555   2.544
 13        2.635   2.604   2.577   2.554   2.533
2.515   2.499   2.484   2.471   2.459
```

14	2.565	2.534	2.507	2.484	2.463
2.445	2.428	2.413	2.400	2.388	
15	2.507	2.475	2.448	2.424	2.403
2.385	2.368	2.353	2.340	2.328	
16	2.456	2.425	2.397	2.373	2.352
2.333	2.317	2.302	2.288	2.276	
17	2.413	2.381	2.353	2.329	2.308
2.289	2.272	2.257	2.243	2.230	
18	2.374	2.342	2.314	2.290	2.269
2.250	2.233	2.217	2.203	2.191	
19	2.340	2.308	2.280	2.256	2.234
2.215	2.198	2.182	2.168	2.155	
20	2.310	2.278	2.250	2.225	2.203
2.184	2.167	2.151	2.137	2.124	
21	2.283	2.250	2.222	2.197	2.176
2.156	2.139	2.123	2.109	2.096	
22	2.259	2.226	2.198	2.173	2.151
2.131	2.114	2.098	2.084	2.071	
23	2.236	2.204	2.175	2.150	2.128
2.109	2.091	2.075	2.061	2.048	
24	2.216	2.183	2.155	2.130	2.108
2.088	2.070	2.054	2.040	2.027	
25	2.198	2.165	2.136	2.111	2.089
2.069	2.051	2.035	2.021	2.007	
26	2.181	2.148	2.119	2.094	2.072
2.052	2.034	2.018	2.003	1.990	
27	2.166	2.132	2.103	2.078	2.056
2.036	2.018	2.002	1.987	1.974	
28	2.151	2.118	2.089	2.064	2.041
2.021	2.003	1.987	1.972	1.959	
29	2.138	2.104	2.075	2.050	2.027
2.007	1.989	1.973	1.958	1.945	
30	2.126	2.092	2.063	2.037	2.015
1.995	1.976	1.960	1.945	1.932	
31	2.114	2.080	2.051	2.026	2.003
1.983	1.965	1.948	1.933	1.920	
32	2.103	2.070	2.040	2.015	1.992
1.972	1.953	1.937	1.922	1.908	
33	2.093	2.060	2.030	2.004	1.982
1.961	1.943	1.926	1.911	1.898	
34	2.084	2.050	2.021	1.995	1.972
1.952	1.933	1.917	1.902	1.888	
35	2.075	2.041	2.012	1.986	1.963
1.942	1.924	1.907	1.892	1.878	
36	2.067	2.033	2.003	1.977	1.954
1.934	1.915	1.899	1.883	1.870	
37	2.059	2.025	1.995	1.969	1.946
1.926	1.907	1.890	1.875	1.861	
38	2.051	2.017	1.988	1.962	1.939
1.918	1.899	1.883	1.867	1.853	
39	2.044	2.010	1.981	1.954	1.931
1.911	1.892	1.875	1.860	1.846	
40	2.038	2.003	1.974	1.948	1.924
1.904	1.885	1.868	1.853	1.839	
41	2.031	1.997	1.967	1.941	1.918
1.897	1.879	1.862	1.846	1.832	
42	2.025	1.991	1.961	1.935	1.912
1.891	1.872	1.855	1.840	1.826	
43	2.020	1.985	1.955	1.929	1.906
1.885	1.866	1.849	1.834	1.820	
44	2.014	1.980	1.950	1.924	1.900
1.879	1.861	1.844	1.828	1.814	
45	2.009	1.974	1.945	1.918	1.895
1.874	1.855	1.838	1.823	1.808	

46	2.004	1.969	1.940	1.913	1.890
1.869	1.850	1.833	1.817	1.803	
47	1.999	1.965	1.935	1.908	1.885
1.864	1.845	1.828	1.812	1.798	
48	1.995	1.960	1.930	1.904	1.880
1.859	1.840	1.823	1.807	1.793	
49	1.990	1.956	1.926	1.899	1.876
1.855	1.836	1.819	1.803	1.789	
50	1.986	1.952	1.921	1.895	1.871
1.850	1.831	1.814	1.798	1.784	
51	1.982	1.947	1.917	1.891	1.867
1.846	1.827	1.810	1.794	1.780	
52	1.978	1.944	1.913	1.887	1.863
1.842	1.823	1.806	1.790	1.776	
53	1.975	1.940	1.910	1.883	1.859
1.838	1.819	1.802	1.786	1.772	
54	1.971	1.936	1.906	1.879	1.856
1.835	1.816	1.798	1.782	1.768	
55	1.968	1.933	1.903	1.876	1.852
1.831	1.812	1.795	1.779	1.764	
56	1.964	1.930	1.899	1.873	1.849
1.828	1.809	1.791	1.775	1.761	
57	1.961	1.926	1.896	1.869	1.846
1.824	1.805	1.788	1.772	1.757	
58	1.958	1.923	1.893	1.866	1.842
1.821	1.802	1.785	1.769	1.754	
59	1.955	1.920	1.890	1.863	1.839
1.818	1.799	1.781	1.766	1.751	
60	1.952	1.917	1.887	1.860	1.836
1.815	1.796	1.778	1.763	1.748	
61	1.949	1.915	1.884	1.857	1.834
1.812	1.793	1.776	1.760	1.745	
62	1.947	1.912	1.882	1.855	1.831
1.809	1.790	1.773	1.757	1.742	
63	1.944	1.909	1.879	1.852	1.828
1.807	1.787	1.770	1.754	1.739	
64	1.942	1.907	1.876	1.849	1.826
1.804	1.785	1.767	1.751	1.737	
65	1.939	1.904	1.874	1.847	1.823
1.802	1.782	1.765	1.749	1.734	
66	1.937	1.902	1.871	1.845	1.821
1.799	1.780	1.762	1.746	1.732	
67	1.935	1.900	1.869	1.842	1.818
1.797	1.777	1.760	1.744	1.729	
68	1.932	1.897	1.867	1.840	1.816
1.795	1.775	1.758	1.742	1.727	
69	1.930	1.895	1.865	1.838	1.814
1.792	1.773	1.755	1.739	1.725	
70	1.928	1.893	1.863	1.836	1.812
1.790	1.771	1.753	1.737	1.722	
71	1.926	1.891	1.861	1.834	1.810
1.788	1.769	1.751	1.735	1.720	
72	1.924	1.889	1.859	1.832	1.808
1.786	1.767	1.749	1.733	1.718	
73	1.922	1.887	1.857	1.830	1.806
1.784	1.765	1.747	1.731	1.716	
74	1.921	1.885	1.855	1.828	1.804
1.782	1.763	1.745	1.729	1.714	
75	1.919	1.884	1.853	1.826	1.802
1.780	1.761	1.743	1.727	1.712	
76	1.917	1.882	1.851	1.824	1.800
1.778	1.759	1.741	1.725	1.710	
77	1.915	1.880	1.849	1.822	1.798
1.777	1.757	1.739	1.723	1.708	

```
 78        1.914   1.878   1.848   1.821   1.797
1.775   1.755   1.738   1.721   1.707
 79        1.912   1.877   1.846   1.819   1.795
1.773   1.754   1.736   1.720   1.705
 80        1.910   1.875   1.845   1.817   1.793
1.772   1.752   1.734   1.718   1.703
 81        1.909   1.874   1.843   1.816   1.792
1.770   1.750   1.733   1.716   1.702
 82        1.907   1.872   1.841   1.814   1.790
1.768   1.749   1.731   1.715   1.700
 83        1.906   1.871   1.840   1.813   1.789
1.767   1.747   1.729   1.713   1.698
 84        1.905   1.869   1.838   1.811   1.787
1.765   1.746   1.728   1.712   1.697
 85        1.903   1.868   1.837   1.810   1.786
1.764   1.744   1.726   1.710   1.695
 86        1.902   1.867   1.836   1.808   1.784
1.762   1.743   1.725   1.709   1.694
 87        1.900   1.865   1.834   1.807   1.783
1.761   1.741   1.724   1.707   1.692
 88        1.899   1.864   1.833   1.806   1.782
1.760   1.740   1.722   1.706   1.691
 89        1.898   1.863   1.832   1.804   1.780
1.758   1.739   1.721   1.705   1.690
 90        1.897   1.861   1.830   1.803   1.779
1.757   1.737   1.720   1.703   1.688
 91        1.895   1.860   1.829   1.802   1.778
1.756   1.736   1.718   1.702   1.687
 92        1.894   1.859   1.828   1.801   1.776
1.755   1.735   1.717   1.701   1.686
 93        1.893   1.858   1.827   1.800   1.775
1.753   1.734   1.716   1.699   1.684
 94        1.892   1.857   1.826   1.798   1.774
1.752   1.733   1.715   1.698   1.683
 95        1.891   1.856   1.825   1.797   1.773
1.751   1.731   1.713   1.697   1.682
 96        1.890   1.854   1.823   1.796   1.772
1.750   1.730   1.712   1.696   1.681
 97        1.889   1.853   1.822   1.795   1.771
1.749   1.729   1.711   1.695   1.680
 98        1.888   1.852   1.821   1.794   1.770
1.748   1.728   1.710   1.694   1.679
 99        1.887   1.851   1.820   1.793   1.769
1.747   1.727   1.709   1.693   1.678
100        1.886   1.850   1.819   1.792   1.768
1.746   1.726   1.708   1.691   1.676
```

Upper critical values of the F distribution for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

10% significance level

$$F_{.10}(\nu_1, \nu_2)$$

```
     \ ν1    1       2       3       4       5
6       7       8       9       10
```

ν_2

1	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.858	60.195
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
31	2.875	2.482	2.270	2.136	2.042	1.973	1.920	1.877	1.842	1.812

32	2.869	2.477	2.263	2.129	2.036
1.967	1.913	1.870	1.835	1.805	
33	2.864	2.471	2.258	2.123	2.030
1.961	1.907	1.864	1.828	1.799	
34	2.859	2.466	2.252	2.118	2.024
1.955	1.901	1.858	1.822	1.793	
35	2.855	2.461	2.247	2.113	2.019
1.950	1.896	1.852	1.817	1.787	
36	2.850	2.456	2.243	2.108	2.014
1.945	1.891	1.847	1.811	1.781	
37	2.846	2.452	2.238	2.103	2.009
1.940	1.886	1.842	1.806	1.776	
38	2.842	2.448	2.234	2.099	2.005
1.935	1.881	1.838	1.802	1.772	
39	2.839	2.444	2.230	2.095	2.001
1.931	1.877	1.833	1.797	1.767	
40	2.835	2.440	2.226	2.091	1.997
1.927	1.873	1.829	1.793	1.763	
41	2.832	2.437	2.222	2.087	1.993
1.923	1.869	1.825	1.789	1.759	
42	2.829	2.434	2.219	2.084	1.989
1.919	1.865	1.821	1.785	1.755	
43	2.826	2.430	2.216	2.080	1.986
1.916	1.861	1.817	1.781	1.751	
44	2.823	2.427	2.213	2.077	1.983
1.913	1.858	1.814	1.778	1.747	
45	2.820	2.425	2.210	2.074	1.980
1.909	1.855	1.811	1.774	1.744	
46	2.818	2.422	2.207	2.071	1.977
1.906	1.852	1.808	1.771	1.741	
47	2.815	2.419	2.204	2.068	1.974
1.903	1.849	1.805	1.768	1.738	
48	2.813	2.417	2.202	2.066	1.971
1.901	1.846	1.802	1.765	1.735	
49	2.811	2.414	2.199	2.063	1.968
1.898	1.843	1.799	1.763	1.732	
50	2.809	2.412	2.197	2.061	1.966
1.895	1.840	1.796	1.760	1.729	
51	2.807	2.410	2.194	2.058	1.964
1.893	1.838	1.794	1.757	1.727	
52	2.805	2.408	2.192	2.056	1.961
1.891	1.836	1.791	1.755	1.724	
53	2.803	2.406	2.190	2.054	1.959
1.888	1.833	1.789	1.752	1.722	
54	2.801	2.404	2.188	2.052	1.957
1.886	1.831	1.787	1.750	1.719	
55	2.799	2.402	2.186	2.050	1.955
1.884	1.829	1.785	1.748	1.717	
56	2.797	2.400	2.184	2.048	1.953
1.882	1.827	1.782	1.746	1.715	
57	2.796	2.398	2.182	2.046	1.951
1.880	1.825	1.780	1.744	1.713	
58	2.794	2.396	2.181	2.044	1.949
1.878	1.823	1.779	1.742	1.711	
59	2.793	2.395	2.179	2.043	1.947
1.876	1.821	1.777	1.740	1.709	
60	2.791	2.393	2.177	2.041	1.946
1.875	1.819	1.775	1.738	1.707	
61	2.790	2.392	2.176	2.039	1.944
1.873	1.818	1.773	1.736	1.705	
62	2.788	2.390	2.174	2.038	1.942
1.871	1.816	1.771	1.735	1.703	
63	2.787	2.389	2.173	2.036	1.941
1.870	1.814	1.770	1.733	1.702	

64	2.786	2.387	2.171	2.035	1.939
1.868	1.813	1.768	1.731	1.700	
65	2.784	2.386	2.170	2.033	1.938
1.867	1.811	1.767	1.730	1.699	
66	2.783	2.385	2.169	2.032	1.937
1.865	1.810	1.765	1.728	1.697	
67	2.782	2.384	2.167	2.031	1.935
1.864	1.808	1.764	1.727	1.696	
68	2.781	2.382	2.166	2.029	1.934
1.863	1.807	1.762	1.725	1.694	
69	2.780	2.381	2.165	2.028	1.933
1.861	1.806	1.761	1.724	1.693	
70	2.779	2.380	2.164	2.027	1.931
1.860	1.804	1.760	1.723	1.691	
71	2.778	2.379	2.163	2.026	1.930
1.859	1.803	1.758	1.721	1.690	
72	2.777	2.378	2.161	2.025	1.929
1.858	1.802	1.757	1.720	1.689	
73	2.776	2.377	2.160	2.024	1.928
1.856	1.801	1.756	1.719	1.687	
74	2.775	2.376	2.159	2.022	1.927
1.855	1.800	1.755	1.718	1.686	
75	2.774	2.375	2.158	2.021	1.926
1.854	1.798	1.754	1.716	1.685	
76	2.773	2.374	2.157	2.020	1.925
1.853	1.797	1.752	1.715	1.684	
77	2.772	2.373	2.156	2.019	1.924
1.852	1.796	1.751	1.714	1.683	
78	2.771	2.372	2.155	2.018	1.923
1.851	1.795	1.750	1.713	1.682	
79	2.770	2.371	2.154	2.017	1.922
1.850	1.794	1.749	1.712	1.681	
80	2.769	2.370	2.154	2.016	1.921
1.849	1.793	1.748	1.711	1.680	
81	2.769	2.369	2.153	2.016	1.920
1.848	1.792	1.747	1.710	1.679	
82	2.768	2.368	2.152	2.015	1.919
1.847	1.791	1.746	1.709	1.678	
83	2.767	2.368	2.151	2.014	1.918
1.846	1.790	1.745	1.708	1.677	
84	2.766	2.367	2.150	2.013	1.917
1.845	1.790	1.744	1.707	1.676	
85	2.765	2.366	2.149	2.012	1.916
1.845	1.789	1.744	1.706	1.675	
86	2.765	2.365	2.149	2.011	1.915
1.844	1.788	1.743	1.705	1.674	
87	2.764	2.365	2.148	2.011	1.915
1.843	1.787	1.742	1.705	1.673	
88	2.763	2.364	2.147	2.010	1.914
1.842	1.786	1.741	1.704	1.672	
89	2.763	2.363	2.146	2.009	1.913
1.841	1.785	1.740	1.703	1.671	
90	2.762	2.363	2.146	2.008	1.912
1.841	1.785	1.739	1.702	1.670	
91	2.761	2.362	2.145	2.008	1.912
1.840	1.784	1.739	1.701	1.670	
92	2.761	2.361	2.144	2.007	1.911
1.839	1.783	1.738	1.701	1.669	
93	2.760	2.361	2.144	2.006	1.910
1.838	1.782	1.737	1.700	1.668	
94	2.760	2.360	2.143	2.006	1.910
1.838	1.782	1.736	1.699	1.667	
95	2.759	2.359	2.142	2.005	1.909
1.837	1.781	1.736	1.698	1.667	

96	2.759	2.359	2.142	2.004	1.908
1.836	1.780	1.735	1.698	1.666	
97	2.758	2.358	2.141	2.004	1.908
1.836	1.780	1.734	1.697	1.665	
98	2.757	2.358	2.141	2.003	1.907
1.835	1.779	1.734	1.696	1.665	
99	2.757	2.357	2.140	2.003	1.906
1.835	1.778	1.733	1.696	1.664	
100	2.756	2.356	2.139	2.002	1.906
1.834	1.778	1.732	1.695	1.663	

ν_2 \ ν_1	11	12	13	14	15
16	17	18	19	20	
1	60.473	60.705	60.903	61.073	61.220
61.350	61.464	61.566	61.658	61.740	
2	9.401	9.408	9.415	9.420	9.425
9.429	9.433	9.436	9.439	9.441	
3	5.222	5.216	5.210	5.205	5.200
5.196	5.193	5.190	5.187	5.184	
4	3.907	3.896	3.886	3.878	3.870
3.864	3.858	3.853	3.849	3.844	
5	3.282	3.268	3.257	3.247	3.238
3.230	3.223	3.217	3.212	3.207	
6	2.920	2.905	2.892	2.881	2.871
2.863	2.855	2.848	2.842	2.836	
7	2.684	2.668	2.654	2.643	2.632
2.623	2.615	2.607	2.601	2.595	
8	2.519	2.502	2.488	2.475	2.464
2.455	2.446	2.438	2.431	2.425	
9	2.396	2.379	2.364	2.351	2.340
2.329	2.320	2.312	2.305	2.298	
10	2.302	2.284	2.269	2.255	2.244
2.233	2.224	2.215	2.208	2.201	
11	2.227	2.209	2.193	2.179	2.167
2.156	2.147	2.138	2.130	2.123	
12	2.166	2.147	2.131	2.117	2.105
2.094	2.084	2.075	2.067	2.060	
13	2.116	2.097	2.080	2.066	2.053
2.042	2.032	2.023	2.014	2.007	
14	2.073	2.054	2.037	2.022	2.010
1.998	1.988	1.978	1.970	1.962	
15	2.037	2.017	2.000	1.985	1.972
1.961	1.950	1.941	1.932	1.924	
16	2.005	1.985	1.968	1.953	1.940
1.928	1.917	1.908	1.899	1.891	
17	1.978	1.958	1.940	1.925	1.912
1.900	1.889	1.879	1.870	1.862	
18	1.954	1.933	1.916	1.900	1.887
1.875	1.864	1.854	1.845	1.837	
19	1.932	1.912	1.894	1.878	1.865
1.852	1.841	1.831	1.822	1.814	
20	1.913	1.892	1.875	1.859	1.845
1.833	1.821	1.811	1.802	1.794	
21	1.896	1.875	1.857	1.841	1.827
1.815	1.803	1.793	1.784	1.776	
22	1.880	1.859	1.841	1.825	1.811
1.798	1.787	1.777	1.768	1.759	
23	1.866	1.845	1.827	1.811	1.796
1.784	1.772	1.762	1.753	1.744	

24	1.853	1.832	1.814	1.797	1.783
1.770	1.759	1.748	1.739	1.730	
25	1.841	1.820	1.802	1.785	1.771
1.758	1.746	1.736	1.726	1.718	
26	1.830	1.809	1.790	1.774	1.760
1.747	1.735	1.724	1.715	1.706	
27	1.820	1.799	1.780	1.764	1.749
1.736	1.724	1.714	1.704	1.695	
28	1.811	1.790	1.771	1.754	1.740
1.726	1.715	1.704	1.694	1.685	
29	1.802	1.781	1.762	1.745	1.731
1.717	1.705	1.695	1.685	1.676	
30	1.794	1.773	1.754	1.737	1.722
1.709	1.697	1.686	1.676	1.667	
31	1.787	1.765	1.746	1.729	1.714
1.701	1.689	1.678	1.668	1.659	
32	1.780	1.758	1.739	1.722	1.707
1.694	1.682	1.671	1.661	1.652	
33	1.773	1.751	1.732	1.715	1.700
1.687	1.675	1.664	1.654	1.645	
34	1.767	1.745	1.726	1.709	1.694
1.680	1.668	1.657	1.647	1.638	
35	1.761	1.739	1.720	1.703	1.688
1.674	1.662	1.651	1.641	1.632	
36	1.756	1.734	1.715	1.697	1.682
1.669	1.656	1.645	1.635	1.626	
37	1.751	1.729	1.709	1.692	1.677
1.663	1.651	1.640	1.630	1.620	
38	1.746	1.724	1.704	1.687	1.672
1.658	1.646	1.635	1.624	1.615	
39	1.741	1.719	1.700	1.682	1.667
1.653	1.641	1.630	1.619	1.610	
40	1.737	1.715	1.695	1.678	1.662
1.649	1.636	1.625	1.615	1.605	
41	1.733	1.710	1.691	1.673	1.658
1.644	1.632	1.620	1.610	1.601	
42	1.729	1.706	1.687	1.669	1.654
1.640	1.628	1.616	1.606	1.596	
43	1.725	1.703	1.683	1.665	1.650
1.636	1.624	1.612	1.602	1.592	
44	1.721	1.699	1.679	1.662	1.646
1.632	1.620	1.608	1.598	1.588	
45	1.718	1.695	1.676	1.658	1.643
1.629	1.616	1.605	1.594	1.585	
46	1.715	1.692	1.672	1.655	1.639
1.625	1.613	1.601	1.591	1.581	
47	1.712	1.689	1.669	1.652	1.636
1.622	1.609	1.598	1.587	1.578	
48	1.709	1.686	1.666	1.648	1.633
1.619	1.606	1.594	1.584	1.574	
49	1.706	1.683	1.663	1.645	1.630
1.616	1.603	1.591	1.581	1.571	
50	1.703	1.680	1.660	1.643	1.627
1.613	1.600	1.588	1.578	1.568	
51	1.700	1.677	1.658	1.640	1.624
1.610	1.597	1.586	1.575	1.565	
52	1.698	1.675	1.655	1.637	1.621
1.607	1.594	1.583	1.572	1.562	
53	1.695	1.672	1.652	1.635	1.619
1.605	1.592	1.580	1.570	1.560	
54	1.693	1.670	1.650	1.632	1.616
1.602	1.589	1.578	1.567	1.557	
55	1.691	1.668	1.648	1.630	1.614
1.600	1.587	1.575	1.564	1.555	

56	1.688	1.666	1.645	1.628	1.612
1.597	1.585	1.573	1.562	1.552	
57	1.686	1.663	1.643	1.625	1.610
1.595	1.582	1.571	1.560	1.550	
58	1.684	1.661	1.641	1.623	1.607
1.593	1.580	1.568	1.558	1.548	
59	1.682	1.659	1.639	1.621	1.605
1.591	1.578	1.566	1.555	1.546	
60	1.680	1.657	1.637	1.619	1.603
1.589	1.576	1.564	1.553	1.543	
61	1.679	1.656	1.635	1.617	1.601
1.587	1.574	1.562	1.551	1.541	
62	1.677	1.654	1.634	1.616	1.600
1.585	1.572	1.560	1.549	1.540	
63	1.675	1.652	1.632	1.614	1.598
1.583	1.570	1.558	1.548	1.538	
64	1.673	1.650	1.630	1.612	1.596
1.582	1.569	1.557	1.546	1.536	
65	1.672	1.649	1.628	1.610	1.594
1.580	1.567	1.555	1.544	1.534	
66	1.670	1.647	1.627	1.609	1.593
1.578	1.565	1.553	1.542	1.532	
67	1.669	1.646	1.625	1.607	1.591
1.577	1.564	1.552	1.541	1.531	
68	1.667	1.644	1.624	1.606	1.590
1.575	1.562	1.550	1.539	1.529	
69	1.666	1.643	1.622	1.604	1.588
1.574	1.560	1.548	1.538	1.527	
70	1.665	1.641	1.621	1.603	1.587
1.572	1.559	1.547	1.536	1.526	
71	1.663	1.640	1.619	1.601	1.585
1.571	1.557	1.545	1.535	1.524	
72	1.662	1.639	1.618	1.600	1.584
1.569	1.556	1.544	1.533	1.523	
73	1.661	1.637	1.617	1.599	1.583
1.568	1.555	1.543	1.532	1.522	
74	1.659	1.636	1.616	1.597	1.581
1.567	1.553	1.541	1.530	1.520	
75	1.658	1.635	1.614	1.596	1.580
1.565	1.552	1.540	1.529	1.519	
76	1.657	1.634	1.613	1.595	1.579
1.564	1.551	1.539	1.528	1.518	
77	1.656	1.632	1.612	1.594	1.578
1.563	1.550	1.538	1.527	1.516	
78	1.655	1.631	1.611	1.593	1.576
1.562	1.548	1.536	1.525	1.515	
79	1.654	1.630	1.610	1.592	1.575
1.561	1.547	1.535	1.524	1.514	
80	1.653	1.629	1.609	1.590	1.574
1.559	1.546	1.534	1.523	1.513	
81	1.652	1.628	1.608	1.589	1.573
1.558	1.545	1.533	1.522	1.512	
82	1.651	1.627	1.607	1.588	1.572
1.557	1.544	1.532	1.521	1.511	
83	1.650	1.626	1.606	1.587	1.571
1.556	1.543	1.531	1.520	1.509	
84	1.649	1.625	1.605	1.586	1.570
1.555	1.542	1.530	1.519	1.508	
85	1.648	1.624	1.604	1.585	1.569
1.554	1.541	1.529	1.518	1.507	
86	1.647	1.623	1.603	1.584	1.568
1.553	1.540	1.528	1.517	1.506	
87	1.646	1.622	1.602	1.583	1.567
1.552	1.539	1.527	1.516	1.505	

```
 88        1.645   1.622   1.601   1.583   1.566
1.551   1.538   1.526   1.515   1.504
 89        1.644   1.621   1.600   1.582   1.565
1.550   1.537   1.525   1.514   1.503
 90        1.643   1.620   1.599   1.581   1.564
1.550   1.536   1.524   1.513   1.503
 91        1.643   1.619   1.598   1.580   1.564
1.549   1.535   1.523   1.512   1.502
 92        1.642   1.618   1.598   1.579   1.563
1.548   1.534   1.522   1.511   1.501
 93        1.641   1.617   1.597   1.578   1.562
1.547   1.534   1.521   1.510   1.500
 94        1.640   1.617   1.596   1.578   1.561
1.546   1.533   1.521   1.509   1.499
 95        1.640   1.616   1.595   1.577   1.560
1.545   1.532   1.520   1.509   1.498
 96        1.639   1.615   1.594   1.576   1.560
1.545   1.531   1.519   1.508   1.497
 97        1.638   1.614   1.594   1.575   1.559
1.544   1.530   1.518   1.507   1.497
 98        1.637   1.614   1.593   1.575   1.558
1.543   1.530   1.517   1.506   1.496
 99        1.637   1.613   1.592   1.574   1.557
1.542   1.529   1.517   1.505   1.495
100        1.636   1.612   1.592   1.573   1.557
1.542   1.528   1.516   1.505   1.494
```

Upper critical values of the F distribution for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

1% significance level

$$F_{.01}(\nu_1, \nu_2)$$

```
     \ ν1    1        2        3        4        5
6        7        8        9       10
   ν2

  1      4052.19 4999.52 5403.34 5624.62 5763.65
5858.97 5928.33 5981.10 6022.50 6055.85
  2       98.502  99.000  99.166  99.249  99.300
99.333  99.356  99.374  99.388  99.399
  3       34.116  30.816  29.457  28.710  28.237
27.911  27.672  27.489  27.345  27.229
  4       21.198  18.000  16.694  15.977  15.522
15.207  14.976  14.799  14.659  14.546
  5       16.258  13.274  12.060  11.392  10.967
10.672  10.456  10.289  10.158  10.051
  6       13.745  10.925   9.780   9.148   8.746
8.466   8.260   8.102   7.976   7.874
  7       12.246   9.547   8.451   7.847   7.460
7.191   6.993   6.840   6.719   6.620
  8       11.259   8.649   7.591   7.006   6.632
6.371   6.178   6.029   5.911   5.814
  9       10.561   8.022   6.992   6.422   6.057
5.802   5.613   5.467   5.351   5.257
```

10	10.044	7.559	6.552	5.994	5.636
5.386	5.200	5.057	4.942	4.849	
11	9.646	7.206	6.217	5.668	5.316
5.069	4.886	4.744	4.632	4.539	
12	9.330	6.927	5.953	5.412	5.064
4.821	4.640	4.499	4.388	4.296	
13	9.074	6.701	5.739	5.205	4.862
4.620	4.441	4.302	4.191	4.100	
14	8.862	6.515	5.564	5.035	4.695
4.456	4.278	4.140	4.030	3.939	
15	8.683	6.359	5.417	4.893	4.556
4.318	4.142	4.004	3.895	3.805	
16	8.531	6.226	5.292	4.773	4.437
4.202	4.026	3.890	3.780	3.691	
17	8.400	6.112	5.185	4.669	4.336
4.102	3.927	3.791	3.682	3.593	
18	8.285	6.013	5.092	4.579	4.248
4.015	3.841	3.705	3.597	3.508	
19	8.185	5.926	5.010	4.500	4.171
3.939	3.765	3.631	3.523	3.434	
20	8.096	5.849	4.938	4.431	4.103
3.871	3.699	3.564	3.457	3.368	
21	8.017	5.780	4.874	4.369	4.042
3.812	3.640	3.506	3.398	3.310	
22	7.945	5.719	4.817	4.313	3.988
3.758	3.587	3.453	3.346	3.258	
23	7.881	5.664	4.765	4.264	3.939
3.710	3.539	3.406	3.299	3.211	
24	7.823	5.614	4.718	4.218	3.895
3.667	3.496	3.363	3.256	3.168	
25	7.770	5.568	4.675	4.177	3.855
3.627	3.457	3.324	3.217	3.129	
26	7.721	5.526	4.637	4.140	3.818
3.591	3.421	3.288	3.182	3.094	
27	7.677	5.488	4.601	4.106	3.785
3.558	3.388	3.256	3.149	3.062	
28	7.636	5.453	4.568	4.074	3.754
3.528	3.358	3.226	3.120	3.032	
29	7.598	5.420	4.538	4.045	3.725
3.499	3.330	3.198	3.092	3.005	
30	7.562	5.390	4.510	4.018	3.699
3.473	3.305	3.173	3.067	2.979	
31	7.530	5.362	4.484	3.993	3.675
3.449	3.281	3.149	3.043	2.955	
32	7.499	5.336	4.459	3.969	3.652
3.427	3.258	3.127	3.021	2.934	
33	7.471	5.312	4.437	3.948	3.630
3.406	3.238	3.106	3.000	2.913	
34	7.444	5.289	4.416	3.927	3.611
3.386	3.218	3.087	2.981	2.894	
35	7.419	5.268	4.396	3.908	3.592
3.368	3.200	3.069	2.963	2.876	
36	7.396	5.248	4.377	3.890	3.574
3.351	3.183	3.052	2.946	2.859	
37	7.373	5.229	4.360	3.873	3.558
3.334	3.167	3.036	2.930	2.843	
38	7.353	5.211	4.343	3.858	3.542
3.319	3.152	3.021	2.915	2.828	
39	7.333	5.194	4.327	3.843	3.528
3.305	3.137	3.006	2.901	2.814	
40	7.314	5.179	4.313	3.828	3.514
3.291	3.124	2.993	2.888	2.801	
41	7.296	5.163	4.299	3.815	3.501
3.278	3.111	2.980	2.875	2.788	

42	7.280	5.149	4.285	3.802	3.488
3.266	3.099	2.968	2.863	2.776	
43	7.264	5.136	4.273	3.790	3.476
3.254	3.087	2.957	2.851	2.764	
44	7.248	5.123	4.261	3.778	3.465
3.243	3.076	2.946	2.840	2.754	
45	7.234	5.110	4.249	3.767	3.454
3.232	3.066	2.935	2.830	2.743	
46	7.220	5.099	4.238	3.757	3.444
3.222	3.056	2.925	2.820	2.733	
47	7.207	5.087	4.228	3.747	3.434
3.213	3.046	2.916	2.811	2.724	
48	7.194	5.077	4.218	3.737	3.425
3.204	3.037	2.907	2.802	2.715	
49	7.182	5.066	4.208	3.728	3.416
3.195	3.028	2.898	2.793	2.706	
50	7.171	5.057	4.199	3.720	3.408
3.186	3.020	2.890	2.785	2.698	
51	7.159	5.047	4.191	3.711	3.400
3.178	3.012	2.882	2.777	2.690	
52	7.149	5.038	4.182	3.703	3.392
3.171	3.005	2.874	2.769	2.683	
53	7.139	5.030	4.174	3.695	3.384
3.163	2.997	2.867	2.762	2.675	
54	7.129	5.021	4.167	3.688	3.377
3.156	2.990	2.860	2.755	2.668	
55	7.119	5.013	4.159	3.681	3.370
3.149	2.983	2.853	2.748	2.662	
56	7.110	5.006	4.152	3.674	3.363
3.143	2.977	2.847	2.742	2.655	
57	7.102	4.998	4.145	3.667	3.357
3.136	2.971	2.841	2.736	2.649	
58	7.093	4.991	4.138	3.661	3.351
3.130	2.965	2.835	2.730	2.643	
59	7.085	4.984	4.132	3.655	3.345
3.124	2.959	2.829	2.724	2.637	
60	7.077	4.977	4.126	3.649	3.339
3.119	2.953	2.823	2.718	2.632	
61	7.070	4.971	4.120	3.643	3.333
3.113	2.948	2.818	2.713	2.626	
62	7.062	4.965	4.114	3.638	3.328
3.108	2.942	2.813	2.708	2.621	
63	7.055	4.959	4.109	3.632	3.323
3.103	2.937	2.808	2.703	2.616	
64	7.048	4.953	4.103	3.627	3.318
3.098	2.932	2.803	2.698	2.611	
65	7.042	4.947	4.098	3.622	3.313
3.093	2.928	2.798	2.693	2.607	
66	7.035	4.942	4.093	3.618	3.308
3.088	2.923	2.793	2.689	2.602	
67	7.029	4.937	4.088	3.613	3.304
3.084	2.919	2.789	2.684	2.598	
68	7.023	4.932	4.083	3.608	3.299
3.080	2.914	2.785	2.680	2.593	
69	7.017	4.927	4.079	3.604	3.295
3.075	2.910	2.781	2.676	2.589	
70	7.011	4.922	4.074	3.600	3.291
3.071	2.906	2.777	2.672	2.585	
71	7.006	4.917	4.070	3.596	3.287
3.067	2.902	2.773	2.668	2.581	
72	7.001	4.913	4.066	3.591	3.283
3.063	2.898	2.769	2.664	2.578	
73	6.995	4.908	4.062	3.588	3.279
3.060	2.895	2.765	2.660	2.574	

74	6.990	4.904	4.058	3.584	3.275
3.056	2.891	2.762	2.657	2.570	
75	6.985	4.900	4.054	3.580	3.272
3.052	2.887	2.758	2.653	2.567	
76	6.981	4.896	4.050	3.577	3.268
3.049	2.884	2.755	2.650	2.563	
77	6.976	4.892	4.047	3.573	3.265
3.046	2.881	2.751	2.647	2.560	
78	6.971	4.888	4.043	3.570	3.261
3.042	2.877	2.748	2.644	2.557	
79	6.967	4.884	4.040	3.566	3.258
3.039	2.874	2.745	2.640	2.554	
80	6.963	4.881	4.036	3.563	3.255
3.036	2.871	2.742	2.637	2.551	
81	6.958	4.877	4.033	3.560	3.252
3.033	2.868	2.739	2.634	2.548	
82	6.954	4.874	4.030	3.557	3.249
3.030	2.865	2.736	2.632	2.545	
83	6.950	4.870	4.027	3.554	3.246
3.027	2.863	2.733	2.629	2.542	
84	6.947	4.867	4.024	3.551	3.243
3.025	2.860	2.731	2.626	2.539	
85	6.943	4.864	4.021	3.548	3.240
3.022	2.857	2.728	2.623	2.537	
86	6.939	4.861	4.018	3.545	3.238
3.019	2.854	2.725	2.621	2.534	
87	6.935	4.858	4.015	3.543	3.235
3.017	2.852	2.723	2.618	2.532	
88	6.932	4.855	4.012	3.540	3.233
3.014	2.849	2.720	2.616	2.529	
89	6.928	4.852	4.010	3.538	3.230
3.012	2.847	2.718	2.613	2.527	
90	6.925	4.849	4.007	3.535	3.228
3.009	2.845	2.715	2.611	2.524	
91	6.922	4.846	4.004	3.533	3.225
3.007	2.842	2.713	2.609	2.522	
92	6.919	4.844	4.002	3.530	3.223
3.004	2.840	2.711	2.606	2.520	
93	6.915	4.841	3.999	3.528	3.221
3.002	2.838	2.709	2.604	2.518	
94	6.912	4.838	3.997	3.525	3.218
3.000	2.835	2.706	2.602	2.515	
95	6.909	4.836	3.995	3.523	3.216
2.998	2.833	2.704	2.600	2.513	
96	6.906	4.833	3.992	3.521	3.214
2.996	2.831	2.702	2.598	2.511	
97	6.904	4.831	3.990	3.519	3.212
2.994	2.829	2.700	2.596	2.509	
98	6.901	4.829	3.988	3.517	3.210
2.992	2.827	2.698	2.594	2.507	
99	6.898	4.826	3.986	3.515	3.208
2.990	2.825	2.696	2.592	2.505	
100	6.895	4.824	3.984	3.513	3.206
2.988	2.823	2.694	2.590	2.503	

ν_2 \ ν_1	11	12	13	14	15
16	17	18	19	20	
1.	6083.35	6106.35	6125.86	6142.70	6157.28
6170.12	6181.42	6191.52	6200.58	6208.74	

```
  2.      99.408  99.416  99.422  99.428  99.432
99.437  99.440  99.444  99.447  99.449
  3.      27.133  27.052  26.983  26.924  26.872
26.827  26.787  26.751  26.719  26.690
  4.      14.452  14.374  14.307  14.249  14.198
14.154  14.115  14.080  14.048  14.020
  5.       9.963   9.888   9.825   9.770   9.722
9.680   9.643   9.610   9.580   9.553
  6.       7.790   7.718   7.657   7.605   7.559
7.519   7.483   7.451   7.422   7.396
  7.       6.538   6.469   6.410   6.359   6.314
6.275   6.240   6.209   6.181   6.155
  8.       5.734   5.667   5.609   5.559   5.515
5.477   5.442   5.412   5.384   5.359
  9.       5.178   5.111   5.055   5.005   4.962
4.924   4.890   4.860   4.833   4.808
 10.       4.772   4.706   4.650   4.601   4.558
4.520   4.487   4.457   4.430   4.405
 11.       4.462   4.397   4.342   4.293   4.251
4.213   4.180   4.150   4.123   4.099
 12.       4.220   4.155   4.100   4.052   4.010
3.972   3.939   3.909   3.883   3.858
 13.       4.025   3.960   3.905   3.857   3.815
3.778   3.745   3.716   3.689   3.665
 14.       3.864   3.800   3.745   3.698   3.656
3.619   3.586   3.556   3.529   3.505
 15.       3.730   3.666   3.612   3.564   3.522
3.485   3.452   3.423   3.396   3.372
 16.       3.616   3.553   3.498   3.451   3.409
3.372   3.339   3.310   3.283   3.259
 17.       3.519   3.455   3.401   3.353   3.312
3.275   3.242   3.212   3.186   3.162
 18.       3.434   3.371   3.316   3.269   3.227
3.190   3.158   3.128   3.101   3.077
 19.       3.360   3.297   3.242   3.195   3.153
3.116   3.084   3.054   3.027   3.003
 20.       3.294   3.231   3.177   3.130   3.088
3.051   3.018   2.989   2.962   2.938
 21.       3.236   3.173   3.119   3.072   3.030
2.993   2.960   2.931   2.904   2.880
 22.       3.184   3.121   3.067   3.019   2.978
2.941   2.908   2.879   2.852   2.827
 23.       3.137   3.074   3.020   2.973   2.931
2.894   2.861   2.832   2.805   2.781
 24.       3.094   3.032   2.977   2.930   2.889
2.852   2.819   2.789   2.762   2.738
 25.       3.056   2.993   2.939   2.892   2.850
2.813   2.780   2.751   2.724   2.699
 26.       3.021   2.958   2.904   2.857   2.815
2.778   2.745   2.715   2.688   2.664
 27.       2.988   2.926   2.871   2.824   2.783
2.746   2.713   2.683   2.656   2.632
 28.       2.959   2.896   2.842   2.795   2.753
2.716   2.683   2.653   2.626   2.602
 29.       2.931   2.868   2.814   2.767   2.726
2.689   2.656   2.626   2.599   2.574
 30.       2.906   2.843   2.789   2.742   2.700
2.663   2.630   2.600   2.573   2.549
 31.       2.882   2.820   2.765   2.718   2.677
2.640   2.606   2.577   2.550   2.525
 32.       2.860   2.798   2.744   2.696   2.655
2.618   2.584   2.555   2.527   2.503
 33.       2.840   2.777   2.723   2.676   2.634
2.597   2.564   2.534   2.507   2.482
```

34.	2.821	2.758	2.704	2.657	2.615
2.578	2.545	2.515	2.488	2.463	
35.	2.803	2.740	2.686	2.639	2.597
2.560	2.527	2.497	2.470	2.445	
36.	2.786	2.723	2.669	2.622	2.580
2.543	2.510	2.480	2.453	2.428	
37.	2.770	2.707	2.653	2.606	2.564
2.527	2.494	2.464	2.437	2.412	
38.	2.755	2.692	2.638	2.591	2.549
2.512	2.479	2.449	2.421	2.397	
39.	2.741	2.678	2.624	2.577	2.535
2.498	2.465	2.434	2.407	2.382	
40.	2.727	2.665	2.611	2.563	2.522
2.484	2.451	2.421	2.394	2.369	
41.	2.715	2.652	2.598	2.551	2.509
2.472	2.438	2.408	2.381	2.356	
42.	2.703	2.640	2.586	2.539	2.497
2.460	2.426	2.396	2.369	2.344	
43.	2.691	2.629	2.575	2.527	2.485
2.448	2.415	2.385	2.357	2.332	
44.	2.680	2.618	2.564	2.516	2.475
2.437	2.404	2.374	2.346	2.321	
45.	2.670	2.608	2.553	2.506	2.464
2.427	2.393	2.363	2.336	2.311	
46.	2.660	2.598	2.544	2.496	2.454
2.417	2.384	2.353	2.326	2.301	
47.	2.651	2.588	2.534	2.487	2.445
2.408	2.374	2.344	2.316	2.291	
48.	2.642	2.579	2.525	2.478	2.436
2.399	2.365	2.335	2.307	2.282	
49.	2.633	2.571	2.517	2.469	2.427
2.390	2.356	2.326	2.299	2.274	
50.	2.625	2.562	2.508	2.461	2.419
2.382	2.348	2.318	2.290	2.265	
51.	2.617	2.555	2.500	2.453	2.411
2.374	2.340	2.310	2.282	2.257	
52.	2.610	2.547	2.493	2.445	2.403
2.366	2.333	2.302	2.275	2.250	
53.	2.602	2.540	2.486	2.438	2.396
2.359	2.325	2.295	2.267	2.242	
54.	2.595	2.533	2.479	2.431	2.389
2.352	2.318	2.288	2.260	2.235	
55.	2.589	2.526	2.472	2.424	2.382
2.345	2.311	2.281	2.253	2.228	
56.	2.582	2.520	2.465	2.418	2.376
2.339	2.305	2.275	2.247	2.222	
57.	2.576	2.513	2.459	2.412	2.370
2.332	2.299	2.268	2.241	2.215	
58.	2.570	2.507	2.453	2.406	2.364
2.326	2.293	2.262	2.235	2.209	
59.	2.564	2.502	2.447	2.400	2.358
2.320	2.287	2.256	2.229	2.203	
60.	2.559	2.496	2.442	2.394	2.352
2.315	2.281	2.251	2.223	2.198	
61.	2.553	2.491	2.436	2.389	2.347
2.309	2.276	2.245	2.218	2.192	
62.	2.548	2.486	2.431	2.384	2.342
2.304	2.270	2.240	2.212	2.187	
63.	2.543	2.481	2.426	2.379	2.337
2.299	2.265	2.235	2.207	2.182	
64.	2.538	2.476	2.421	2.374	2.332
2.294	2.260	2.230	2.202	2.177	
65.	2.534	2.471	2.417	2.369	2.327
2.289	2.256	2.225	2.198	2.172	

66.	2.529	2.466	2.412	2.365	2.322
2.285	2.251	2.221	2.193	2.168	
67.	2.525	2.462	2.408	2.360	2.318
2.280	2.247	2.216	2.188	2.163	
68.	2.520	2.458	2.403	2.356	2.314
2.276	2.242	2.212	2.184	2.159	
69.	2.516	2.454	2.399	2.352	2.310
2.272	2.238	2.208	2.180	2.155	
70.	2.512	2.450	2.395	2.348	2.306
2.268	2.234	2.204	2.176	2.150	
71.	2.508	2.446	2.391	2.344	2.302
2.264	2.230	2.200	2.172	2.146	
72.	2.504	2.442	2.388	2.340	2.298
2.260	2.226	2.196	2.168	2.143	
73.	2.501	2.438	2.384	2.336	2.294
2.256	2.223	2.192	2.164	2.139	
74.	2.497	2.435	2.380	2.333	2.290
2.253	2.219	2.188	2.161	2.135	
75.	2.494	2.431	2.377	2.329	2.287
2.249	2.215	2.185	2.157	2.132	
76.	2.490	2.428	2.373	2.326	2.284
2.246	2.212	2.181	2.154	2.128	
77.	2.487	2.424	2.370	2.322	2.280
2.243	2.209	2.178	2.150	2.125	
78.	2.484	2.421	2.367	2.319	2.277
2.239	2.206	2.175	2.147	2.122	
79.	2.481	2.418	2.364	2.316	2.274
2.236	2.202	2.172	2.144	2.118	
80.	2.478	2.415	2.361	2.313	2.271
2.233	2.199	2.169	2.141	2.115	
81.	2.475	2.412	2.358	2.310	2.268
2.230	2.196	2.166	2.138	2.112	
82.	2.472	2.409	2.355	2.307	2.265
2.227	2.193	2.163	2.135	2.109	
83.	2.469	2.406	2.352	2.304	2.262
2.224	2.191	2.160	2.132	2.106	
84.	2.466	2.404	2.349	2.302	2.259
2.222	2.188	2.157	2.129	2.104	
85.	2.464	2.401	2.347	2.299	2.257
2.219	2.185	2.154	2.126	2.101	
86.	2.461	2.398	2.344	2.296	2.254
2.216	2.182	2.152	2.124	2.098	
87.	2.459	2.396	2.342	2.294	2.252
2.214	2.180	2.149	2.121	2.096	
88.	2.456	2.393	2.339	2.291	2.249
2.211	2.177	2.147	2.119	2.093	
89.	2.454	2.391	2.337	2.289	2.247
2.209	2.175	2.144	2.116	2.091	
90.	2.451	2.389	2.334	2.286	2.244
2.206	2.172	2.142	2.114	2.088	
91.	2.449	2.386	2.332	2.284	2.242
2.204	2.170	2.139	2.111	2.086	
92.	2.447	2.384	2.330	2.282	2.240
2.202	2.168	2.137	2.109	2.083	
93.	2.444	2.382	2.327	2.280	2.237
2.200	2.166	2.135	2.107	2.081	
94.	2.442	2.380	2.325	2.277	2.235
2.197	2.163	2.133	2.105	2.079	
95.	2.440	2.378	2.323	2.275	2.233
2.195	2.161	2.130	2.102	2.077	
96.	2.438	2.375	2.321	2.273	2.231
2.193	2.159	2.128	2.100	2.075	
97.	2.436	2.373	2.319	2.271	2.229
2.191	2.157	2.126	2.098	2.073	

98.	2.434	2.371	2.317	2.269	2.227
2.189	2.155	2.124	2.096	2.071	
99.	2.432	2.369	2.315	2.267	2.225
2.187	2.153	2.122	2.094	2.069	
100.	2.430	2.368	2.313	2.265	2.223
2.185	2.151	2.120	2.092	2.067	

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions
1.3.6.7. Tables for Probability Distributions

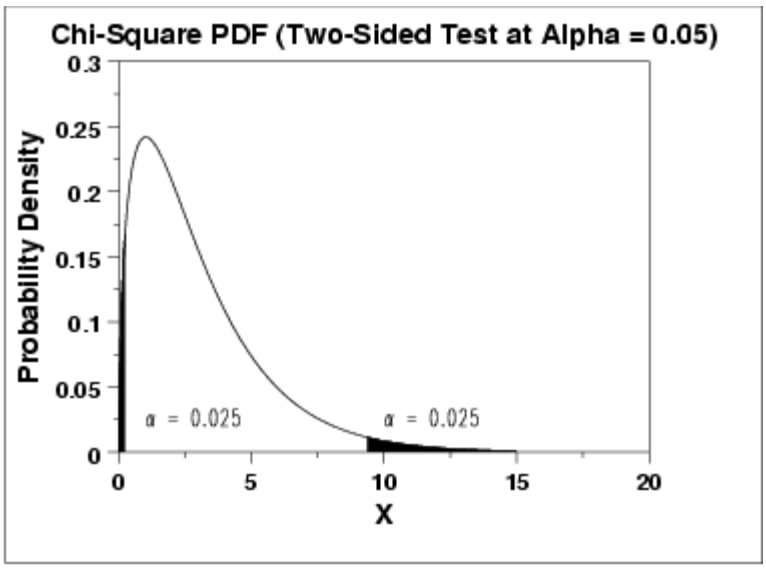
1.3.6.7.4. Critical Values of the Chi-Square Distribution

How to Use This Table

This table contains the critical values of the chi-square distribution. Because of the lack of symmetry of the chi-square distribution, separate tables are provided for the upper and lower tails of the distribution.

A test statistic with v degrees of freedom is computed from the data. For upper-tail one-sided tests, the test statistic is compared with a value from the table of upper-tail critical values. For two-sided tests, the test statistic is compared with values from both the table for the upper-tail critical values and the table for the lower-tail critical values.

The significance level, α, is demonstrated with the graph below which shows a chi-square distribution with 3 degrees of freedom for a two-sided test at significance level $\alpha = 0.05$. If the test statistic is greater than the upper-tail critical value or less than the lower-tail critical value, we reject the null hypothesis. Specific instructions are given below.



Given a specified value of α:

1. For a two-sided test, find the column corresponding to $1-\alpha/2$ in the table for upper-tail critical values and reject

the null hypothesis if the test statistic is greater than the tabled value. Similarly, find the column corresponding to $\alpha/2$ in the table for lower-tail critical values and reject the null hypothesis if the test statistic is less than the tabled value.

2. For an upper-tail one-sided test, find the column corresponding to 1-α in the table containing upper-tail critical and reject the null hypothesis if the test statistic is greater than the tabled value.
3. For a lower-tail one-sided test, find the column corresponding to α in the lower-tail critical values table and reject the null hypothesis if the computed test statistic is less than the tabled value.

Upper-tail critical values of chi-square distribution with ν degrees of freedom

	Probability less than the critical value			
ν	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252

17	24.769	27.587	30.191	33.409
40.790				
18	25.989	28.869	31.526	34.805
42.312				
19	27.204	30.144	32.852	36.191
43.820				
20	28.412	31.410	34.170	37.566
45.315				
21	29.615	32.671	35.479	38.932
46.797				
22	30.813	33.924	36.781	40.289
48.268				
23	32.007	35.172	38.076	41.638
49.728				
24	33.196	36.415	39.364	42.980
51.179				
25	34.382	37.652	40.646	44.314
52.620				
26	35.563	38.885	41.923	45.642
54.052				
27	36.741	40.113	43.195	46.963
55.476				
28	37.916	41.337	44.461	48.278
56.892				
29	39.087	42.557	45.722	49.588
58.301				
30	40.256	43.773	46.979	50.892
59.703				
31	41.422	44.985	48.232	52.191
61.098				
32	42.585	46.194	49.480	53.486
62.487				
33	43.745	47.400	50.725	54.776
63.870				
34	44.903	48.602	51.966	56.061
65.247				
35	46.059	49.802	53.203	57.342
66.619				
36	47.212	50.998	54.437	58.619
67.985				
37	48.363	52.192	55.668	59.893
69.347				
38	49.513	53.384	56.896	61.162
70.703				
39	50.660	54.572	58.120	62.428
72.055				
40	51.805	55.758	59.342	63.691
73.402				
41	52.949	56.942	60.561	64.950
74.745				
42	54.090	58.124	61.777	66.206
76.084				
43	55.230	59.304	62.990	67.459
77.419				
44	56.369	60.481	64.201	68.710
78.750				
45	57.505	61.656	65.410	69.957
80.077				
46	58.641	62.830	66.617	71.201
81.400				
47	59.774	64.001	67.821	72.443
82.720				
48	60.907	65.171	69.023	73.683
84.037				

49	62.038	66.339	70.222	74.919
85.351				
50	63.167	67.505	71.420	76.154
86.661				
51	64.295	68.669	72.616	77.386
87.968				
52	65.422	69.832	73.810	78.616
89.272				
53	66.548	70.993	75.002	79.843
90.573				
54	67.673	72.153	76.192	81.069
91.872				
55	68.796	73.311	77.380	82.292
93.168				
56	69.919	74.468	78.567	83.513
94.461				
57	71.040	75.624	79.752	84.733
95.751				
58	72.160	76.778	80.936	85.950
97.039				
59	73.279	77.931	82.117	87.166
98.324				
60	74.397	79.082	83.298	88.379
99.607				
61	75.514	80.232	84.476	89.591
100.888				
62	76.630	81.381	85.654	90.802
102.166				
63	77.745	82.529	86.830	92.010
103.442				
64	78.860	83.675	88.004	93.217
104.716				
65	79.973	84.821	89.177	94.422
105.988				
66	81.085	85.965	90.349	95.626
107.258				
67	82.197	87.108	91.519	96.828
108.526				
68	83.308	88.250	92.689	98.028
109.791				
69	84.418	89.391	93.856	99.228
111.055				
70	85.527	90.531	95.023	100.425
112.317				
71	86.635	91.670	96.189	101.621
113.577				
72	87.743	92.808	97.353	102.816
114.835				
73	88.850	93.945	98.516	104.010
116.092				
74	89.956	95.081	99.678	105.202
117.346				
75	91.061	96.217	100.839	106.393
118.599				
76	92.166	97.351	101.999	107.583
119.850				
77	93.270	98.484	103.158	108.771
121.100				
78	94.374	99.617	104.316	109.958
122.348				
79	95.476	100.749	105.473	111.144
123.594				
80	96.578	101.879	106.629	112.329
124.839				

81	97.680	103.010	107.783	113.512	126.083
82	98.780	104.139	108.937	114.695	127.324
83	99.880	105.267	110.090	115.876	128.565
84	100.980	106.395	111.242	117.057	129.804
85	102.079	107.522	112.393	118.236	131.041
86	103.177	108.648	113.544	119.414	132.277
87	104.275	109.773	114.693	120.591	133.512
88	105.372	110.898	115.841	121.767	134.746
89	106.469	112.022	116.989	122.942	135.978
90	107.565	113.145	118.136	124.116	137.208
91	108.661	114.268	119.282	125.289	138.438
92	109.756	115.390	120.427	126.462	139.666
93	110.850	116.511	121.571	127.633	140.893
94	111.944	117.632	122.715	128.803	142.119
95	113.038	118.752	123.858	129.973	143.344
96	114.131	119.871	125.000	131.141	144.567
97	115.223	120.990	126.141	132.309	145.789
98	116.315	122.108	127.282	133.476	147.010
99	117.407	123.225	128.422	134.642	148.230
100	118.498	124.342	129.561	135.807	149.449
100	118.498	124.342	129.561	135.807	149.449

Lower-tail critical values of chi-square distribution with ν degrees of freedom

	Probability less than the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1.	.016	.004	.001	.000	.000
2.	.211	.103	.051	.020	.002

3.	.584	.352	.216	.115
.024				
4.	1.064	.711	.484	.297
.091				
5.	1.610	1.145	.831	.554
.210				
6.	2.204	1.635	1.237	.872
.381				
7.	2.833	2.167	1.690	1.239
.598				
8.	3.490	2.733	2.180	1.646
.857				
9.	4.168	3.325	2.700	2.088
1.152				
10.	4.865	3.940	3.247	2.558
1.479				
11.	5.578	4.575	3.816	3.053
1.834				
12.	6.304	5.226	4.404	3.571
2.214				
13.	7.042	5.892	5.009	4.107
2.617				
14.	7.790	6.571	5.629	4.660
3.041				
15.	8.547	7.261	6.262	5.229
3.483				
16.	9.312	7.962	6.908	5.812
3.942				
17.	10.085	8.672	7.564	6.408
4.416				
18.	10.865	9.390	8.231	7.015
4.905				
19.	11.651	10.117	8.907	7.633
5.407				
20.	12.443	10.851	9.591	8.260
5.921				
21.	13.240	11.591	10.283	8.897
6.447				
22.	14.041	12.338	10.982	9.542
6.983				
23.	14.848	13.091	11.689	10.196
7.529				
24.	15.659	13.848	12.401	10.856
8.085				
25.	16.473	14.611	13.120	11.524
8.649				
26.	17.292	15.379	13.844	12.198
9.222				
27.	18.114	16.151	14.573	12.879
9.803				
28.	18.939	16.928	15.308	13.565
10.391				
29.	19.768	17.708	16.047	14.256
10.986				
30.	20.599	18.493	16.791	14.953
11.588				
31.	21.434	19.281	17.539	15.655
12.196				
32.	22.271	20.072	18.291	16.362
12.811				
33.	23.110	20.867	19.047	17.074
13.431				
34.	23.952	21.664	19.806	17.789
14.057				

35.	24.797	22.465	20.569	18.509
14.688				
36.	25.643	23.269	21.336	19.233
15.324				
37.	26.492	24.075	22.106	19.960
15.965				
38.	27.343	24.884	22.878	20.691
16.611				
39.	28.196	25.695	23.654	21.426
17.262				
40.	29.051	26.509	24.433	22.164
17.916				
41.	29.907	27.326	25.215	22.906
18.575				
42.	30.765	28.144	25.999	23.650
19.239				
43.	31.625	28.965	26.785	24.398
19.906				
44.	32.487	29.787	27.575	25.148
20.576				
45.	33.350	30.612	28.366	25.901
21.251				
46.	34.215	31.439	29.160	26.657
21.929				
47.	35.081	32.268	29.956	27.416
22.610				
48.	35.949	33.098	30.755	28.177
23.295				
49.	36.818	33.930	31.555	28.941
23.983				
50.	37.689	34.764	32.357	29.707
24.674				
51.	38.560	35.600	33.162	30.475
25.368				
52.	39.433	36.437	33.968	31.246
26.065				
53.	40.308	37.276	34.776	32.018
26.765				
54.	41.183	38.116	35.586	32.793
27.468				
55.	42.060	38.958	36.398	33.570
28.173				
56.	42.937	39.801	37.212	34.350
28.881				
57.	43.816	40.646	38.027	35.131
29.592				
58.	44.696	41.492	38.844	35.913
30.305				
59.	45.577	42.339	39.662	36.698
31.020				
60.	46.459	43.188	40.482	37.485
31.738				
61.	47.342	44.038	41.303	38.273
32.459				
62.	48.226	44.889	42.126	39.063
33.181				
63.	49.111	45.741	42.950	39.855
33.906				
64.	49.996	46.595	43.776	40.649
34.633				
65.	50.883	47.450	44.603	41.444
35.362				
66.	51.770	48.305	45.431	42.240
36.093				

67.	52.659	49.162	46.261	43.038
36.826				
68.	53.548	50.020	47.092	43.838
37.561				
69.	54.438	50.879	47.924	44.639
38.298				
70.	55.329	51.739	48.758	45.442
39.036				
71.	56.221	52.600	49.592	46.246
39.777				
72.	57.113	53.462	50.428	47.051
40.519				
73.	58.006	54.325	51.265	47.858
41.264				
74.	58.900	55.189	52.103	48.666
42.010				
75.	59.795	56.054	52.942	49.475
42.757				
76.	60.690	56.920	53.782	50.286
43.507				
77.	61.586	57.786	54.623	51.097
44.258				
78.	62.483	58.654	55.466	51.910
45.010				
79.	63.380	59.522	56.309	52.725
45.764				
80.	64.278	60.391	57.153	53.540
46.520				
81.	65.176	61.261	57.998	54.357
47.277				
82.	66.076	62.132	58.845	55.174
48.036				
83.	66.976	63.004	59.692	55.993
48.796				
84.	67.876	63.876	60.540	56.813
49.557				
85.	68.777	64.749	61.389	57.634
50.320				
86.	69.679	65.623	62.239	58.456
51.085				
87.	70.581	66.498	63.089	59.279
51.850				
88.	71.484	67.373	63.941	60.103
52.617				
89.	72.387	68.249	64.793	60.928
53.386				
90.	73.291	69.126	65.647	61.754
54.155				
91.	74.196	70.003	66.501	62.581
54.926				
92.	75.100	70.882	67.356	63.409
55.698				
93.	76.006	71.760	68.211	64.238
56.472				
94.	76.912	72.640	69.068	65.068
57.246				
95.	77.818	73.520	69.925	65.898
58.022				
96.	78.725	74.401	70.783	66.730
58.799				
97.	79.633	75.282	71.642	67.562
59.577				
98.	80.541	76.164	72.501	68.396
60.356				

99.	81.449	77.046	73.361	69.230
61.137				
100.	82.358	77.929	74.222	70.065
61.918				

1.3.6.7.5. Critical Values of the t* Distribution

How to Use This Table

This table contains upper critical values of the t* distribution that are appropriate for determining whether or not a calibration line is in a state of statistical control from measurements on a check standard at three points in the calibration interval. A test statistic with ν degrees of freedom is compared with the critical value. If the absolute value of the test statistic exceeds the tabled value, the calibration of the instrument is judged to be out of control.

Upper critical values of t* distribution at significance level 0.05 for testing the output of a linear calibration line at 3 points

ν	$t^*_{.05}(\nu)$	ν	$t^*_{.05}(\nu)$
1	37.544	61	2.455
2	7.582	62	2.454
3	4.826	63	2.453
4	3.941	64	2.452
5	3.518	65	2.451
6	3.274	66	2.450
7	3.115	67	2.449
8	3.004	68	2.448
9	2.923	69	2.447
10	2.860	70	2.446
11	2.811	71	2.445
12	2.770	72	2.445
13	2.737	73	2.444
14	2.709	74	2.443
15	2.685	75	2.442
16	2.665	76	2.441
17	2.647	77	2.441
18	2.631	78	2.440
19	2.617	79	2.439
20	2.605	80	2.439
21	2.594	81	2.438
22	2.584	82	2.437
23	2.574	83	2.437
24	2.566	84	2.436
25	2.558	85	2.436
26	2.551	86	2.435
27	2.545	87	2.435
28	2.539	88	2.434
29	2.534	89	2.434
30	2.528	90	2.433
31	2.524	91	2.432
32	2.519	92	2.432
33	2.515	93	2.431
34	2.511	94	2.431
35	2.507	95	2.431

36	2.504	96	2.430
37	2.501	97	2.430
38	2.498	98	2.429
39	2.495	99	2.429
40	2.492	100	2.428
41	2.489	101	2.428
42	2.487	102	2.428
43	2.484	103	2.427
44	2.482	104	2.427
45	2.480	105	2.426
46	2.478	106	2.426
47	2.476	107	2.426
48	2.474	108	2.425
49	2.472	109	2.425
50	2.470	110	2.425
51	2.469	111	2.424
52	2.467	112	2.424
53	2.466	113	2.424
54	2.464	114	2.423
55	2.463	115	2.423
56	2.461	116	2.423
57	2.460	117	2.422
58	2.459	118	2.422
59	2.457	119	2.422
60	2.456	120	2.422

1.3.6.7.6. Critical Values of the Normal PPCC Distribution

How to Use This Table

This table contains the critical values of the normal probability plot correlation coefficient (PPCC) distribution that are appropriate for determining whether or not a data set came from a population with approximately a normal distribution. It is used in conjuction with a normal probability plot. The test statistic is the correlation coefficient of the points that make up a normal probability plot. This test statistic is compared with the critical value below. If the test statistic is less than the tabulated value, the null hypothesis that the data came from a population with a normal distribution is rejected.

For example, suppose a set of 50 data points had a correlation coefficient of 0.985 from the normal probability plot. At the 5% significance level, the critical value is 0.9761. Since 0.985 is greater than 0.9761, we cannot reject the null hypothesis that the data came from a population with a normal distribution.

Since perferct normality implies perfect correlation (i.e., a correlation value of 1), we are only interested in rejecting normality for correlation values that are too low. That is, this is a lower one-tailed test.

The values in this table were determined from simulation studies by Filliben and Devaney.

Critical values of the normal PPCC for testing if data come from a normal distribution

N	0.01	0.05
3	0.8687	0.8790
4	0.8234	0.8666
5	0.8240	0.8786
6	0.8351	0.8880
7	0.8474	0.8970
8	0.8590	0.9043
9	0.8689	0.9115
10	0.8765	0.9173
11	0.8838	0.9223
12	0.8918	0.9267
13	0.8974	0.9310
14	0.9029	0.9343

15	0.9080	0.9376
16	0.9121	0.9405
17	0.9160	0.9433
18	0.9196	0.9452
19	0.9230	0.9479
20	0.9256	0.9498
21	0.9285	0.9515
22	0.9308	0.9535
23	0.9334	0.9548
24	0.9356	0.9564
25	0.9370	0.9575
26	0.9393	0.9590
27	0.9413	0.9600
28	0.9428	0.9615
29	0.9441	0.9622
30	0.9462	0.9634
31	0.9476	0.9644
32	0.9490	0.9652
33	0.9505	0.9661
34	0.9521	0.9671
35	0.9530	0.9678
36	0.9540	0.9686
37	0.9551	0.9693
38	0.9555	0.9700
39	0.9568	0.9704
40	0.9576	0.9712
41	0.9589	0.9719
42	0.9593	0.9723
43	0.9609	0.9730
44	0.9611	0.9734
45	0.9620	0.9739
46	0.9629	0.9744
47	0.9637	0.9748
48	0.9640	0.9753
49	0.9643	0.9758
50	0.9654	0.9761
55	0.9683	0.9781
60	0.9706	0.9797
65	0.9723	0.9809
70	0.9742	0.9822
75	0.9758	0.9831
80	0.9771	0.9841
85	0.9784	0.9850
90	0.9797	0.9857
95	0.9804	0.9864
100	0.9814	0.9869
110	0.9830	0.9881
120	0.9841	0.9889
130	0.9854	0.9897
140	0.9865	0.9904
150	0.9871	0.9909
160	0.9879	0.9915
170	0.9887	0.9919
180	0.9891	0.9923
190	0.9897	0.9927
200	0.9903	0.9930
210	0.9907	0.9933
220	0.9910	0.9936
230	0.9914	0.9939
240	0.9917	0.9941
250	0.9921	0.9943
260	0.9924	0.9945
270	0.9926	0.9947
280	0.9929	0.9949
290	0.9931	0.9951
300	0.9933	0.9952
310	0.9936	0.9954
320	0.9937	0.9955
330	0.9939	0.9956
340	0.9941	0.9957
350	0.9942	0.9958
360	0.9944	0.9959
370	0.9945	0.9960
380	0.9947	0.9961

390	0.9948	0.9962
400	0.9949	0.9963
410	0.9950	0.9964
420	0.9951	0.9965
430	0.9953	0.9966
440	0.9954	0.9966
450	0.9954	0.9967
460	0.9955	0.9968
470	0.9956	0.9968
480	0.9957	0.9969
490	0.9958	0.9969
500	0.9959	0.9970
525	0.9961	0.9972
550	0.9963	0.9973
575	0.9964	0.9974
600	0.9965	0.9975
625	0.9967	0.9976
650	0.9968	0.9977
675	0.9969	0.9977
700	0.9970	0.9978
725	0.9971	0.9979
750	0.9972	0.9980
775	0.9973	0.9980
800	0.9974	0.9981
825	0.9975	0.9981
850	0.9975	0.9982
875	0.9976	0.9982
900	0.9977	0.9983
925	0.9977	0.9983
950	0.9978	0.9984
975	0.9978	0.9984
1000	0.9979	0.9984

1. Exploratory Data Analysis

1.4. EDA Case Studies

Summary This section presents a series of case studies that demonstrate the application of EDA methods to specific problems. In some cases, we have focused on just one EDA technique that uncovers virtually all there is to know about the data. For other case studies, we need several EDA techniques, the selection of which is dictated by the outcome of the previous step in the analaysis sequence. Note in these case studies how the flow of the analysis is motivated by the focus on underlying assumptions and general EDA principles.

Table of Contents for Section 4

1. Introduction
2. By Problem Category

1.4.1. Case Studies Introduction

Purpose

The purpose of the first eight case studies is to show how EDA graphics and quantitative measures and tests are applied to data from scientific processes and to critique those data with regard to the following assumptions that typically underlie a measurement process; namely, that the data behave like:

- random drawings
- from a fixed distribution
- with a fixed location
- with a fixed standard deviation

Case studies 9 and 10 show the use of EDA techniques in distributional modeling and the analysis of a designed experiment, respectively.

$Y_i = C + E_i$

If the above assumptions are satisfied, the process is said to be statistically "in control" with the core characteristic of having "predictability". That is, probability statements can be made about the process, not only in the past, but also in the future.

An appropriate model for an "in control" process is

$$\boldsymbol{Y_i = C + E_i}$$

where $\boldsymbol{C}$ is a constant (the "deterministic" or "structural" component), and where $\boldsymbol{E_i}$ is the error term (or "random" component).

The constant $\boldsymbol{C}$ is the average value of the process--it is the primary summary number which shows up on any report. Although $\boldsymbol{C}$ is (assumed) fixed, it is unknown, and so a primary analysis objective of the engineer is to arrive at an estimate of $\boldsymbol{C}$.

This goal partitions into 4 sub-goals:

1. Is the most common estimator of $\boldsymbol{C}$, $\bar{\boldsymbol{Y}}$, the best estimator for $\boldsymbol{C}$? What does "best" mean?

2. If $\bar{\boldsymbol{Y}}$ is best, what is the uncertainty $s_{\bar{Y}}$ for $\bar{\boldsymbol{Y}}$. In particular, is the usual formula for the uncertainty of

$\bar{Y}$:

$$s_{\bar{Y}} = s/\sqrt{N}$$

valid? Here, ***s*** is the standard deviation of the data and ***N*** is the sample size.

3. If $\bar{Y}$ is **not** the best estimator for ***C***, what is a better estimator for ***C*** (for example, median, midrange, midmean)?

4. If there is a better estimator, $\hat{C}$, what is its uncertainty? That is, what is $s_{\hat{C}}$?

EDA and the routine checking of underlying assumptions provides insight into all of the above.

1. Location and variation checks provide information as to whether ***C*** is really constant.

2. Distributional checks indicate whether $\bar{Y}$ is the best estimator. Techniques for distributional checking include histograms, normal probability plots, and probability plot correlation coefficient plots.

3. Randomness checks ascertain whether the usual

   $$s_{\bar{Y}} = s/\sqrt{N}$$

   is valid.

4. Distributional tests assist in determining a better estimator, if needed.

5. Simulator tools (namely bootstrapping) provide values for the uncertainty of alternative estimators.

Assumptions not satisfied

If one or more of the above assumptions is not satisfied, then we use EDA techniques, or some mix of EDA and classical techniques, to find a more appropriate model for the data. That is,

$$\boldsymbol{Y_i = D + E_i}$$

where ***D*** is the deterministic part and ***E*** is an error component.

If the data are not random, then we may investigate fitting some simple time series models to the data. If the constant location and scale assumptions are violated, we may need to investigate the measurement process to see if there is an explanation.

The assumptions on the error term are still quite relevant in the sense that for an appropriate model the error

component should follow the assumptions. The criterion for validating the model, or comparing competing models, is framed in terms of these assumptions.

Multivariable data

Although the case studies in this chapter utilize univariate data, the assumptions above are relevant for multivariable data as well.

If the data are not univariate, then we are trying to find a model

$$Y_i = F(X_1, ..., X_k) + E_i$$

where F is some function based on one or more variables. The error component, which is a univariate data set, of a good model should satisfy the assumptions given above. The criterion for validating and comparing models is based on how well the error component follows these assumptions.

The load cell calibration case study in the process modeling chapter shows an example of this in the regression context.

First three case studies utilize data with known characteristics

The first three case studies utilize data that are randomly generated from the following distributions:

- normal distribution with mean 0 and standard deviation 1
- uniform distribution with mean 0 and standard deviation $\sqrt{1/12}$ (uniform over the interval (0,1))
- random walk

The other univariate case studies utilize data from scientific processes. The goal is to determine if

$$Y_i = C + E_i$$

is a reasonable model. This is done by testing the underlying assumptions. If the assumptions are satisfied, then an estimate of C and an estimate of the uncertainty of C are computed. If the assumptions are not satisfied, we attempt to find a model where the error component does satisfy the underlying assumptions.

Graphical methods that are applied to the data

To test the underlying assumptions, each data set is analyzed using four graphical methods that are particularly suited for this purpose:

1. run sequence plot which is useful for detecting shifts of location or scale

2. lag plot which is useful for detecting non-randomness in the data

3. histogram which is useful for trying to determine the underlying distribution

4. normal probability plot for deciding whether the data follow the normal distribution

There are a number of other techniques for addressing the underlying assumptions. However, the four plots listed above provide an excellent opportunity for addressing all of the assumptions on a single page of graphics.

Additional graphical techniques are used in certain case studies to develop models that do have error components that satisfy the underlying assumptions.

Quantitative methods that are applied to the data

The normal and uniform random number data sets are also analyzed with the following quantitative techniques, which are explained in more detail in an earlier section:

1. Summary statistics which include:
   - mean
   - standard deviation
   - autocorrelation coefficient to test for randomness
   - normal and uniform probability plot correlation coefficients (ppcc) to test for a normal or uniform distribution, respectively
   - Wilk-Shapiro test for a normal distribution

2. Linear fit of the data as a function of time to assess drift (test for fixed location)

3. Bartlett test for fixed variance

4. Autocorrelation plot and coefficient to test for randomness

5. Runs test to test for lack of randomness

6. Anderson-Darling test for a normal distribution

7. Grubbs test for outliers

8. Summary report

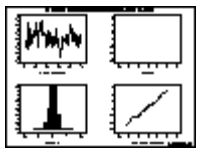
Although the graphical methods applied to the normal and uniform random numbers are sufficient to assess the validity of the underlying assumptions, the quantitative techniques are used to show the different flavor of the graphical and quantitative approaches.

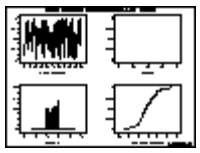
The remaining case studies intermix one or more of these

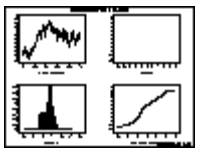
quantitative techniques into the analysis where appropriate.

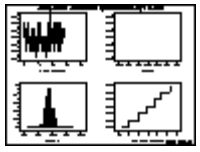
1.4.2. Case Studies

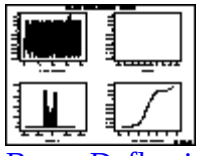
Univariate $Y_i = C + E_i$

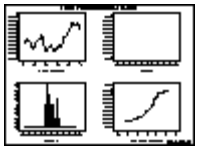

Normal Random Numbers

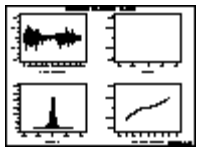

Uniform Random Numbers

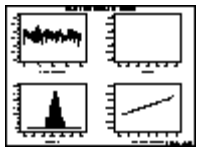

Random Walk


Josephson Junction Cryothermometry

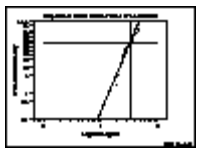

Beam Deflections


Filter Transmittance

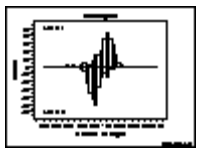

Standard Resistor


Heat Flow Meter 1

Reliability


Fatigue Life of Aluminum Alloy Specimens

Multi-Factor


Ceramic Strength

1.4.2.1. Normal Random Numbers

Normal Random Numbers

This example illustrates the univariate analysis of a set of normal random numbers.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.1.1. Background and Data

Generation The normal random numbers used in this case study are from a Rand Corporation publication.

The motivation for studying a set of normal random numbers is to illustrate the ideal case where all four underlying assumptions hold.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following is the set of normal random numbers used for this case study.

```
-1.2760 -1.2180 -0.4530 -0.3500  0.7230
 0.6760 -1.0990 -0.3140 -0.3940 -0.6330
-0.3180 -0.7990 -1.6640  1.3910  0.3820
 0.7330  0.6530  0.2190 -0.6810  1.1290
-1.3770 -1.2570  0.4950 -0.1390 -0.8540
 0.4280 -1.3220 -0.3150 -0.7320 -1.3480
 2.3340 -0.3370 -1.9550 -0.6360 -1.3180
-0.4330  0.5450  0.4280 -0.2970  0.2760
-1.1360  0.6420  3.4360 -1.6670  0.8470
-1.1730 -0.3550  0.0350  0.3590  0.9300
 0.4140 -0.0110  0.6660 -1.1320 -0.4100
-1.0770  0.7340  1.4840 -0.3400  0.7890
-0.4940  0.3640 -1.2370 -0.0440 -0.1110
-0.2100  0.9310  0.6160 -0.3770 -0.4330
 1.0480  0.0370  0.7590  0.6090 -2.0430
-0.2900  0.4040 -0.5430  0.4860  0.8690
 0.3470  2.8160 -0.4640 -0.6320 -1.6140
 0.3720 -0.0740 -0.9160  1.3140 -0.0380
 0.6370  0.5630 -0.1070  0.1310 -1.8080
-1.1260  0.3790  0.6100 -0.3640 -2.6260
 2.1760  0.3930 -0.9240  1.9110 -1.0400
-1.1680  0.4850  0.0760 -0.7690  1.6070
-1.1850 -0.9440 -1.6040  0.1850 -0.2580
-0.3000 -0.5910 -0.5450  0.0180 -0.4850
 0.9720  1.7100  2.6820  2.8130 -1.5310
-0.4900  2.0710  1.4440 -1.0920  0.4780
 1.2100  0.2940 -0.2480  0.7190  1.1030
 1.0900  0.2120 -1.1850 -0.3380 -1.1340
 2.6470  0.7770  0.4500  2.2470  1.1510
-1.6760  0.3840  1.1330  1.3930  0.8140
 0.3980  0.3180 -0.9280  2.4160 -0.9360
 1.0360  0.0240 -0.5600  0.2030 -0.8710
 0.8460 -0.6990 -0.3680  0.3440 -0.9260
-0.7970 -1.4040 -1.4720 -0.1180  1.4560
 0.6540 -0.9550  2.9070  1.6880  0.7520
-0.4340  0.7460  0.1490 -0.1700 -0.4790
 0.5220  0.2310 -0.6190 -0.2650  0.4190
 0.5580 -0.5490  0.1920 -0.3340  1.3730
-1.2880 -0.5390 -0.8240  0.2440 -1.0700
 0.0100  0.4820 -0.4690 -0.0900  1.1710
 1.3720  1.7690 -1.0570  1.6460  0.4810
-0.6000 -0.5920  0.6100 -0.0960 -1.3750
```

```
 0.8540 -0.5350  1.6070  0.4280 -0.6150
 0.3310 -0.3360 -1.1520  0.5330 -0.8330
-0.1480 -1.1440  0.9130  0.6840  1.0430
 0.5540 -0.0510 -0.9440 -0.4400 -0.2120
-1.1480 -1.0560  0.6350 -0.3280 -1.2210
 0.1180 -2.0450 -1.9770 -1.1330  0.3380
 0.3480  0.9700 -0.0170  1.2170 -0.9740
-1.2910 -0.3990 -1.2090 -0.2480  0.4800
 0.2840  0.4580  1.3070 -1.6250 -0.6290
-0.5040 -0.0560 -0.1310  0.0480  1.8790
-1.0160  0.3600 -0.1190  2.3310  1.6720
-1.0530  0.8400 -0.2460  0.2370 -1.3120
 1.6030 -0.9520 -0.5660  1.6000  0.4650
 1.9510  0.1100  0.2510  0.1160 -0.9570
-0.1900  1.4790 -0.9860  1.2490  1.9340
 0.0700 -1.3580 -1.2460 -0.9590 -1.2970
-0.7220  0.9250  0.7830 -0.4020  0.6190
 1.8260  1.2720 -0.9450  0.4940  0.0500
-1.6960  1.8790  0.0630  0.1320  0.6820
 0.5440 -0.4170 -0.6660 -0.1040 -0.2530
-2.5430 -1.3330  1.9870  0.6680  0.3600
 1.9270  1.1830  1.2110  1.7650  0.3500
-0.3590  0.1930 -1.0230 -0.2220 -0.6160
-0.0600 -1.3190  0.7850 -0.4300 -0.2980
 0.2480 -0.0880 -1.3790  0.2950 -0.1150
-0.6210 -0.6180  0.2090  0.9790  0.9060
-0.0990 -1.3760  1.0470 -0.8720 -2.2000
-1.3840  1.4250 -0.8120  0.7480 -1.0930
-0.4630 -1.2810 -2.5140  0.6750  1.1450
 1.0830 -0.6670 -0.2230 -1.5920 -1.2780
 0.5030  1.4340  0.2900  0.3970 -0.8370
-0.9730 -0.1200 -1.5940 -0.9960 -1.2440
-0.8570 -0.3710 -0.2160  0.1480 -2.1060
-1.4530  0.6860 -0.0750 -0.2430 -0.1700
-0.1220  1.1070 -1.0390 -0.6360 -0.8600
-0.8950 -1.4580 -0.5390 -0.1590 -0.4200
 1.6320  0.5860 -0.4680 -0.3860 -0.3540
 0.2030 -1.2340  2.3810 -0.3880 -0.0630
 2.0720 -1.4450 -0.6800  0.2240 -0.1200
 1.7530 -0.5710  1.2230 -0.1260  0.0340
-0.4350 -0.3750 -0.9850 -0.5850 -0.2030
-0.5560  0.0240  0.1260  1.2500 -0.6150
 0.8760 -1.2270 -2.6470 -0.7450  1.7970
-1.2310  0.5470 -0.6340 -0.8360 -0.7190
 0.8330  1.2890 -0.0220 -0.4310  0.5820
 0.7660 -0.5740 -1.1530  0.5200 -1.0180
-0.8910  0.3320 -0.4530 -1.1270  2.0850
-0.7220 -1.5080  0.4890 -0.4960 -0.0250
 0.6440 -0.2330 -0.1530  1.0980  0.7570
-0.0390 -0.4600  0.3930  2.0120  1.3560
 0.1050 -0.1710 -0.1100 -1.1450  0.8780
-0.9090 -0.3280  1.0210 -1.6130  1.5600
-1.1920  1.7700 -0.0030  0.3690  0.0520
 0.6470  1.0290  1.5260  0.2370 -1.3280
-0.0420  0.5530  0.7700  0.3240 -0.4890
-0.3670  0.3780  0.6010 -1.9960 -0.7380
 0.4980  1.0720  1.5670  0.3020  1.1570
-0.7200  1.4030  0.6980 -0.3700 -0.5510
```

1.4.2.1.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

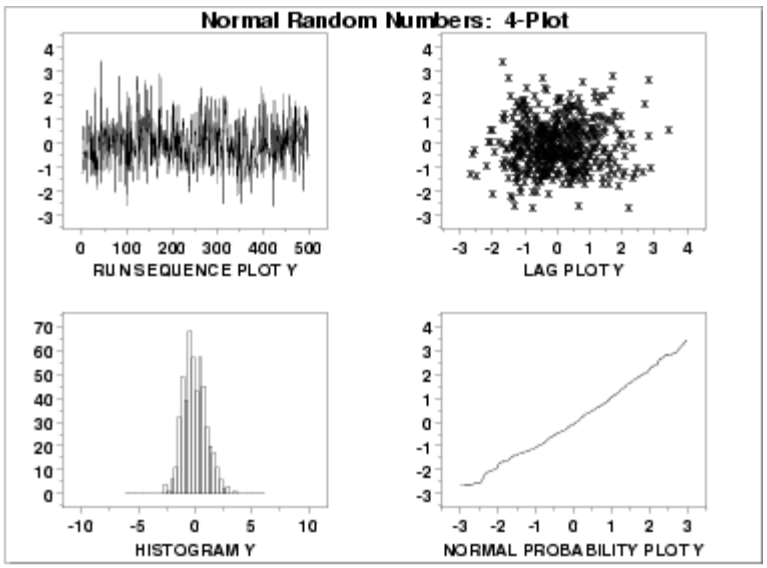
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time. The run sequence plot does not show any obvious outliers.

2. The lag plot (upper right) does not indicate any non-random pattern in the data.

3. The histogram (lower left) shows that the data are reasonably symmetric, there do not appear to be significant outliers in the tails, and that it is reasonable to assume that the data are from approximately a normal distribution.

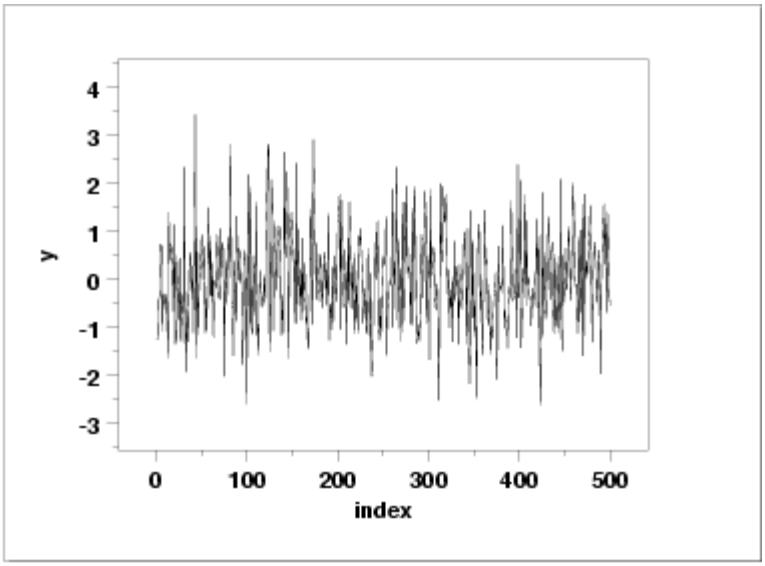
4. The normal probability plot (lower right) verifies that an assumption of normality is in fact reasonable.

From the above plots, we conclude that the underlying assumptions are valid and the data follow approximately a normal distribution. Therefore, the confidence interval form given previously is appropriate for quantifying the uncertainty of the population mean. The numerical values for this model are given in the Quantitative Output and Interpretation section.

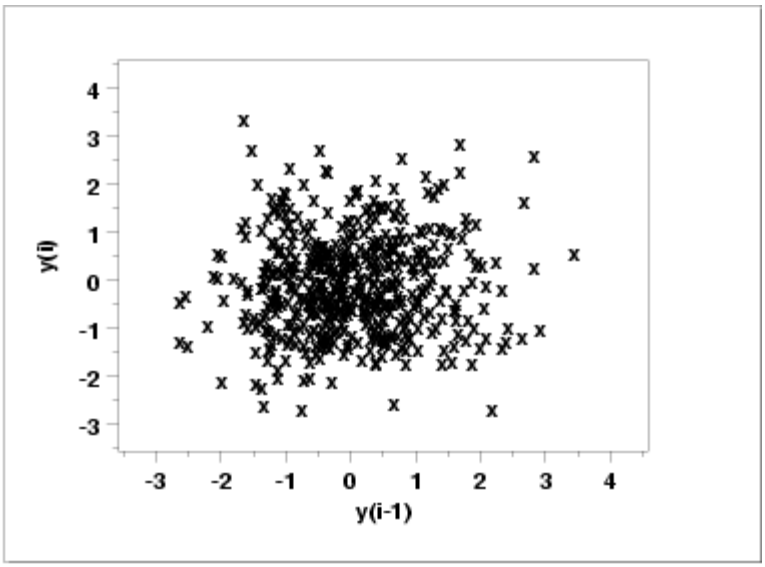
Individual Plots

Although it is usually not necessary, the plots can be generated individually to give more detail.

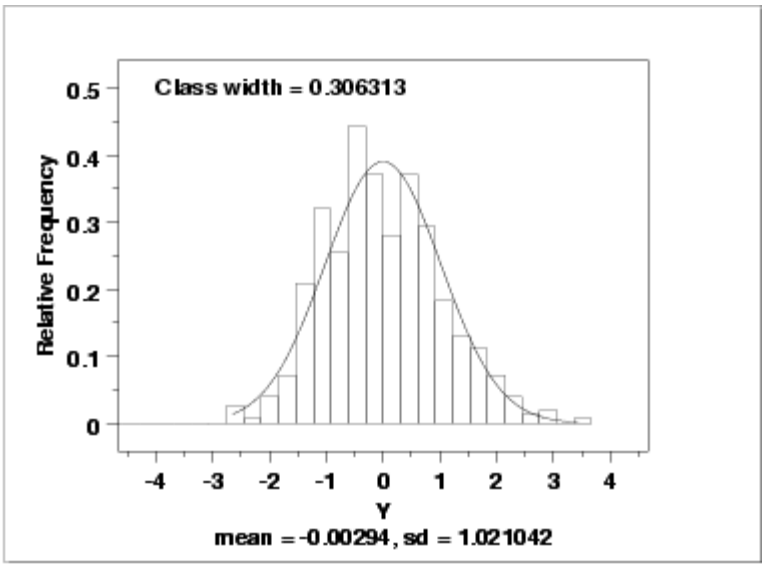
Run Sequence Plot



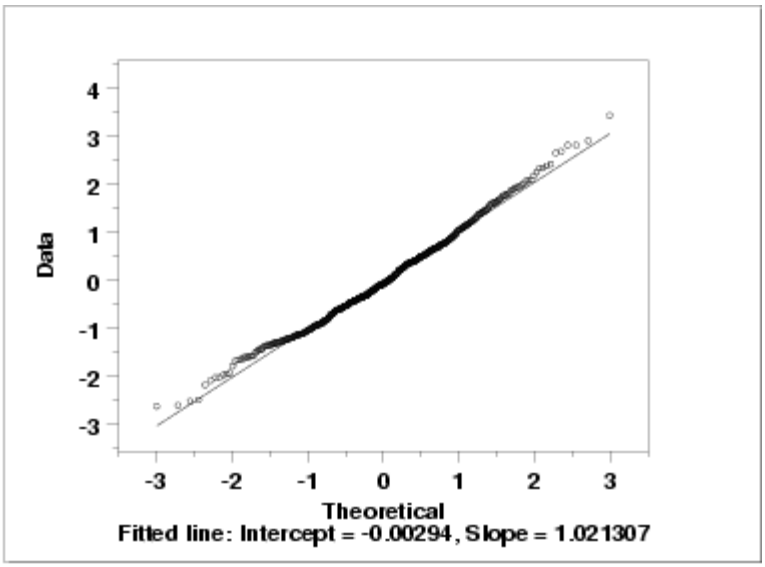
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot



1.4.2.1.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics are computed from the data.

```
Sample size  = 500
Mean         =  -0.2935997E-02
Median       =  -0.9300000E-01
Minimum      =  -0.2647000E+01
Maximum      =   0.3436000E+01
Range        =   0.6083000E+01
Stan. Dev.   =   0.1021041E+01
```

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

```
      Coefficient     Estimate      Stan. Error
t-Value
           B0        0.699127E-02     0.9155E-01
0.0764
           B1       -0.396298E-04     0.3167E-03
-0.1251

      Residual Standard Deviation = 1.02205
      Residual Degrees of Freedom = 498
```

The absolute value of the *t*-value for the slope parameter is smaller than the critical value of $t_{0.975,498} = 1.96$. Thus, we conclude that the slope is not different from zero at the 0.05 significance level.

Variation

One simple way to detect a change in variation is with Bartlett's test, after dividing the data set into several equal-sized intervals. The choice of the number of intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

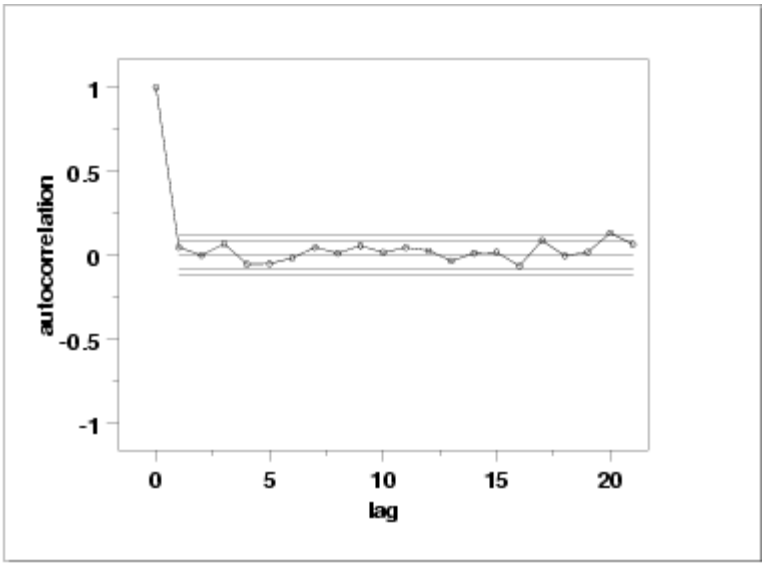
Test statistic: $T = 2.373660$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $X^2_{1-\alpha,k-1} = 7.814728$
Critical region: Reject H_0 if $T > 7.814728$

In this case, Bartlett's test indicates that the variances are not significantly different in the four intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests including the lag plot shown on the previous page.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95 % and 99 % confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of most interest, is 0.045. The critical values at the 5% significance level are -0.087 and 0.087. Since 0.045 is within the critical region, the lag 1 autocorrelation is not statistically significant, so there is no evidence of non-randomness.

A common test for randomness is the runs test.

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a random manner

Test statistic: $Z = -1.0744$
Significance level: $\alpha = 0.05$
Critical value: $Z_{1-\alpha/2} = 1.96$
Critical region: Reject H_0 if $|Z| > 1.96$

The runs test fails to reject the null hypothesis that the data were produced in a random manner.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot, or PPCC. For this data set the PPCC based on a normal distribution is 0.996. Since the PPCC is greater than the critical value of 0.987 (this is a tabulated value), the normality assumption is not rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. The results of the Anderson-Darling test follow.

H_0: the data are normally distributed
H_a: the data are not normally distributed

Adjusted test statistic: $A^2 = 1.0612$
Significance level: $\alpha = 0.05$
Critical value: 0.787
Critical region: Reject H_0 if $A^2 > 0.787$

The Anderson-Darling test rejects the normality assumption at the 0.05 significance level.

Outlier Analysis

A test for outliers is the Grubbs test.

H_0: there are no outliers in the data
H_a: the maximum value is an outlier

Test statistic: $G = 3.368068$
Significance level: $\alpha = 0.05$
Critical value for an upper one-tailed test: 3.863087
Critical region: Reject H_0 if $G > 3.863087$

For this data set, Grubbs' test does not detect any outliers at the 0.05 significance level.

Model

Since the underlying assumptions were validated both graphically and analytically, we conclude that a reasonable model for the data is:

$$Y_i = C + E_i$$

where C is the estimated value of the mean, -0.00294. We can express the uncertainty for C as a 95 % confidence interval (-0.09266, 0.08678).

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
Analysis of 500 normal random numbers

1: Sample Size                          = 500

2: Location
   Mean                                 = -
```

```
0.00294
    Standard Deviation of Mean            =
0.045663
    95% Confidence Interval for Mean      = (-
0.09266,0.086779)
    Drift with respect to location?       = NO

 3: Variation
    Standard Deviation                    =
1.021042
    95% Confidence Interval for SD        =
(0.961437,1.088585)
    Drift with respect to variation?
    (based on Bartletts test on quarters
    of the data)                           = NO

 4: Data are Normal?
      (as tested by Anderson-Darling)     = YES

 5: Randomness
    Autocorrelation                       =
0.045059
    Data are Random?
      (as measured by autocorrelation)    = YES

 6: Statistical Control
    (i.e., no drift in location or scale,
    data are random, distribution is
    fixed, here we are testing only for
    fixed normal)
    Data Set is in Statistical Control?   = YES

 7: Outliers?
    (as determined by Grubbs' test)       = NO
```

1.4.2.1.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale, and the data seem to follow a normal distribution.
3. Generate the individual plots.	

1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf. 4. Generate a normal probability plot.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is a good distribution for these data. 4. The normal probability plot verifies that the normal distribution is a reasonable distribution for these data.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Barltett's test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Check for normality by computing the normal probability plot correlation coefficient. 6. Check for outliers using Grubbs' test. 7. Print a univariate report (this	1. The summary statistics table displays 25+ statistics. 2. The mean is -0.00294 and a 95% confidence interval is (-0.093,0.087). The linear fit indicates no drift in location since the slope parameter is statistically not significant. 3. The standard deviation is 1.02 with a 95% confidence interval of (0.96,1.09). Bartlett's test indicates no significant change in variation. 4. The lag 1 autocorrelation is 0.04. From the autocorrelation plot, this is within the 95% confidence interval bands.

assumes steps 2 thru 6 have already been run).	5. The normal probability plot correlation coefficient is 0.996. At the 5% level, we cannot reject the normality assumption. 6. Grubbs' test detects no outliers at the 5% level. 7. The results are summarized in a convenient report.

1.4.2.2. Uniform Random Numbers

Uniform Random Numbers

This example illustrates the univariate analysis of a set of uniform random numbers.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.2.1. Background and Data

Generation The uniform random numbers used in this case study are from a Rand Corporation publication.

The motivation for studying a set of uniform random numbers is to illustrate the effects of a known underlying non-normal distribution.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following is the set of uniform random numbers used for this case study.

.100973	.253376	.520135	.863467	.354876
.809590	.911739	.292749	.375420	.480564
.894742	.962480	.524037	.206361	.040200
.822916	.084226	.895319	.645093	.032320
.902560	.159533	.476435	.080336	.990190
.252909	.376707	.153831	.131165	.886767
.439704	.436276	.128079	.997080	.157361
.476403	.236653	.989511	.687712	.171768
.660657	.471734	.072768	.503669	.736170
.658133	.988511	.199291	.310601	.080545
.571824	.063530	.342614	.867990	.743923
.403097	.852697	.760202	.051656	.926866
.574818	.730538	.524718	.623885	.635733
.213505	.325470	.489055	.357548	.284682
.870983	.491256	.737964	.575303	.529647
.783580	.834282	.609352	.034435	.273884
.985201	.776714	.905686	.072210	.940558
.609709	.343350	.500739	.118050	.543139
.808277	.325072	.568248	.294052	.420152
.775678	.834529	.963406	.288980	.831374
.670078	.184754	.061068	.711778	.886854
.020086	.507584	.013676	.667951	.903647
.649329	.609110	.995946	.734887	.517649
.699182	.608928	.937856	.136823	.478341
.654811	.767417	.468509	.505804	.776974
.730395	.718640	.218165	.801243	.563517
.727080	.154531	.822374	.211157	.825314
.385537	.743509	.981777	.402772	.144323
.600210	.455216	.423796	.286026	.699162
.680366	.252291	.483693	.687203	.766211
.399094	.400564	.098932	.050514	.225685
.144642	.756788	.962977	.882254	.382145
.914991	.452368	.479276	.864616	.283554
.947508	.992337	.089200	.803369	.459826
.940368	.587029	.734135	.531403	.334042
.050823	.441048	.194985	.157479	.543297
.926575	.576004	.088122	.222064	.125507
.374211	.100020	.401286	.074697	.966448
.943928	.707258	.636064	.932916	.505344
.844021	.952563	.436517	.708207	.207317
.611969	.044626	.457477	.745192	.433729
.653945	.959342	.582605	.154744	.526695

.270799	.535936	.783848	.823961	.011833
.211594	.945572	.857367	.897543	.875462
.244431	.911904	.259292	.927459	.424811
.621397	.344087	.211686	.848767	.030711
.205925	.701466	.235237	.831773	.208898
.376893	.591416	.262522	.966305	.522825
.044935	.249475	.246338	.244586	.251025
.619627	.933565	.337124	.005499	.765464
.051881	.599611	.963896	.546928	.239123
.287295	.359631	.530726	.898093	.543335
.135462	.779745	.002490	.103393	.598080
.839145	.427268	.428360	.949700	.130212
.489278	.565201	.460588	.523601	.390922
.867728	.144077	.939108	.364770	.617429
.321790	.059787	.379252	.410556	.707007
.867431	.715785	.394118	.692346	.140620
.117452	.041595	.660000	.187439	.242397
.118963	.195654	.143001	.758753	.794041
.921585	.666743	.680684	.962852	.451551
.493819	.476072	.464366	.794543	.590479
.003320	.826695	.948643	.199436	.168108
.513488	.881553	.015403	.545605	.014511
.980862	.482645	.240284	.044499	.908896
.390947	.340735	.441318	.331851	.623241
.941509	.498943	.548581	.886954	.199437
.548730	.809510	.040696	.382707	.742015
.123387	.250162	.529894	.624611	.797524
.914071	.961282	.966986	.102591	.748522
.053900	.387595	.186333	.253798	.145065
.713101	.024674	.054556	.142777	.938919
.740294	.390277	.557322	.709779	.017119
.525275	.802180	.814517	.541784	.561180
.993371	.430533	.512969	.561271	.925536
.040903	.116644	.988352	.079848	.275938
.171539	.099733	.344088	.461233	.483247
.792831	.249647	.100229	.536870	.323075
.754615	.020099	.690749	.413887	.637919
.763558	.404401	.105182	.161501	.848769
.091882	.009732	.825395	.270422	.086304
.833898	.737464	.278580	.900458	.549751
.981506	.549493	.881997	.918707	.615068
.476646	.731895	.020747	.677262	.696229
.064464	.271246	.701841	.361827	.757687
.649020	.971877	.499042	.912272	.953750
.587193	.823431	.540164	.405666	.281310
.030068	.227398	.207145	.329507	.706178
.083586	.991078	.542427	.851366	.158873
.046189	.755331	.223084	.283060	.326481
.333105	.914051	.007893	.326046	.047594
.119018	.538408	.623381	.594136	.285121
.590290	.284666	.879577	.762207	.917575
.374161	.613622	.695026	.390212	.557817
.651483	.483470	.894159	.269400	.397583
.911260	.717646	.489497	.230694	.541374
.775130	.382086	.864299	.016841	.482774
.519081	.398072	.893555	.195023	.717469
.979202	.885521	.029773	.742877	.525165
.344674	.218185	.931393	.278817	.570568

1.4.2.2.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

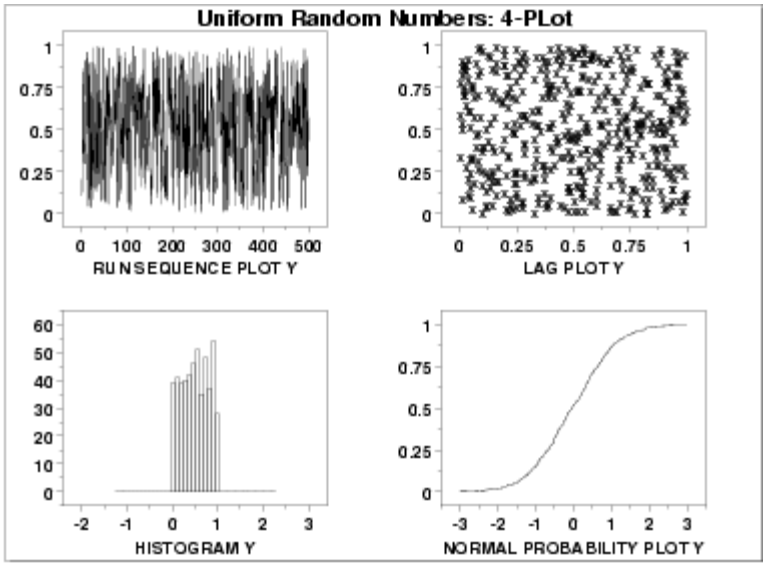
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.

2. The lag plot (upper right) does not indicate any non-random pattern in the data.

3. The histogram shows that the frequencies are relatively flat across the range of the data. This suggests that the uniform distribution might provide a better distributional fit than the normal distribution.

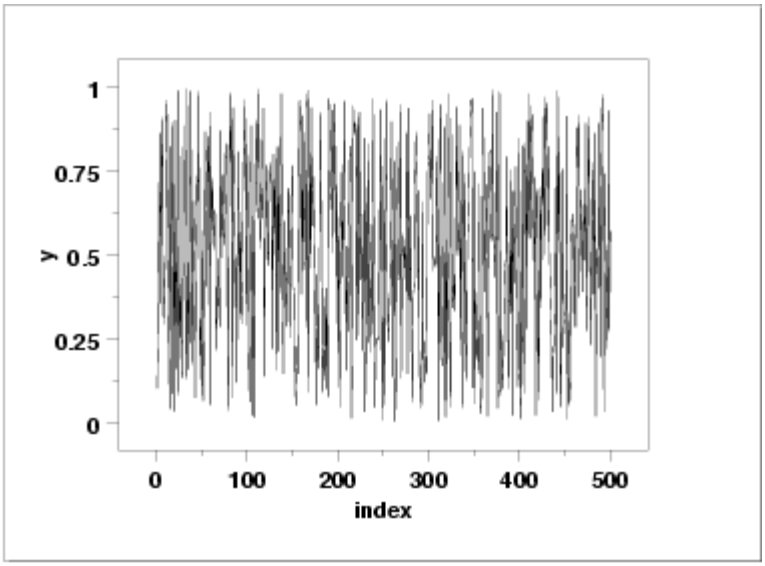
4. The normal probability plot verifies that an assumption of normality is not reasonable. In this case, the 4-plot should be followed up by a uniform probability plot to determine if it provides a better fit to the data. This is shown below.

From the above plots, we conclude that the underlying assumptions are valid. Therefore, the model $\boldsymbol{Y_i = C + E_i}$ is valid. However, since the data are not normally distributed, using the mean as an estimate of C and the confidence interval cited above for quantifying its uncertainty are not valid or appropriate.

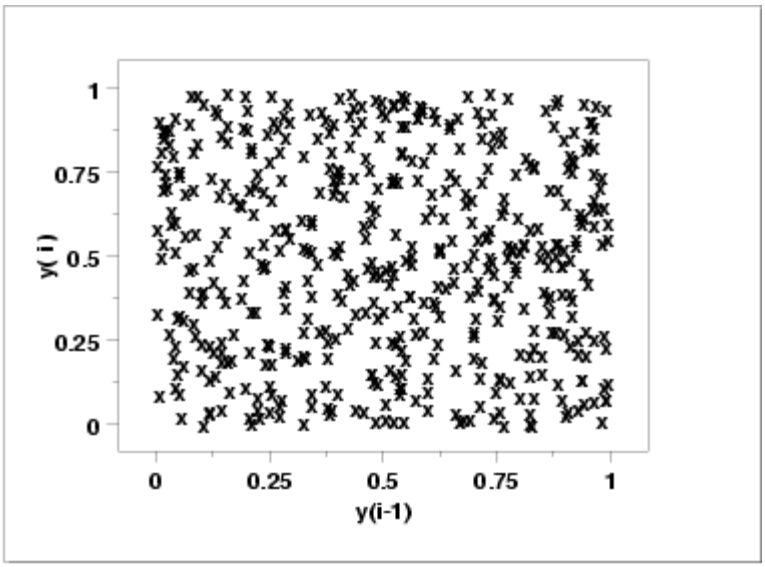
Individual Plots

Although it is usually not necessary, the plots can be generated individually to give more detail.

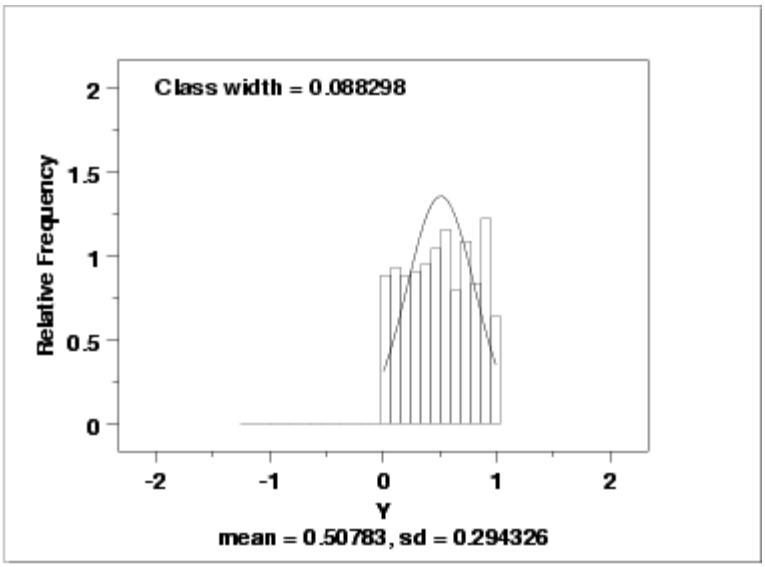
Run Sequence Plot



Lag Plot

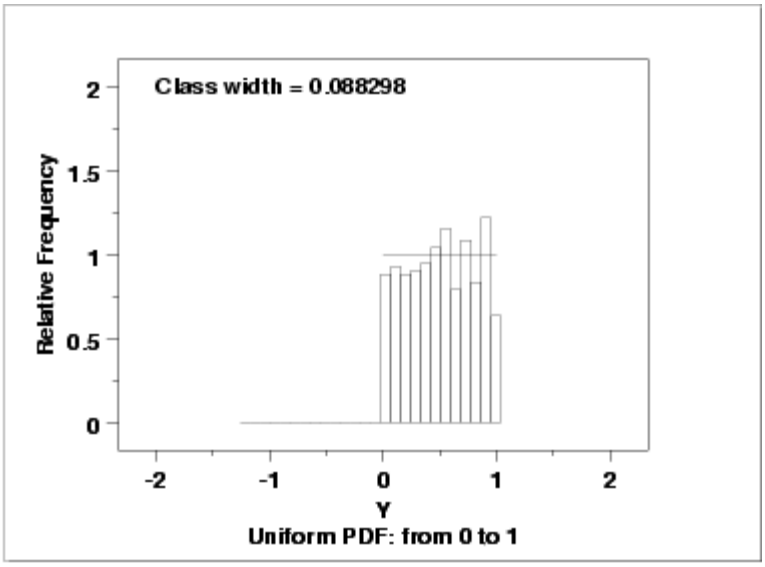


Histogram (with overlaid Normal PDF)



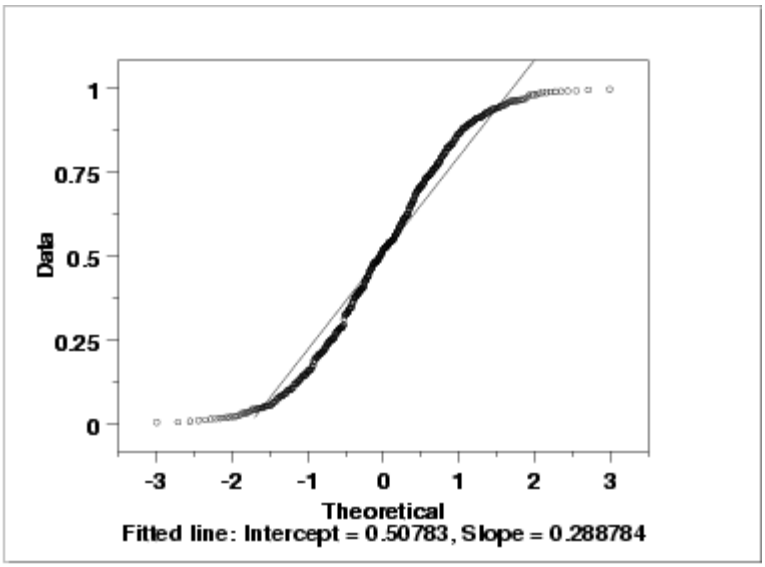
This plot shows that a normal distribution is a poor fit. The flatness of the histogram suggests that a uniform distribution might be a better fit.

Histogram (with overlaid Uniform PDF)



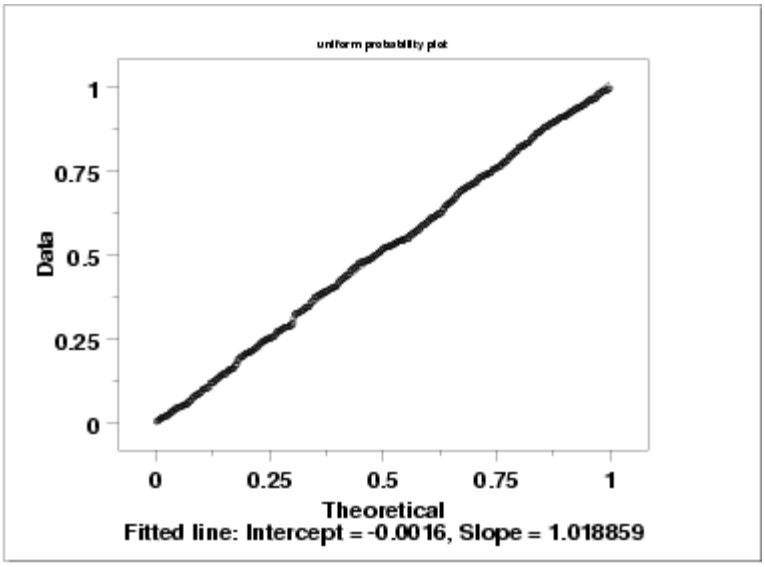
Since the histogram from the 4-plot suggested that the uniform distribution might be a good fit, we overlay a uniform distribution on top of the histogram. This indicates a much better fit than a normal distribution.

Normal Probability Plot



As with the histogram, the normal probability plot shows that the normal distribution does not fit these data well.

Uniform Probability Plot



Since the above plots suggested that a uniform distribution might be appropriate, we generate a uniform probability plot. This plot shows that the uniform distribution provides an excellent fit to the data.

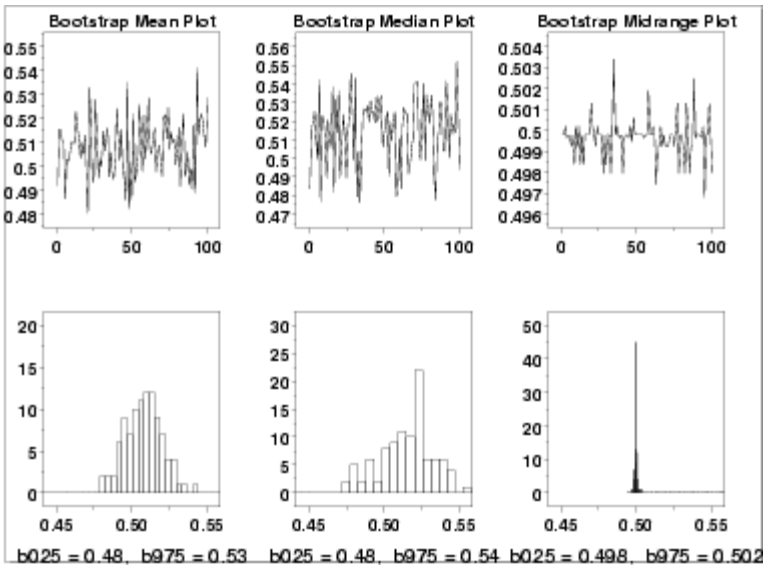
Better Model

Since the data follow the underlying assumptions, but with a uniform distribution rather than a normal distribution, we would still like to characterize ***C*** by a typical value plus or minus a confidence interval. In this case, we would like to find a location estimator with the smallest variability.

The bootstrap plot is an ideal tool for this purpose. The following plots show the bootstrap plot, with the

corresponding histogram, for the mean, median, mid-range, and median absolute deviation.

Bootstrap Plots



Mid-Range is Best

From the above histograms, it is obvious that for these data, the mid-range is far superior to the mean or median as an estimate for location.

Using the mean, the location estimate is 0.507 and a 95% confidence interval for the mean is (0.482,0.534). Using the mid-range, the location estimate is 0.499 and the 95% confidence interval for the mid-range is (0.497,0.503).

Although the values for the location are similar, the difference in the uncertainty intervals is quite large.

Note that in the case of a uniform distribution it is known theoretically that the mid-range is the best linear unbiased estimator for location. However, in many applications, the most appropriate estimator will not be known or it will be mathematically intractable to determine a valid condfidence interval. The bootstrap provides a method for determining (and comparing) confidence intervals in these cases.

1.4.2.2.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics are computed for the data.

```
Sample size  = 500
Mean         =   0.5078304
Median       =   0.5183650
Minimum      =   0.0024900
Maximum      =   0.9970800
Range        =   0.9945900
Stan. Dev.   =   0.2943252
```

Because the graphs of the data indicate the data may not be normally distributed, we also compute two other statistics for the data, the normal PPCC and the uniform PPCC.

```
Normal PPCC  =   0.9771602
Uniform PPCC =   0.9995682
```

The uniform probability plot correlation coefficient (PPCC) value is larger than the normal PPCC value. This is evidence that the uniform distribution fits these data better than does a normal distribution.

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

```
      Coefficient      Estimate      Stan. Error
t-Value
            B0         0.522923         0.2638E-01
19.82
            B1        -0.602478E-04     0.9125E-04
-0.66

      Residual Standard Deviation = 0.2944917
      Residual Degrees of Freedom = 498
```

The *t*-value of the slope parameter, -0.66, is smaller than the critical value of $t_{0.975,498} = 1.96$. Thus, we conclude that the slope is not different from zero at the 0.05

significance level.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since we know this data set is not approximated well by the normal distribution, we use the alternative Levene test. In particular, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

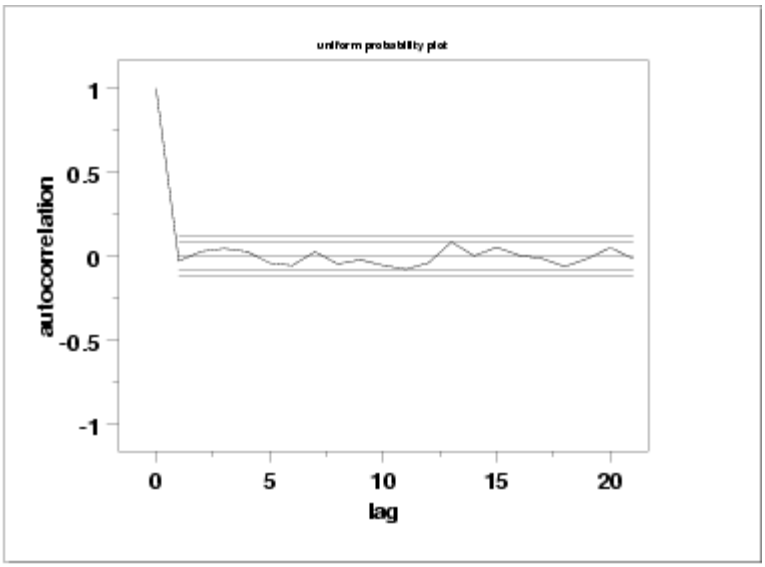
Test statistic: $W = 0.07983$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.623$
Critical region: Reject H_0 if $W > 2.623$

In this case, the Levene test indicates that the variances are not significantly different in the four intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests including the lag plot shown on the previous page.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted using 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of most interest, is 0.03. The critical values at the 5 % significance level are -0.087 and 0.087. This indicates that

the lag 1 autocorrelation is not statistically significant, so there is no evidence of non-randomness.

A common test for randomness is the runs test.

```
      H0:  the sequence was produced in a random
manner
      Ha:  the sequence was not produced in a
random manner

      Test statistic:  Z = 0.2686
      Significance level:  α = 0.05
      Critical value:  Z1-α/2 = 1.96
      Critical region:  Reject H0 if |Z| > 1.96
```

The runs test fails to reject the null hypothesis that the data were produced in a random manner.

Distributional Analysis

Probability plots are a graphical test of assessing whether a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot, or PPCC. For this data set the PPCC based on a normal distribution is 0.977. Since the PPCC is less than the critical value of 0.987 (this is a tabulated value), the normality assumption is rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. The results of the Anderson-Darling test follow.

```
      H0:  the data are normally distributed
      Ha:  the data are not normally distributed

      Adjusted test statistic:  A2 = 5.765
      Significance level:  α = 0.05
      Critical value:  0.787
      Critical region:  Reject H0 if A2 > 0.787
```

The Anderson-Darling test rejects the normality assumption because the value of the test statistic, 5.765, is larger than the critical value of 0.787 at the 0.05 significance level.

Model

Based on the graphical and quantitative analysis, we use the model

$$\boldsymbol{Y_i = C + E_i}$$

where $\boldsymbol{C}$ is estimated by the mid-range and the uncertainty interval for $\boldsymbol{C}$ is based on a bootstrap analysis. Specifically,

$\boldsymbol{C}$ = 0.499
95% confidence limit for $\boldsymbol{C}$ = (0.497,0.503)

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
 Analysis for 500 uniform random numbers

 1: Sample Size                              = 500

 2: Location
    Mean                                     =
0.50783
    Standard Deviation of Mean               =
0.013163
    95% Confidence Interval for Mean         =
(0.48197,0.533692)
    Drift with respect to location?          = NO

 3: Variation
    Standard Deviation                       =
0.294326
    95% Confidence Interval for SD           =
(0.277144,0.313796)
    Drift with respect to variation?
    (based on Levene's test on quarters
    of the data)                              = NO

 4: Distribution
    Normal PPCC                              =
0.9771602
    Data are Normal?
      (as measured by Normal PPCC)           = NO

    Uniform PPCC                             =
0.9995682
    Data are Uniform?
      (as measured by Uniform PPCC)          = YES

 5: Randomness
    Autocorrelation                          = -
0.03099
    Data are Random?
      (as measured by autocorrelation)       = YES

 6: Statistical Control
    (i.e., no drift in location or scale,
    data is random, distribution is
    fixed, here we are testing only for
    fixed uniform)
    Data Set is in Statistical Control?      = YES
```

1.4.2.2.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale, and the data do not seem to follow a normal distribution.
3. Generate the individual plots.	

1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf. 4. Generate a histogram with an overlaid uniform pdf. 5. Generate a normal probability plot. 6. Generate a uniform probability plot.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is not a good distribution for these data. 4. The histogram indicates that a uniform distribution is a good distribution for these data. 5. The normal probability plot verifies that the normal distribution is not a reasonable distribution for these data. 6. The uniform probability plot verifies that the uniform distribution is a reasonable distribution for these data.
4. Generate the bootstrap plot. 1. Generate a bootstrap plot.	1. The bootstrap plot clearly shows the superiority of the mid-range over the mean and median as the location estimator of choice for this problem.
5. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the	1. The summary statistics table displays 25+ statistics. 2. The mean is 0.5078 and a 95% confidence interval is (0.482,0.534). The linear fit indicates no drift in location since the slope parameter

standard deviation, and detect drift in variation by dividing the data into quarters and computing Barltetts test for equal standard deviations.	is statistically not significant.
4. Check for randomness by generating an autocorrelation plot and a runs test.	3. The standard deviation is 0.29 with a 95% confidence interval of (0.277,0.314). Levene's test indicates no significant drift in variation.
5. Check for normality by computing the normal probability plot correlation coefficient.	4. The lag 1 autocorrelation is -0.03. From the autocorrelation plot, this is within the 95% confidence interval bands.
6. Print a univariate report (this assumes steps 2 thru 6 have already been run).	5. The uniform probability plot correlation coefficient is 0.9995. This indicates that the uniform distribution is a good fit.
	6. The results are summarized in a convenient report.

1.4.2.3. Random Walk

Random Walk This example illustrates the univariate analysis of a set of numbers derived from a random walk.

1. Background and Data
2. Test Underlying Assumptions
3. Develop Better Model
4. Validate New Model
5. Work This Example Yourself

1.4.2.3.1. Background and Data

Generation A random walk can be generated from a set of uniform random numbers by the formula:

$$R_i = \sum_{j=1}^{i}(U_j - 0.5)$$

where U is a set of uniform random numbers.

The motivation for studying a set of random walk data is to illustrate the effects of a known underlying autocorrelation structure (i.e., non-randomness) in the data.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following is the set of random walk numbers used for this case study.

```
-0.399027
-0.645651
-0.625516
-0.262049
-0.407173
-0.097583
 0.314156
 0.106905
-0.017675
-0.037111
 0.357631
 0.820111
 0.844148
 0.550509
 0.090709
 0.413625
-0.002149
 0.393170
 0.538263
 0.070583
 0.473143
 0.132676
 0.109111
-0.310553
 0.179637
-0.067454
-0.190747
-0.536916
-0.905751
-0.518984
-0.579280
-0.643004
-1.014925
-0.517845
```

```
-0.860484
-0.884081
-1.147428
-0.657917
-0.470205
-0.798437
-0.637780
-0.666046
-1.093278
-1.089609
-0.853439
-0.695306
-0.206795
-0.507504
-0.696903
-1.116358
-1.044534
-1.481004
-1.638390
-1.270400
-1.026477
-1.123380
-0.770683
-0.510481
-0.958825
-0.531959
-0.457141
-0.226603
-0.201885
-0.078000
 0.057733
-0.228762
-0.403292
-0.414237
-0.556689
-0.772007
-0.401024
-0.409768
-0.171804
-0.096501
-0.066854
 0.216726
 0.551008
 0.660360
 0.194795
-0.031321
 0.453880
 0.730594
 1.136280
 0.708490
 1.149048
 1.258757
 1.102107
 1.102846
 0.720896
 0.764035
 1.072312
 0.897384
 0.965632
 0.759684
 0.679836
 0.955514
 1.290043
 1.753449
 1.542429
 1.873803
 2.043881
 1.728635
 1.289703
 1.501481
 1.888335
 1.408421
 1.416005
 0.929681
 1.097632
 1.501279
 1.650608
 1.759718
 2.255664
 2.490551
 2.508200
```

2.707382
2.816310
3.254166
2.890989
2.869330
3.024141
3.291558
3.260067
3.265871
3.542845
3.773240
3.991880
3.710045
4.011288
4.074805
4.301885
3.956416
4.278790
3.989947
4.315261
4.200798
4.444307
4.926084
4.828856
4.473179
4.573389
4.528605
4.452401
4.238427
4.437589
4.617955
4.370246
4.353939
4.541142
4.807353
4.706447
4.607011
4.205943
3.756457
3.482142
3.126784
3.383572
3.846550
4.228803
4.110948
4.525939
4.478307
4.457582
4.822199
4.605752
5.053262
5.545598
5.134798
5.438168
5.397993
5.838361
5.925389
6.159525
6.190928
6.024970
5.575793
5.516840
5.211826
4.869306
4.912601
5.339177
5.415182
5.003303
4.725367
4.350873
4.225085
3.825104
3.726391
3.301088
3.767535
4.211463
4.418722
4.554786
4.987701
4.993045
5.337067

```
5.789629
5.726147
5.934353
5.641670
5.753639
5.298265
5.255743
5.500935
5.434664
5.588610
6.047952
6.130557
5.785299
5.811995
5.582793
5.618730
5.902576
6.226537
5.738371
5.449965
5.895537
6.252904
6.650447
7.025909
6.770340
7.182244
6.941536
7.368996
7.293807
7.415205
7.259291
6.970976
7.319743
6.850454
6.556378
6.757845
6.493083
6.824855
6.533753
6.410646
6.502063
6.264585
6.730889
6.753715
6.298649
6.048126
5.794463
5.539049
5.290072
5.409699
5.843266
5.680389
5.185889
5.451353
5.003233
5.102844
5.566741
5.613668
5.352791
5.140087
4.999718
5.030444
5.428537
5.471872
5.107334
5.387078
4.889569
4.492962
4.591042
4.930187
4.857455
4.785815
5.235515
4.865727
4.855005
4.920206
4.880794
4.904395
4.795317
5.163044
4.807122
```

```
5.246230
5.111000
5.228429
5.050220
4.610006
4.489258
4.399814
4.606821
4.974252
5.190037
5.084155
5.276501
4.917121
4.534573
4.076168
4.236168
3.923607
3.666004
3.284967
2.980621
2.623622
2.882375
3.176416
3.598001
3.764744
3.945428
4.408280
4.359831
4.353650
4.329722
4.294088
4.588631
4.679111
4.182430
4.509125
4.957768
4.657204
4.325313
4.338800
4.720353
4.235756
4.281361
3.795872
4.276734
4.259379
3.999663
3.544163
3.953058
3.844006
3.684740
3.626058
3.457909
3.581150
4.022659
4.021602
4.070183
4.457137
4.156574
4.205304
4.514814
4.055510
3.938217
4.180232
3.803619
3.553781
3.583675
3.708286
4.005810
4.419880
4.881163
5.348149
4.950740
5.199262
4.753162
4.640757
4.327090
4.080888
3.725953
3.939054
3.463728
3.018284
```

2.661061
3.099980
3.340274
3.230551
3.287873
3.497652
3.014771
3.040046
3.342226
3.656743
3.698527
3.759707
4.253078
4.183611
4.196580
4.257851
4.683387
4.224290
3.840934
4.329286
3.909134
3.685072
3.356611
2.956344
2.800432
2.761665
2.744913
3.037743
2.787390
2.387619
2.424489
2.247564
2.502179
2.022278
2.213027
2.126914
2.264833
2.528391
2.432792
2.037974
1.699475
2.048244
1.640126
1.149858
1.475253
1.245675
0.831979
1.165877
1.403341
1.181921
1.582379
1.632130
2.113636
2.163129
2.545126
2.963833
3.078901
3.055547
3.287442
2.808189
2.985451
3.181679
2.746144
2.517390
2.719231
2.581058
2.838745
2.987765
3.459642
3.458684
3.870956
4.324706
4.411899
4.735330
4.775494
4.681160
4.462470
3.992538
3.719936
3.427081
3.256588

```
3.462766
3.046353
3.537430
3.579857
3.931223
3.590096
3.136285
3.391616
3.114700
2.897760
2.724241
2.557346
2.971397
2.479290
2.305336
1.852930
1.471948
1.510356
1.633737
1.727873
1.512994
1.603284
1.387950
1.767527
2.029734
2.447309
2.321470
2.435092
2.630118
2.520330
2.578147
2.729630
2.713100
3.107260
2.876659
2.774242
3.185503
3.403148
3.392646
3.123339
3.164713
3.439843
3.321929
3.686229
3.203069
3.185843
3.204924
3.102996
3.496552
3.191575
3.409044
3.888246
4.273767
3.803540
4.046417
4.071581
3.916256
3.634441
4.065834
3.844651
3.915219
```

1.4.2.3.2. Test Underlying Assumptions

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

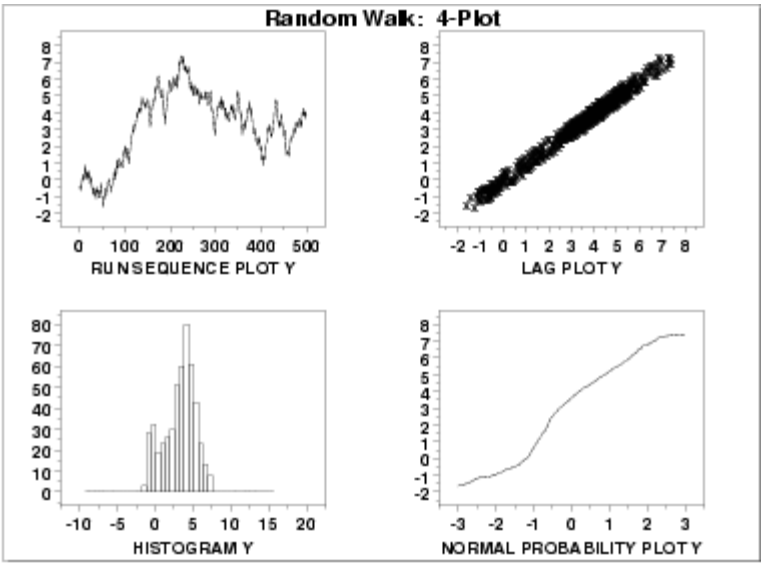
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid, with ***s*** denoting the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

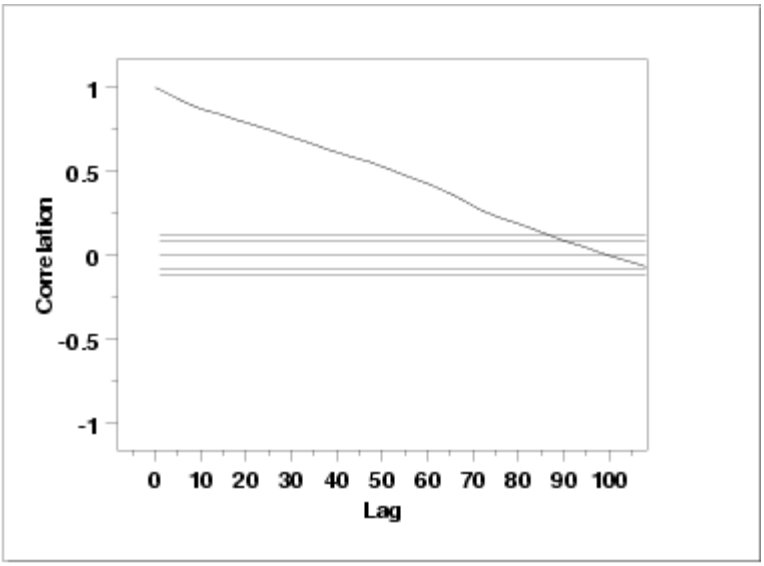
1. The run sequence plot (upper left) indicates significant shifts in location over time.
2. The lag plot (upper right) indicates significant non-randomness in the data.
3. When the assumptions of randomness and constant location and scale are not satisfied, the distributional assumptions are not meaningful. Therefore we do not attempt to make any interpretation of the histogram (lower left) or the normal probability plot (lower right).

From the above plots, we conclude that the underlying assumptions are seriously violated. Therefore the $\boldsymbol{Y_i = C + E_i}$ model is not valid.

When the randomness assumption is seriously violated, a time series model may be appropriate. The lag plot often suggests a reasonable model. For example, in this case the strongly linear appearance of the lag plot suggests a model fitting $\boldsymbol{Y_i}$ versus $\boldsymbol{Y_{i-1}}$ might be appropriate. When the data are non-random, it is helpful to supplement the lag plot with an autocorrelation plot and a spectral plot. Although in this case the lag plot is enough to suggest an appropriate model, we provide the autocorrelation and spectral plots for comparison.

Autocorrelation Plot

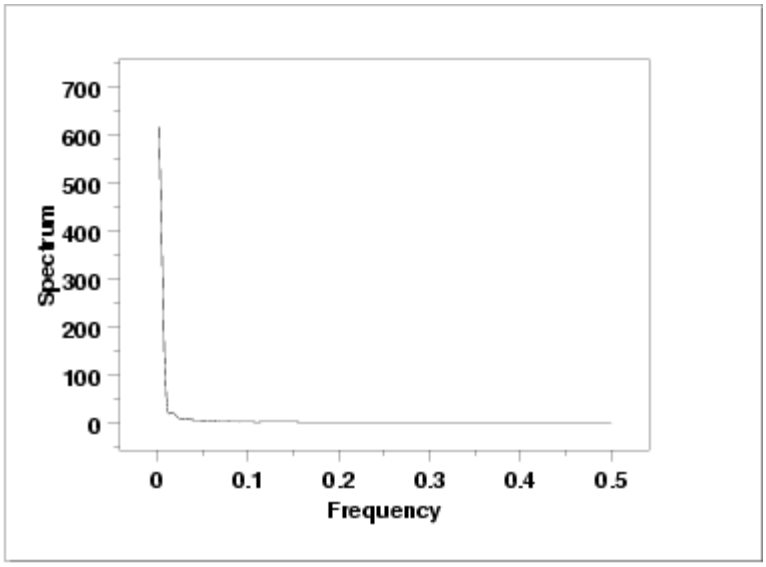
When the lag plot indicates significant non-randomness, it can be helpful to follow up with a an autocorrelation plot.



This autocorrelation plot shows significant autocorrelation at lags 1 through 100 in a linearly decreasing fashion.

Spectral Plot

Another useful plot for non-random data is the spectral plot.



This spectral plot shows a single dominant low frequency peak.

Quantitative Output

Although the 4-plot above clearly shows the violation of the assumptions, we supplement the graphical output with some quantitative measures.

Summary Statistics

As a first step in the analysis, common summary statistics are computed from the data.

```
Sample size  = 500
Mean         =   3.216681
Median       =   3.612030
Minimum      =  -1.638390
Maximum      =   7.415205
Range        =   9.053595
Stan. Dev.   =   2.078675
```

We also computed the autocorrelation to be 0.987, which is evidence of a very strong autocorrelation.

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

```
      Coefficient      Estimate      Stan. Error
t-Value
          B0          1.83351          0.1721
10.650
          B1          0.552164E-02     0.5953E-03
9.275

      Residual Standard Deviation = 1.9214
      Residual Degrees of Freedom = 498
```

The *t*-value of the slope parameter, 9.275, is larger than the critical value of $t_{0.975,498} = 1.96$. Thus, we conclude that the slope is different from zero at the 0.05 significance level.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since we know this data set is not approximated well by the normal distribution, we use the alternative Levene test. In particular, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

Test statistic: $W = 10.459$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.623$
Critical region: Reject H_0 if $W > 2.623$

In this case, the Levene test indicates that the variances are significantly different in the four intervals since the test statistic of 10.459 is greater than the 95 % critical value of 2.623. Therefore we conclude that the scale is not constant.

Randomness

Although the lag 1 autocorrelation coefficient above clearly shows the non-randomness, we show the output from a runs test as well.

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a random manner

Test statistic: $Z = -20.3239$
Significance level: $\alpha = 0.05$
Critical value: $Z_{1-\alpha/2} = 1.96$
Critical region: Reject H_0 if $|Z| > 1.96$

The runs test rejects the null hypothesis that the data were produced in a random manner at the 0.05 significance level.

Distributional Assumptions

Since the quantitative tests show that the assumptions of randomness and constant location and scale are not met, the distributional measures will not be meaningful. Therefore these quantitative tests are omitted.

1.4.2.3.3. Develop A Better Model

Lag Plot Suggests Better Model

Since the underlying assumptions did not hold, we need to develop a better model.

The lag plot showed a distinct linear pattern. Given the definition of the lag plot, Y_i versus Y_{i-1}, a good candidate model is a model of the form

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

Fit Output

The results of a linear fit of this model generated the following results.

Coefficient	Estimate	Stan. Error	*t*-Value
A_0	0.050165	0.024171	2.075
A_1	0.987087	0.006313	156.350

Residual Standard Deviation = 0.2931
Residual Degrees of Freedom = 497

The slope parameter, A_1, has a t value of 156.350 which is statistically significant. Also, the residual standard deviation is 0.2931. This can be compared to the standard deviation shown in the summary table, which is 2.078675. That is, the fit to the autoregressive model has reduced the variability by a factor of 7.

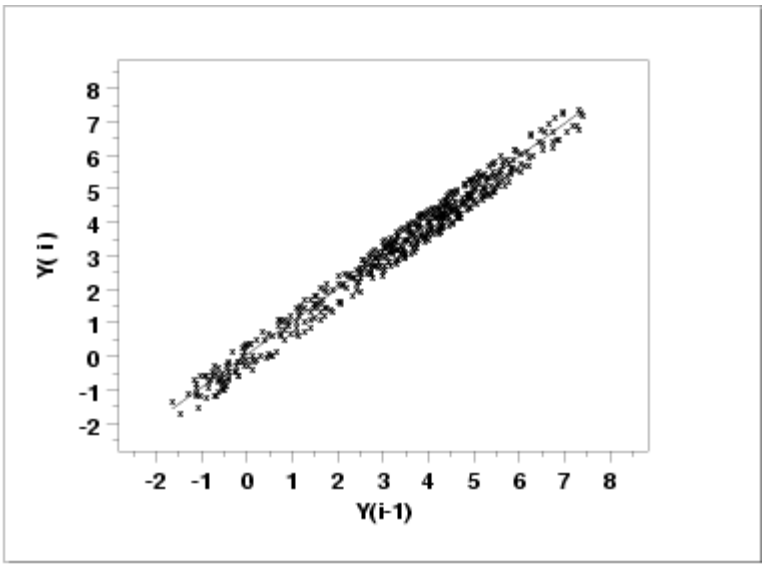
Time Series Model

This model is an example of a time series model. More extensive discussion of time series is given in the Process Monitoring chapter.

1.4.2.3.4. Validate New Model

Plot Predicted with Original Data

The first step in verifying the model is to plot the predicted values from the fit with the original data.

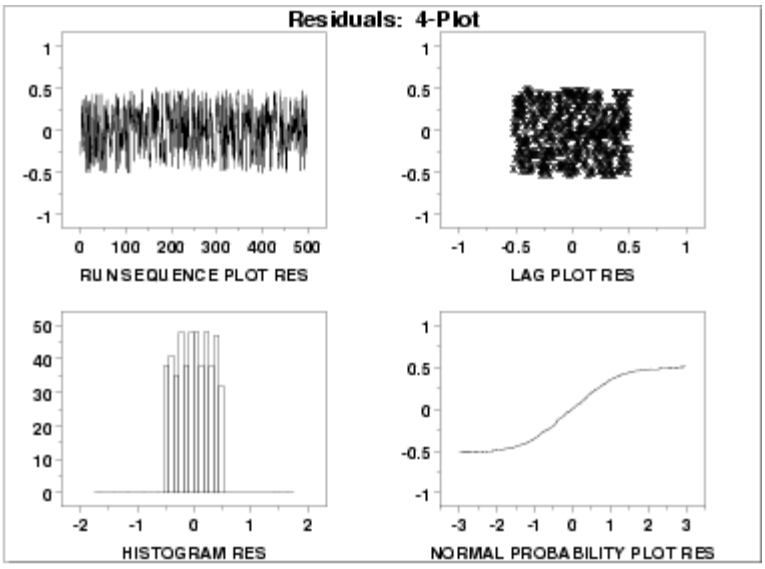


This plot indicates a reasonably good fit.

Test Underlying Assumptions on the Residuals

In addition to the plot of the predicted values, the residual standard deviation from the fit also indicates a significant improvement for the model. The next step is to validate the underlying assumptions for the error component, or residuals, from this model.

4-Plot of Residuals



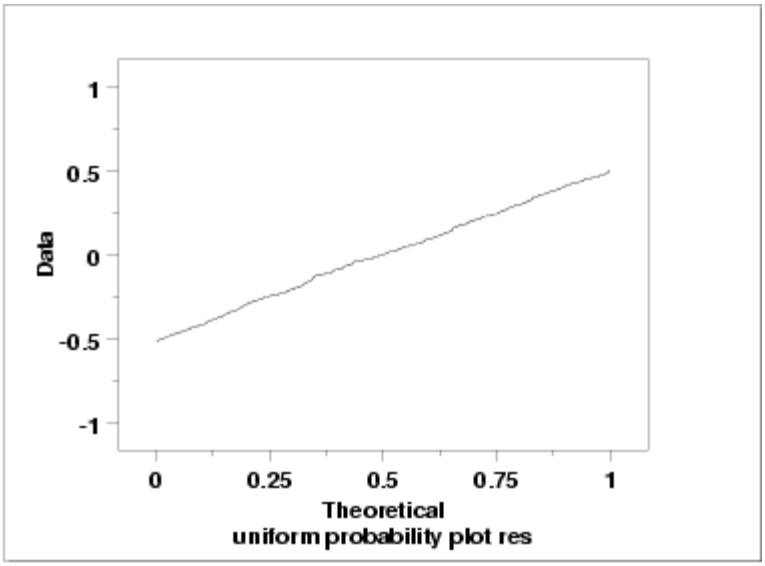
Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates no significant shifts in location or scale over time.
2. The lag plot (upper right) exhibits a random appearance.
3. The histogram shows a relatively flat appearance. This indicates that a uniform probability distribution may be an appropriate model for the error component (or residuals).
4. The normal probability plot clearly shows that the normal distribution is not an appropriate model for the error component.

A uniform probability plot can be used to further test the suggestion that a uniform distribution might be a good model for the error component.

Uniform Probability Plot of Residuals



Since the uniform probability plot is nearly linear, this verifies that a uniform distribution is a good model for the error component.

Conclusions

Since the residuals from our model satisfy the underlying assumptions, we conlude that

$$Y_i = 0.0502 + 0.987 * Y_{i-1} + E_i$$

where the E_i follow a uniform distribution is a good model for this data set. We could simplify this model to

$$Y_i = 1.0 * Y_{i-1} + E_i$$

This has the advantage of simplicity (the current point is simply the previous point plus a uniformly distributed error term).

Using Scientific and Engineering Knowledge

In this case, the above model makes sense based on our definition of the random walk. That is, a random walk is the cumulative sum of uniformly distributed data points. It makes sense that modeling the current point as the previous point plus a uniformly distributed error term is about as good as we can do. Although this case is a bit artificial in that we knew how the data were constructed, it is common and desirable to use scientific and engineering knowledge of the process that generated the data in formulating and testing models for the data. Quite often, several competing models will produce nearly equivalent mathematical results. In this case, selecting the model that best approximates the scientific understanding of the process is a reasonable choice.

Time Series Model

This model is an example of a time series model. More extensive discussion of time series is given in the Process Monitoring chapter.

1.4.2.3.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. Validate assumptions. 1. 4-plot of Y. 2. Generate a table of summary statistics. 3. Generate a linear fit to detect drift in location. 4. Detect drift in variation by	1. Based on the 4-plot, there are shifts in location and scale and the data are not random. 2. The summary statistics table displays 25+ statistics.

dividing the data into quarters and computing Levene's test for equal standard deviations. 5. Check for randomness by generating a runs test.	3. The linear fit indicates drift in location since the slope parameter is statistically significant. 4. Levene's test indicates significant drift in variation. 5. The runs test indicates significant non-randomness.
3. Generate the randomness plots. 1. Generate an autocorrelation plot. 2. Generate a spectral plot.	1. The autocorrelation plot shows significant autocorrelation at lag 1. 2. The spectral plot shows a single dominant low frequency peak.
4. Fit $Y_i = A0 + A1*Y_{i-1} + E_i$ and validate. 1. Generate the fit. 2. Plot fitted line with original data. 3. Generate a 4-plot of the residuals from the fit. 4. Generate a uniform probability plot of the residuals.	1. The residual standard deviation from the fit is 0.29 (compared to the standard deviation of 2.08 from the original data). 2. The plot of the predicted values with the original data indicates a good fit. 3. The 4-plot indicates that the assumptions of constant location and scale are valid. The lag plot indicates that the data are random. However, the histogram and normal probability plot indicate that the uniform disribution might be a better model for the residuals than the normal distribution. 4. The uniform

	probability plot verifies that the residuals can be fit by a uniform distribution.

1.4.2.4. Josephson Junction Cryothermometry

Josephson Junction Cryothermometry

This example illustrates the univariate analysis of Josephson junction cyrothermometry.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.4.1. Background and Data

Generation This data set was collected by Bob Soulen of NIST in October, 1971 as a sequence of observations collected equi-spaced in time from a volt meter to ascertain the process temperature in a Josephson junction cryothermometry (low temperature) experiment. The response variable is voltage counts.

Motivation The motivation for studying this data set is to illustrate the case where there is discreteness in the measurements, but the underlying assumptions hold. In this case, the discreteness is due to the data being integers.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following are the data used for this case study.

```
2899 2898 2898 2900 2898
2901 2899 2901 2900 2898
2898 2898 2898 2900 2898
2897 2899 2897 2899 2899
2900 2897 2900 2900 2899
2898 2898 2899 2899 2899
2899 2899 2898 2899 2899
2899 2902 2899 2900 2898
2899 2899 2899 2899 2899
2899 2900 2899 2900 2898
2901 2900 2899 2899 2899
2899 2899 2900 2899 2898
2898 2898 2900 2896 2897
2899 2899 2900 2898 2900
2901 2898 2899 2901 2900
2898 2900 2899 2899 2897
2899 2898 2899 2899 2898
2899 2897 2899 2899 2897
2899 2897 2899 2897 2897
2899 2897 2898 2898 2899
2897 2898 2897 2899 2899
2898 2898 2897 2898 2895
2897 2898 2898 2896 2898
2898 2897 2896 2898 2898
2897 2897 2898 2898 2896
2898 2898 2896 2899 2898
2898 2898 2899 2899 2898
2898 2899 2899 2899 2900
2900 2901 2899 2898 2898
2900 2899 2898 2901 2897
2898 2898 2900 2899 2899
2898 2898 2899 2898 2901
2900 2897 2897 2898 2898
2900 2898 2899 2898 2898
2898 2896 2895 2898 2898
2898 2898 2897 2897 2895
```

```
2897 2897 2900 2898 2896
2897 2898 2898 2899 2898
2897 2898 2898 2896 2900
2899 2898 2896 2898 2896
2896 2896 2897 2897 2896
2897 2897 2896 2898 2896
2898 2896 2897 2896 2897
2897 2898 2897 2896 2895
2898 2896 2896 2898 2896
2898 2898 2897 2897 2898
2897 2899 2896 2897 2899
2900 2898 2898 2897 2898
2899 2899 2900 2900 2900
2900 2899 2899 2899 2898
2900 2901 2899 2898 2900
2901 2901 2900 2899 2898
2901 2899 2901 2900 2901
2898 2900 2900 2898 2900
2900 2898 2899 2901 2900
2899 2899 2900 2900 2899
2900 2901 2899 2898 2898
2899 2896 2898 2897 2898
2898 2897 2897 2897 2898
2897 2899 2900 2899 2897
2898 2900 2900 2898 2898
2899 2900 2898 2900 2900
2898 2900 2898 2898 2898
2898 2898 2899 2898 2900
2897 2899 2898 2899 2898
2897 2900 2901 2899 2898
2898 2901 2898 2899 2897
2899 2897 2896 2898 2898
2899 2900 2896 2897 2897
2898 2899 2899 2898 2898
2897 2897 2898 2897 2897
2898 2898 2898 2896 2895
2898 2898 2898 2896 2898
2898 2898 2897 2897 2899
2896 2900 2897 2897 2898
2896 2897 2898 2898 2898
2897 2897 2898 2899 2897
2898 2899 2897 2900 2896
2899 2897 2898 2897 2900
2899 2900 2897 2897 2898
2897 2899 2899 2898 2897
2901 2900 2898 2901 2899
2900 2899 2898 2900 2900
2899 2898 2897 2900 2898
2898 2897 2899 2898 2900
2899 2898 2899 2897 2900
2898 2902 2897 2898 2899
2899 2899 2898 2897 2898
2897 2898 2899 2900 2900
2899 2898 2899 2900 2899
2900 2899 2899 2899 2899
2899 2898 2899 2899 2900
2902 2899 2900 2900 2901
2899 2901 2899 2899 2902
2898 2898 2898 2898 2899
2899 2900 2900 2900 2898
2899 2899 2900 2899 2900
2899 2900 2898 2898 2898
2900 2898 2899 2900 2899
2899 2900 2898 2898 2899
2899 2899 2899 2898 2898
2897 2898 2899 2897 2897
2901 2898 2897 2898 2899
2898 2897 2899 2898 2897
2898 2898 2897 2898 2899
2899 2899 2899 2900 2899
2899 2897 2898 2899 2900
2898 2897 2901 2899 2901
2898 2899 2901 2900 2900
2899 2900 2900 2900 2900
2901 2900 2901 2899 2897
2900 2900 2901 2899 2898
2900 2899 2899 2900 2899
2900 2899 2900 2899 2901
2900 2900 2899 2899 2898
2899 2900 2898 2899 2899
2901 2898 2898 2900 2899
```

```
2899 2898 2897 2898 2897
2899 2899 2899 2898 2898
2897 2898 2899 2897 2897
2899 2898 2898 2899 2899
2901 2899 2899 2899 2897
2900 2896 2898 2898 2900
2897 2899 2897 2896 2898
2897 2898 2899 2896 2899
2901 2898 2898 2896 2897
2899 2897 2898 2899 2898
2898 2898 2898 2898 2898
2899 2900 2899 2901 2898
2899 2899 2898 2900 2898
2899 2899 2901 2900 2901
2899 2901 2899 2901 2899
2900 2902 2899 2898 2899
2900 2899 2900 2900 2901
2900 2899 2901 2901 2899
2898 2901 2897 2898 2901
2900 2902 2899 2900 2898
2900 2899 2900 2899 2899
2899 2898 2900 2898 2899
2899 2899 2899 2898 2900
```

1.4.2.4.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

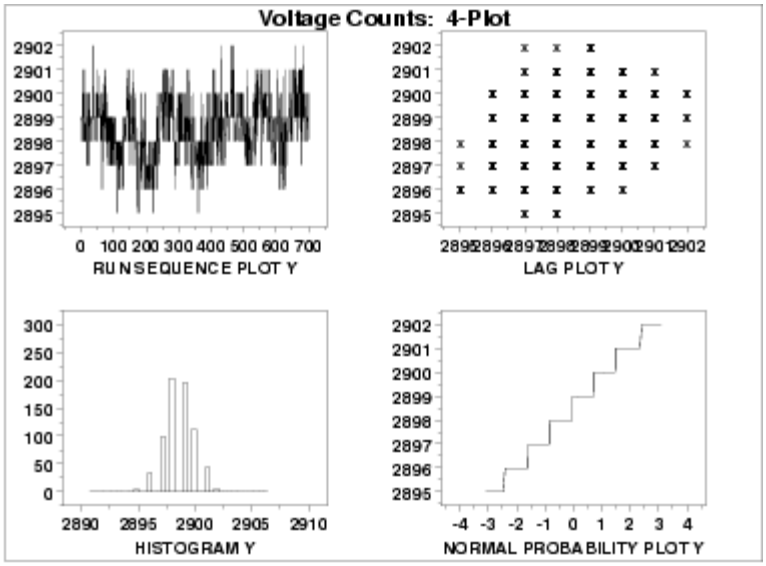
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram (lower left) shows that the data are reasonably symmetric, there does not appear to be significant outliers in the tails, and that it is reasonable to assume that the data can be fit with a normal distribution.
4. The normal probability plot (lower right) is difficult to interpret due to the fact that there are only a few distinct values with many repeats.

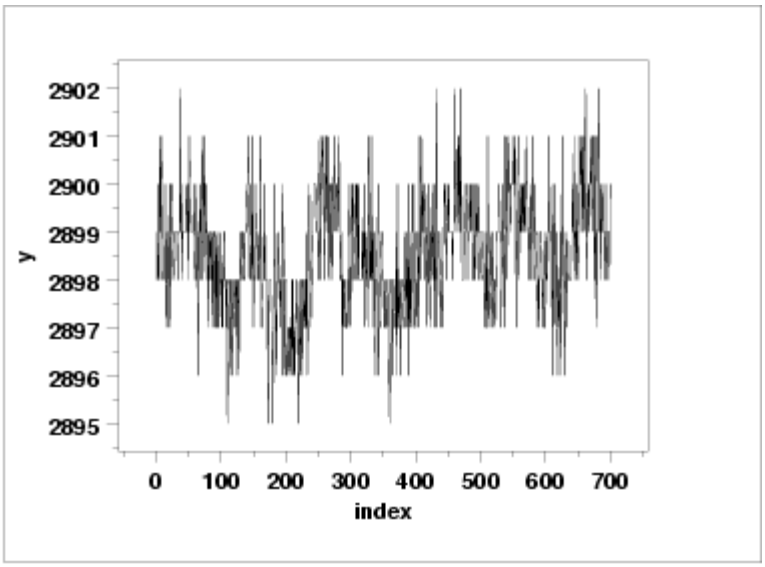
The integer data with only a few distinct values and many repeats accounts for the discrete appearance of several of the plots (e.g., the lag plot and the normal probability plot). In this case, the nature of the data makes the normal probability plot difficult to interpret, especially since each number is repeated many times. However, the histogram indicates that a normal distribution should provide an adequate model for the data.

From the above plots, we conclude that the underlying assumptions are valid and the data can be reasonably approximated with a normal distribution. Therefore, the commonly used uncertainty standard is valid and appropriate. The numerical values for this model are given in the Quantitative Output and Interpretation section.

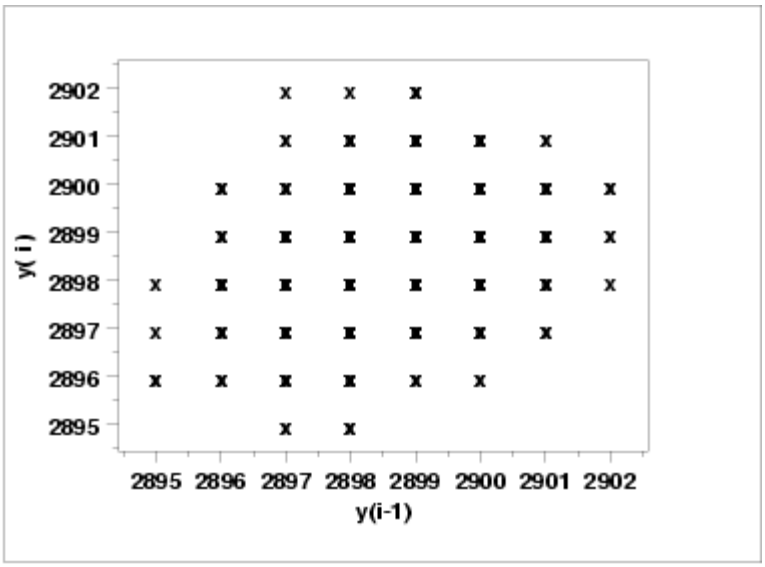
Individual Plots

Although it is normally not necessary, the plots can be generated individually to give more detail.

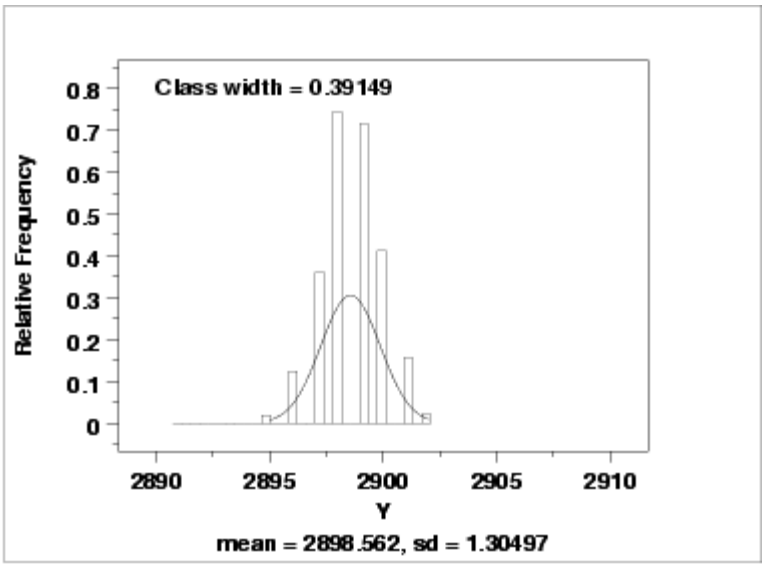
Run Sequence Plot



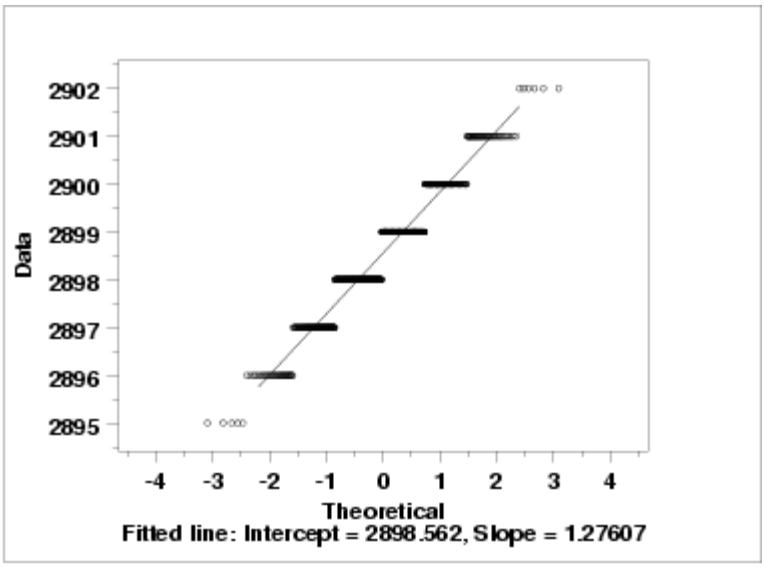
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot


2902
2901
2900
2899
2898
2897
2896
2895
Data
-4
-3
-2
-1
0
1
2
3
4
Theoretical
Fitted line: Intercept = 2898.562, Slope = 1.27607

1.4.2.4.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics were computed from the data.

```
Sample size  = 700
Mean         =   2898.562
Median       =   2899.000
Minimum      =   2895.000
Maximum      =   2902.000
Range        =      7.000
Stan. Dev.   =      1.305
```

Because of the discrete nature of the data, we also compute the normal PPCC.

```
Normal PPCC = 0.97484
```

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

```
      Coefficient     Estimate     Stan. Error
t-Value
           B0         2.898E+03       9.745E-02
29739.288
           B1         1.071E-03       2.409e-04
4.445

      Residual Standard Deviation = 1.288
      Residual Degrees of Freedom = 698
```

The slope parameter, B_1, has a *t* value of 4.445 which is statistically significant (the critical value is 1.96). However, the value of the slope is 1.071E-03. Given that the slope is nearly zero, the assumption of constant location is not seriously violated even though it is statistically significant.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since the nature of the data (a few distinct

points repeated many times) makes the normality assumption questionable, we use the alternative Levene test. In particular, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

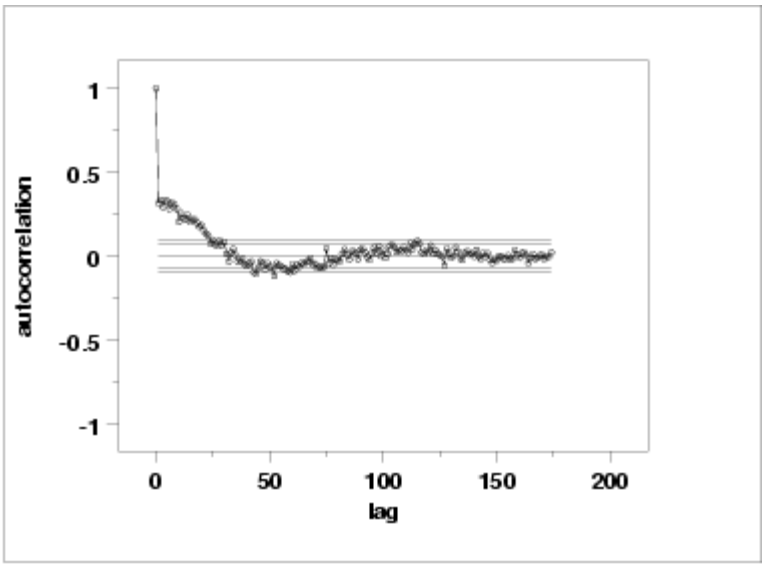
Test statistic: $W = 1.43$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.618$
Critical region: Reject H_0 if $W > 2.618$

Since the Levene test statistic value of 1.43 is less than the 95 % critical value of 2.618, we conclude that the variances are not significantly different in the four intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the previous section is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95 % and 99 % confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of most interest, is 0.31. The critical values at the 5 % level of significance are -0.087 and 0.087. This indicates that the lag 1 autocorrelation is statistically significant, so there is some evidence for non-randomness.

A common test for randomness is the runs test.

```
      H0:  the sequence was produced in a random
manner
      Ha:  the sequence was not produced in a
random manner

      Test statistic:  Z = -13.4162
      Significance level:  α = 0.05
      Critical value:  Z1-α/2 = 1.96
      Critical region:  Reject H0 if |Z| > 1.96
```

The runs test indicates non-randomness.

Although the runs test and lag 1 autocorrelation indicate some mild non-randomness, it is not sufficient to reject the $\boldsymbol{Y_i = C + E_i}$ model. At least part of the non-randomness can be explained by the discrete nature of the data.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot, or PPCC. For this data set the PPCC based on a normal distribution is 0.975. Since the PPCC is less than the critical value of 0.987 (this is a tabulated value), the normality assumption is rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. The results of the Anderson-Darling test follow.

```
      H0:  the data are normally distributed
      Ha:  the data are not normally distributed

      Adjusted test statistic:  A2 = 16.858
      Significance level:  α = 0.05
      Critical value:  0.787
      Critical region:  Reject H0 if A2 > 0.787
```

The Anderson-Darling test rejects the normality assumption because the test statistic, 16.858, is greater than the 95 % critical value 0.787.

Although the data are not strictly normal, the violation of the normality assumption is not severe enough to conclude that the $\boldsymbol{Y_i = C + E_i}$ model is unreasonable. At least part of the non-normality can be explained by the discrete nature of the data.

Outlier Analysis

A test for outliers is the Grubbs test.

```
      H0:  there are no outliers in the data
      Ha:  the maximum value is an outlier

      Test statistic:  G = 2.729201
      Significance level:  α = 0.05
      Critical value for a one-tailed test:
```

```
3.950619
      Critical region:  Reject H0 if G > 3.950619
```

For this data set, Grubbs' test does not detect any outliers at the 0.05 significance level.

Model

Although the randomness and normality assumptions were mildly violated, we conclude that a reasonable model for the data is:

$$Y_i = 2898.7 + E_i$$

In addition, a 95 % confidence interval for the mean value is (2898.515, 2898.928).

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
 Analysis for Josephson Junction Cryothermometry
Data

 1: Sample Size                              = 700

 2: Location
    Mean                                     =
2898.562
    Standard Deviation of Mean               =
0.049323
    95% Confidence Interval for Mean         =
(2898.465,2898.658)
    Drift with respect to location?          = YES
    (Further analysis indicates that
    the drift, while statistically
    significant, is not practically
    significant)

 3: Variation
    Standard Deviation                       =
1.30497
    95% Confidence Interval for SD           =
(1.240007,1.377169)
    Drift with respect to variation?
    (based on Levene's test on quarters
    of the data)                              = NO

 4: Distribution
    Normal PPCC                              =
0.97484
    Data are Normal?
      (as measured by Normal PPCC)           = NO

 5: Randomness
    Autocorrelation                          =
0.314802
    Data are Random?
      (as measured by autocorrelation)      = NO

 6: Statistical Control
    (i.e., no drift in location or scale,
    data are random, distribution is
    fixed, here we are testing only for
    fixed normal)
    Data Set is in Statistical Control?    = NO

    Note: Although we have violations of
    the assumptions, they are mild enough,
    and at least partially explained by the
    discrete nature of the data, so we may model
    the data as if it were in statistical
    control

 7: Outliers?
    (as determined by Grubbs test)           = NO
```

1.4.2.4.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale. Due to the nature of the data (a few distinct points with many repeats), the normality assumption is

	questionable.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf. 4. Generate a normal probability plot.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is a good distribution for these data. 4. The discrete nature of the data masks the normality or non-normality of the data somewhat. The plot indicates that a normal distribution provides a rough approximation for the data.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Levene's test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Check for normality by computing the normal probability plot correlation coefficient.	1. The summary statistics table displays 25+ statistics. 2. The mean is 2898.56 and a 95% confidence interval is (2898.46,2898.66). The linear fit indicates no meaningful drift in location since the value of the slope parameter is near zero. 3. The standard devaition is 1.30 with a 95% confidence interval of (1.24,1.38). Levene's test indicates no significant drift in variation.

6. Check for outliers using Grubbs' test. 7. Print a univariate report (this assumes steps 2 thru 6 have already been run).	4. The lag 1 autocorrelation is 0.31. This indicates some mild non-randomness. 5. The normal probability plot correlation coefficient is 0.975. At the 5% level, we reject the normality assumption. 6. Grubbs' test detects no outliers at the 5% level. 7. The results are summarized in a convenient report.

1. Exploratory Data Analysis
1.4. EDA Case Studies
1.4.2. Case Studies

1.4.2.5. Beam Deflections

Beam Deflection This example illustrates the univariate analysis of beam deflection data.

1. Background and Data
2. Test Underlying Assumptions
3. Develop a Better Model
4. Validate New Model
5. Work This Example Yourself

1.4.2.5.1. Background and Data

Generation This data set was collected by H. S. Lew of NIST in 1969 to measure steel-concrete beam deflections. The response variable is the deflection of a beam from the center point.

The motivation for studying this data set is to show how the underlying assumptions are affected by periodic data.

Data The following are the data used for this case study.

```
-213
-564
 -35
 -15
 141
 115
-420
-360
 203
-338
-431
 194
-220
-513
 154
-125
-559
  92
 -21
-579
 -52
  99
-543
-175
 162
-457
-346
 204
-300
-474
 164
-107
-572
  -8
  83
-541
-224
 180
-420
-374
 201
-236
-531
  83
  27
-564
-112
 131
```

```
-507
-254
 199
-311
-495
 143
 -46
-579
 -90
 136
-472
-338
 202
-287
-477
 169
-124
-568
  17
  48
-568
-135
 162
-430
-422
 172
 -74
-577
 -13
  92
-534
-243
 194
-355
-465
 156
 -81
-578
 -64
 139
-449
-384
 193
-198
-538
 110
 -44
-577
  -6
  66
-552
-164
 161
-460
-344
 205
-281
-504
 134
 -28
-576
-118
 156
-437
-381
 200
-220
-540
  83
  11
-568
-160
 172
-414
-408
 188
-125
-572
 -32
 139
-492
```

```
-321
 205
-262
-504
 142
 -83
-574
   0
  48
-571
-106
 137
-501
-266
 190
-391
-406
 194
-186
-553
  83
 -13
-577
 -49
 103
-515
-280
 201
 300
-506
 131
 -45
-578
 -80
 138
-462
-361
 201
-211
-554
  32
  74
-533
-235
 187
-372
-442
 182
-147
-566
  25
  68
-535
-244
 194
-351
-463
 174
-125
-570
  15
  72
-550
-190
 172
-424
-385
 198
-218
-536
  96
```

1.4.2.5.2. Test Underlying Assumptions

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

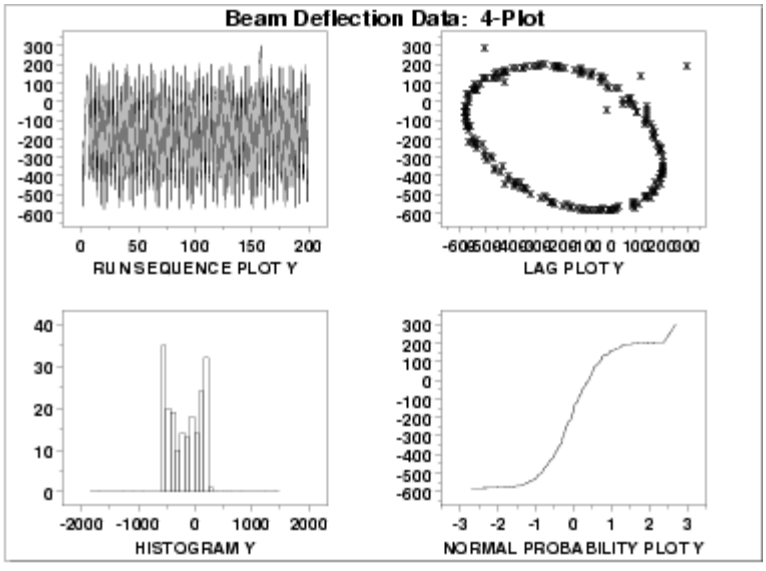
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) shows that the data are not random. The lag plot further indicates the presence of a few outliers.
3. When the randomness assumption is thus seriously violated, the histogram (lower left) and normal probability plot (lower right) are ignored since determining the distribution of the data is only meaningful when the data are random.

From the above plots we conclude that the underlying randomness assumption is not valid. Therefore, the model

$$Y_i = C + E_i$$

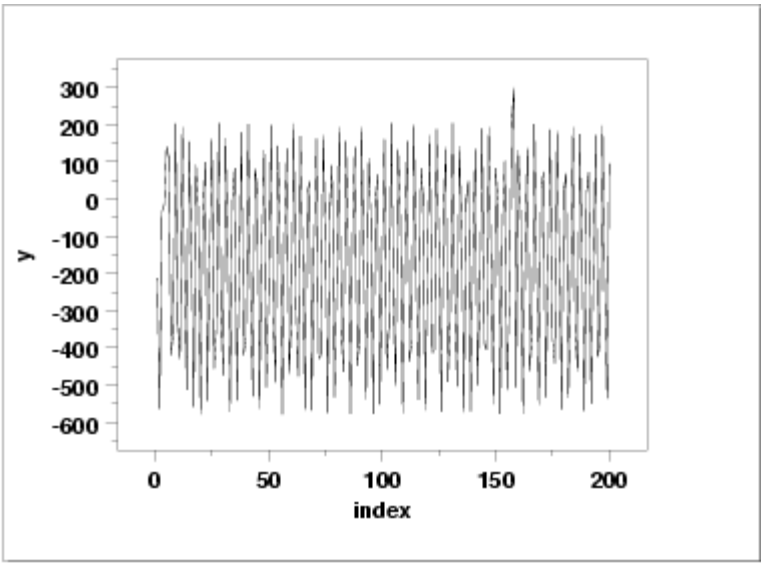
is not appropriate.

We need to develop a better model. Non-random data can frequently be modeled using time series mehtodology. Specifically, the circular pattern in the lag plot indicates that a sinusoidal model might be appropriate. The sinusoidal model will be developed in the next section.

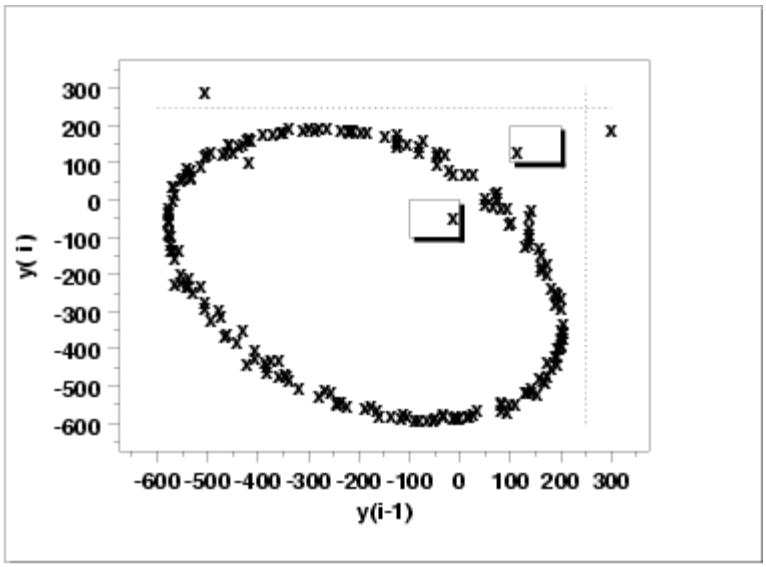
Individual Plots

The plots can be generated individually for more detail. In this case, only the run sequence plot and the lag plot are drawn since the distributional plots are not meaningful.

Run Sequence Plot



Lag Plot



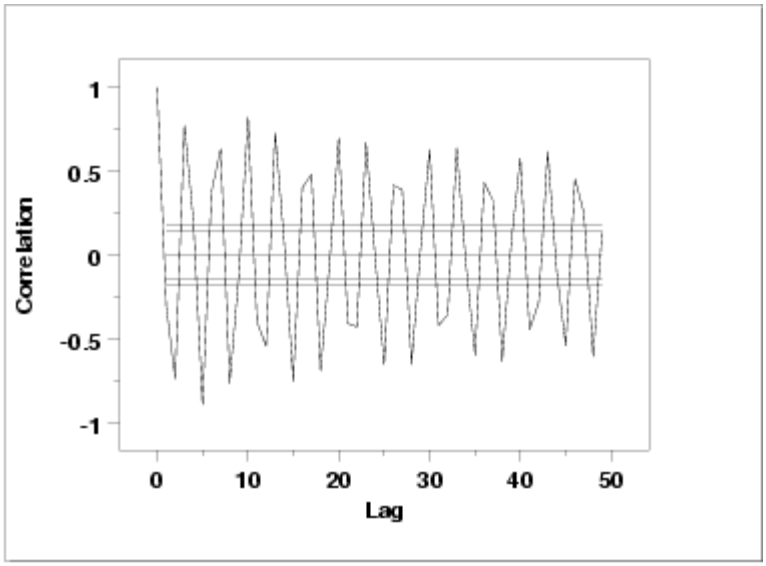
We have drawn some lines and boxes on the plot to better isolate the outliers. The following data points appear to be outliers based on the lag plot.

INDEX	Y(i-1)	Y(i)
158	-506.00	300.00
157	300.00	201.00
3	-15.00	-35.00
5	115.00	141.00

That is, the third, fifth, 157th, and 158th points appear to be outliers.

Autocorrelation Plot

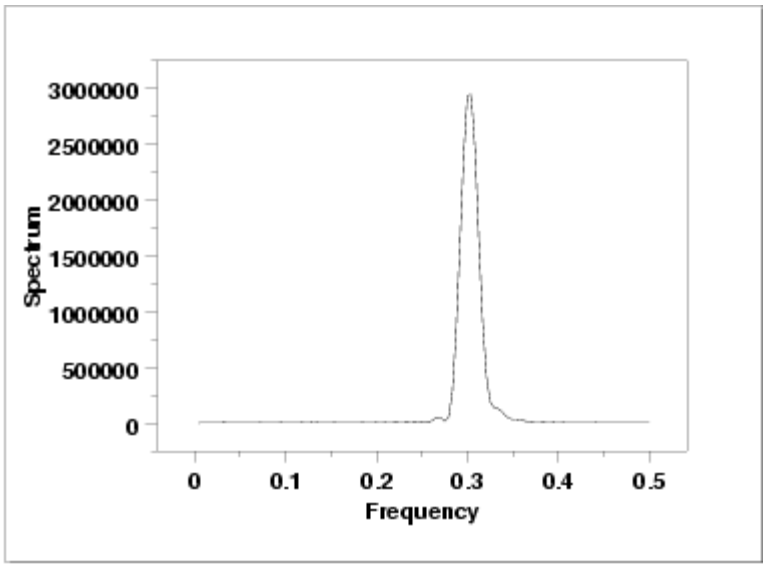
When the lag plot indicates significant non-randomness, it can be helpful to follow up with a an autocorrelation plot.



This autocorrelation plot shows a distinct cyclic pattern. As with the lag plot, this suggests a sinusoidal model.

Spectral Plot

Another useful plot for non-random data is the spectral plot.



This spectral plot shows a single dominant peak at a frequency of 0.3. This frequency of 0.3 will be used in fitting the sinusoidal model in the next section.

Quantitative Results

Although the lag plot, autocorrelation plot, and spectral plot clearly show the violation of the randomness assumption, we supplement the graphical output with some quantitative measures.

Summary Statistics

As a first step in the analysis, summary statistics are computed from the data.

```
Sample size  =  200
Mean         = -177.4350
Median       = -162.0000
Minimum      = -579.0000
Maximum      =  300.0000
Range        =  879.0000
Stan. Dev.   =  277.3322
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero.

Coefficient	Estimate	Stan. Error	*t*-Value
A_0	-178.175	39.47	-4.514
A_1	0.7366E-02	0.34	0.022

```
Residual Standard Deviation = 278.0313
Residual Degrees of Freedom = 198
```

The slope parameter, A1, has a *t* value of 0.022 which is statistically not significant. This indicates that the slope can in fact be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-

sized intervals. However, the Bartlett the non-randomness of this data does not allows us to assume normality, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

Test statistic: $W = 0.09378$
Degrees of freedom: $k - 1 = 3$
Sample size: $N = 200$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.651$
Critical region: Reject H_0 if $W > 2.651$

In this case, the Levene test indicates that the variances are not significantly different in the four intervals since the test statistic value, 0.9378, is less than the critical value of 2.651.

Randomness

A runs test is used to check for randomness

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a random manner

Test statistic: $Z = 2.6938$
Significance level: $\alpha = 0.05$
Critical value: $Z_{1-\alpha/2} = 1.96$
Critical region: Reject H_0 if $|Z| > 1.96$

The absolute value of the test statistic is larger than the critical value at the 5 % significance level, so we conclude that the data are not random.

Distributional Assumptions

Since the quantitative tests show that the assumptions of constant scale and non-randomness are not met, the distributional measures will not be meaningful. Therefore these quantitative tests are omitted.

1.4.2.5.3. Develop a Better Model

Sinusoidal Model

The lag plot and autocorrelation plot in the previous section strongly suggested a sinusoidal model might be appropriate. The basic sinusoidal model is:

$$Y_i = C + \alpha \sin\left(2\pi\omega T_i + \phi\right) + E_i$$

where ***C*** is constant defining a mean level, α is an amplitude for the sine function, ω is the frequency, T_i is a time variable, and ϕ is the phase. This sinusoidal model can be fit using non-linear least squares.

To obtain a good fit, sinusoidal models require good starting values for ***C***, the amplitude, and the frequency.

*Good Starting Value for **C***

A good starting value for ***C*** can be obtained by calculating the mean of the data. If the data show a trend, i.e., the assumption of constant location is violated, we can replace ***C*** with a linear or quadratic least squares fit. That is, the model becomes

$$Y_i = (B_0 + B_1 * T_i) + \alpha \sin\left(2\pi\omega T_i + \phi\right) + E_i$$

or

$$Y_i = (B_0 + B_1 * T_i + B2 * T_i^2) + \alpha \sin\left(2\pi\omega T_i + \phi\right) + E_i$$

Since our data did not have any meaningful change of location, we can fit the simpler model with ***C*** equal to the mean. From the summary output in the previous page, the mean is -177.44.

Good Starting Value for Frequency

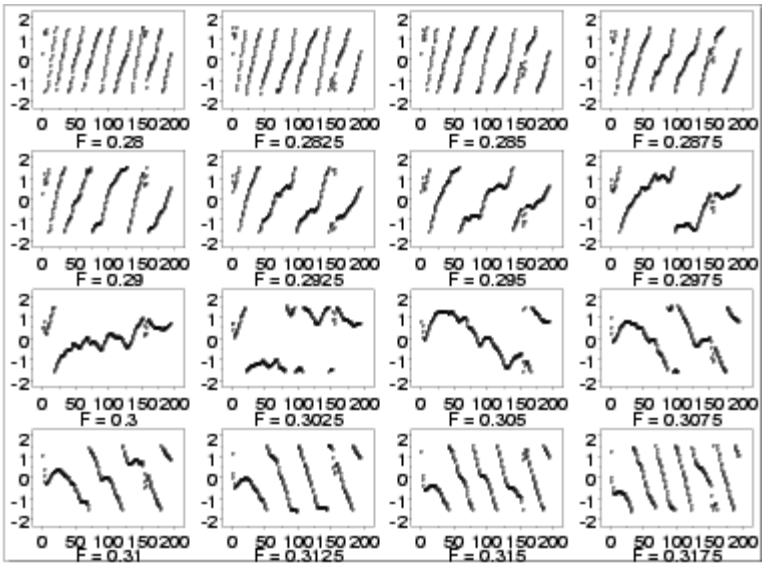
The starting value for the frequency can be obtained from the spectral plot, which shows the dominant frequency is about 0.3.

Complex Demodulation Phase Plot

The complex demodulation phase plot can be used to refine this initial estimate for the frequency.

For the complex demodulation plot, if the lines slope from left to right, the frequency should be increased. If the lines slope from right to left, it should be decreased. A relatively flat (i.e., horizontal) slope indicates a good frequency. We could generate the demodulation phase plot for 0.3 and then use trial and error to obtain a better estimate for the frequency. To simplify this, we generate 16 of these plots on a single page starting

with a frequency of 0.28, increasing in increments of 0.0025, and stopping at 0.3175.



Interpretation

The plots start with lines sloping from left to right but gradually change to a right to left slope. The relatively flat slope occurs for frequency 0.3025 (third row, second column). The complex demodulation phase plot restricts the range from $\pi/2$ to $-\pi/2$. This is why the plot appears to show some breaks.

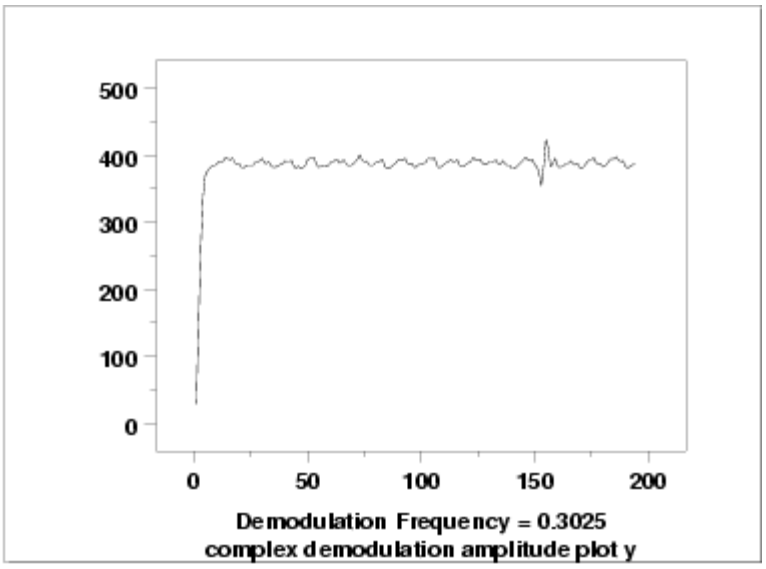
Good Starting Values for Amplitude

The complex demodulation amplitude plot is used to find a good starting value for the amplitude. In addition, this plot indicates whether or not the amplitude is constant over the entire range of the data or if it varies. If the plot is essentially flat, i.e., zero slope, then it is reasonable to assume a constant amplitude in the non-linear model. However, if the slope varies over the range of the plot, we may need to adjust the model to be:

$$Y_i = C + (B_0 + B_1 * T_i)\sin(2\pi\omega T_i + \phi) + E_i$$

That is, we replace α with a function of time. A linear fit is specified in the model above, but this can be replaced with a more elaborate function if needed.

Complex Demodulation Amplitude Plot



The complex demodulation amplitude plot for this data shows that:

1. The amplitude is fixed at approximately 390.
2. There is a short start-up effect.
3. There is a change in amplitude at around x=160 that should be investigated for an outlier.

In terms of a non-linear model, the plot indicates that fitting a single constant for α should be adequate for this data set.

Fit Results

Using starting estimates of 0.3025 for the frequency, 390 for the amplitude, and -177.44 for C, the following parameters were estimated.

Coefficient	Estimate	Stan. Error	*t*-Value
C	-178.786	11.02	-16.22
AMP	-361.766	26.19	-13.81
FREQ	0.302596	0.1510E-03	2005.00
PHASE	1.46536	0.4909E-01	29.85

Residual Standard Deviation = 155.8484
Residual Degrees of Freedom = 196

Model

From the fit results, our proposed model is:

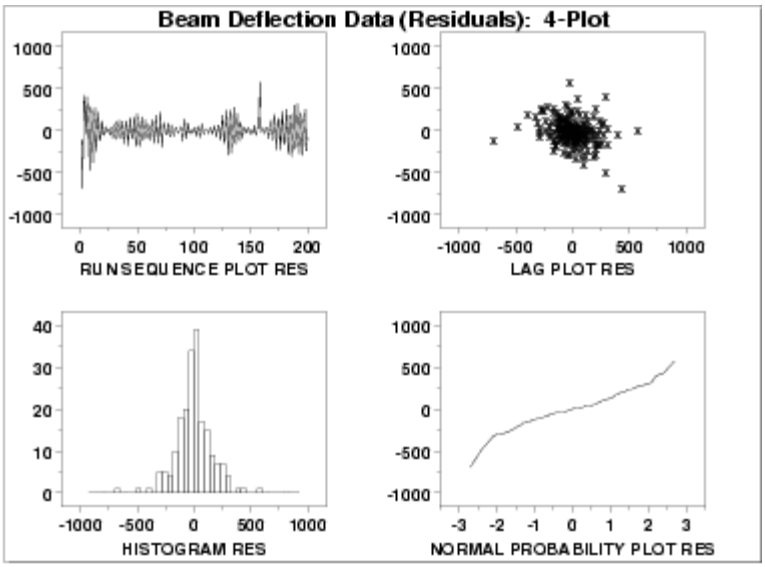
$$\hat{Y}_i = -178.786 - 361.766[2\pi(0.302596)T_i + 1.46536]$$

We will evaluate the adequacy of this model in the next section.

1.4.2.5.4. Validate New Model

4-Plot of Residuals

The first step in evaluating the fit is to generate a 4-plot of the residuals.



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location. There does seem to be some shifts in scale. A start-up effect was detected previously by the complex demodulation amplitude plot. There does appear to be a few outliers.

2. The lag plot (upper right) shows that the data are random. The outliers also appear in the lag plot.

3. The histogram (lower left) and the normal probability plot (lower right) do not show any serious non-normality in the residuals. However, the bend in the left portion of the normal probability plot shows some cause for concern.

The 4-plot indicates that this fit is reasonably good. However, we will attempt to improve the fit by removing the outliers.

Fit Results with Outliers Removed

The following parameter estimates were obtained after removing three outliers.

Coefficient	Estimate	Stan. Error	*t*-Value

```
C             -178.788        10.57          -16.91
AMP           -361.759        25.45          -14.22
FREQ          0.302597    0.1457E-03       2077.00
PHASE          1.46533    0.4715E-01         31.08

Residual Standard Deviation = 148.3398
Residual Degrees of Freedom = 193
```

New Fit to Edited Data

The original fit, with a residual standard deviation of 155.84, was:

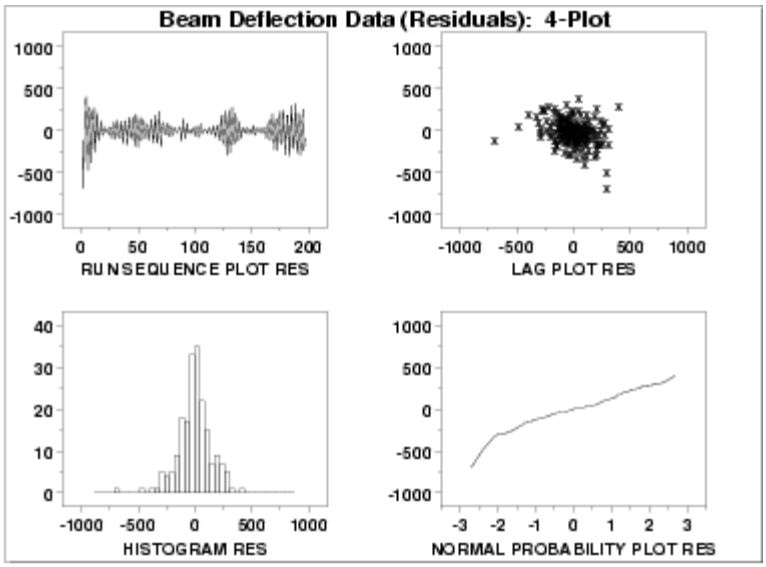
$$\hat{Y}_i = -178.786 - 361.766[2\pi(0.302596)T_i + 1.46536]$$

The new fit, with a residual standard deviation of 148.34, is:

$$\hat{Y}_i = -178.788 - 361.759[2\pi(0.302597)T_i + 1.46533]$$

There is minimal change in the parameter estimates and about a 5 % reduction in the residual standard deviation. In this case, removing the residuals has a modest benefit in terms of reducing the variability of the model.

4-Plot for New Fit



This plot shows that the underlying assumptions are satisfied and therefore the new fit is a good descriptor of the data.

In this case, it is a judgment call whether to use the fit with or without the outliers removed.

1.4.2.5.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. Validate assumptions. 1. 4-plot of Y. 2. Generate a run sequence plot. 3. Generate a lag plot.	1. Based on the 4-plot, there are no obvious shifts in location and scale, but the data are not random. 2. Based on the run sequence plot, there are no obvious shifts in location and

4. Generate an autocorrelation plot. 5. Generate a spectral plot. 6. Generate a table of summary statistics. 7. Generate a linear fit to detect drift in location. 8. Detect drift in variation by dividing the data into quarters and computing Levene's test statistic for equal standard deviations. 9. Check for randomness by generating a runs test.	scale. 3. Based on the lag plot, the data are not random. 4. The autocorrelation plot shows significant autocorrelation at lag 1. 5. The spectral plot shows a single dominant low frequency peak. 6. The summary statistics table displays 25+ statistics. 7. The linear fit indicates no drift in location since the slope parameter is not statistically significant. 8. Levene's test indicates no significant drift in variation. 9. The runs test indicates significant non-randomness.
3. Fit $Y_i = C + A*SIN(2*PI*omega*t_i+phi)$. 1. Generate a complex demodulation phase plot. 2. Generate a complex demodulation amplitude plot. 3. Fit the non-linear model.	1. Complex demodulation phase plot indicates a starting frequency of 0.3025. 2. Complex demodulation amplitude plot indicates an amplitude of 390 (but there is a short start-up effect). 3. Non-linear fit generates final parameter estimates. The residual standard deviation from the fit is 155.85 (compared to the standard deviation of 277.73 from the original data).

4. Validate fit.	
1. Generate a 4-plot of the residuals from the fit.	1. The 4-plot indicates that the assumptions of constant location and scale are valid. The lag plot indicates that the data are random. The histogram and normal probability plot indicate that the residuals that the normality assumption for the residuals are not seriously violated, although there is a bend on the probablity plot that warrants attention.
2. Generate a nonlinear fit with outliers removed.	2. The fit after removing 3 outliers shows some marginal improvement in the model (a 5% reduction in the residual standard deviation).
3. Generate a 4-plot of the residuals from the fit with the outliers removed.	3. The 4-plot of the model fit after 3 outliers removed shows marginal improvement in satisfying model assumptions.

1.4.2.6. Filter Transmittance

Filter Transmittance This example illustrates the univariate analysis of filter transmittance data.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.6.1. Background and Data

Generation This data set was collected by NIST chemist Radu Mavrodineaunu in the 1970's from an automatic data acquisition system for a filter transmittance experiment. The response variable is transmittance.

The motivation for studying this data set is to show how the underlying autocorrelation structure in a relatively small data set helped the scientist detect problems with his automatic data acquisition system.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following are the data used for this case study.

```
2.00180
2.00170
2.00180
2.00190
2.00180
2.00170
2.00150
2.00140
2.00150
2.00150
2.00170
2.00180
2.00180
2.00190
2.00190
2.00210
2.00200
2.00160
2.00140
2.00130
2.00130
2.00150
2.00150
2.00160
2.00150
2.00140
2.00130
2.00140
2.00150
2.00140
2.00150
2.00160
2.00150
2.00160
2.00190
2.00200
2.00200
2.00210
2.00220
```

```
2.00230
2.00240
2.00250
2.00270
2.00260
2.00260
2.00260
2.00270
2.00260
2.00250
2.00240
```

1.4.2.6.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

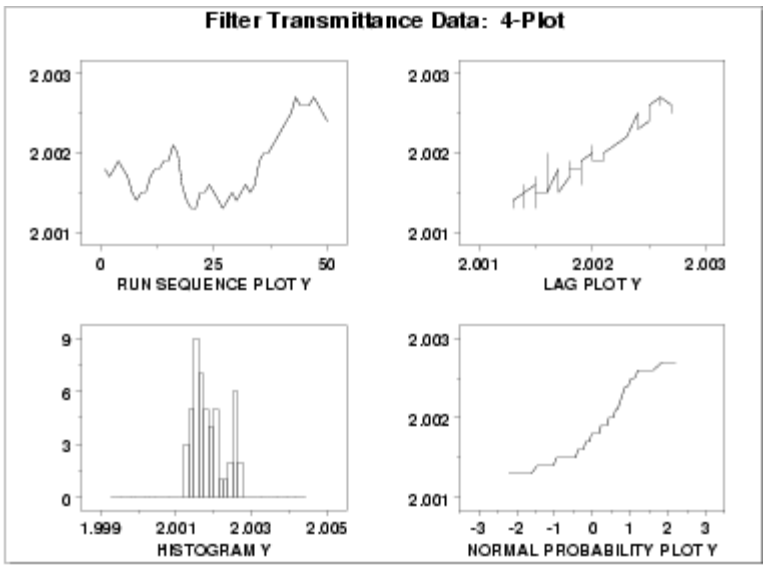
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates a significant shift in location around x=35.
2. The linear appearance in the lag plot (upper right) indicates a non-random pattern in the data.
3. Since the lag plot indicates significant non-randomness, we do not make any interpretation of either the histogram (lower left) or the normal probability plot (lower right).

The serious violation of the non-randomness assumption means that the univariate model

$$Y_i = C + E_i$$

is not valid. Given the linear appearance of the lag plot, the first step might be to consider a model of the type

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

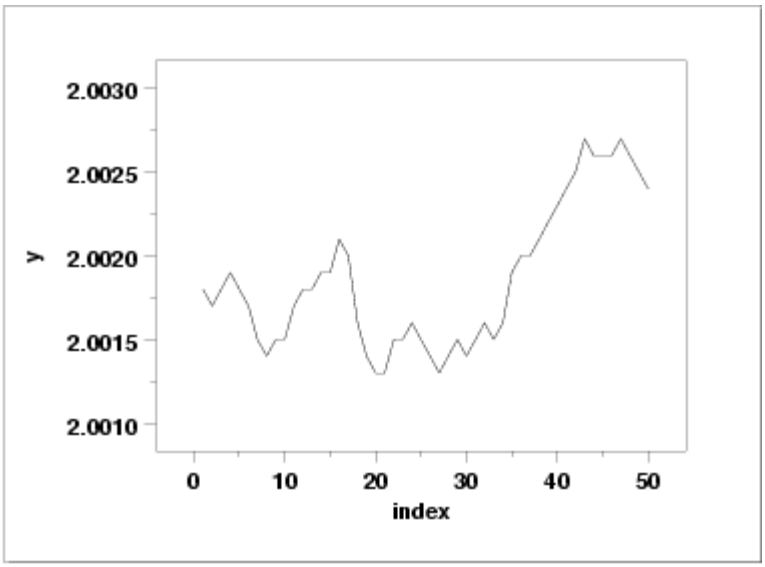
However, in this case discussions with the scientist revealed that non-randomness was entirely unexpected. An examination of the experimental process revealed that the sampling rate for the automatic data acquisition system was too fast. That is, the equipment did not have sufficient time to reset before the next sample started, resulting in the current measurement being contaminated by the previous measurement. The solution was to rerun the experiment allowing more time between samples.

Simple graphical techniques can be quite effective in revealing unexpected results in the data. When this occurs, it is important to investigate whether the unexpected result is due to problems in the experiment and data collection or is indicative of unexpected underlying structure in the data. This determination cannot be made on the basis of statistics alone. The role of the graphical and statistical analysis is to detect problems or unexpected results in the data. Resolving the issues requires the knowledge of the scientist or engineer.

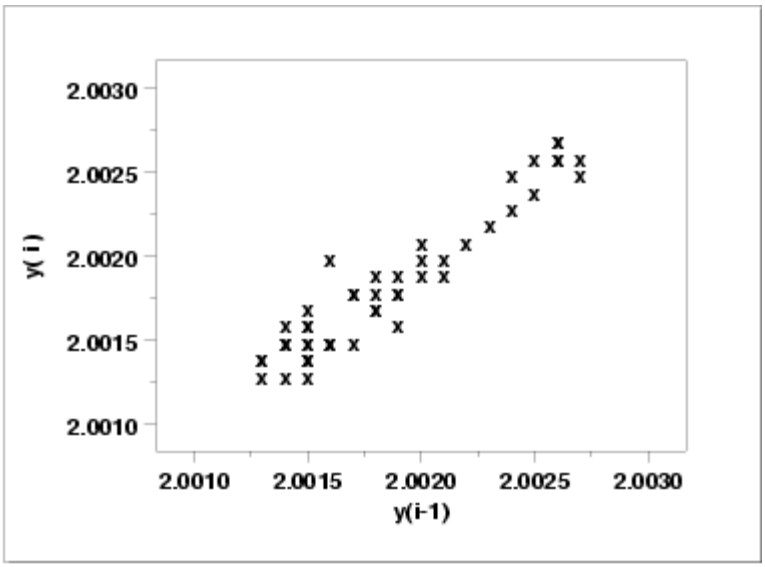
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail. Since the lag plot indicates significant non-randomness, we omit the distributional plots.

Run Sequence Plot


2.0030
2.0025
2.0020
2.0015
2.0010
y
0
10
20
30
40
50
index

Lag Plot


2.0030
2.0025
2.0020
2.0015
2.0010
y(i)
2.0010
2.0015
2.0020
2.0025
2.0030
y(i-1)

1.4.2.6.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics are computed from the data.

```
Sample size  = 50
Mean         =  2.0019
Median       =  2.0018
Minimum      =  2.0013
Maximum      =  2.0027
Range        =  0.0014
Stan. Dev.   =  0.0004
```

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

Coefficient	Estimate	Stan. Error	*t*-Value
B_0	2.00138	0.9695E-04	0.2064E+05
B_1	0.185E-04	0.3309E-05	5.582

```
Residual Standard Deviation = 0.3376404E-03
Residual Degrees of Freedom = 48
```

The slope parameter, B_1, has a *t* value of 5.582, which is statistically significant. Although the estimated slope, 0.185E-04, is nearly zero, the range of data (2.0013 to 2.0027) is also very small. In this case, we conclude that there is drift in location, although it is relatively small.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal sized intervals. However, the Bartlett test is not robust for non-normality. Since the normality assumption is questionable for these data, we use the alternative Levene test. In particular, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of four or

eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

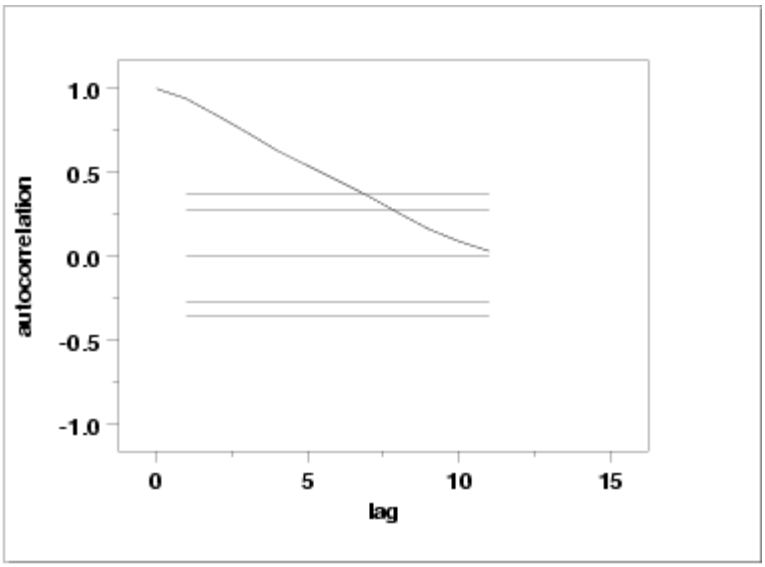
Test statistic: $W = 0.971$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.806$
Critical region: Reject H_0 if $W > 2.806$

In this case, since the Levene test statistic value of 0.971 is less than the critical value of 2.806 at the 5 % level, we conclude that there is no evidence of a change in variation.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot in the previous seciton is a simple graphical technique.

One check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95 % and 99 % confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of most interest, is 0.93. The critical values at the 5 % level are -0.277 and 0.277. This indicates that the lag 1 autocorrelation is statistically significant, so there is strong evidence of non-randomness.

A common test for randomness is the runs test.

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a random manner

Test statistic: $Z = -5.3246$

Significance level: $\alpha = 0.05$
Critical value: $Z_{1-\alpha/2} = 1.96$
Critical region: Reject H_0 if $|Z| > 1.96$

Because the test statistic is outside of the critical region, we reject the null hypothesis and conclude that the data are not random.

Distributional Analysis

Since we rejected the randomness assumption, the distributional tests are not meaningful. Therefore, these quantitative tests are omitted. We also omit Grubbs' outlier test since it also assumes the data are approximately normally distributed.

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
 Analysis for filter transmittance data

 1: Sample Size                            = 50

 2: Location
    Mean                                   =
2.001857
    Standard Deviation of Mean             =
0.00006
    95% Confidence Interval for Mean       =
(2.001735,2.001979)
    Drift with respect to location?        = NO

 3: Variation
    Standard Deviation                     =
0.00043
    95% Confidence Interval for SD         =
(0.000359,0.000535)
    Change in variation?
    (based on Levene's test on quarters
    of the data)                            = NO

 4: Distribution
    Distributional tests omitted due to
    non-randomness of the data

 5: Randomness
    Lag One Autocorrelation                =
0.937998
    Data are Random?
      (as measured by autocorrelation)    = NO

 6: Statistical Control
    (i.e., no drift in location or scale,
    data are random, distribution is
    fixed, here we are testing only for
    normal)
    Data Set is in Statistical Control?    = NO

 7: Outliers?
    (Grubbs' test omitted)                 = NO
```

1.4.2.6.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there is a shift in location and the data are not random.
3. Generate the individual plots. 1. Generate a run sequence plot.	1. The run sequence

2. Generate a lag plot.	plot indicates that there is a shift in location. 2. The strong linear pattern of the lag plot indicates significant non-randomness.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Compute a linear fit based on quarters of the data to detect drift in location. 3. Compute Levene's test based on quarters of the data to detect changes in variation. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Print a univariate report (this assumes steps 2 thru 4 have already been run).	1. The summary statistics table displays 25+ statistics. 2. The linear fit indicates a slight drift in location since the slope parameter is statistically significant, but small. 3. Levene's test indicates no significant drift in variation. 4. The lag 1 autocorrelation is 0.94. This is outside the 95% confidence interval bands which indicates significant non-randomness. 5. The results are summarized in a convenient report.

1.4.2.7. Standard Resistor

Standard Resistor This example illustrates the univariate analysis of standard resistor data.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.7.1. Background and Data

Generation This data set was collected by Ron Dziuba of NIST over a 5-year period from 1980 to 1985. The response variable is resistor values.

The motivation for studying this data set is to illustrate data that violate the assumptions of constant location and scale.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following are the data used for this case study.

```
27.8680
27.8929
27.8773
27.8530
27.8876
27.8725
27.8743
27.8879
27.8728
27.8746
27.8863
27.8716
27.8818
27.8872
27.8885
27.8945
27.8797
27.8627
27.8870
27.8895
27.9138
27.8931
27.8852
27.8788
27.8827
27.8939
27.8558
27.8814
27.8479
27.8479
27.8848
27.8809
27.8479
27.8611
27.8630
27.8679
27.8637
27.8985
27.8900
27.8577
27.8848
27.8869
27.8976
```

```
27.8610
27.8567
27.8417
27.8280
27.8555
27.8639
27.8702
27.8582
27.8605
27.8900
27.8758
27.8774
27.9008
27.8988
27.8897
27.8990
27.8958
27.8830
27.8967
27.9105
27.9028
27.8977
27.8953
27.8970
27.9190
27.9180
27.8997
27.9204
27.9234
27.9072
27.9152
27.9091
27.8882
27.9035
27.9267
27.9138
27.8955
27.9203
27.9239
27.9199
27.9646
27.9411
27.9345
27.8712
27.9145
27.9259
27.9317
27.9239
27.9247
27.9150
27.9444
27.9457
27.9166
27.9066
27.9088
27.9255
27.9312
27.9439
27.9210
27.9102
27.9083
27.9121
27.9113
27.9091
27.9235
27.9291
27.9253
27.9092
27.9117
27.9194
27.9039
27.9515
27.9143
27.9124
27.9128
27.9260
27.9339
27.9500
27.9530
27.9430
27.9400
```

27.8850
27.9350
27.9120
27.9260
27.9660
27.9280
27.9450
27.9390
27.9429
27.9207
27.9205
27.9204
27.9198
27.9246
27.9366
27.9234
27.9125
27.9032
27.9285
27.9561
27.9616
27.9530
27.9280
27.9060
27.9380
27.9310
27.9347
27.9339
27.9410
27.9397
27.9472
27.9235
27.9315
27.9368
27.9403
27.9529
27.9263
27.9347
27.9371
27.9129
27.9549
27.9422
27.9423
27.9750
27.9339
27.9629
27.9587
27.9503
27.9573
27.9518
27.9527
27.9589
27.9300
27.9629
27.9630
27.9660
27.9730
27.9660
27.9630
27.9570
27.9650
27.9520
27.9820
27.9560
27.9670
27.9520
27.9470
27.9720
27.9610
27.9437
27.9660
27.9580
27.9660
27.9700
27.9600
27.9660
27.9770
27.9110
27.9690
27.9698
27.9616

```
27.9371
27.9700
27.9265
27.9964
27.9842
27.9667
27.9610
27.9943
27.9616
27.9397
27.9799
28.0086
27.9709
27.9741
27.9675
27.9826
27.9676
27.9703
27.9789
27.9786
27.9722
27.9831
28.0043
27.9548
27.9875
27.9495
27.9549
27.9469
27.9744
27.9744
27.9449
27.9837
27.9585
28.0096
27.9762
27.9641
27.9854
27.9877
27.9839
27.9817
27.9845
27.9877
27.9880
27.9822
27.9836
28.0030
27.9678
28.0146
27.9945
27.9805
27.9785
27.9791
27.9817
27.9805
27.9782
27.9753
27.9792
27.9704
27.9794
27.9814
27.9794
27.9795
27.9881
27.9772
27.9796
27.9736
27.9772
27.9960
27.9795
27.9779
27.9829
27.9829
27.9815
27.9811
27.9773
27.9778
27.9724
27.9756
27.9699
27.9724
27.9666
```

27.9666
27.9739
27.9684
27.9861
27.9901
27.9879
27.9865
27.9876
27.9814
27.9842
27.9868
27.9834
27.9892
27.9864
27.9843
27.9838
27.9847
27.9860
27.9872
27.9869
27.9602
27.9852
27.9860
27.9836
27.9813
27.9623
27.9843
27.9802
27.9863
27.9813
27.9881
27.9850
27.9850
27.9830
27.9866
27.9888
27.9841
27.9863
27.9903
27.9961
27.9905
27.9945
27.9878
27.9929
27.9914
27.9914
27.9997
28.0006
27.9999
28.0004
28.0020
28.0029
28.0008
28.0040
28.0078
28.0065
27.9959
28.0073
28.0017
28.0042
28.0036
28.0055
28.0007
28.0066
28.0011
27.9960
28.0083
27.9978
28.0108
28.0088
28.0088
28.0139
28.0092
28.0092
28.0049
28.0111
28.0120
28.0093
28.0116
28.0102
28.0139

28.0113
28.0158
28.0156
28.0137
28.0236
28.0171
28.0224
28.0184
28.0199
28.0190
28.0204
28.0170
28.0183
28.0201
28.0182
28.0183
28.0175
28.0127
28.0211
28.0057
28.0180
28.0183
28.0149
28.0185
28.0182
28.0192
28.0213
28.0216
28.0169
28.0162
28.0167
28.0167
28.0169
28.0169
28.0161
28.0152
28.0179
28.0215
28.0194
28.0115
28.0174
28.0178
28.0202
28.0240
28.0198
28.0194
28.0171
28.0134
28.0121
28.0121
28.0141
28.0101
28.0114
28.0122
28.0124
28.0171
28.0165
28.0166
28.0159
28.0181
28.0200
28.0116
28.0144
28.0141
28.0116
28.0107
28.0169
28.0105
28.0136
28.0138
28.0114
28.0122
28.0122
28.0116
28.0025
28.0097
28.0066
28.0072
28.0066
28.0068
28.0067

28.0130
28.0091
28.0088
28.0091
28.0091
28.0115
28.0087
28.0128
28.0139
28.0095
28.0115
28.0101
28.0121
28.0114
28.0121
28.0122
28.0121
28.0168
28.0212
28.0219
28.0221
28.0204
28.0169
28.0141
28.0142
28.0147
28.0159
28.0165
28.0144
28.0182
28.0155
28.0155
28.0192
28.0204
28.0185
28.0248
28.0185
28.0226
28.0271
28.0290
28.0240
28.0302
28.0243
28.0288
28.0287
28.0301
28.0273
28.0313
28.0293
28.0300
28.0344
28.0308
28.0291
28.0287
28.0358
28.0309
28.0286
28.0308
28.0291
28.0380
28.0411
28.0420
28.0359
28.0368
28.0327
28.0361
28.0334
28.0300
28.0347
28.0359
28.0344
28.0370
28.0355
28.0371
28.0318
28.0390
28.0390
28.0390
28.0376
28.0376
28.0377

28.0345
28.0333
28.0429
28.0379
28.0401
28.0401
28.0423
28.0393
28.0382
28.0424
28.0386
28.0386
28.0373
28.0397
28.0412
28.0565
28.0419
28.0456
28.0426
28.0423
28.0391
28.0403
28.0388
28.0408
28.0457
28.0455
28.0460
28.0456
28.0464
28.0442
28.0416
28.0451
28.0432
28.0434
28.0448
28.0448
28.0373
28.0429
28.0392
28.0469
28.0443
28.0356
28.0474
28.0446
28.0348
28.0368
28.0418
28.0445
28.0533
28.0439
28.0474
28.0435
28.0419
28.0538
28.0538
28.0463
28.0491
28.0441
28.0411
28.0507
28.0459
28.0519
28.0554
28.0512
28.0507
28.0582
28.0471
28.0539
28.0530
28.0502
28.0422
28.0431
28.0395
28.0177
28.0425
28.0484
28.0693
28.0490
28.0453
28.0494
28.0522

28.0393
28.0443
28.0465
28.0450
28.0539
28.0566
28.0585
28.0486
28.0427
28.0548
28.0616
28.0298
28.0726
28.0695
28.0629
28.0503
28.0493
28.0537
28.0613
28.0643
28.0678
28.0564
28.0703
28.0647
28.0579
28.0630
28.0716
28.0586
28.0607
28.0601
28.0611
28.0606
28.0611
28.0066
28.0412
28.0558
28.0590
28.0750
28.0483
28.0599
28.0490
28.0499
28.0565
28.0612
28.0634
28.0627
28.0519
28.0551
28.0696
28.0581
28.0568
28.0572
28.0529
28.0421
28.0432
28.0211
28.0363
28.0436
28.0619
28.0573
28.0499
28.0340
28.0474
28.0534
28.0589
28.0466
28.0448
28.0576
28.0558
28.0522
28.0480
28.0444
28.0429
28.0624
28.0610
28.0461
28.0564
28.0734
28.0565
28.0503
28.0581

28.0519
28.0625
28.0583
28.0645
28.0642
28.0535
28.0510
28.0542
28.0677
28.0416
28.0676
28.0596
28.0635
28.0558
28.0623
28.0718
28.0585
28.0552
28.0684
28.0646
28.0590
28.0465
28.0594
28.0303
28.0533
28.0561
28.0585
28.0497
28.0582
28.0507
28.0562
28.0715
28.0468
28.0411
28.0587
28.0456
28.0705
28.0534
28.0558
28.0536
28.0552
28.0461
28.0598
28.0598
28.0650
28.0423
28.0442
28.0449
28.0660
28.0506
28.0655
28.0512
28.0407
28.0475
28.0411
28.0512
28.1036
28.0641
28.0572
28.0700
28.0577
28.0637
28.0534
28.0461
28.0701
28.0631
28.0575
28.0444
28.0592
28.0684
28.0593
28.0677
28.0512
28.0644
28.0660
28.0542
28.0768
28.0515
28.0579
28.0538
28.0526

28.0833
28.0637
28.0529
28.0535
28.0561
28.0736
28.0635
28.0600
28.0520
28.0695
28.0608
28.0608
28.0590
28.0290
28.0939
28.0618
28.0551
28.0757
28.0698
28.0717
28.0529
28.0644
28.0613
28.0759
28.0745
28.0736
28.0611
28.0732
28.0782
28.0682
28.0756
28.0857
28.0739
28.0840
28.0862
28.0724
28.0727
28.0752
28.0732
28.0703
28.0849
28.0795
28.0902
28.0874
28.0971
28.0638
28.0877
28.0751
28.0904
28.0971
28.0661
28.0711
28.0754
28.0516
28.0961
28.0689
28.1110
28.1062
28.0726
28.1141
28.0913
28.0982
28.0703
28.0654
28.0760
28.0727
28.0850
28.0877
28.0967
28.1185
28.0945
28.0834
28.0764
28.1129
28.0797
28.0707
28.1008
28.0971
28.0826
28.0857
28.0984

```
28.0869
28.0795
28.0875
28.1184
28.0746
28.0816
28.0879
28.0888
28.0924
28.0979
28.0702
28.0847
28.0917
28.0834
28.0823
28.0917
28.0779
28.0852
28.0863
28.0942
28.0801
28.0817
28.0922
28.0914
28.0868
28.0832
28.0881
28.0910
28.0886
28.0961
28.0857
28.0859
28.1086
28.0838
28.0921
28.0945
28.0839
28.0877
28.0803
28.0928
28.0885
28.0940
28.0856
28.0849
28.0955
28.0955
28.0846
28.0871
28.0872
28.0917
28.0931
28.0865
28.0900
28.0915
28.0963
28.0917
28.0950
28.0898
28.0902
28.0867
28.0843
28.0939
28.0902
28.0911
28.0909
28.0949
28.0867
28.0932
28.0891
28.0932
28.0887
28.0925
28.0928
28.0883
28.0946
28.0977
28.0914
28.0959
28.0926
28.0923
28.0950
```

```
28.1006
28.0924
28.0963
28.0893
28.0956
28.0980
28.0928
28.0951
28.0958
28.0912
28.0990
28.0915
28.0957
28.0976
28.0888
28.0928
28.0910
28.0902
28.0950
28.0995
28.0965
28.0972
28.0963
28.0946
28.0942
28.0998
28.0911
28.1043
28.1002
28.0991
28.0959
28.0996
28.0926
28.1002
28.0961
28.0983
28.0997
28.0959
28.0988
28.1029
28.0989
28.1000
28.0944
28.0979
28.1005
28.1012
28.1013
28.0999
28.0991
28.1059
28.0961
28.0981
28.1045
28.1047
28.1042
28.1146
28.1113
28.1051
28.1065
28.1065
28.0985
28.1000
28.1066
28.1041
28.0954
28.1090
```

1.4.2.7.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

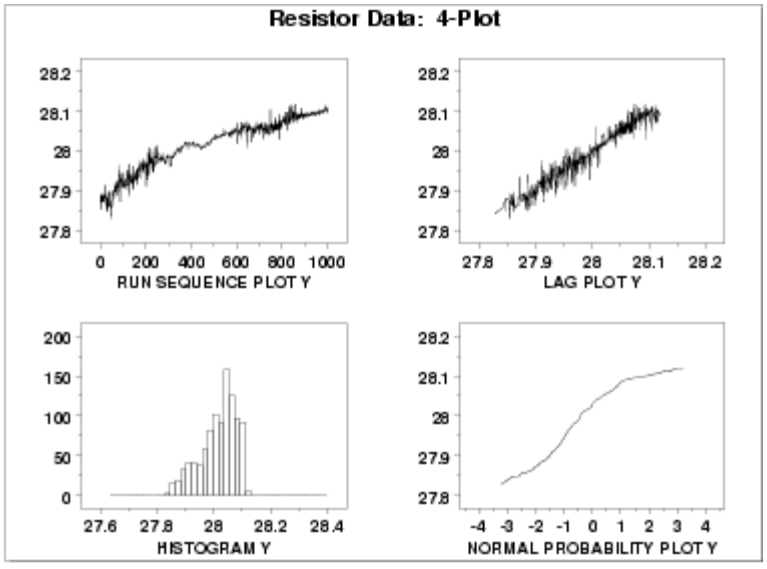
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates significant shifts in both location and variation. Specifically, the location is increasing with time. The variability seems greater in the first and last third of the data than it does in the middle third.

2. The lag plot (upper right) shows a significant non-random pattern in the data. Specifically, the strong linear appearance of this plot is indicative of a model that relates Y_t to Y_{t-1}.

3. The distributional plots, the histogram (lower left) and the normal probability plot (lower right), are not interpreted since the randomness assumption is so clearly violated.

The serious violation of the non-randomness assumption means that the univariate model

$$Y_i = C + E_i$$

is not valid. Given the linear appearance of the lag plot, the first step might be to consider a model of the type

$$Y_i = A_0 + A_1 * Y_{i-1} + E_i$$

However, discussions with the scientist revealed the following:

1. the drift with respect to location was expected.

2. the non-constant variability was not expected.

The scientist examined the data collection device and determined that the non-constant variation was a seasonal effect. The high variability data in the first and last thirds was collected in winter while the more stable middle third was collected in the summer. The seasonal effect was determined to be caused by the amount of humidity affecting the measurement equipment. In this case, the solution was to modify the test equipment to be less sensitive to enviromental factors.

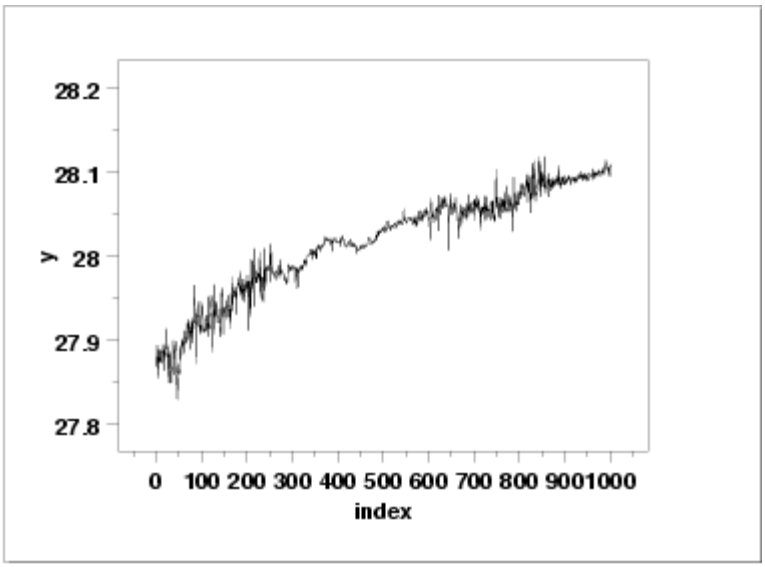
Simple graphical techniques can be quite effective in revealing unexpected results in the data. When this occurs, it is important to investigate whether the unexpected result is due to problems in the experiment and data collection, or is it in fact indicative of an unexpected underlying structure in the data. This determination cannot be made on the basis of statistics alone. The role of the graphical and statistical analysis is to detect problems or unexpected results in the data. Resolving the issues requires the knowledge of the

scientist or engineer.

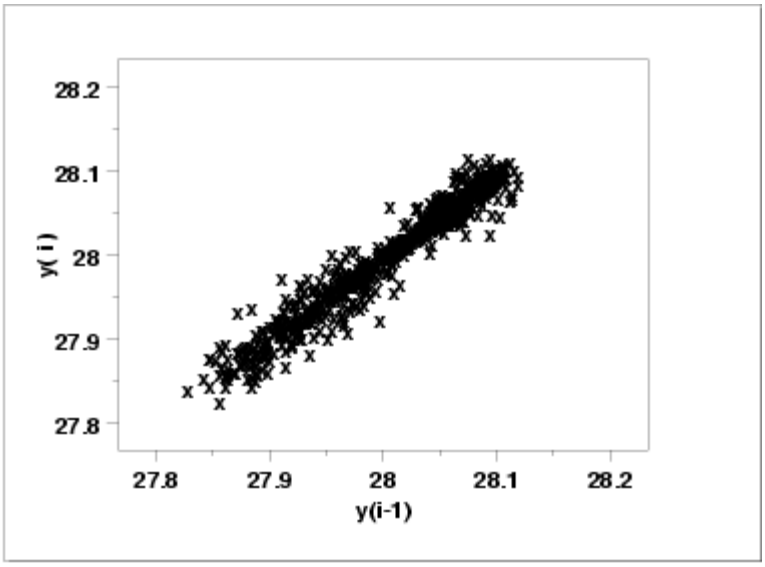
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail. Since the lag plot indicates significant non-randomness, we omit the distributional plots.

Run Sequence Plot



Lag Plot



1.4.2.7.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics are computed from the data.

```
Sample size  = 1000
Mean         =   28.01634
Median       =   28.02910
Minimum      =   27.82800
Maximum      =   28.11850
Range        =    0.29050
Stan. Dev.   =    0.06349
```

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

Coefficient	Estimate	Stan. Error	*t*-Value
B_0	27.9114	0.1209E-02	0.2309E+05
B_1	0.20967E-03	0.2092E-05	100.2

```
Residual Standard Deviation = 0.1909796E-01
Residual Degrees of Freedom = 998
```

The slope parameter, B_1, has a *t* value of 100.2 which is statistically significant. The value of the slope parameter estimate is 0.00021. Although this number is nearly zero, we need to take into account that the original scale of the data is from about 27.8 to 28.2. In this case, we conclude that there is a drift in location.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since the normality assumption is questionable for these data, we use the alternative Levene test. In particular, we use the Levene test based on the median rather the mean. The choice of the number of

intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$
H_a: At least one σ_i^2 is not equal to the others.

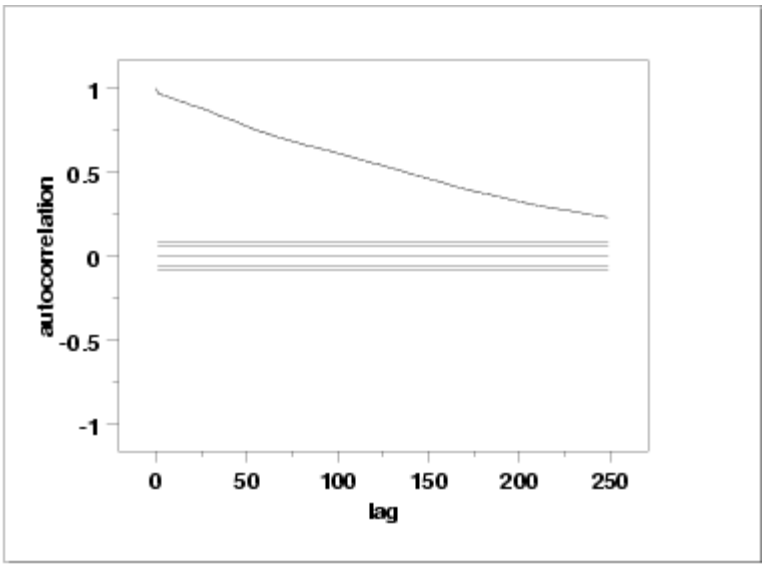
Test statistic: $W = 140.85$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $F_{\alpha,k-1,N-k} = 2.614$
Critical region: Reject H_0 if $W > 2.614$

In this case, since the Levene test statistic value of 140.85 is greater than the 5 % significance level critical value of 2.614, we conclude that there is significant evidence of nonconstant variation.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot in the previous section is a simple graphical technique.

One check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95 % and 99 % confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of greatest interest, is 0.97. The critical values at the 5 % significance level are -0.062 and 0.062. This indicates that the lag 1 autocorrelation is statistically significant, so there is strong evidence of non-randomness.

A common test for randomness is the runs test.

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a

random manner

Test statistic: $Z = -30.5629$
Significance level: $\alpha = 0.05$
Critical value: $Z_{1-\alpha/2} = 1.96$
Critical region: Reject H_0 if $|Z| > 1.96$

Because the test statistic is outside of the critical region, we reject the null hypothesis and conclude that the data are not random.

Distributional Analysis

Since we rejected the randomness assumption, the distributional tests are not meaningful. Therefore, these quantitative tests are omitted. Since the Grubbs' test for outliers also assumes the approximate normality of the data, we omit Grubbs' test as well.

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
 Analysis for resistor case study

 1: Sample Size                              = 1000

 2: Location
    Mean                                     =
28.01635
    Standard Deviation of Mean               =
0.002008
    95% Confidence Interval for Mean         =
(28.0124,28.02029)
    Drift with respect to location?          = NO

 3: Variation
    Standard Deviation                       =
0.063495
    95% Confidence Interval for SD           =
(0.060829,0.066407)
    Change in variation?
    (based on Levene's test on quarters
    of the data)                              = YES

 4: Randomness
    Autocorrelation                          =
0.972158
    Data Are Random?
      (as measured by autocorrelation)      = NO

 5: Distribution
    Distributional test omitted due to
    non-randomness of the data

 6: Statistical Control
    (i.e., no drift in location or scale,
    data are random, distribution is
    fixed)
    Data Set is in Statistical Control?   = NO

 7: Outliers?
    (Grubbs' test omitted due to
    non-randomness of the data)
```

1.4.2.7.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step. *NOTE: This case study has 1,000 points. For better performance, it is highly recommended that you check the "No Update" box on the Spreadsheet window for this case study. This will suppress subsequent updating of the Spreadsheet window as the data are created or modified.*	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are shifts

	in location and variation and the data are not random.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot.	1. The run sequence plot indicates that there are shifts of location and variation. 2. The lag plot shows a strong linear pattern, which indicates significant non-randomness.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the sample mean, a confidence interval for the population mean, and compute a linear fit to detect drift in location. 3. Generate the sample standard deviation, a confidence interval for the population standard deviation, and detect drift in variation by dividing the data into quarters and computing Levene's test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Print a univariate report (this assumes steps 2 thru 5 have already been run).	1. The summary statistics table displays 25+ statistics. 2. The mean is 28.0163 and a 95% confidence interval is (28.0124,28.02029). The linear fit indicates drift in location since the slope parameter estimate is statistically significant. 3. The standard deviation is 0.0635 with a 95% confidence interval of (0.060829,0.066407). Levene's test indicates significant change in variation. 4. The lag 1 autocorrelation is 0.97. From the autocorrelation plot, this is outside the 95% confidence interval bands, indicating significant non-randomness. 5. The results are

	summarized in a convenient report.

1.4.2.8. Heat Flow Meter 1

Heat Flow Meter Calibration and Stability	This example illustrates the univariate analysis of standard resistor data.

1. Background and Data
2. Graphical Output and Interpretation
3. Quantitative Output and Interpretation
4. Work This Example Yourself

1.4.2.8.1. Background and Data

Generation This data set was collected by Bob Zarr of NIST in January, 1990 from a heat flow meter calibration and stability analysis. The response variable is a calibration factor.

The motivation for studying this data set is to illustrate a well-behaved process where the underlying assumptions hold and the process is in statistical control.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

Data The following are the data used for this case study.

```
9.206343
9.299992
9.277895
9.305795
9.275351
9.288729
9.287239
9.260973
9.303111
9.275674
9.272561
9.288454
9.255672
9.252141
9.297670
9.266534
9.256689
9.277542
9.248205
9.252107
9.276345
9.278694
9.267144
9.246132
9.238479
9.269058
9.248239
9.257439
9.268481
9.288454
9.258452
9.286130
9.251479
9.257405
9.268343
9.291302
9.219460
9.270386
9.218808
9.241185
9.269989
9.226585
```

```
9.258556
9.286184
9.320067
9.327973
9.262963
9.248181
9.238644
9.225073
9.220878
9.271318
9.252072
9.281186
9.270624
9.294771
9.301821
9.278849
9.236680
9.233988
9.244687
9.221601
9.207325
9.258776
9.275708
9.268955
9.257269
9.264979
9.295500
9.292883
9.264188
9.280731
9.267336
9.300566
9.253089
9.261376
9.238409
9.225073
9.235526
9.239510
9.264487
9.244242
9.277542
9.310506
9.261594
9.259791
9.253089
9.245735
9.284058
9.251122
9.275385
9.254619
9.279526
9.275065
9.261952
9.275351
9.252433
9.230263
9.255150
9.268780
9.290389
9.274161
9.255707
9.261663
9.250455
9.261952
9.264041
9.264509
9.242114
9.239674
9.221553
9.241935
9.215265
9.285930
9.271559
9.266046
9.285299
9.268989
9.267987
9.246166
9.231304
9.240768
9.260506
```

```
9.274355
9.292376
9.271170
9.267018
9.308838
9.264153
9.278822
9.255244
9.229221
9.253158
9.256292
9.262602
9.219793
9.258452
9.267987
9.267987
9.248903
9.235153
9.242933
9.253453
9.262671
9.242536
9.260803
9.259825
9.253123
9.240803
9.238712
9.263676
9.243002
9.246826
9.252107
9.261663
9.247311
9.306055
9.237646
9.248937
9.256689
9.265777
9.299047
9.244814
9.287205
9.300566
9.256621
9.271318
9.275154
9.281834
9.253158
9.269024
9.282077
9.277507
9.284910
9.239840
9.268344
9.247778
9.225039
9.230750
9.270024
9.265095
9.284308
9.280697
9.263032
9.291851
9.252072
9.244031
9.283269
9.196848
9.231372
9.232963
9.234956
9.216746
9.274107
9.273776
```

1.4.2.8.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

   $$Y_i = C + E_i$$

   is appropriate and valid.

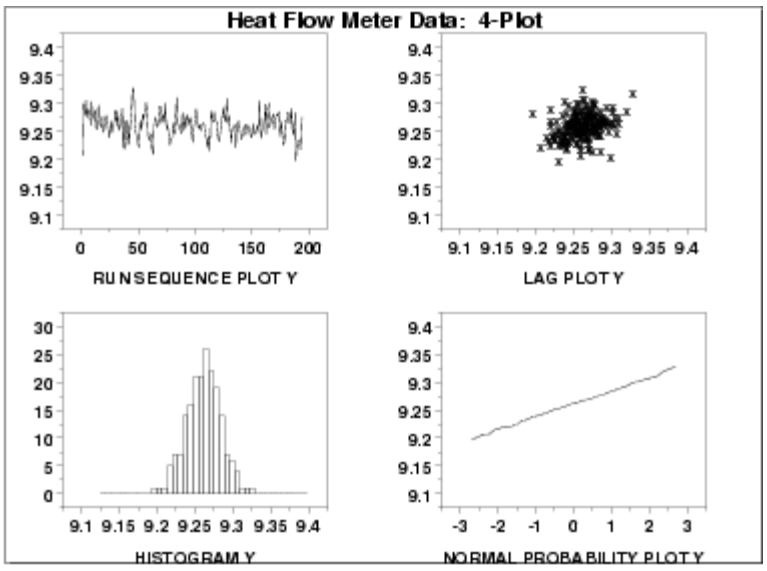
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

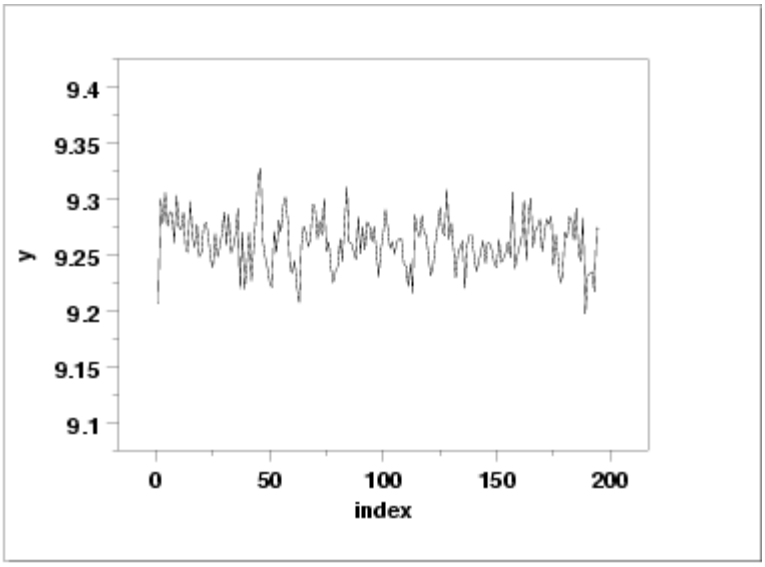
The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram (lower left) shows that the data are reasonably symmetric, there does not appear to be significant outliers in the tails, and it seems reasonable to assume that the data are from approximately a normal distribution.
4. The normal probability plot (lower right) verifies that an assumption of normality is in fact reasonable.

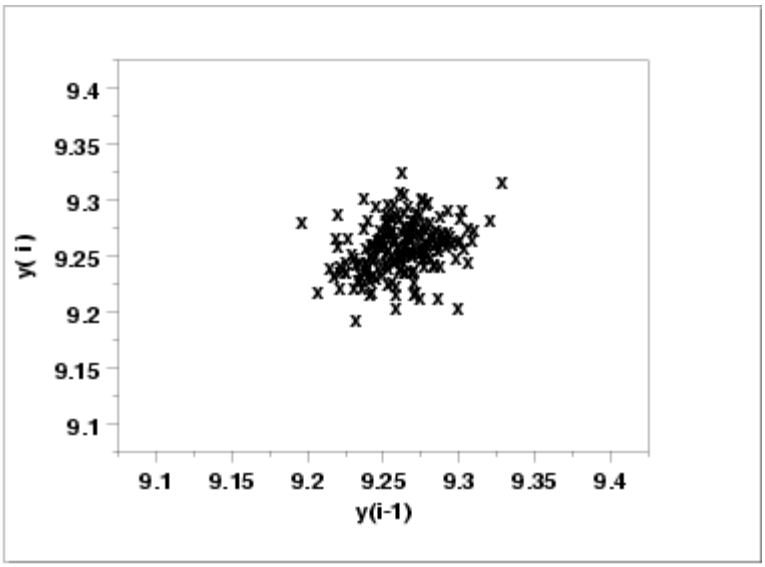
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail.

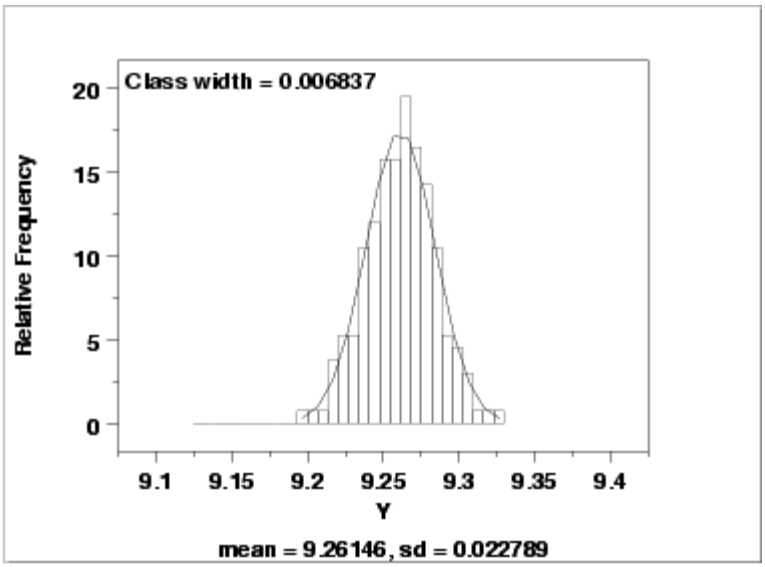
Run Sequence Plot



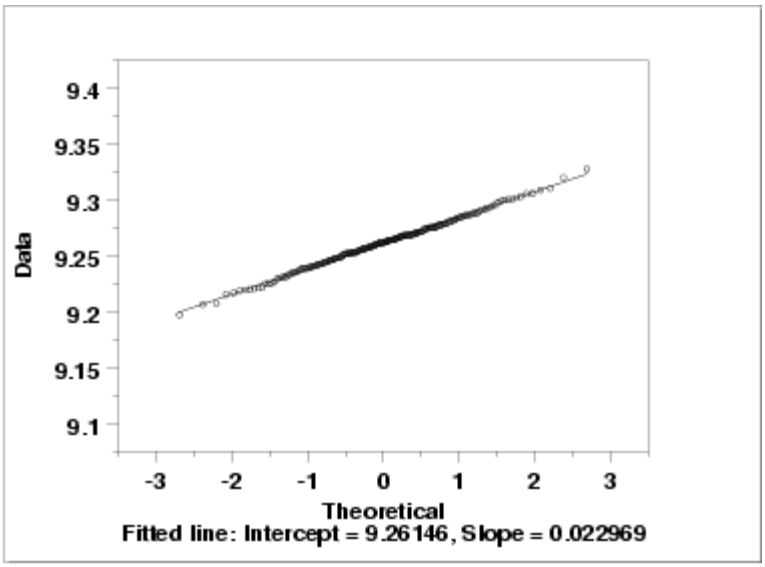
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot



1.4.2.8.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, common summary statistics are computed from the data.

```
Sample size  = 195
Mean         =   9.261460
Median       =   9.261952
Minimum      =   9.196848
Maximum      =   9.327973
Range        =   0.131126
Stan. Dev.   =   0.022789
```

Location

One way to quantify a change in location over time is to fit a straight line to the data using an index variable as the independent variable in the regression. For our data, we assume that data are in sequential run order and that the data were collected at equally spaced time intervals. In our regression, we use the index variable X = 1, 2, ..., N, where N is the number of observations. If there is no significant drift in the location over time, the slope parameter should be zero.

Coefficient	Estimate	Stan. Error	*t*-Value
B_0	9.26699	0.3253E-02	2849.
B_1	-0.56412E-04	0.2878E-04	-1.960

```
Residual Standard Deviation = 0.2262372E-01
Residual Degrees of Freedom = 193
```

The slope parameter, B_1, has a *t* value of -1.96 which is (barely) statistically significant since it is essentially equal to the 95 % level cutoff of -1.96. However, notice that the value of the slope parameter estimate is -0.00056. This slope, even though statistically significant, can essentially be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. The choice of the number of intervals is somewhat arbitrary, although values of four or eight are reasonable. We will divide our data into four intervals.

$$2 \quad 2 \quad 2 \quad 2$$

H_0: $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$
H_a: At least one σ_i^2 is not equal to the others.

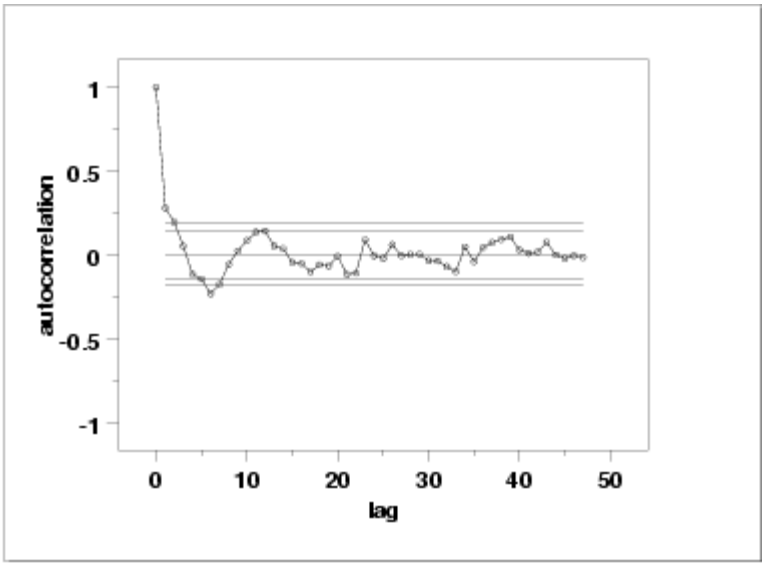
Test statistic: $T = 3.147$
Degrees of freedom: $k - 1 = 3$
Significance level: $\alpha = 0.05$
Critical value: $X^2_{1-\alpha,k-1} = 7.815$
Critical region: Reject H_0 if $T > 7.815$

In this case, since the Bartlett test statistic of 3.147 is less than the critical value at the 5 % significance level of 7.815, we conclude that the variances are not significantly different in the four intervals. That is, the assumption of constant scale is valid.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the previous section is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95 % and 99 % confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1).



The lag 1 autocorrelation, which is generally the one of greatest interest, is 0.281. The critical values at the 5 % significance level are -0.087 and 0.087. This indicates that the lag 1 autocorrelation is statistically significant, so there is evidence of non-randomness.

A common test for randomness is the runs test.

H_0: the sequence was produced in a random manner
H_a: the sequence was not produced in a random manner

Test statistic: $Z = -3.2306$
Significance level: $\alpha = 0.05$

```
Critical value:  Z1-α/2 = 1.96
Critical region:  Reject H0 if |Z| > 1.96
```

The value of the test statistic is less than -1.96, so we reject the null hypothesis at the 0.05 significant level and conclude that the data are not random.

Although the autocorrelation plot and the runs test indicate some mild non-randomness, the violation of the randomness assumption is not serious enough to warrant developing a more sophisticated model. It is common in practice that some of the assumptions are mildly violated and it is a judgement call as to whether or not the violations are serious enough to warrant developing a more sophisticated model for the data.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot. For this data set the correlation coefficient is 0.996. Since this is greater than the critical value of 0.987 (this is a tabulated value), the normality assumption is not rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. The results of the Anderson-Darling test follow.

```
H0:  the data are normally distributed
Ha:  the data are not normally distributed

Adjusted test statistic:  A 2 = 0.129
Significance level:  α = 0.05
Critical value:  0.787
Critical region:  Reject H0 if A 2 > 0.787
```

The Anderson-Darling test also does not reject the normality assumption because the test statistic, 0.129, is less than the critical value at the 5 % significance level of 0.787.

Outlier Analysis

A test for outliers is the Grubbs' test.

```
H0:  there are no outliers in the data
Ha:  the maximum value is an outlier

Test statistic:  G = 2.918673
Significance level:  α = 0.05
Critical value for an upper one-tailed
test:  3.597898
Critical region:  Reject H0 if G > 3.597898
```

For this data set, Grubbs' test does not detect any outliers at the 0.05 significance level.

Model

Since the underlying assumptions were validated both

graphically and analytically, with a mild violation of the randomness assumption, we conclude that a reasonable model for the data is:

$$Y_i = 9.26146 + E_i$$

We can express the uncertainty for ***C***, here estimated by 9.26146, as the 95 % confidence interval (9.258242,9.26479).

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report. The report for the heat flow meter data follows.

```
 Analysis for heat flow meter data

 1: Sample Size                              = 195

 2: Location
    Mean                                     =
9.26146
    Standard Deviation of Mean               =
0.001632
    95 % Confidence Interval for Mean        =
(9.258242,9.264679)
    Drift with respect to location?          = NO

 3: Variation
    Standard Deviation                       =
0.022789
    95 % Confidence Interval for SD          =
(0.02073,0.025307)
    Drift with respect to variation?
    (based on Bartlett's test on quarters
    of the data)                              = NO

 4: Randomness
    Autocorrelation                          =
0.280579
    Data are Random?
      (as measured by autocorrelation)      = NO

 5: Data are Normal?
      (as tested by Anderson-Darling)       = YES

 6: Statistical Control
    (i.e., no drift in location or scale,
    data are random, distribution is
    fixed, here we are testing only for
    fixed normal)
    Data Set is in Statistical Control?    = YES

 7: Outliers?
    (as determined by Grubbs' test)          = NO
```

1.4.2.8.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale, and the data seem to follow a normal distribution.
3. Generate the individual plots.	

1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf. 4. Generate a normal probability plot.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is a good distribution for these data. 4. The normal probability plot verifies that the normal distribution is a reasonable distribution for these data.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Bartlett's test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Check for normality by computing the normal probability plot correlation coefficient. 6. Check for outliers using Grubbs' test. 7. Print a univariate report (this	1. The summary statistics table displays 25+ statistics. 2. The mean is 9.261 and a 95% confidence interval is (9.258,9.265). The linear fit indicates no drift in location since the slope parameter estimate is essentially zero. 3. The standard deviation is 0.023 with a 95% confidence interval of (0.0207,0.0253). Bartlett's test indicates no significant change in variation. 4. The lag 1 autocorrelation is 0.28. From the autocorrelation plot, this is statistically significant at the 95% level.

assumes steps 2 thru 6 have already been run).	5. The normal probability plot correlation coefficient is 0.999. At the 5% level, we cannot reject the normality assumption. 6. Grubbs' test detects no outliers at the 5% level. 7. The results are summarized in a convenient report.

1.4.2.9. Fatigue Life of Aluminum Alloy Specimens

Fatigue Life of Aluminum Alloy Specimens

This example illustrates the univariate analysis of the fatigue life of aluminum alloy specimens.

1. Background and Data
2. Graphical Output and Interpretation

1.4.2.9.1. Background and Data

Generation

This data set comprises measurements of fatigue life (thousands of cycles until rupture) of rectangular strips of 6061-T6 aluminum sheeting, subjected to periodic loading with maximum stress of 21,000 psi (pounds per square inch), as reported by Birnbaum and Saunders (1958).

Purpose of Analysis

The goal of this case study is to select a probabilistic model, from among several reasonable alternatives, to describe the dispersion of the resulting measured values of life-length.

The original study, in the field of statistical reliability analysis, was concerned with the prediction of failure times of a material subjected to a load varying in time. It was well-known that a structure designed to withstand a particular static load may fail sooner than expected under a dynamic load.

If a realistic model for the probability distribution of lifetime can be found, then it can be used to estimate the time by which a part or structure needs to be replaced to guarantee that the probability of failure does not exceed some maximum acceptable value, for example 0.1 %, while it is in service.

The chapter of this eHandbook that is concerned with the assessment of product reliability contains additional material on statistical methods used in reliability analysis. This case study is meant to complement that chapter by showing the use of graphical and other techniques in the model selection stage of such analysis.

When there is no cogent reason to adopt a particular model, or when none of the models under consideration seems adequate for the purpose, one may opt for a non-parametric statistical method, for example to produce tolerance bounds or confidence intervals.

A non-parametric method does not rely on the assumption that the data are like a sample from a particular probability distribution that is fully specified up to the values of some adjustable parameters. For example, the Gaussian probability distribution is a parametric model with two adjustable parameters.

The price to be paid when using non-parametric methods is loss of efficiency, meaning that they may require more data for statistical inference than a parametric counterpart would, if applicable. For example, non-parametric confidence intervals for model parameters may be considerably wider than what a confidence interval would need to be if the underlying distribution could be identified correctly. Such identification is what we will attempt in this case study.

It should be noted --- a point that we will stress later in the development of this case study --- that the very exercise of selecting a model often contributes substantially to the uncertainty of the conclusions derived after the selection has been made.

Software The analyses used in this case study can be generated using R code.

Data The following data are used for this case study.

```
  370 1016 1235 1419 1567 1820
  706 1018 1238 1420 1578 1868
  716 1020 1252 1420 1594 1881
  746 1055 1258 1450 1602 1890
  785 1085 1262 1452 1604 1893
  797 1102 1269 1475 1608 1895
  844 1102 1270 1478 1630 1910
  855 1108 1290 1481 1642 1923
  858 1115 1293 1485 1674 1940
  886 1120 1300 1502 1730 1945
  886 1134 1310 1505 1750 2023
  930 1140 1313 1513 1750 2100
  960 1199 1315 1522 1763 2130
  988 1200 1330 1522 1768 2215
  990 1200 1355 1530 1781 2268
 1000 1203 1390 1540 1782 2440
 1010 1222 1416 1560 1792
```

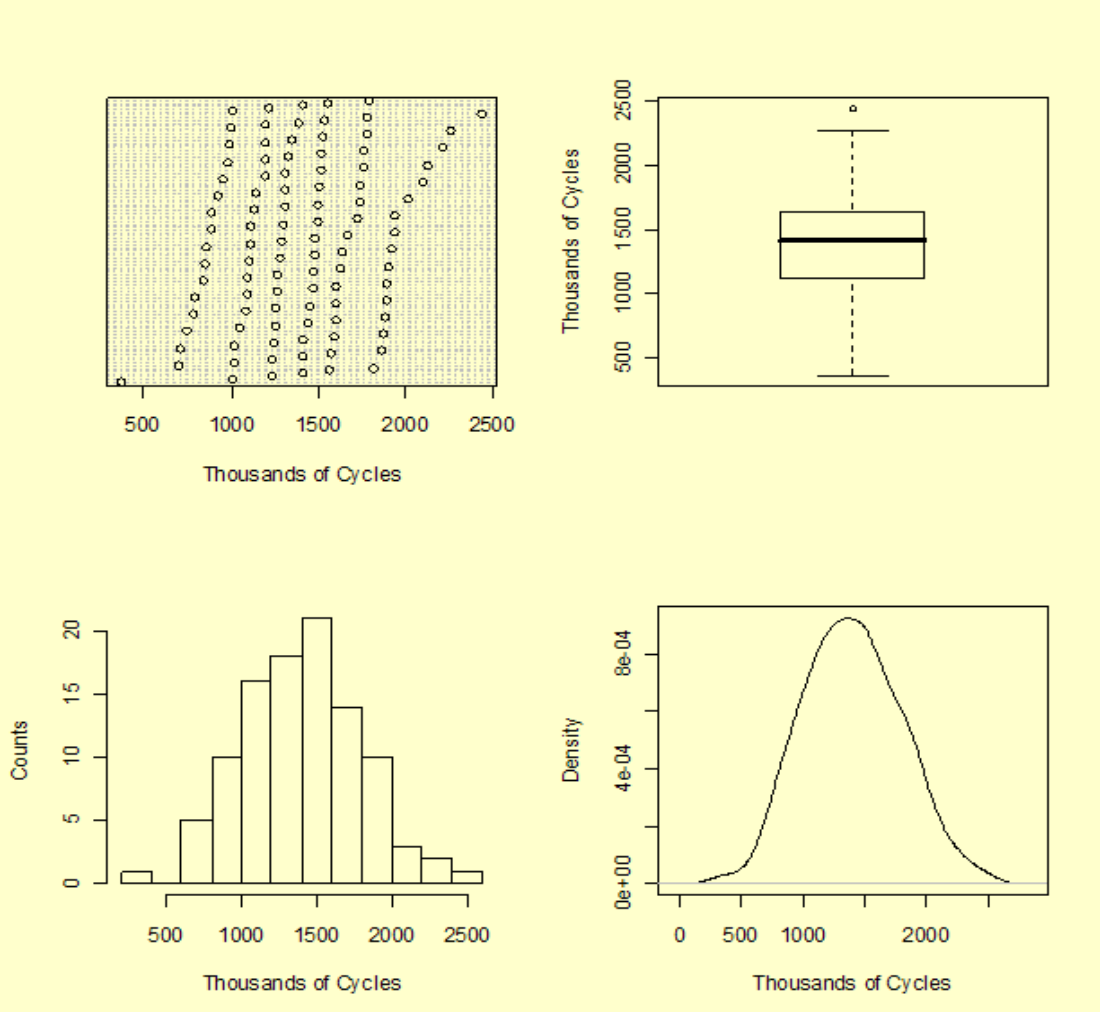
1.4.2.9.2. Graphical Output and Interpretation

Goal

The goal of this analysis is to select a probabilistic model to describe the dispersion of the measured values of fatigue life of specimens of an aluminum alloy described in [1.4.2.9.1], from among several reasonable alternatives.

Initial Plots of the Data

Simple diagrams can be very informative about location, spread, and to detect possibly anomalous data values or particular patterns (clustering, for example). These include dot-charts, boxplots, and histograms. Since building an effective histogram requires that a choice be made of bin size, and this choice can be influential, one may wish to examine a non-parametric estimate of the underlying probability density.



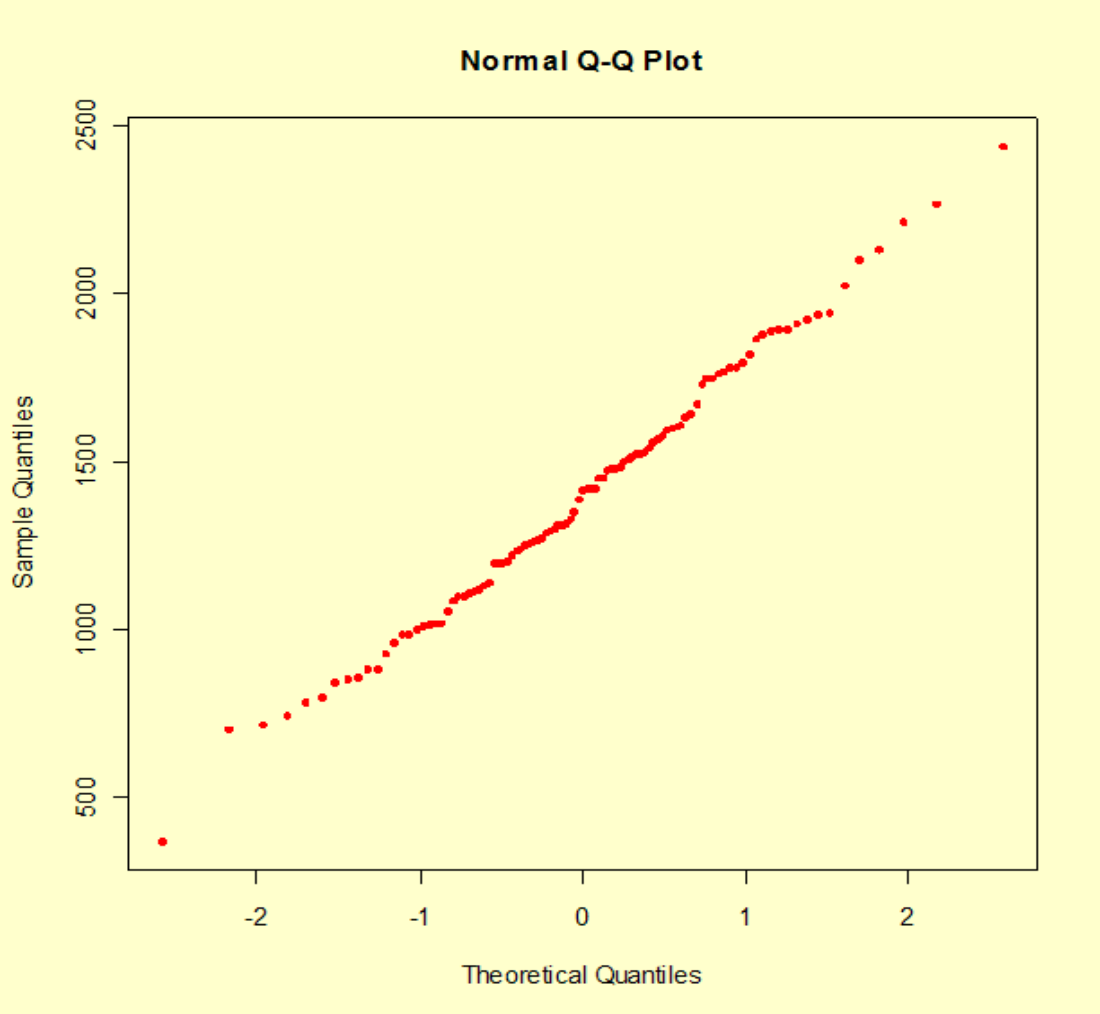
These several plots variously show that the measurements range from a value

slightly greater than 350,000 to slightly less than 2,500,000 cycles. The boxplot suggests that the largest measured value may be an outlier.

A recommended first step is to check consistency between the data and what is to be expected if the data were a sample from a particular probability distribution. Knowledge about the underlying properties of materials and of relevant industrial processes typically offer clues as to the models that should be entertained. Graphical diagnostic techniques can be very useful at this exploratory stage: foremost among these, for univariate data, is the quantile-quantile plot, or QQ-plot (Wilk and Gnanadesikan, 1968).

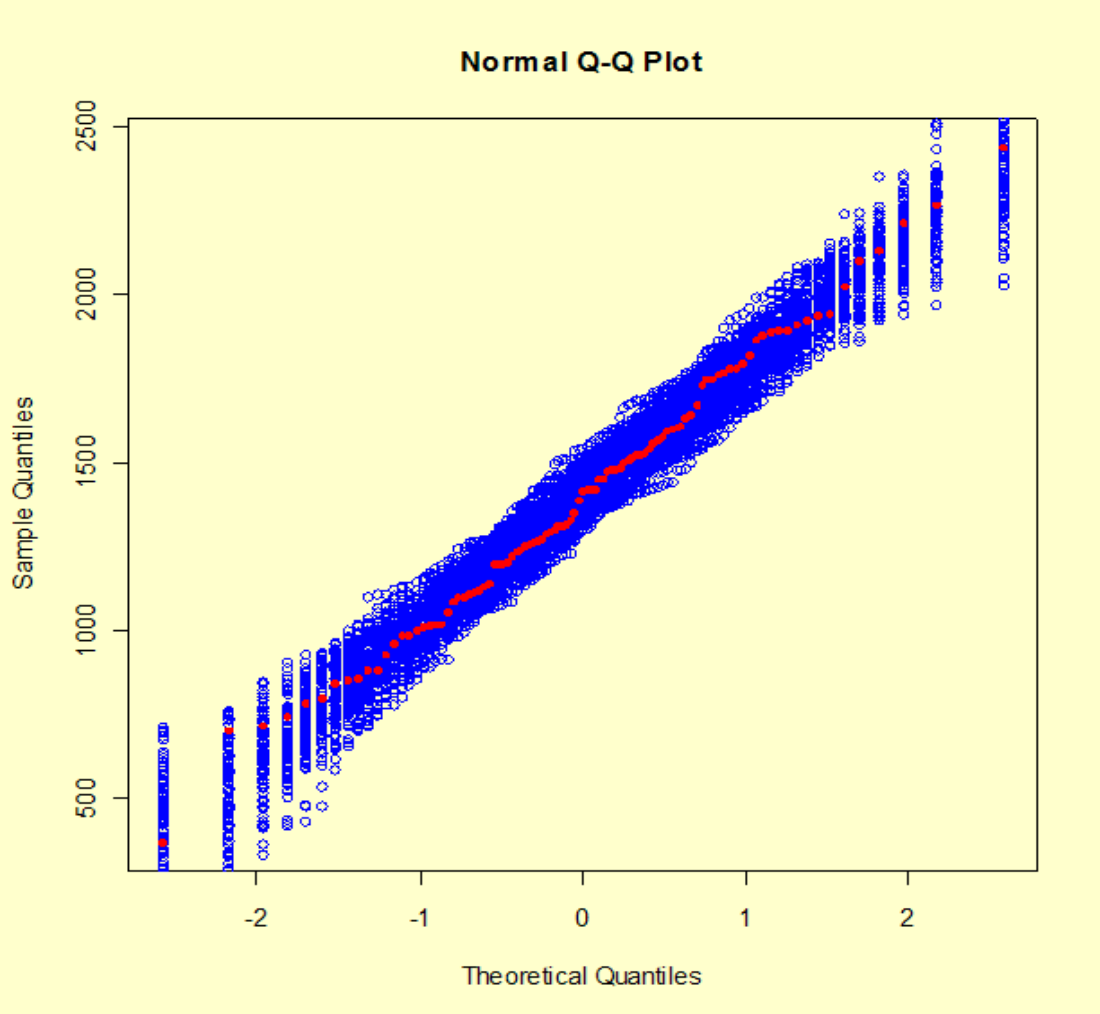
Each data point is represented by one point in the QQ-plot. The ordinate of each of these points is one data value; if this data value happens to be the *k*th order statistic in the sample (that is, the *k*th largest value), then the corresponding abscissa is the "typical" value that the *k*th largest value should have in a sample of the same size as the data, drawn from a particular distribution. If *F* denotes the cumulative probability distribution function of interest, and the sample comprises *n* values, then $F^{-1}[(k - 1/2) / (n + 1/2)]$ is a reasonable choice for that "typical" value, because it is an approximation to the median of the *k*th order statistic in a sample of size *n* from this distribution.

The following figure shows a QQ-plot of our data relative to the Gaussian (or, normal) probability distribution. If the data matched expectations perfectly, then the points would all fall on a straight line.



In practice, one needs to gauge whether the deviations from such perfect alignment are commensurate with the natural variability associated with sampling. This can easily be done by examining how variable QQ-plots of samples from the target distribution may be.

The following figure shows, superimposed on the QQ-plot of the data, the QQ-plots of 99 samples of the same size as the data, drawn from a Gaussian distribution with the same mean and standard deviation as the data.



The fact that the cloud of QQ-plots corresponding to 99 samples from the Gaussian distribution effectively covers the QQ-plot for the data, suggests that the chances are better than 1 in 100 that our data are inconsistent with the Gaussian model.

This proves nothing, of course, because even the rarest of events may happen. However, it is commonly taken to be indicative of an acceptable fit for general purposes. In any case, one may naturally wonder if an alternative model might not provide an even better fit.

Knowing the provenance of the data, that they portray strength of a material, strongly suggests that one may like to examine alternative models, because in many studies of reliability non-Gaussian models tend to be more appropriate than Gaussian models.

Candidate Distributions

There are many probability distributions that could reasonably be entertained as candidate models for the data. However, we will restrict ourselves to consideration

of the following because these have proven to be useful in reliability studies.

- Normal distribution
- Gamma distribution
- Birnbaum-Saunders distribution
- 3-parameter Weibull distribution

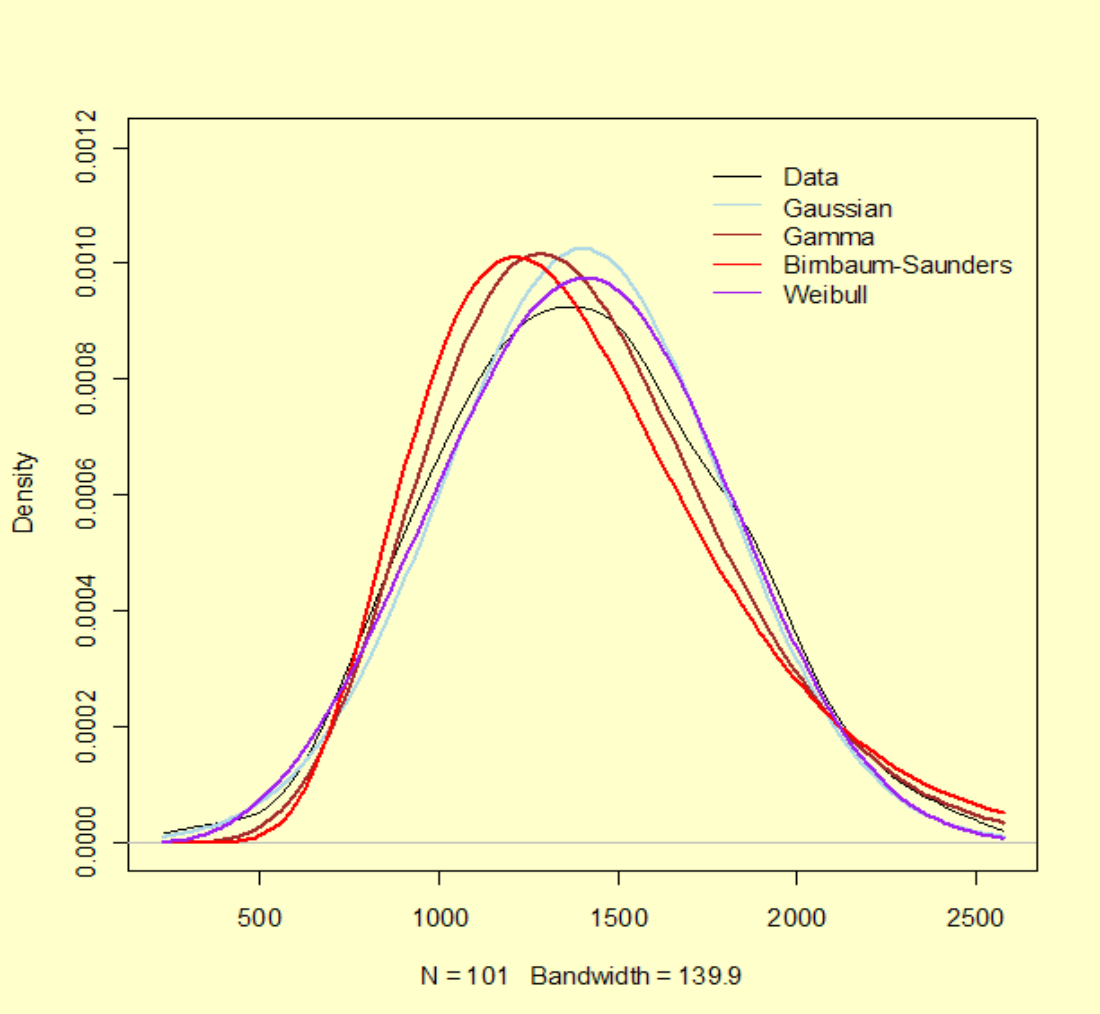
Approach

A very simple approach amounts to comparing QQ-plots of the data for the candidate models under consideration. This typically involves first fitting the models to the data, for example employing the method of maximum likelihood [1.3.6.5.2].

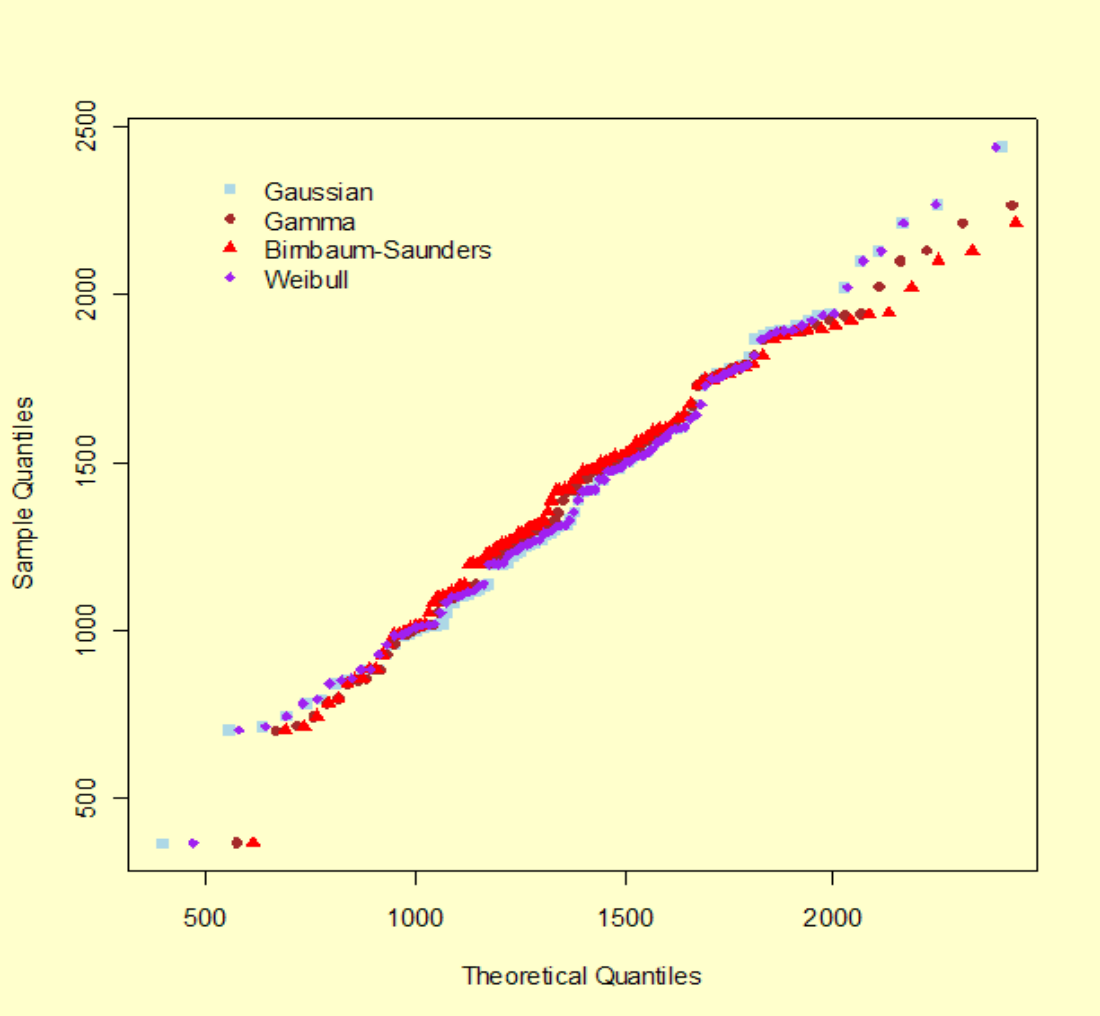
The maximum likelihood estimates are the following:

- Gaussian: mean 1401, standard deviation 389
- Gamma: shape 11.85, rate 0.00846
- Birnbaum-Saunders: shape 0.310, scale 1337
- 3-parameter Weibull: location 181, shape 3.43, scale 1357

The following figure shows how close (or how far) the best fitting probability densities of the four distributions approximate the non-parametric probability density estimate. This comparison, however, takes into account neither the fact that our sample is fairly small (101 measured values), nor that the fitted models themselves have been estimated from the same data that the non-parametric estimate was derived from.



These limitations notwithstanding, it is worth examining the corresponding QQ-plots, shown below, which suggest that the Gaussian and the 3-parameter Weibull may be the best models.



Model Selection

A more careful comparison of the merits of the alternative models needs to take into account the fact that the 3-parameter Weibull model (precisely because it has three parameters), may be intrinsically more flexible than the others, which all have two adjustable parameters only.

Two criteria can be employed for a formal comparison: Akaike's Information Criterion (AIC), and the Bayesian Information Criterion (BIC) (Hastie et. al., 2001). The smaller the value of either model selection criterion, the better the model:

```
     AIC  BIC
GAU 1495 1501
GAM 1499 1504
BS  1507 1512
WEI 1498 1505
```

On this basis (and according both to AIC and BIC), there seems to be no cogent reason to replace the Gaussian model by any of the other three. The values of BIC can also be used to derive an approximate answer to the question of how strongly the data may support each of these models. Doing this involves the application of Bayesian statistical methods [8.1.10].

We start from an *a priori* assignment of equal probabilities to all four models,

indicating that we have no reason to favor one over another at the outset, and then update these probabilities based on the measured values of lifetime. The updated probabilities of the four models, called their *posterior probabilities*, are approximately proportional to exp(-BIC(GAU)/2), exp(-BIC(GAM)/2), exp(-BIC(BS)/2), and exp(-BIC(WEI)/2). The values are 76 % for GAU, 16 % for GAM, 0.27 % for BS, and 7.4 % for WEI.

One possible use for the selected model is to answer the question of the age in service by which a part or structure needs to be replaced to guarantee that the probability of failure does not exceed some maximum acceptable value, for example 0.1 %.The answer to this question is the 0.1st percentile of the fitted distribution, that is $G^{-1}(0.001)$ = 198 thousand cycles, where, in this case, G^{-1} denotes the inverse of the fitted, Gaussian probability distribution.

To assess the uncertainty of this estimate one may employ the statistical bootstrap [1.3.3.4]. In this case, this involves drawing a suitably large number of bootstrap samples from the data, and for each of them applying the model fitting and model selection exercise described above, ending with the calculation of $G^{-1}(0.001)$ for the best model (which may vary from sample to sample).

The bootstrap samples should be of the same size as the data, with each being drawn uniformly at random from the data, *with* replacement. This process, based on 5,000 bootstrap samples, yielded a 95 % confidence interval for the 0.1st percentile ranging from 40 to 366 thousands of cycles. The large uncertainty is not surprising given that we are attempting to estimate the largest value that is exceeded with probability 99.9 %, based on a sample comprising only 101 measured values.

Prediction Intervals

One more application in this analysis is to evaluate prediction intervals for the fatigue life of the aluminum alloy specimens. For example, if we were to test three new specimens using the same process, we would want to know (with 95 % confidence) the minimum number of cycles for these three specimens. That is, we need to find a statistical interval $[L, \infty]$ that contains the fatigue life of all three future specimens with 95 % confidence. The desired interval is a one-sided, lower 95 % prediction interval. Since tables of factors for constructing L, are widely available for normal models, we use the results corresponding to the normal model here for illustration. Specifically, L is computed as

$$L = \bar{x} + rs$$
$$L = 1400.91 - 2.16(391.32) = 555.66 \text{ cycles} \times 1000$$

where factor r is given in Table A.14 of Hahn and Meeker (1991) or can be obtained from an R program.

1.4.2.10. Ceramic Strength

Ceramic Strength

This case study analyzes the effect of machining factors on the strength of ceramics.

1. Background and Data
2. Analysis of the Response Variable
3. Analysis of Batch Effect
4. Analysis of Lab Effect
5. Analysis of Primary Factors
6. Work This Example Yourself

1.4.2.10.1. Background and Data

Generation

The data for this case study were collected by Said Jahanmir of the NIST Ceramics Division in 1996 in connection with a NIST/industry ceramics consortium for strength optimization of ceramic strength

The motivation for studying this data set is to illustrate the analysis of multiple factors from a designed experiment

This case study will utilize only a subset of a full study that was conducted by Lisa Gill and James Filliben of the NIST Statistical Engineering Division

The response variable is a measure of the strength of the ceramic material (bonded S_i nitrate). The complete data set contains the following variables:

1. Factor 1 = Observation ID, i.e., run number (1 to 960)
2. Factor 2 = Lab (1 to 8)
3. Factor 3 = Bar ID within lab (1 to 30)
4. Factor 4 = Test number (1 to 4)
5. Response Variable = Strength of Ceramic
6. Factor 5 = Table speed (2 levels: 0.025 and 0.125)
7. Factor 6 = Down feed rate (2 levels: 0.050 and 0.125)
8. Factor 7 = Wheel grit size (2 levels: 150 and 80)
9. Factor 8 = Direction (2 levels: longitudinal and transverse)
10. Factor 9 = Treatment (1 to 16)
11. Factor 10 = Set of 15 within lab (2 levels: 1 and 2)
12. Factor 11 = Replication (2 levels: 1 and 2)
13. Factor 12 = Bar Batch (1 and 2)

The four primary factors of interest are:

1. Table speed (X1)
2. Down feed rate (X2)
3. Wheel grit size (X3)
4. Direction (X4)

For this case study, we are using only half the data. Specifically, we are using the data with the direction longitudinal. Therefore, we have only three primary factors

In addition, we are interested in the nuisance factors

1. Lab
2. Batch

Purpose of Analysis

The goals of this case study are:

1. Determine which of the four primary factors has the strongest effect on the strength of the ceramic material
2. Estimate the magnitude of the effects
3. Determine the optimal settings for the primary factors
4. Determine if the nuisance factors (lab and batch) have an effect on the ceramic strength

This case study is an example of a designed experiment. The Process Improvement chapter contains a detailed discussion of the construction and analysis of designed experiments. This case study is meant to complement the material in that chapter by showing how an EDA approach (emphasizing the use of graphical techniques) can be used in the analysis of designed experiments

Software

The analyses used in this case study can be generated using both Dataplot code and R code.

Data

The following are the data used for this case study

Run	Lab	Batch	Y	X1	X2	X3
1	1	1	608.781	-1	-1	-1
2	1	2	569.670	-1	-1	-1
3	1	1	689.556	-1	-1	-1
4	1	2	747.541	-1	-1	-1
5	1	1	618.134	-1	-1	-1
6	1	2	612.182	-1	-1	-1
7	1	1	680.203	-1	-1	-1
8	1	2	607.766	-1	-1	-1
9	1	1	726.232	-1	-1	-1
10	1	2	605.380	-1	-1	-1
11	1	1	518.655	-1	-1	-1
12	1	2	589.226	-1	-1	-1
13	1	1	740.447	-1	-1	-1
14	1	2	588.375	-1	-1	-1
15	1	1	666.830	-1	-1	-1
16	1	2	531.384	-1	-1	-1
17	1	1	710.272	-1	-1	-1
18	1	2	633.417	-1	-1	-1
19	1	1	751.669	-1	-1	-1
20	1	2	619.060	-1	-1	-1
21	1	1	697.979	-1	-1	-1
22	1	2	632.447	-1	-1	-1
23	1	1	708.583	-1	-1	-1
24	1	2	624.256	-1	-1	-1
25	1	1	624.972	-1	-1	-1
26	1	2	575.143	-1	-1	-1
27	1	1	695.070	-1	-1	-1
28	1	2	549.278	-1	-1	-1
29	1	1	769.391	-1	-1	-1
30	1	2	624.972	-1	-1	-1
61	1	1	720.186	-1	1	1
62	1	2	587.695	-1	1	1
63	1	1	723.657	-1	1	1
64	1	2	569.207	-1	1	1
65	1	1	703.700	-1	1	1
66	1	2	613.257	-1	1	1
67	1	1	697.626	-1	1	1
68	1	2	565.737	-1	1	1

69	1	1	714.980	-1	1	1
70	1	2	662.131	-1	1	1
71	1	1	657.712	-1	1	1
72	1	2	543.177	-1	1	1
73	1	1	609.989	-1	1	1
74	1	2	512.394	-1	1	1
75	1	1	650.771	-1	1	1
76	1	2	611.190	-1	1	1
77	1	1	707.977	-1	1	1
78	1	2	659.982	-1	1	1
79	1	1	712.199	-1	1	1
80	1	2	569.245	-1	1	1
81	1	1	709.631	-1	1	1
82	1	2	725.792	-1	1	1
83	1	1	703.160	-1	1	1
84	1	2	608.960	-1	1	1
85	1	1	744.822	-1	1	1
86	1	2	586.060	-1	1	1
87	1	1	719.217	-1	1	1
88	1	2	617.441	-1	1	1
89	1	1	619.137	-1	1	1
90	1	2	592.845	-1	1	1
151	2	1	753.333	1	1	1
152	2	2	631.754	1	1	1
153	2	1	677.933	1	1	1
154	2	2	588.113	1	1	1
155	2	1	735.919	1	1	1
156	2	2	555.724	1	1	1
157	2	1	695.274	1	1	1
158	2	2	702.411	1	1	1
159	2	1	504.167	1	1	1
160	2	2	631.754	1	1	1
161	2	1	693.333	1	1	1
162	2	2	698.254	1	1	1
163	2	1	625.000	1	1	1
164	2	2	616.791	1	1	1
165	2	1	596.667	1	1	1
166	2	2	551.953	1	1	1
167	2	1	640.898	1	1	1
168	2	2	636.738	1	1	1
169	2	1	720.506	1	1	1
170	2	2	571.551	1	1	1
171	2	1	700.748	1	1	1
172	2	2	521.667	1	1	1
173	2	1	691.604	1	1	1
174	2	2	587.451	1	1	1
175	2	1	636.738	1	1	1
176	2	2	700.422	1	1	1
177	2	1	731.667	1	1	1
178	2	2	595.819	1	1	1
179	2	1	635.079	1	1	1
180	2	2	534.236	1	1	1
181	2	1	716.926	1	-1	-1
182	2	2	606.188	1	-1	-1
183	2	1	759.581	1	-1	-1
184	2	2	575.303	1	-1	-1
185	2	1	673.903	1	-1	-1
186	2	2	590.628	1	-1	-1
187	2	1	736.648	1	-1	-1
188	2	2	729.314	1	-1	-1
189	2	1	675.957	1	-1	-1
190	2	2	619.313	1	-1	-1
191	2	1	729.230	1	-1	-1
192	2	2	624.234	1	-1	-1
193	2	1	697.239	1	-1	-1
194	2	2	651.304	1	-1	-1
195	2	1	728.499	1	-1	-1
196	2	2	724.175	1	-1	-1
197	2	1	797.662	1	-1	-1
198	2	2	583.034	1	-1	-1
199	2	1	668.530	1	-1	-1
200	2	2	620.227	1	-1	-1
201	2	1	815.754	1	-1	-1
202	2	2	584.861	1	-1	-1
203	2	1	777.392	1	-1	-1
204	2	2	565.391	1	-1	-1
205	2	1	712.140	1	-1	-1
206	2	2	622.506	1	-1	-1
207	2	1	663.622	1	-1	-1
208	2	2	628.336	1	-1	-1
209	2	1	684.181	1	-1	-1

210	2	2	587.145	1	-1	-1
271	3	1	629.012	1	-1	1
272	3	2	584.319	1	-1	1
273	3	1	640.193	1	-1	1
274	3	2	538.239	1	-1	1
275	3	1	644.156	1	-1	1
276	3	2	538.097	1	-1	1
277	3	1	642.469	1	-1	1
278	3	2	595.686	1	-1	1
279	3	1	639.090	1	-1	1
280	3	2	648.935	1	-1	1
281	3	1	439.418	1	-1	1
282	3	2	583.827	1	-1	1
283	3	1	614.664	1	-1	1
284	3	2	534.905	1	-1	1
285	3	1	537.161	1	-1	1
286	3	2	569.858	1	-1	1
287	3	1	656.773	1	-1	1
288	3	2	617.246	1	-1	1
289	3	1	659.534	1	-1	1
290	3	2	610.337	1	-1	1
291	3	1	695.278	1	-1	1
292	3	2	584.192	1	-1	1
293	3	1	734.040	1	-1	1
294	3	2	598.853	1	-1	1
295	3	1	687.665	1	-1	1
296	3	2	554.774	1	-1	1
297	3	1	710.858	1	-1	1
298	3	2	605.694	1	-1	1
299	3	1	701.716	1	-1	1
300	3	2	627.516	1	-1	1
301	3	1	382.133	1	1	-1
302	3	2	574.522	1	1	-1
303	3	1	719.744	1	1	-1
304	3	2	582.682	1	1	-1
305	3	1	756.820	1	1	-1
306	3	2	563.872	1	1	-1
307	3	1	690.978	1	1	-1
308	3	2	715.962	1	1	-1
309	3	1	670.864	1	1	-1
310	3	2	616.430	1	1	-1
311	3	1	670.308	1	1	-1
312	3	2	778.011	1	1	-1
313	3	1	660.062	1	1	-1
314	3	2	604.255	1	1	-1
315	3	1	790.382	1	1	-1
316	3	2	571.906	1	1	-1
317	3	1	714.750	1	1	-1
318	3	2	625.925	1	1	-1
319	3	1	716.959	1	1	-1
320	3	2	682.426	1	1	-1
321	3	1	603.363	1	1	-1
322	3	2	707.604	1	1	-1
323	3	1	713.796	1	1	-1
324	3	2	617.400	1	1	-1
325	3	1	444.963	1	1	-1
326	3	2	689.576	1	1	-1
327	3	1	723.276	1	1	-1
328	3	2	676.678	1	1	-1
329	3	1	745.527	1	1	-1
330	3	2	563.290	1	1	-1
361	4	1	778.333	-1	-1	1
362	4	2	581.879	-1	-1	1
363	4	1	723.349	-1	-1	1
364	4	2	447.701	-1	-1	1
365	4	1	708.229	-1	-1	1
366	4	2	557.772	-1	-1	1
367	4	1	681.667	-1	-1	1
368	4	2	593.537	-1	-1	1
369	4	1	566.085	-1	-1	1
370	4	2	632.585	-1	-1	1
371	4	1	687.448	-1	-1	1
372	4	2	671.350	-1	-1	1
373	4	1	597.500	-1	-1	1
374	4	2	569.530	-1	-1	1
375	4	1	637.410	-1	-1	1
376	4	2	581.667	-1	-1	1
377	4	1	755.864	-1	-1	1
378	4	2	643.449	-1	-1	1
379	4	1	692.945	-1	-1	1
380	4	2	581.593	-1	-1	1

381	4	1	766.532	-1	-1	1
382	4	2	494.122	-1	-1	1
383	4	1	725.663	-1	-1	1
384	4	2	620.948	-1	-1	1
385	4	1	698.818	-1	-1	1
386	4	2	615.903	-1	-1	1
387	4	1	760.000	-1	-1	1
388	4	2	606.667	-1	-1	1
389	4	1	775.272	-1	-1	1
390	4	2	579.167	-1	-1	1
421	4	1	708.885	-1	1	-1
422	4	2	662.510	-1	1	-1
423	4	1	727.201	-1	1	-1
424	4	2	436.237	-1	1	-1
425	4	1	642.560	-1	1	-1
426	4	2	644.223	-1	1	-1
427	4	1	690.773	-1	1	-1
428	4	2	586.035	-1	1	-1
429	4	1	688.333	-1	1	-1
430	4	2	620.833	-1	1	-1
431	4	1	743.973	-1	1	-1
432	4	2	652.535	-1	1	-1
433	4	1	682.461	-1	1	-1
434	4	2	593.516	-1	1	-1
435	4	1	761.430	-1	1	-1
436	4	2	587.451	-1	1	-1
437	4	1	691.542	-1	1	-1
438	4	2	570.964	-1	1	-1
439	4	1	643.392	-1	1	-1
440	4	2	645.192	-1	1	-1
441	4	1	697.075	-1	1	-1
442	4	2	540.079	-1	1	-1
443	4	1	708.229	-1	1	-1
444	4	2	707.117	-1	1	-1
445	4	1	746.467	-1	1	-1
446	4	2	621.779	-1	1	-1
447	4	1	744.819	-1	1	-1
448	4	2	585.777	-1	1	-1
449	4	1	655.029	-1	1	-1
450	4	2	703.980	-1	1	-1
541	5	1	715.224	-1	-1	-1
542	5	2	698.237	-1	-1	-1
543	5	1	614.417	-1	-1	-1
544	5	2	757.120	-1	-1	-1
545	5	1	761.363	-1	-1	-1
546	5	2	621.751	-1	-1	-1
547	5	1	716.106	-1	-1	-1
548	5	2	472.125	-1	-1	-1
549	5	1	659.502	-1	-1	-1
550	5	2	612.700	-1	-1	-1
551	5	1	730.781	-1	-1	-1
552	5	2	583.170	-1	-1	-1
553	5	1	546.928	-1	-1	-1
554	5	2	599.771	-1	-1	-1
555	5	1	734.203	-1	-1	-1
556	5	2	549.227	-1	-1	-1
557	5	1	682.051	-1	-1	-1
558	5	2	605.453	-1	-1	-1
559	5	1	701.341	-1	-1	-1
560	5	2	569.599	-1	-1	-1
561	5	1	759.729	-1	-1	-1
562	5	2	637.233	-1	-1	-1
563	5	1	689.942	-1	-1	-1
564	5	2	621.774	-1	-1	-1
565	5	1	769.424	-1	-1	-1
566	5	2	558.041	-1	-1	-1
567	5	1	715.286	-1	-1	-1
568	5	2	583.170	-1	-1	-1
569	5	1	776.197	-1	-1	-1
570	5	2	345.294	-1	-1	-1
571	5	1	547.099	1	-1	1
572	5	2	570.999	1	-1	1
573	5	1	619.942	1	-1	1
574	5	2	603.232	1	-1	1
575	5	1	696.046	1	-1	1
576	5	2	595.335	1	-1	1
577	5	1	573.109	1	-1	1
578	5	2	581.047	1	-1	1
579	5	1	638.794	1	-1	1
580	5	2	455.878	1	-1	1
581	5	1	708.193	1	-1	1

582	5	2	627.880	1	-1	1
583	5	1	502.825	1	-1	1
584	5	2	464.085	1	-1	1
585	5	1	632.633	1	-1	1
586	5	2	596.129	1	-1	1
587	5	1	683.382	1	-1	1
588	5	2	640.371	1	-1	1
589	5	1	684.812	1	-1	1
590	5	2	621.471	1	-1	1
591	5	1	738.161	1	-1	1
592	5	2	612.727	1	-1	1
593	5	1	671.492	1	-1	1
594	5	2	606.460	1	-1	1
595	5	1	709.771	1	-1	1
596	5	2	571.760	1	-1	1
597	5	1	685.199	1	-1	1
598	5	2	599.304	1	-1	1
599	5	1	624.973	1	-1	1
600	5	2	579.459	1	-1	1
601	6	1	757.363	1	1	1
602	6	2	761.511	1	1	1
603	6	1	633.417	1	1	1
604	6	2	566.969	1	1	1
605	6	1	658.754	1	1	1
606	6	2	654.397	1	1	1
607	6	1	664.666	1	1	1
608	6	2	611.719	1	1	1
609	6	1	663.009	1	1	1
610	6	2	577.409	1	1	1
611	6	1	773.226	1	1	1
612	6	2	576.731	1	1	1
613	6	1	708.261	1	1	1
614	6	2	617.441	1	1	1
615	6	1	739.086	1	1	1
616	6	2	577.409	1	1	1
617	6	1	667.786	1	1	1
618	6	2	548.957	1	1	1
619	6	1	674.481	1	1	1
620	6	2	623.315	1	1	1
621	6	1	695.688	1	1	1
622	6	2	621.761	1	1	1
623	6	1	588.288	1	1	1
624	6	2	553.978	1	1	1
625	6	1	545.610	1	1	1
626	6	2	657.157	1	1	1
627	6	1	752.305	1	1	1
628	6	2	610.882	1	1	1
629	6	1	684.523	1	1	1
630	6	2	552.304	1	1	1
631	6	1	717.159	-1	1	-1
632	6	2	545.303	-1	1	-1
633	6	1	721.343	-1	1	-1
634	6	2	651.934	-1	1	-1
635	6	1	750.623	-1	1	-1
636	6	2	635.240	-1	1	-1
637	6	1	776.488	-1	1	-1
638	6	2	641.083	-1	1	-1
639	6	1	750.623	-1	1	-1
640	6	2	645.321	-1	1	-1
641	6	1	600.840	-1	1	-1
642	6	2	566.127	-1	1	-1
643	6	1	686.196	-1	1	-1
644	6	2	647.844	-1	1	-1
645	6	1	687.870	-1	1	-1
646	6	2	554.815	-1	1	-1
647	6	1	725.527	-1	1	-1
648	6	2	620.087	-1	1	-1
649	6	1	658.796	-1	1	-1
650	6	2	711.301	-1	1	-1
651	6	1	690.380	-1	1	-1
652	6	2	644.355	-1	1	-1
653	6	1	737.144	-1	1	-1
654	6	2	713.812	-1	1	-1
655	6	1	663.851	-1	1	-1
656	6	2	696.707	-1	1	-1
657	6	1	766.630	-1	1	-1
658	6	2	589.453	-1	1	-1
659	6	1	625.922	-1	1	-1
660	6	2	634.468	-1	1	-1
721	7	1	694.430	1	1	-1
722	7	2	599.751	1	1	-1

723	7	1	730.217	1	1	-1
724	7	2	624.542	1	1	-1
725	7	1	700.770	1	1	-1
726	7	2	723.505	1	1	-1
727	7	1	722.242	1	1	-1
728	7	2	674.717	1	1	-1
729	7	1	763.828	1	1	-1
730	7	2	608.539	1	1	-1
731	7	1	695.668	1	1	-1
732	7	2	612.135	1	1	-1
733	7	1	688.887	1	1	-1
734	7	2	591.935	1	1	-1
735	7	1	531.021	1	1	-1
736	7	2	676.656	1	1	-1
737	7	1	698.915	1	1	-1
738	7	2	647.323	1	1	-1
739	7	1	735.905	1	1	-1
740	7	2	811.970	1	1	-1
741	7	1	732.039	1	1	-1
742	7	2	603.883	1	1	-1
743	7	1	751.832	1	1	-1
744	7	2	608.643	1	1	-1
745	7	1	618.663	1	1	-1
746	7	2	630.778	1	1	-1
747	7	1	744.845	1	1	-1
748	7	2	623.063	1	1	-1
749	7	1	690.826	1	1	-1
750	7	2	472.463	1	1	-1
811	7	1	666.893	-1	1	1
812	7	2	645.932	-1	1	1
813	7	1	759.860	-1	1	1
814	7	2	577.176	-1	1	1
815	7	1	683.752	-1	1	1
816	7	2	567.530	-1	1	1
817	7	1	729.591	-1	1	1
818	7	2	821.654	-1	1	1
819	7	1	730.706	-1	1	1
820	7	2	684.490	-1	1	1
821	7	1	763.124	-1	1	1
822	7	2	600.427	-1	1	1
823	7	1	724.193	-1	1	1
824	7	2	686.023	-1	1	1
825	7	1	630.352	-1	1	1
826	7	2	628.109	-1	1	1
827	7	1	750.338	-1	1	1
828	7	2	605.214	-1	1	1
829	7	1	752.417	-1	1	1
830	7	2	640.260	-1	1	1
831	7	1	707.899	-1	1	1
832	7	2	700.767	-1	1	1
833	7	1	715.582	-1	1	1
834	7	2	665.924	-1	1	1
835	7	1	728.746	-1	1	1
836	7	2	555.926	-1	1	1
837	7	1	591.193	-1	1	1
838	7	2	543.299	-1	1	1
839	7	1	592.252	-1	1	1
840	7	2	511.030	-1	1	1
901	8	1	740.833	-1	-1	1
902	8	2	583.994	-1	-1	1
903	8	1	786.367	-1	-1	1
904	8	2	611.048	-1	-1	1
905	8	1	712.386	-1	-1	1
906	8	2	623.338	-1	-1	1
907	8	1	738.333	-1	-1	1
908	8	2	679.585	-1	-1	1
909	8	1	741.480	-1	-1	1
910	8	2	665.004	-1	-1	1
911	8	1	729.167	-1	-1	1
912	8	2	655.860	-1	-1	1
913	8	1	795.833	-1	-1	1
914	8	2	715.711	-1	-1	1
915	8	1	723.502	-1	-1	1
916	8	2	611.999	-1	-1	1
917	8	1	718.333	-1	-1	1
918	8	2	577.722	-1	-1	1
919	8	1	768.080	-1	-1	1
920	8	2	615.129	-1	-1	1
921	8	1	747.500	-1	-1	1
922	8	2	540.316	-1	-1	1
923	8	1	775.000	-1	-1	1

924	8	2	711.667	-1	-1	1
925	8	1	760.599	-1	-1	1
926	8	2	639.167	-1	-1	1
927	8	1	758.333	-1	-1	1
928	8	2	549.491	-1	-1	1
929	8	1	682.500	-1	-1	1
930	8	2	684.167	-1	-1	1
931	8	1	658.116	1	-1	-1
932	8	2	672.153	1	-1	-1
933	8	1	738.213	1	-1	-1
934	8	2	594.534	1	-1	-1
935	8	1	681.236	1	-1	-1
936	8	2	627.650	1	-1	-1
937	8	1	704.904	1	-1	-1
938	8	2	551.870	1	-1	-1
939	8	1	693.623	1	-1	-1
940	8	2	594.534	1	-1	-1
941	8	1	624.993	1	-1	-1
942	8	2	602.660	1	-1	-1
943	8	1	700.228	1	-1	-1
944	8	2	585.450	1	-1	-1
945	8	1	611.874	1	-1	-1
946	8	2	555.724	1	-1	-1
947	8	1	579.167	1	-1	-1
948	8	2	574.934	1	-1	-1
949	8	1	720.872	1	-1	-1
950	8	2	584.625	1	-1	-1
951	8	1	690.320	1	-1	-1
952	8	2	555.724	1	-1	-1
953	8	1	677.933	1	-1	-1
954	8	2	611.874	1	-1	-1
955	8	1	674.600	1	-1	-1
956	8	2	698.254	1	-1	-1
957	8	1	611.999	1	-1	-1
958	8	2	748.130	1	-1	-1
959	8	1	530.680	1	-1	-1
960	8	2	689.942	1	-1	-1

1.4.2.10.2. Analysis of the Response Variable

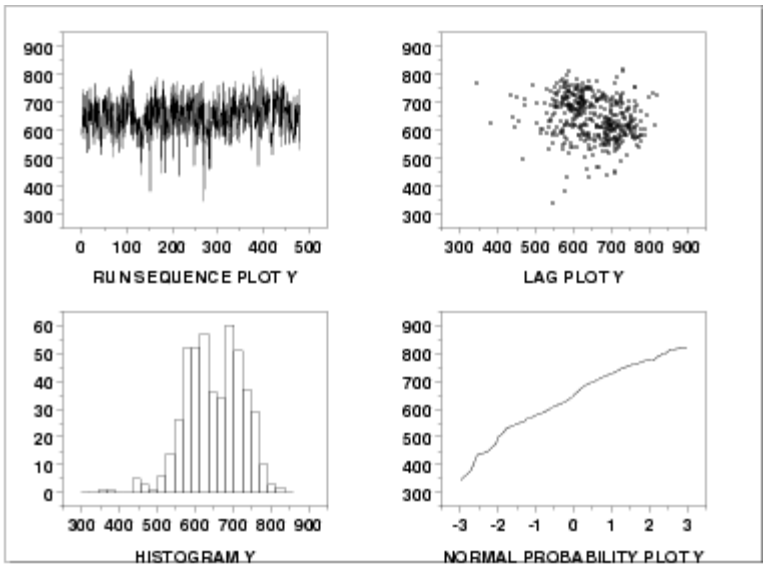
Numerical Summary

As a first step in the analysis, common summary statistics are computed for the response variable.

```
Sample size  = 480
Mean         =   650.0773
Median       =   646.6275
Minimum      =   345.2940
Maximum      =   821.6540
Range        =   476.3600
Stan. Dev.   =    74.6383
```

4-Plot

The next step is generate a 4-plot of the response variable.



This 4-plot shows:

1. The run sequence plot (upper left corner) shows that the location and scale are relatively constant. It also shows a few outliers on the low side. Most of the points are in the range 500 to 750. However, there are about half a dozen points in the 300 to 450 range that may require special attention.

   A run sequence plot is useful for designed experiments in that it can reveal time effects. Time is normally a nuisance factor. That is, the time order on which runs are made should not have a significant effect on the response. If a time effect does appear to exist, this

means that there is a potential bias in the experiment that needs to be investigated and resolved.

2. The lag plot (the upper right corner) does not show any significant structure. This is another tool for detecting any potential time effect.

3. The histogram (the lower left corner) shows the response appears to be reasonably symmetric, but with a bimodal distribution.

4. The normal probability plot (the lower right corner) shows some curvature indicating that distributions other than the normal may provide a better fit.

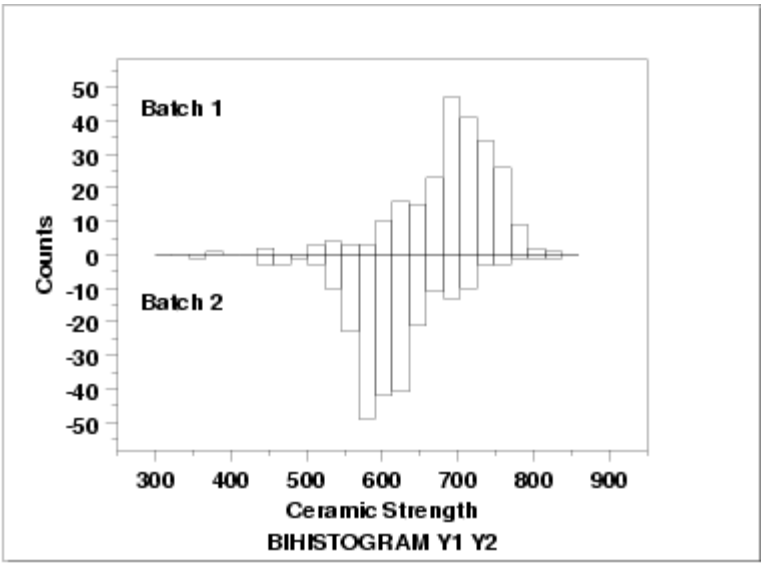
1.4.2.10.3. Analysis of the Batch Effect

Batch is a Nuisance Factor

The two nuisance factors in this experiment are the batch number and the lab. There are two batches and eight labs. Ideally, these factors will have minimal effect on the response variable.

We will investigate the batch factor first.

Bihistogram



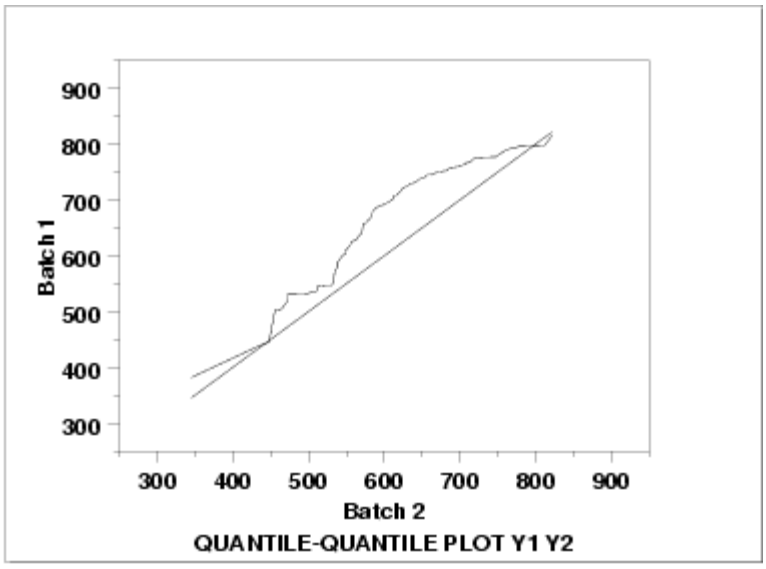
This bihistogram shows the following.

1. There does appear to be a batch effect.
2. The batch 1 responses are centered at 700 while the batch 2 responses are centered at 625. That is, the batch effect is approximately 75 units.
3. The variability is comparable for the 2 batches.
4. Batch 1 has some skewness in the lower tail. Batch 2 has some skewness in the center of the distribution, but not as much in the tails compared to batch 1.
5. Both batches have a few low-lying points.

Although we could stop with the bihistogram, we will show a few other commonly used two-sample graphical techniques

for comparison.

Quantile-Quantile Plot

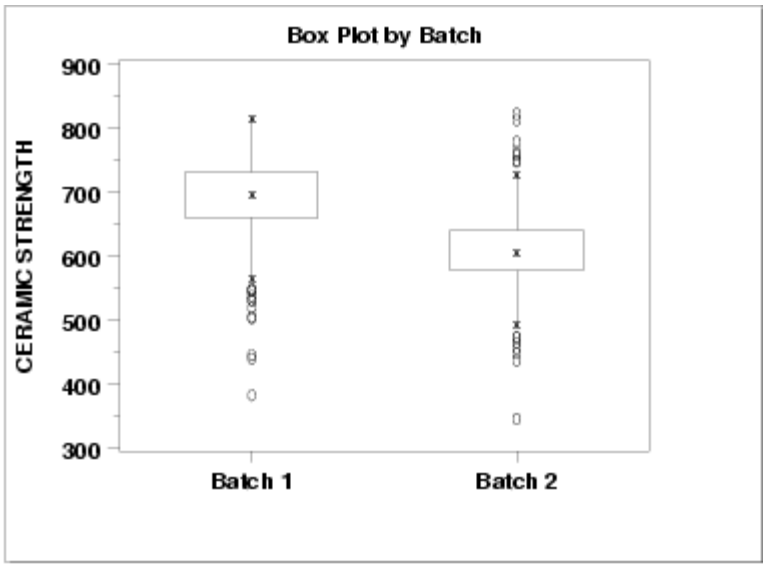


This q-q plot shows the following.

1. Except for a few points in the right tail, the batch 1 values have higher quantiles than the batch 2 values. This implies that batch 1 has a greater location value than batch 2.

2. The q-q plot is not linear. This implies that the difference between the batches is not explained simply by a shift in location. That is, the variation and/or skewness varies as well. From the bihistogram, it appears that the skewness in batch 2 is the most likely explanation for the non-linearity in the q-q plot.

Box Plot



This box plot shows the following.

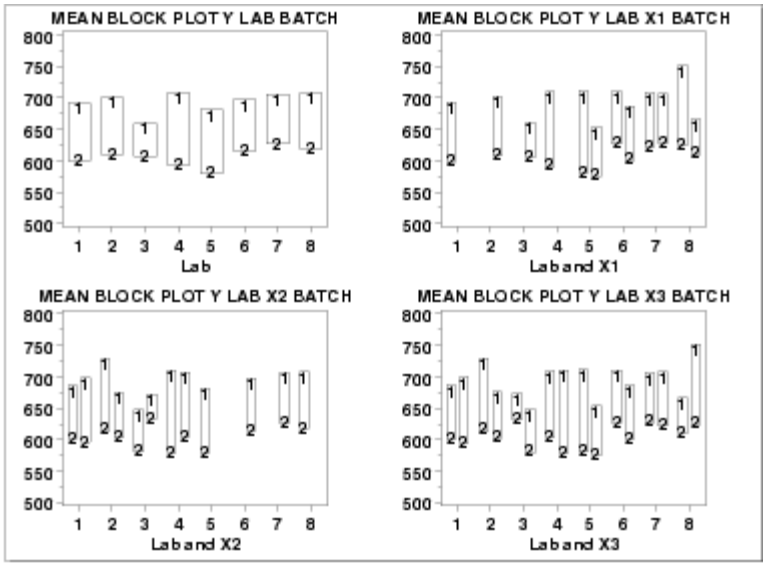
1. The median for batch 1 is approximately 700 while the median for batch 2 is approximately 600.

2. The spread is reasonably similar for both batches, maybe slightly larger for batch 1.

3. Both batches have a number of outliers on the low side. Batch 2 also has a few outliers on the high side. Box plots are a particularly effective method for identifying the presence of outliers.

Block Plots

A block plot is generated for each of the eight labs, with "1" and "2" denoting the batch numbers. In the first plot, we do not include any of the primary factors. The next 3 block plots include one of the primary factors. Note that each of the 3 primary factors (table speed = X1, down feed rate = X2, wheel grit size = X3) has 2 levels. With 8 labs and 2 levels for the primary factor, we would expect 16 separate blocks on these plots. The fact that some of these blocks are missing indicates that some of the combinations of lab and primary factor are empty.



These block plots show the following.

1. The mean for batch 1 is greater than the mean for batch 2 in **all** of the cases above. This is strong evidence that the batch effect is real and consistent across labs and primary factors.

Quantitative Techniques

We can confirm some of the conclusions drawn from the above graphics by using quantitative techniques. The *F*-test can be used to test whether or not the variances from the two batches are equal and the two sample *t*-test can be used to test whether or not the means from the two batches are equal. Summary statistics for each batch are shown below.

```
Batch 1:
    NUMBER OF OBSERVATIONS =  240
    MEAN                   =  688.9987
```

```
    STANDARD DEVIATION     =   65.5491
    VARIANCE               = 4296.6845

Batch 2:
    NUMBER OF OBSERVATIONS =  240
    MEAN                   =  611.1559
    STANDARD DEVIATION     =   61.8543
    VARIANCE               = 3825.9544
```

F-Test The two-sided *F*-test indicates that the variances for the two batches are not significantly different at the 5 % level.

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

Test statistic: $F = 1.123$
Numerator degrees of freedom: $\nu_1 = 239$
Denominator degrees of freedom: $\nu_2 = 239$
Significance level: $\alpha = 0.05$
Critical values: $F_{1-\alpha/2,\nu_1,\nu_2} = 0.845$
$F_{\alpha/2,\nu_1,\nu_2} = 1.289$
Critical region: Reject H_0 if $F < 0.845$ or $F > 1.289$

Two Sample t-Test Since the *F*-test indicates that the two batch variances are equal, we can pool the variances for the two-sided, two-sample *t*-test to compare batch means.

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Test statistic: $T = 13.3806$
Pooled standard deviation: $s_p = 63.7285$
Degrees of freedom: $\nu = 478$
Significance level: $\alpha = 0.05$
Critical value: $t_{1-\alpha/2,\nu} = 1.965$
Critical region: Reject H_0 if $|T| > 1.965$

The *t*-test indicates that the mean for batch 1 is larger than the mean for batch 2 at the 5 % significance level.

Conclusions We can draw the following conclusions from the above analysis.

1. There is in fact a significant batch effect. This batch effect is consistent across labs and primary factors.
2. The magnitude of the difference is on the order of 75 to 100 (with batch 2 being smaller than batch 1). The standard deviations do not appear to be significantly different.
3. There is some skewness in the batches.

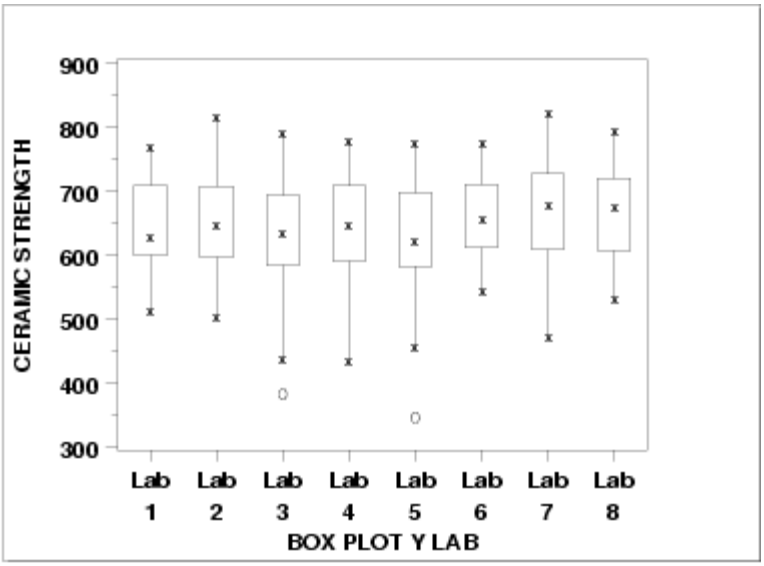
This batch effect was completely unexpected by the scientific investigators in this study.

Note that although the quantitative techniques support the conclusions of unequal means and equal standard deviations, they do not show the more subtle features of the data such as the presence of outliers and the skewness of the batch 2 data.

1.4.2.10.4. Analysis of the Lab Effect

Box Plot

The next matter is to determine if there is a lab effect. The first step is to generate a box plot for the ceramic strength based on the lab.

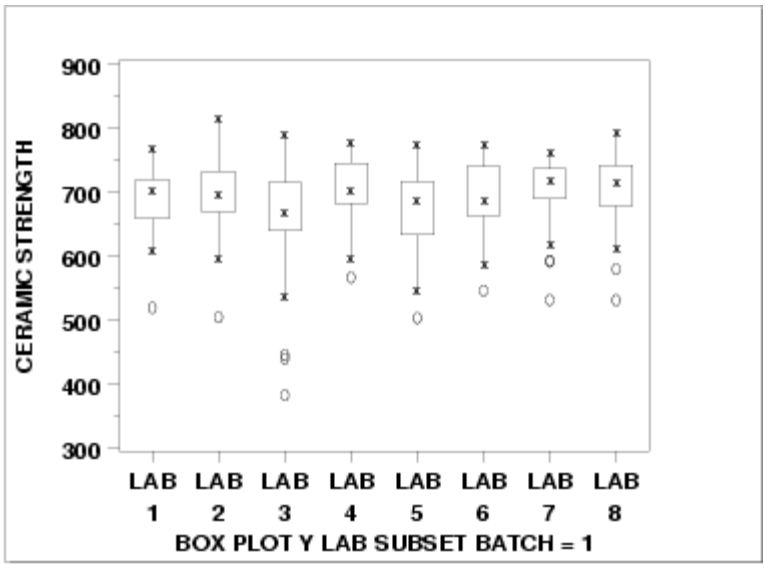


This box plot shows the following.

1. There is minor variation in the medians for the 8 labs.
2. The scales are relatively constant for the labs.
3. Two of the labs (3 and 5) have outliers on the low side.

Box Plot for Batch 1

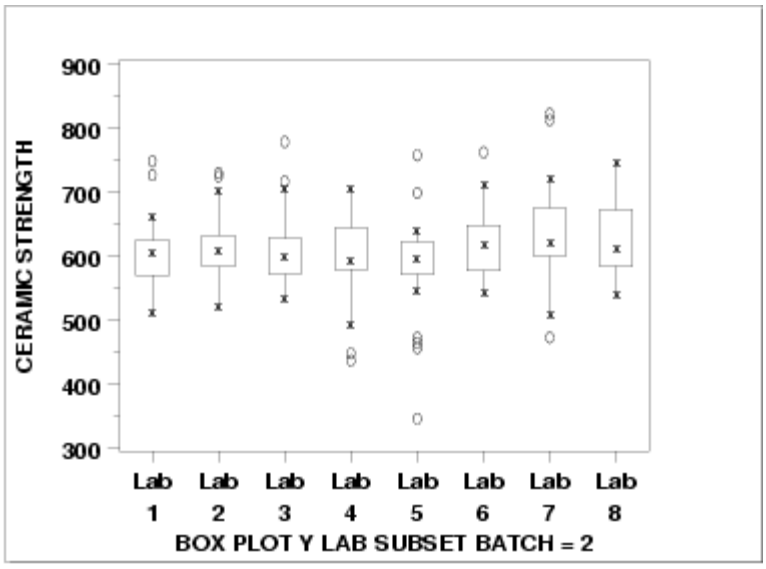
Given that the previous section showed a distinct batch effect, the next step is to generate the box plots for the two batches separately.



This box plot shows the following.

1. Each of the labs has a median in the 650 to 700 range.
2. The variability is relatively constant across the labs.
3. Each of the labs has at least one outlier on the low side.

Box Plot for Batch 2



This box plot shows the following.

1. The medians are in the range 550 to 600.
2. There is a bit more variability, across the labs, for batch2 compared to batch 1.
3. Six of the eight labs show outliers on the high side. Three of the labs show outliers on the low side.

Conclusions We can draw the following conclusions about a possible lab

effect from the above box plots.

1. The batch effect (of approximately 75 to 100 units) on location dominates any lab effects.

2. It is reasonable to treat the labs as homogeneous.

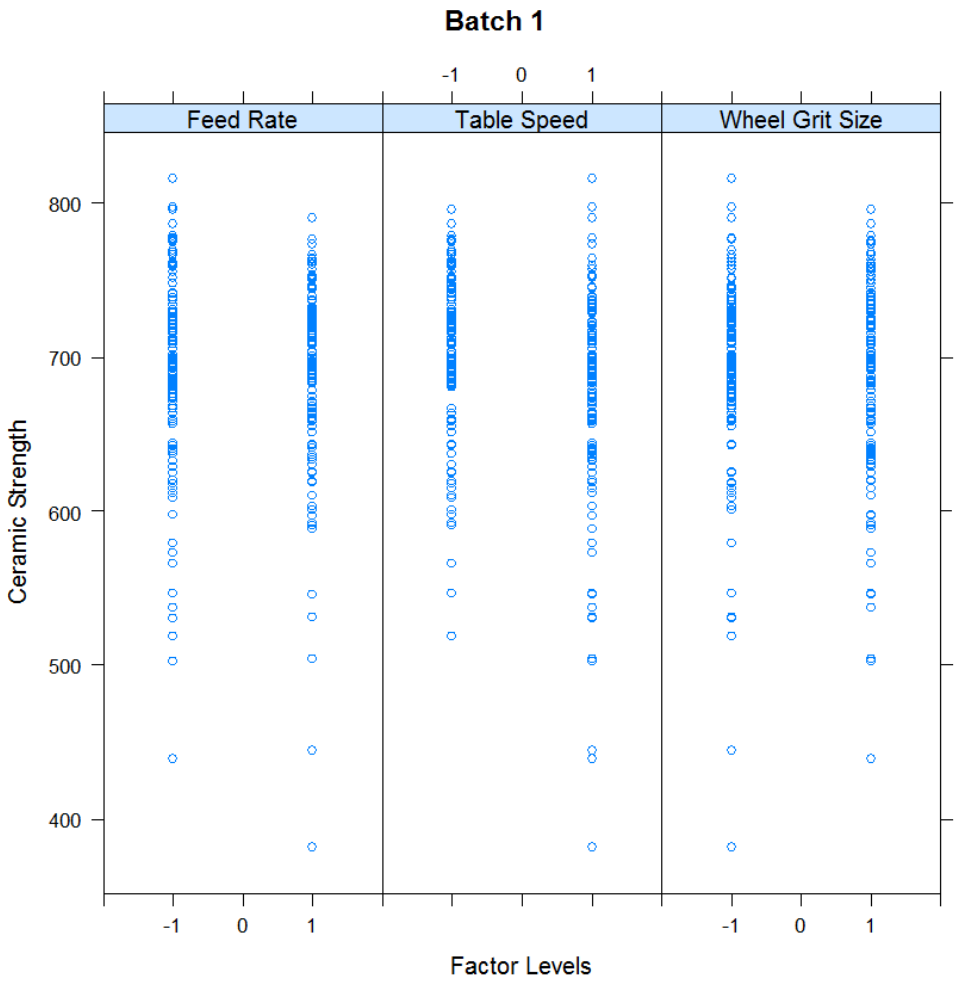
1.4.2.10.5. Analysis of Primary Factors

Main effects

The first step in analyzing the primary factors is to determine which factors are the most significant. The DOE scatter plot, DOE mean plot, and the DOE standard deviation plots will be the primary tools, with "DOE" being short for "design of experiments".

Since the previous pages showed a significant batch effect but a minimal lab effect, we will generate separate plots for batch 1 and batch 2. However, the labs will be treated as equivalent.

DOE Scatter Plot for Batch 1



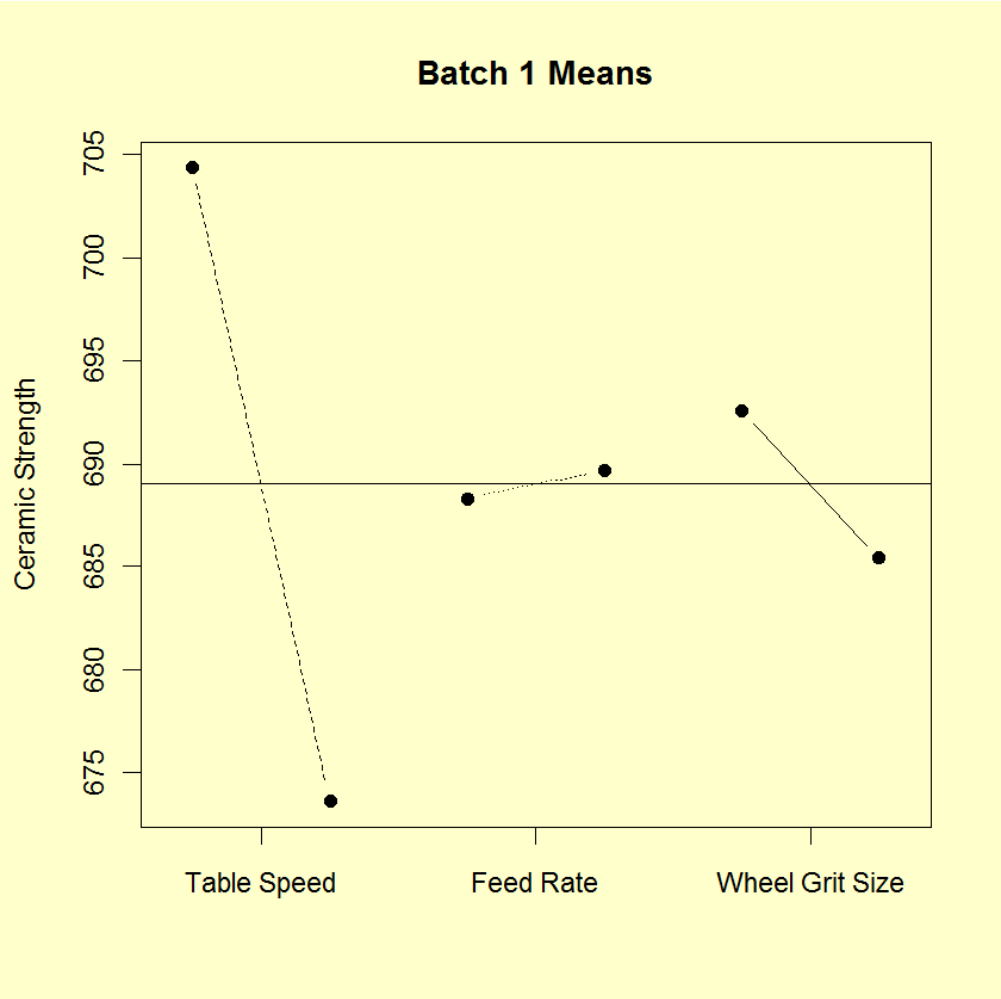
This DOE scatter plot shows the following for batch 1.

1. Most of the points are between 500 and 800.

2. There are about a dozen or so points between 300 and 500.

3. Except for the outliers on the low side (i.e., the points between 300 and 500), the distribution of the points is comparable for the 3 primary factors in terms of location and spread.

DOE Mean Plot for Batch 1



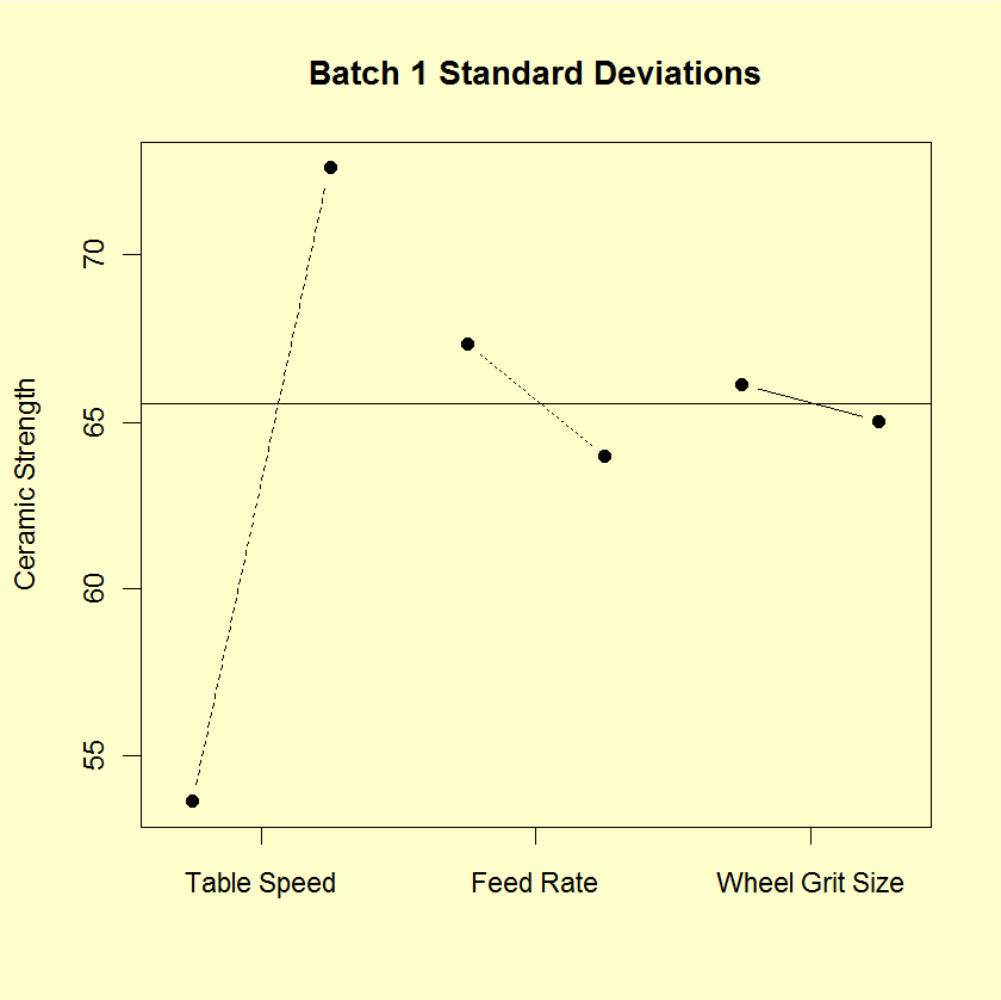
This DOE mean plot shows the following for batch 1.

1. The table speed factor (X1) is the most significant factor with an effect, the difference between the two points, of approximately 35 units.

2. The wheel grit factor (X3) is the next most significant factor with an effect of approximately 10 units.

3. The feed rate factor (X2) has minimal effect.

DOE SD Plot for Batch 1

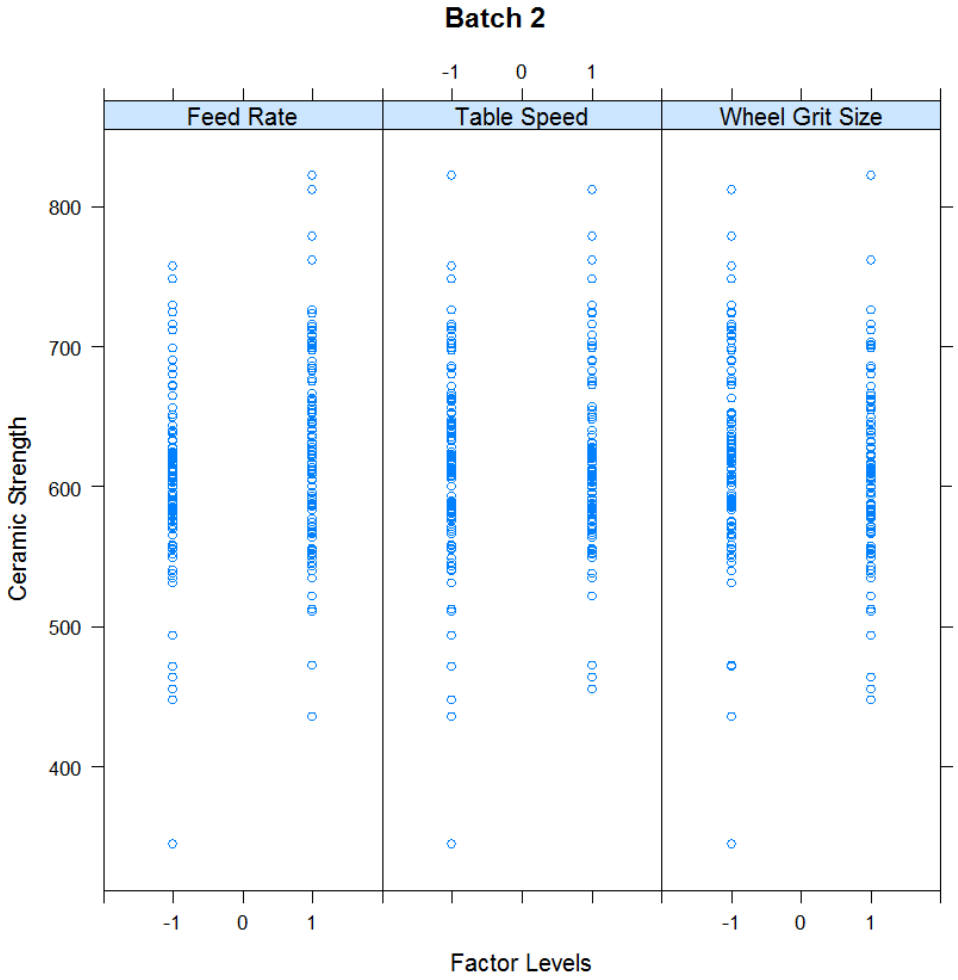


This DOE standard deviation plot shows the following for batch 1.

1. The table speed factor (X1) has a significant difference in variability between the levels of the factor. The difference is approximately 20 units.

2. The wheel grit factor (X3) and the feed rate factor (X2) have minimal differences in variability.

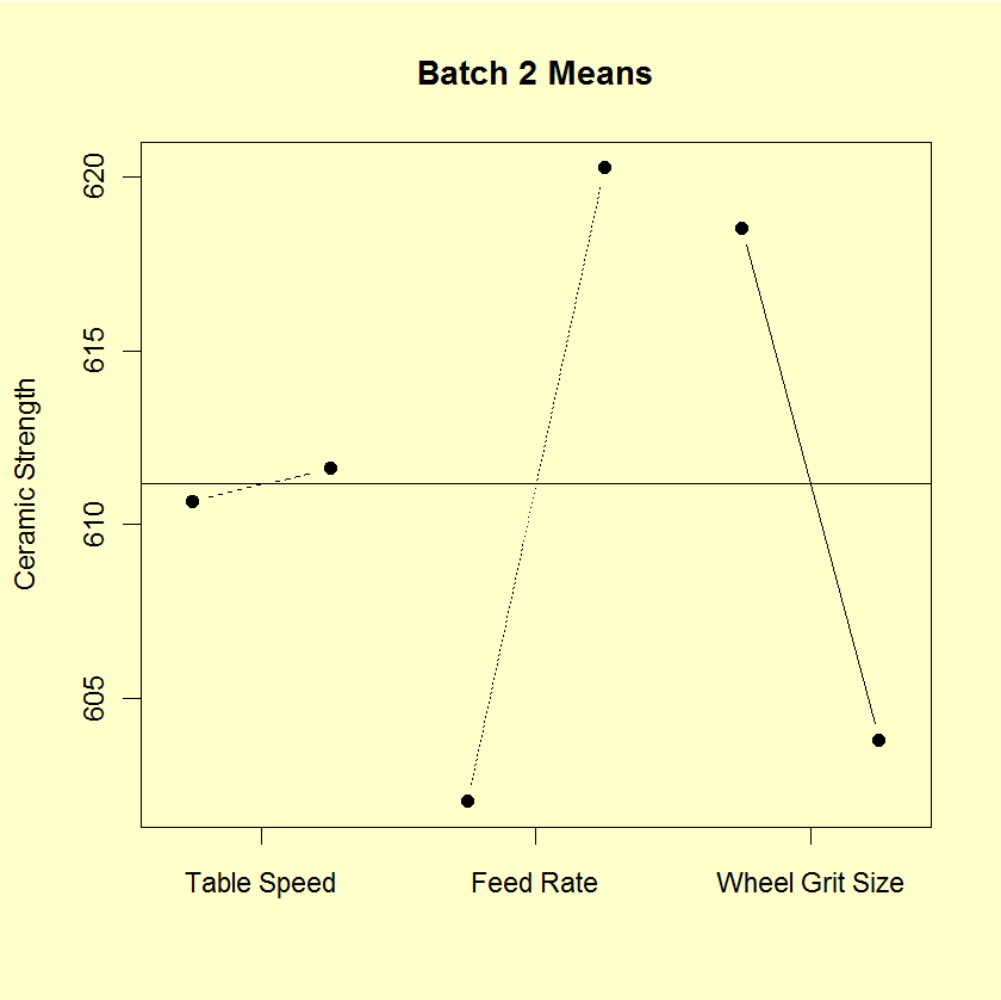
DOE
Scatter Plot
for Batch 2



This DOE scatter plot shows the following for batch 2.

1. Most of the points are between 450 and 750.
2. There are a few outliers on both the low side and the high side.
3. Except for the outliers (i.e., the points less than 450 or greater than 750), the distribution of the points is comparable for the 3 primary factors in terms of location and spread.

DOE Mean Plot for Batch 2

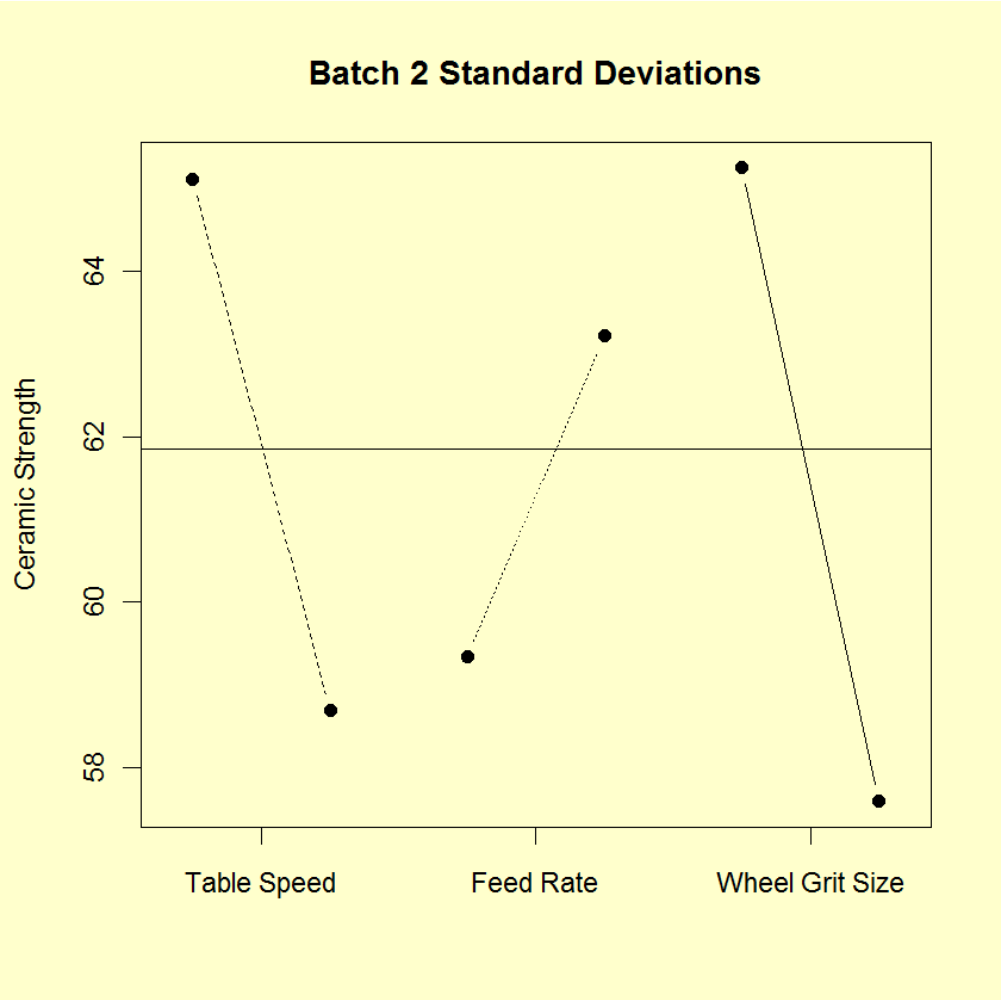


This DOE mean plot shows the following for batch 2.

1. The feed rate (X2) and wheel grit (X3) factors have an approximately equal effect of about 15 or 20 units.

2. The table speed factor (X1) has a minimal effect.

DOE SD Plot for Batch 2



This DOE standard deviation plot shows the following for batch 2.

1. The difference in the standard deviations is roughly comparable for the three factors (slightly less for the feed rate factor).

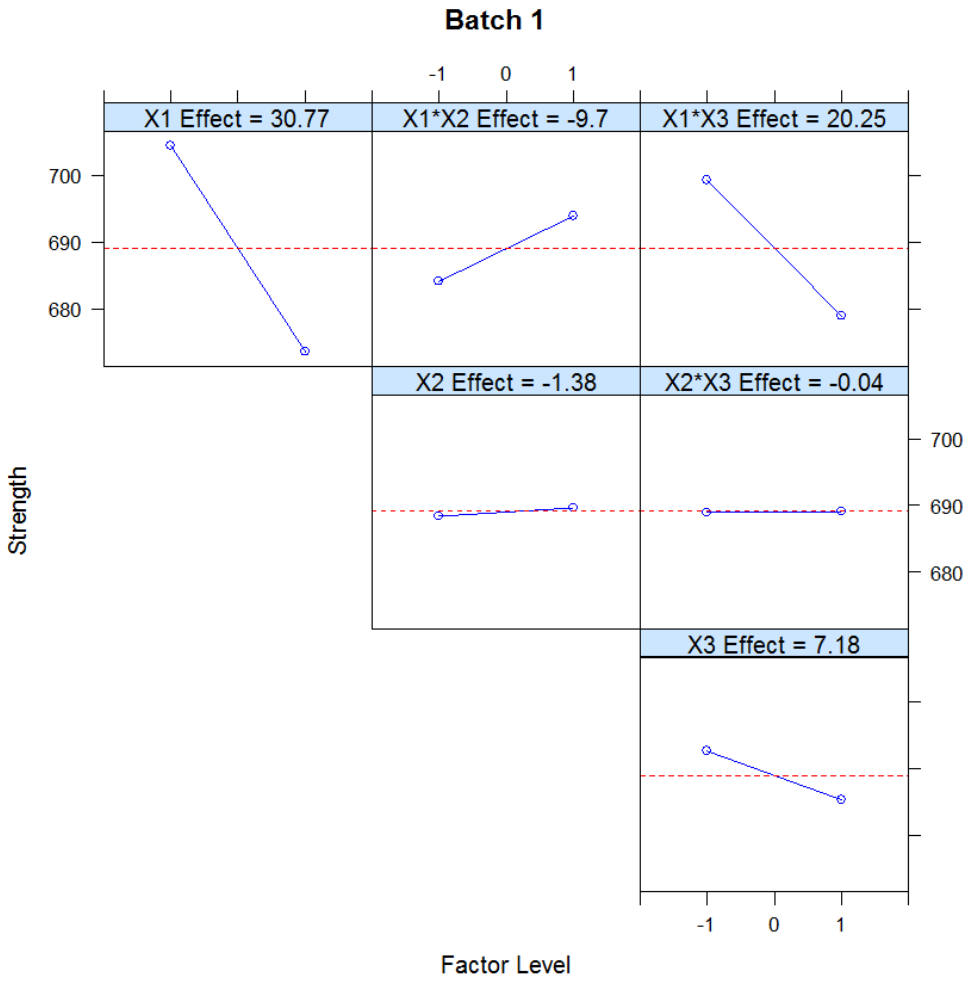
Interaction Effects

The above plots graphically show the main effects. An additonal concern is whether or not there any significant interaction effects.

Main effects and 2-term interaction effects are discussed in the chapter on Process Improvement.

In the following DOE interaction plots, the labels on the plot give the variables and the estimated effect. For example, factor 1 is table speed and it has an estimated effect of 30.77 (it is actually -30.77 if the direction is taken into account).

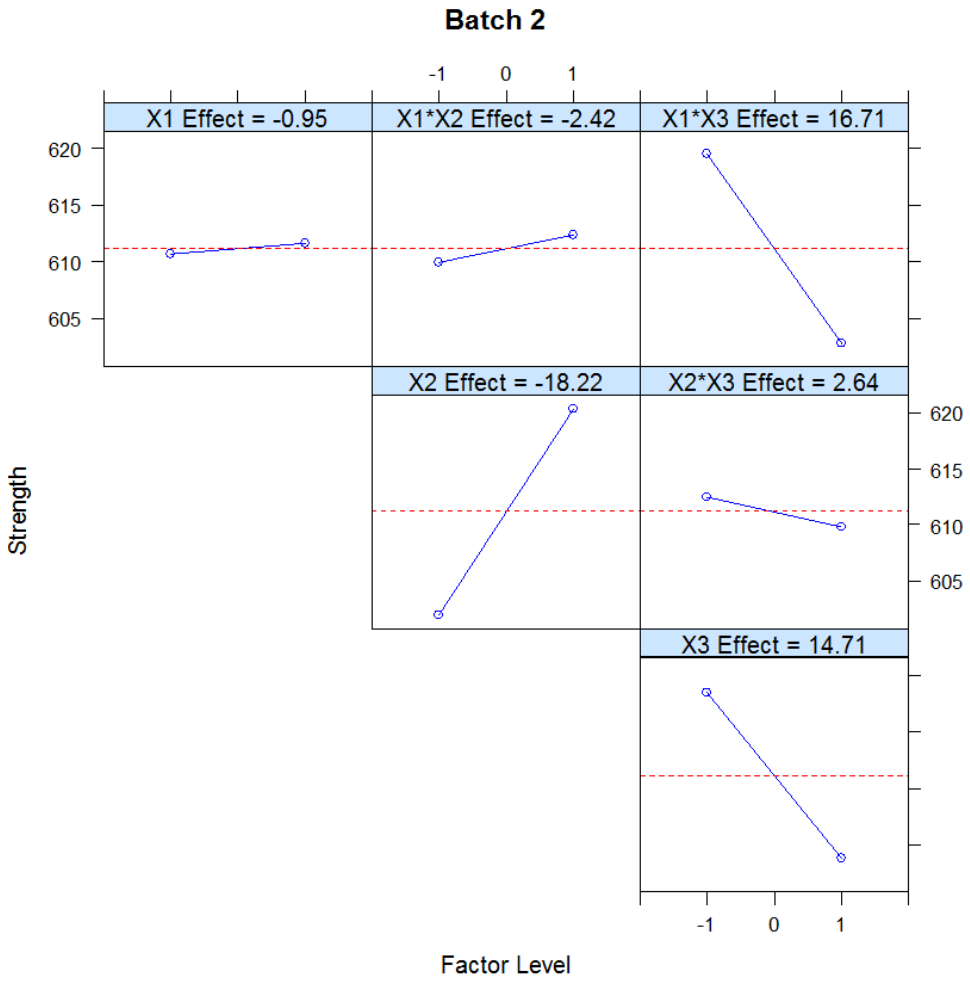
DOE Interaction Plot for Batch 1



The ranked list of factors for batch 1 is:

1. Table speed (X1) with an estimated effect of -30.77.
2. The interaction of table speed (X1) and wheel grit (X3) with an estimated effect of -20.25.
3. The interaction of table speed (X1) and feed rate (X2) with an estimated effect of 9.7.
4. Wheel grit (X3) with an estimated effect of -7.18.
5. Down feed (X2) and the down feed interaction with wheel grit (X3) are essentially zero.

DOE Interaction Plot for Batch 2



The ranked list of factors for batch 2 is:

1. Down feed (X2) with an estimated effect of 18.22.
2. The interaction of table speed (X1) and wheel grit (X3) with an estimated effect of -16.71.
3. Wheel grit (X3) with an estimated effect of -14.71
4. Remaining main effect and 2-factor interaction effects are essentially zero.

Conclusions From the above plots, we can draw the following overall conclusions.

1. The batch effect (of approximately 75 units) is the dominant primary factor.
2. The most important factors differ from batch to batch. See the above text for the ranked list of factors with the estimated effects.

1.4.2.10.6. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to use Dataplot to repeat the analysis outlined in the case study description on the previous page. It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. Plot of the response variable 1. Numerical summary of Y. 2. 4-plot of Y.	1. The summary shows the mean strength is 650.08 and the standard deviation of the strength is 74.64. 2. The 4-plot shows no drift in the location and scale and a

	bimodal distribution.
3. Determine if there is a batch effect. 1. Generate a bihistogram based on the 2 batches. 2. Generate a q-q plot. 3. Generate a box plot. 4. Generate block plots. 5. Perform a 2-sample t-test for equal means. 6. Perform an F-test for equal standard deviations.	1. The bihistogram shows a distinct batch effect of approximately 75 units. 2. The q-q plot shows that batch 1 and batch 2 do not come from a common distribution. 3. The box plot shows that there is a batch effect of approximately 75 to 100 units and there are some outliers. 4. The block plot shows that the batch effect is consistent across labs and levels of the primary factor. 5. The t-test confirms the batch effect with respect to the means. 6. The F-test does not indicate any significant batch effect with respect to the standard deviations.
4. Determine if there is a lab effect. 1. Generate a box plot for the labs with the 2 batches combined. 2. Generate a box plot for the labs for batch 1 only. 3. Generate a box plot for the labs for batch 2 only.	1. The box plot does not show a significant lab effect. 2. The box plot does not show a significant lab effect for batch 1. 3. The box plot does not show a significant lab effect for batch 2.
5. Analysis of primary factors. 1. Generate a DOE scatter plot for batch 1.	1. The DOE scatter plot shows the range of the points and the

2. Generate a DOE mean plot for batch 1.	presence of outliers.
3. Generate a DOE sd plot for batch 1.	2. The DOE mean plot shows that table speed is the most significant factor for batch 1.
4. Generate a DOE scatter plot for batch 2.	3. The DOE sd plot shows that table speed has the most variability for batch 1.
5. Generate a DOE mean plot for batch 2.	4. The DOE scatter plot shows the range of the points and the presence of outliers.
6. Generate a DOE sd plot for batch 2.	5. The DOE mean plot shows that feed rate and wheel grit are the most significant factors for batch 2.
7. Generate a DOE interaction effects matrix plot for batch 1.	6. The DOE sd plot shows that the variability is comparable for all 3 factors for batch 2.
8. Generate a DOE interaction effects matrix plot for batch 2.	7. The DOE interaction effects matrix plot provides a ranked list of factors with the estimated effects.
	8. The DOE interaction effects matrix plot provides a ranked list of factors with the estimated effects.

1.4.3. References For Chapter 1: Exploratory Data Analysis

Anscombe, F. (1973), Graphs in Statistical Analysis, *The American Statistician*, pp. 195-199.

Anscombe, F. and Tukey, J. W. (1963), The Examination and Analysis of Residuals, *Technometrics*, pp. 141-160.

Barnett and Lewis (1994), *Outliers in Statistical Data*, 3rd. Ed., John Wiley and Sons.

Birnbaum, Z. W. and Saunders, S. C. (1958), A Statistical Model for Life-Length of Materials, *Journal of the American Statistical Association*, 53(281), pp. 151-160.

Bloomfield, Peter (1976), *Fourier Analysis of Time Series*, John Wiley and Sons.

Box, G. E. P. and Cox, D. R. (1964), An Analysis of Transformations, *Journal of the Royal Statistical Society*, pp. 211-243, discussion pp. 244-252.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley and Sons.

Box, G. E. P., and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day.

Bradley, (1968). *Distribution-Free Statistical Tests*, Chapter 12.

Brown, M. B. and Forsythe, A. B. (1974), *Journal of the American Statistical Association*, 69, pp. 364-367.

Chakravarti, Laha, and Roy, (1967). *Handbook of Methods of Applied Statistics, Volume I*, John Wiley and Sons, pp. 392-394.

Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), *Graphical Methods for Data Analysis*, Wadsworth.

Chatfield, C. (1989). *The Analysis of Time Series: An Introduction*, Fourth Edition, Chapman & Hall, New York, NY.

Cleveland, William (1985), *Elements of Graphing Data*, Wadsworth.

Cleveland, William and Marylyn McGill, Editors (1988), *Dynamic Graphics for Statistics*, Wadsworth.

Cleveland, William (1993), *Visualizing Data*, Hobart Press.

Devaney, Judy (1997), *Equation Discovery Through Global Self-Referenced Geometric Intervals and Machine Learning*, Ph.d thesis, George Mason University, Fairfax, VA.

Draper and Smith, (1981). *Applied Regression Analysis*, 2nd ed., John Wiley and Sons.

du Toit, Steyn, and Stumpf (1986), *Graphical Exploratory Data Analysis*, Springer-Verlag.

Efron and Gong (February 1983), A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation, *The American Statistician*.

Evans, Hastings, and Peacock (2000), *Statistical Distributions*, 3rd. Ed., John Wiley and Sons.

Everitt, Brian (1978), *Multivariate Techniques for Multivariate Data*, North-Holland.

Filliben, J. J. (February 1975), The Probability Plot Correlation Coefficient Test for Normality, *Technometrics*, pp. 111-117.

Fuller Jr., E. R., Frieman, S. W., Quinn, J. B., Quinn, G. D., and Carter, W. C. (1994), Fracture Mechanics Approach to the Design of Glass Aircraft Windows: A Case Study, *SPIE Proceedings*, Vol. 2286, (Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, WA).

Gill, Lisa (April 1997), *Summary Analysis: High Performance Ceramics Experiment to Characterize the Effect of Grinding Parameters on Sintered Reaction Bonded Silicon Nitride, Reaction Bonded Silicon Nitride, and Sintered Silicon Nitride* , presented at the NIST - Ceramic Machining Consortium, 10th Program Review Meeting, April 10, 1997.

Granger and Hatanaka (1964), *Spectral Analysis of Economic Time Series*, Princeton University Press.

Grubbs, Frank (1950), Sample Criteria for Testing Outlying Observations, *Annals of Mathematical Statistics*, 21(1) pp. 27-58.

Grubbs, Frank (February 1969), Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 11(1), pp. 1-21.

Hahn, G. J. and Meeker, W. Q. (1991), *Statistical Intervals*, John Wiley and Sons.

Harris, Robert L. (1996), *Information Graphics*, Management Graphics.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Hawkins, D. M. (1980), *Identification of Outliers*, Chapman and Hall.

Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", *The ASQC Basic References in Quality Control: Statistical Techniques*, Edward F. Mykytka, Ph.D., Editor.

Jenkins and Watts, (1968), *Spectral Analysis and Its Applications*, Holden-Day.

Johnson, Kotz, and Balakrishnan, (1994), *Continuous Univariate Distributions, Volumes I and II*, 2nd. Ed., John Wiley and Sons.

Johnson, Kotz, and Kemp, (1992), *Univariate Discrete Distributions*, 2nd. Ed., John Wiley and Sons.

Kuo, Way and Pierson, Marcia Martens, Eds. (1993), *Quality Through Engineering Design"*, specifically, the article Filliben, Cetinkunt, Yu, and Dommenz (1993), *Exploratory Data Analysis Techniques as Applied to a High-Precision Turning Machine*, Elsevier, New York, pp. 199-223.

Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

McNeil, Donald (1977), *Interactive Data Analysis*, John Wiley and Sons.

Mendenhall, William and Reinmuth, James (1982), *Statistics for Management and Ecomonics, Fourth Edition*, Duxbury Press.

Mosteller, Frederick and Tukey, John (1977), *Data Analysis and Regression*, Addison-Wesley.

Natrella, Mary (1963), *Experimental Statistics*, National Bureau of Standards Handbook 91.

Nelson, Wayne (1982), *Applied Life Data Analysis*, Addison-Wesley.

Nelson, Wayne and Doganaksoy, Necip (1992), A Computer Program POWNOR for Fitting the Power-Normal and -Lognormal Models to Life or Strength Data from Specimens of Various Sizes, *NISTIR 4760*, U.S. Department of Commerce, National Institute of Standards and Technology.

Neter, Wasserman, and Kunter (1990). *Applied Linear Statistical Models*, 3rd ed., Irwin.

Pepi, John W., (1994), Failsafe Design of an All BK-7 Glass Aircraft Window, *SPIE Proceedings*, Vol. 2286, (Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, WA).

The RAND Corporation (1955), *A Million Random Digits with 100,000 Normal Deviates*, Free Press.

Rosner, Bernard (May 1983), Percentage Points for a Generalized ESD Many-Outlier Procedure,*Technometrics*, 25(2), pp. 165-172.

Ryan, Thomas (1997), *Modern Regression Methods*, John Wiley.

Scott, David (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization* , John Wiley and Sons.

Snedecor, George W. and Cochran, William G. (1989), *Statistical Methods*, Eighth Edition, Iowa State University Press.

Stefansky, W. (1972), Rejecting Outliers in Factorial Designs, *Technometrics*, 14, pp. 469-479.

Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, *Journal of the American Statistical Association*, 69, pp. 730-737.

Stephens, M. A. (1976). Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters, *Annals of Statistics*, 4, pp. 357-369.

Stephens, M. A. (1977). Goodness of Fit for the Extreme Value Distribution, *Biometrika*, 64, pp. 583-588.

Stephens, M. A. (1977). *Goodness of Fit with Special Reference to Tests for Exponentiality* , Technical Report No. 262, Department of Statistics, Stanford University, Stanford, CA.

Stephens, M. A. (1979). Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function, *Biometrika*, 66, pp. 591-595.

Tietjen and Moore (August 1972), Some Grubbs-Type Statistics for the Detection of Outliers, *Technometrics*, 14(3), pp. 583-597.

Tufte, Edward (1983), *The Visual Display of Quantitative Information*, Graphics Press.

Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley.

Velleman, Paul and Hoaglin, David (1981), *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury.

Wainer, Howard (1981), *Visual Revelations*, Copernicus.

Wilk, M. B. and Gnanadesikan, R. (1968), Probability Plotting Methods for the Analysis of Data, *Biometrika*, 5(5), pp. 1-19.

2. Measurement Process Characterization

1. Characterization

1. Issues
2. Check standards

2. Control

1. Issues
2. Bias and long-term variability
3. Short-term variability

3. Calibration

1. Issues
2. Artifacts
3. Designs
4. Catalog of designs
5. Artifact control
6. Instruments
7. Instrument control

4. Gauge R & R studies

1. Issues
2. Design
3. Data collection
4. Variability
5. Bias
6. Uncertainty

5. Uncertainty analysis

1. Issues
2. Approach
3. Type A evaluations
4. Type B evaluations
5. Propagation of error
6. Error budget
7. Expanded uncertainties
8. Uncorrected bias

6. Case Studies

1. Gauge study
2. Check standard
3. Type A uncertainty
4. Type B uncertainty

Detailed table of contents

References for Chapter 2

2. Measurement Process Characterization - Detailed Table of Contents

1. Characterization [2.1.]
    1. What are the issues for characterization? [2.1.1.]
        1. Purpose [2.1.1.1.]
        2. Reference base [2.1.1.2.]
        3. Bias and Accuracy [2.1.1.3.]
        4. Variability [2.1.1.4.]
    2. What is a check standard? [2.1.2.]
        1. Assumptions [2.1.2.1.]
        2. Data collection [2.1.2.2.]
        3. Analysis [2.1.2.3.]

2. Statistical control of a measurement process [2.2.]
    1. What are the issues in controlling the measurement process? [2.2.1.]
    2. How are bias and variability controlled? [2.2.2.]
        1. Shewhart control chart [2.2.2.1.]
            1. EWMA control chart [2.2.2.1.1.]
        2. Data collection [2.2.2.2.]
        3. Monitoring bias and long-term variability [2.2.2.3.]
        4. Remedial actions [2.2.2.4.]
    3. How is short-term variability controlled? [2.2.3.]
        1. Control chart for standard deviations [2.2.3.1.]
        2. Data collection [2.2.3.2.]
        3. Monitoring short-term precision [2.2.3.3.]
        4. Remedial actions [2.2.3.4.]

3. Calibration [2.3.]
    1. Issues in calibration [2.3.1.]
        1. Reference base [2.3.1.1.]
        2. Reference standards [2.3.1.2.]
    2. What is artifact (single-point) calibration? [2.3.2.]
    3. What are calibration designs? [2.3.3.]
        1. Elimination of special types of bias [2.3.3.1.]
            1. Left-right (constant instrument) bias [2.3.3.1.1.]
            2. Bias caused by instrument drift [2.3.3.1.2.]
        2. Solutions to calibration designs [2.3.3.2.]
            1. General matrix solutions to calibration designs [2.3.3.2.1.]
        3. Uncertainties of calibrated values [2.3.3.3.]
            1. Type A evaluations for calibration designs [2.3.3.3.1.]
            2. Repeatability and level-2 standard deviations [2.3.3.3.2.]
            3. Combination of repeatability and level-2 standard deviations [2.3.3.3.3.]

4. Calculation of standard deviations for 1,1,1,1 design [2.3.3.3.4.]
5. Type B uncertainty [2.3.3.3.5.]
6. Expanded uncertainties [2.3.3.3.6.]

4. Catalog of calibration designs [2.3.4.]
   1. Mass weights [2.3.4.1.]
      1. Design for 1,1,1 [2.3.4.1.1.]
      2. Design for 1,1,1,1 [2.3.4.1.2.]
      3. Design for 1,1,1,1,1 [2.3.4.1.3.]
      4. Design for 1,1,1,1,1,1 [2.3.4.1.4.]
      5. Design for 2,1,1,1 [2.3.4.1.5.]
      6. Design for 2,2,1,1,1 [2.3.4.1.6.]
      7. Design for 2,2,2,1,1 [2.3.4.1.7.]
      8. Design for 5,2,2,1,1,1 [2.3.4.1.8.]
      9. Design for 5,2,2,1,1,1,1 [2.3.4.1.9.]
      10. Design for 5,3,2,1,1,1 [2.3.4.1.10.]
      11. Design for 5,3,2,1,1,1,1 [2.3.4.1.11.]
      12. Design for 5,3,2,2,1,1,1 [2.3.4.1.12.]
      13. Design for 5,4,4,3,2,2,1,1 [2.3.4.1.13.]
      14. Design for 5,5,2,2,1,1,1,1 [2.3.4.1.14.]
      15. Design for 5,5,3,2,1,1,1 [2.3.4.1.15.]
      16. Design for 1,1,1,1,1,1,1,1 weights [2.3.4.1.16.]
      17. Design for 3,2,1,1,1 weights [2.3.4.1.17.]
      18. Design for 10 and 20 pound weights [2.3.4.1.18.]
   2. Drift-elimination designs for gage blocks [2.3.4.2.]
      1. Doiron 3-6 Design [2.3.4.2.1.]
      2. Doiron 3-9 Design [2.3.4.2.2.]
      3. Doiron 4-8 Design [2.3.4.2.3.]
      4. Doiron 4-12 Design [2.3.4.2.4.]
      5. Doiron 5-10 Design [2.3.4.2.5.]
      6. Doiron 6-12 Design [2.3.4.2.6.]
      7. Doiron 7-14 Design [2.3.4.2.7.]
      8. Doiron 8-16 Design [2.3.4.2.8.]
      9. Doiron 9-18 Design [2.3.4.2.9.]
      10. Doiron 10-20 Design [2.3.4.2.10.]
      11. Doiron 11-22 Design [2.3.4.2.11.]
   3. Designs for electrical quantities [2.3.4.3.]
      1. Left-right balanced design for 3 standard cells [2.3.4.3.1.]
      2. Left-right balanced design for 4 standard cells [2.3.4.3.2.]
      3. Left-right balanced design for 5 standard cells [2.3.4.3.3.]
      4. Left-right balanced design for 6 standard cells [2.3.4.3.4.]
      5. Left-right balanced design for 4 references and 4 test items [2.3.4.3.5.]
      6. Design for 8 references and 8 test items [2.3.4.3.6.]
      7. Design for 4 reference zeners and 2 test zeners [2.3.4.3.7.]
      8. Design for 4 reference zeners and 3 test zeners [2.3.4.3.8.]
      9. Design for 3 references and 1 test resistor [2.3.4.3.9.]
      10. Design for 4 references and 1 test resistor [2.3.4.3.10.]
   4. Roundness measurements [2.3.4.4.]
      1. Single trace roundness design [2.3.4.4.1.]
      2. Multiple trace roundness designs [2.3.4.4.2.]
   5. Designs for angle blocks [2.3.4.5.]
      1. Design for 4 angle blocks [2.3.4.5.1.]
      2. Design for 5 angle blocks [2.3.4.5.2.]
      3. Design for 6 angle blocks [2.3.4.5.3.]
   6. Thermometers in a bath [2.3.4.6.]

7. Humidity standards [2.3.4.7.]
    1. Drift-elimination design for 2 reference weights and 3 cylinders [2.3.4.7.1.]
5. Control of artifact calibration [2.3.5.]
    1. Control of precision [2.3.5.1.]
        1. Example of control chart for precision [2.3.5.1.1.]
    2. Control of bias and long-term variability [2.3.5.2.]
        1. Example of Shewhart control chart for mass calibrations [2.3.5.2.1.]
        2. Example of EWMA control chart for mass calibrations [2.3.5.2.2.]
6. Instrument calibration over a regime [2.3.6.]
    1. Models for instrument calibration [2.3.6.1.]
    2. Data collection [2.3.6.2.]
    3. Assumptions for instrument calibration [2.3.6.3.]
    4. What can go wrong with the calibration procedure [2.3.6.4.]
        1. Example of day-to-day changes in calibration [2.3.6.4.1.]
    5. Data analysis and model validation [2.3.6.5.]
        1. Data on load cell #32066 [2.3.6.5.1.]
    6. Calibration of future measurements [2.3.6.6.]
    7. Uncertainties of calibrated values [2.3.6.7.]
        1. Uncertainty for quadratic calibration using propagation of error [2.3.6.7.1.]
        2. Uncertainty for linear calibration using check standards [2.3.6.7.2.]
        3. Comparison of check standard analysis and propagation of error [2.3.6.7.3.]
7. Instrument control for linear calibration [2.3.7.]
    1. Control chart for a linear calibration line [2.3.7.1.]

4. Gauge R & R studies [2.4.]
    1. What are the important issues? [2.4.1.]
    2. Design considerations [2.4.2.]
    3. Data collection for time-related sources of variability [2.4.3.]
        1. Simple design [2.4.3.1.]
        2. 2-level nested design [2.4.3.2.]
        3. 3-level nested design [2.4.3.3.]
    4. Analysis of variability [2.4.4.]
        1. Analysis of repeatability [2.4.4.1.]
        2. Analysis of reproducibility [2.4.4.2.]
        3. Analysis of stability [2.4.4.3.]
            1. Example of calculations [2.4.4.4.4.]
    5. Analysis of bias [2.4.5.]
        1. Resolution [2.4.5.1.]
        2. Linearity of the gauge [2.4.5.2.]
        3. Drift [2.4.5.3.]
        4. Differences among gauges [2.4.5.4.]
        5. Geometry/configuration differences [2.4.5.5.]
        6. Remedial actions and strategies [2.4.5.6.]
    6. Quantifying uncertainties from a gauge study [2.4.6.]

5. Uncertainty analysis [2.5.]
    1. Issues [2.5.1.]
    2. Approach [2.5.2.]
        1. Steps [2.5.2.1.]
    3. Type A evaluations [2.5.3.]
        1. Type A evaluations of random components [2.5.3.1.]
            1. Type A evaluations of time-dependent effects [2.5.3.1.1.]
            2. Measurement configuration within the laboratory [2.5.3.1.2.]
        2. Material inhomogeneity [2.5.3.2.]

1. Data collection and analysis [2.5.3.2.1.]
3. Type A evaluations of bias [2.5.3.3.]
1. Inconsistent bias [2.5.3.3.1.]
2. Consistent bias [2.5.3.3.2.]
3. Bias with sparse data [2.5.3.3.3.]
4. Type B evaluations [2.5.4.]
1. Standard deviations from assumed distributions [2.5.4.1.]
5. Propagation of error considerations [2.5.5.]
1. Formulas for functions of one variable [2.5.5.1.]
2. Formulas for functions of two variables [2.5.5.2.]
3. Propagation of error for many variables [2.5.5.3.]
6. Uncertainty budgets and sensitivity coefficients [2.5.6.]
1. Sensitivity coefficients for measurements on the test item [2.5.6.1.]
2. Sensitivity coefficients for measurements on a check standard [2.5.6.2.]
3. Sensitivity coefficients for measurements from a 2-level design [2.5.6.3.]
4. Sensitivity coefficients for measurements from a 3-level design [2.5.6.4.]
5. Example of uncertainty budget [2.5.6.5.]
7. Standard and expanded uncertainties [2.5.7.]
1. Degrees of freedom [2.5.7.1.]
8. Treatment of uncorrected bias [2.5.8.]
1. Computation of revised uncertainty [2.5.8.1.]

6. Case studies [2.6.]
1. Gauge study of resistivity probes [2.6.1.]
1. Background and data [2.6.1.1.]
1. Database of resistivity measurements [2.6.1.1.1.]
2. Analysis and interpretation [2.6.1.2.]
3. Repeatability standard deviations [2.6.1.3.]
4. Effects of days and long-term stability [2.6.1.4.]
5. Differences among 5 probes [2.6.1.5.]
6. Run gauge study example using Dataplot™ [2.6.1.6.]
7. Dataplot™ macros [2.6.1.7.]
2. Check standard for resistivity measurements [2.6.2.]
1. Background and data [2.6.2.1.]
1. Database for resistivity check standard [2.6.2.1.1.]
2. Analysis and interpretation [2.6.2.2.]
1. Repeatability and level-2 standard deviations [2.6.2.2.1.]
3. Control chart for probe precision [2.6.2.3.]
4. Control chart for bias and long-term variability [2.6.2.4.]
5. Run check standard example yourself [2.6.2.5.]
6. Dataplot™ macros [2.6.2.6.]
3. Evaluation of type A uncertainty [2.6.3.]
1. Background and data [2.6.3.1.]
1. Database of resistivity measurements [2.6.3.1.1.]
2. Measurements on wiring configurations [2.6.3.1.2.]
2. Analysis and interpretation [2.6.3.2.]
1. Difference between 2 wiring configurations [2.6.3.2.1.]
3. Run the type A uncertainty analysis using Dataplot™ [2.6.3.3.]
4. Dataplot™ macros [2.6.3.4.]
4. Evaluation of type B uncertainty and propagation of error [2.6.4.]

7. References [2.7.]

NIST
SEMATECH

2. Measurement Process Characterization

2.1. Characterization

The primary goal of this section is to lay the groundwork for understanding the measurement process in terms of the errors that affect the process.

What are the issues for characterization?

1. Purpose
2. Reference base
3. Bias and Accuracy
4. Variability

What is a check standard?

1. Assumptions
2. Data collection
3. Analysis

2.1.1. What are the issues for characterization?

'Goodness' of measurements

A measurement process can be thought of as a well-run production process in which measurements are the output. The 'goodness' of measurements is the issue, and goodness is characterized in terms of the errors that affect the measurements.

Bias, variability and uncertainty

The goodness of measurements is quantified in terms of

- Bias
- Short-term variability or instrument precision
- Day-to-day or long-term variability
- Uncertainty

Requires ongoing statistical control program

The continuation of goodness is guaranteed by a statistical control program that controls both

- Short-term variability or instrument precision
- Long-term variability which controls bias and day-to-day variability of the process

Scope is limited to ongoing processes

The techniques in this chapter are intended primarily for ongoing processes. One-time tests and special tests or destructive tests are difficult to characterize. Examples of ongoing processes are:

- Calibration where similar test items are measured on a regular basis
- Certification where materials are characterized on a regular basis
- Production where the metrology (tool) errors may be significant
- Special studies where data can be collected over the life of the study

Application to production processes

The material in this chapter is pertinent to the study of production processes for which the size of the metrology (tool) error may be an important consideration. More specific guidance on assessing metrology errors can be found in the section on gauge studies.

2.1.1. What are the issues for characterization?

2.1.1.1. Purpose

Purpose is to understand and quantify the effect of error on reported values

The purpose of characterization is to develop an understanding of the sources of error in the measurement process and how they affect specific measurement results. This section provides the background for:

- identifying sources of error in the measurement process
- understanding and quantifying errors in the measurement process
- codifying the effects of these errors on a specific reported value in a statement of uncertainty

Important concepts

Characterization relies upon the understanding of certain underlying concepts of measurement systems; namely,

- reference base (authority) for the measurement
- bias
- variability
- check standard

Reported value is a generic term that identifies the result that is transmitted to the customer

The reported value is the measurement result for a particular test item. It can be:

- a single measurement
- an average of several measurements
- a least-squares prediction from a model
- a combination of several measurement results that are related by a physical model

2.1.1.2. Reference base

Ultimate authority

The most critical element of any measurement process is the relationship between a single measurement and the reference base for the unit of measurement. The reference base is the ultimate source of authority for the measurement unit.

For fundamental units

Reference bases for fundamental units of measurement (length, mass, temperature, voltage, and time) and some derived units (such as pressure, force, flow rate, etc.) are maintained by national and regional standards laboratories. Consensus values from interlaboratory tests or instrumentation/standards as maintained in specific environments may serve as reference bases for other units of measurement.

For comparison purposes

A reference base, for comparison purposes, may be based on an agreement among participating laboratories or organizations and derived from

- measurements made with a standard test method
- measurements derived from an interlaboratory test

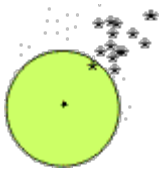
2.1.1.3. Bias and Accuracy

Definition of Accuracy and Bias

Accuracy is a qualitative term referring to whether there is agreement between a measurement made on an object and its true (target or reference) value. Bias is a quantitative term describing the difference between the average of measurements made on the same object and its true value. In particular, for a measurement laboratory, bias is the difference (generally unknown) between a laboratory's average value (over time) for a test item and the average that would be achieved by the reference laboratory if it undertook the same measurements on the same test item.

Depiction of bias and unbiased measurements

Unbiased measurements relative to the target



Biased measurements relative to the target

Identification of bias

Bias in a measurement process can be identified by:

1. Calibration of standards and/or instruments by a reference laboratory, where a value is assigned to the client's standard based on comparisons with the reference laboratory's standards.
2. Check standards , where violations of the control limits on a control chart for the check standard suggest that re-calibration of standards or instruments is needed.
3. Measurement assurance programs, where artifacts from a reference laboratory or other qualified agency are sent to a client and measured in the client's environment as a 'blind' sample.
4. Interlaboratory comparisons, where reference standards or materials are circulated among several laboratories.

Reduction of

Bias can be eliminated or reduced by calibration of

bias

standards and/or instruments. Because of costs and time constraints, the majority of calibrations are performed by secondary or tertiary laboratories and are related to the reference base via a chain of intercomparisons that start at the reference laboratory.

Bias can also be reduced by corrections to in-house measurements based on comparisons with artifacts or instruments circulated for that purpose (reference materials).

Caution

Errors that contribute to bias can be present even where all equipment and standards are properly calibrated and under control. Temperature probably has the most potential for introducing this type of bias into the measurements. For example, a constant heat source will introduce serious errors in dimensional measurements of metal objects. Temperature affects chemical and electrical measurements as well.

Generally speaking, errors of this type can be identified only by those who are thoroughly familiar with the measurement technology. The reader is advised to consult the technical literature and experts in the field for guidance.

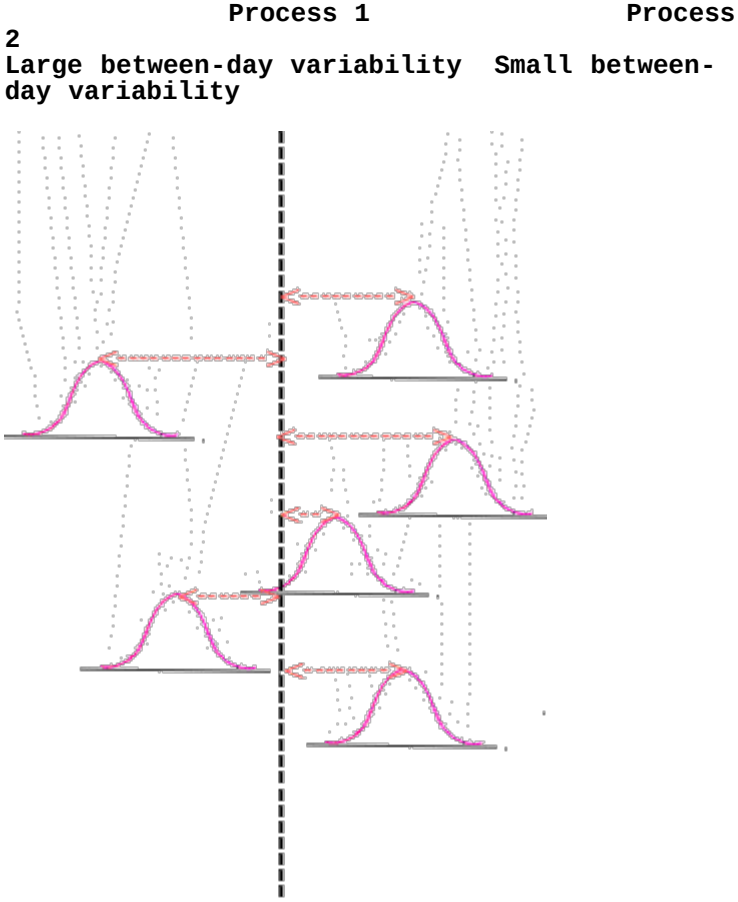
2.1.1.4. Variability

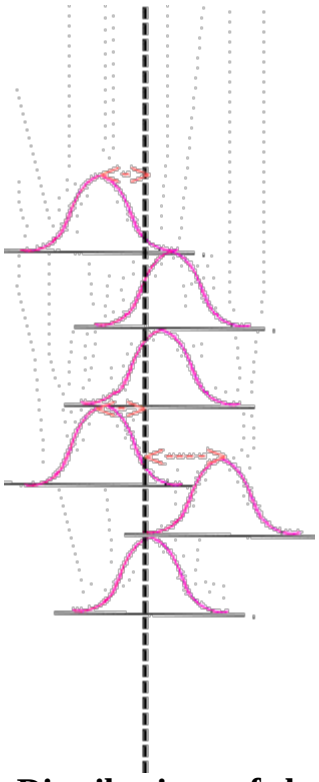
Sources of time-dependent variability

Variability is the tendency of the measurement process to produce slightly different measurements on the same test item, where conditions of measurement are either stable or vary over time, temperature, operators, etc. In this chapter we consider two sources of time-dependent variability:

- Short-term variability ascribed to the precision of the instrument
- Long-term variability related to changes in environment and handling techniques

Depiction of two measurement processes with the same short-term variability over six days where process 1 has large between-day variability and process 2 has negligible between-day variability





Distributions of short-term measurements over 6 days where distances from the centerlines illustrate between-day variability

Short-term variability

Short-term errors affect the precision of the instrument. Even very precise instruments exhibit small changes caused by random errors. It is useful to think in terms of measurements performed with a single instrument over minutes or hours; this is to be understood, normally, as the time that it takes to complete a measurement sequence.

Terminology

Four terms are in common usage to describe short-term phenomena. They are interchangeable.

1. precision
2. repeatability
3. within-time variability
4. short-term variability

Precision is quantified by a standard deviation

The measure of precision is a standard deviation. Good precision implies a small standard deviation. This standard deviation is called the short-term standard deviation of the process or the repeatability standard deviation.

Caution -- long-term variability may be dominant

With very precise instrumentation, it is not unusual to find that the variability exhibited by the measurement process from day-to-day often exceeds the precision of the instrument because of small changes in environmental conditions and handling techniques which cannot be

controlled or corrected in the measurement process. The measurement process is not completely characterized until this source of variability is quantified.

Terminology

Three terms are in common usage to describe long-term phenomena. They are interchangeable.

1. day-to-day variability
2. long-term variability
3. reproducibility

Caution -- regarding term 'reproducibility'

The term 'reproducibility' is given very specific definitions in some national and international standards. However, the definitions are not always in agreement. Therefore, it is used here only in a generic sense to indicate variability across days.

Definitions in this Handbook

We adopt precise definitions and provide data collection and analysis techniques in the sections on check standards and measurement control for estimating:

- Level-1 standard deviation for short-term variability
- Level-2 standard deviation for day-to-day variability

In the section on gauge studies, the concept of variability is extended to include very long-term measurement variability:

- Level-1 standard deviation for short-term variability
- Level-2 standard deviation for day-to-day variability
- Level-3 standard deviation for very long-term variability

We refer to the standard deviations associated with these three kinds of uncertainty as "Level 1, 2, and 3 standard deviations", respectively.

Long-term variability is quantified by a standard deviation

The measure of long-term variability is the standard deviation of measurements taken over several days, weeks or months.

The simplest method for doing this assessment is by analysis of a check standard database. The measurements on the check standards are structured to cover a long time interval and to capture all sources of variation in the measurement process.

2.1.2. What is a check standard?

A check standard is useful for gathering data on the process

Check standard methodology is a tool for collecting data on the measurement process to expose errors that afflict the process over time. Time-dependent sources of error are evaluated and quantified from the database of check standard measurements. It is a device for controlling the bias and long-term variability of the process once a baseline for these quantities has been established from historical data on the check standard.

Think in terms of data

A check standard can be an artifact or defined quantity

The check standard should be thought of in terms of a database of measurements. It can be defined as an artifact or as a characteristic of the measurement process whose value can be replicated from measurements taken over the life of the process. Examples are:

- measurements on a stable artifact
- differences between values of two reference standards as estimated from a calibration experiment
- values of a process characteristic, such as a bias term, which is estimated from measurements on reference standards and/or test items.

An artifact check standard must be close in material content and geometry to the test items that are measured in the workload. If possible, it should be one of the test items from the workload. Obviously, it should be a stable artifact and should be available to the measurement process at all times.

Solves the difficulty of sampling the process

Measurement processes are similar to production processes in that they are continual and are expected to produce identical results (within acceptable limits) over time, instruments, operators, and environmental conditions. However, it is difficult to sample the output of the measurement process because, normally, test items change with each measurement sequence.

Surrogate for unseen measurements

Measurements on the check standard, spaced over time at regular intervals, act as surrogates for measurements that could be made on test items if sufficient time and resources were available.

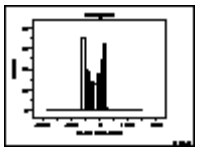
2.1.2.1. Assumptions

Case study:

Resistivity check standard

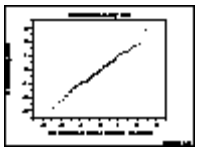
Before applying the quality control procedures recommended in this chapter to check standard data, basic assumptions should be examined. The basic assumptions underlying the quality control procedures are:

1. The data come from a single statistical distribution.
2. The distribution is a normal distribution.
3. The errors are uncorrelated over time.

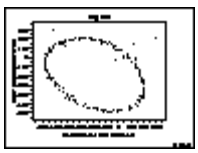


An easy method for checking the assumption of a single normal distribution is to construct a histogram of the check standard data. The histogram should follow a bell-shaped pattern with a single hump. Types of anomalies that indicate a problem with the measurement system are:

1. a double hump indicating that errors are being drawn from two or more distributions;
2. long tails indicating outliers in the process;
3. flat pattern or one with humps at either end indicating that the measurement process in not in control or not properly specified.



Another graphical method for testing the normality assumption is a probability plot. The points are expected to fall approximately on a straight line if the data come from a normal distribution. Outliers, or data from other distributions, will produce an S-shaped curve.



A graphical method for testing for correlation among measurements is a time-lag plot. Correlation will frequently not be a problem if measurements are properly structured over time. Correlation problems generally occur when measurements are taken so close together in time that the instrument cannot properly recover from one measurement to the next. Correlations over time are usually present but are often negligible.

2.1.2.2. Data collection

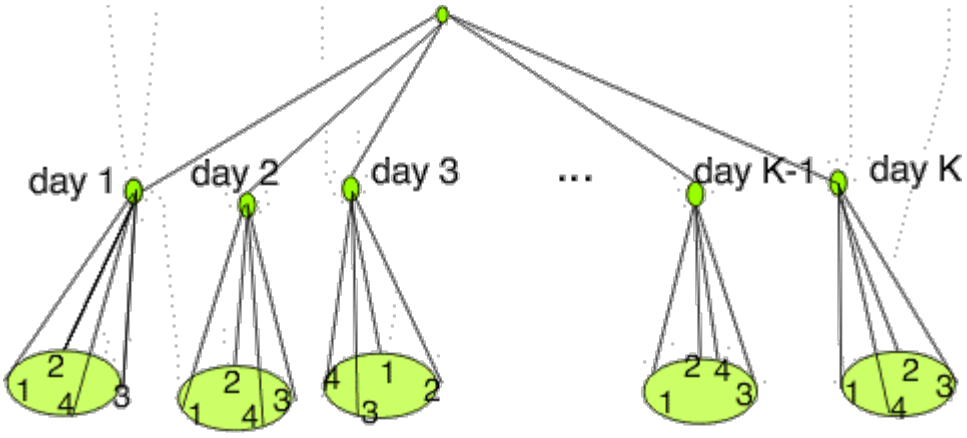
Schedule for making measurements

A schedule for making check standard measurements over time (once a day, twice a week, or whatever is appropriate for sampling all conditions of measurement) should be set up and adhered to. The check standard measurements should be structured in the same way as values reported on the test items. For example, if the reported values are averages of two repetitions made within 5 minutes of each other, the check standard values should be averages of the two measurements made in the same manner.

Exception

One exception to this rule is that there should be at least $J = 2$ repetitions per day. Without this redundancy, there is no way to check on the short-term precision of the measurement system.

Depiction of schedule for making check standard measurements with four repetitions per day over K days on the surface of a silicon wafer with the repetitions randomized at various positions on the wafer



K days - 4 repetitions

2-level design for measurement process

Case study: Resistivity check standard for measurements on silicon wafers

The values for the check standard should be recorded along with pertinent environmental readings and identifications for all other significant factors. The best way to record this information is in one file with one line or row (on a spreadsheet) of information in fixed fields for each check standard measurement. A list of typical entries follows.

1. Identification for check standard
2. Date
3. Identification for the measurement design (if applicable)
4. Identification for the instrument

5. Check standard value
6. Short-term standard deviation from J repetitions
7. Degrees of freedom
8. Operator identification
9. Environmental readings (if pertinent)

2.1.2.3. Analysis

Short-term or level-1 standard deviations from J repetitions

An analysis of the check standard data is the basis for quantifying random errors in the measurement process -- particularly time-dependent errors.

Given that we have a database of check standard measurements as described in data collection where

$$Y_{kj}\left(k=1,\cdots K;j=1,\cdots J\right)$$

represents the jth repetition on the kth day, the mean for the kth day is

$$\bar{Y}_{k\bullet}=\frac{1}{J}\sum_{j=1}^{J}Y_{kj}$$

and the short-term (level-1) standard deviation with $v = J - 1$ degrees of freedom is

$$s_k=\sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left(Y_{kj}-\bar{Y}_{k\bullet}\right)^2}.$$

Drawback of short-term standard deviations

An individual short-term standard deviation will not be a reliable estimate of precision if the degrees of freedom is less than ten, but the individual estimates can be pooled over the K days to obtain a more reliable estimate. The pooled level-1 standard deviation estimate with $v = K(J - 1)$ degrees of freedom is

$$s_1=\sqrt{\frac{1}{K}\sum_{k=1}^{K}s_k^2}.$$

This standard deviation can be interpreted as quantifying the basic precision of the instrumentation used in the measurement process.

Process (level-2)

The level-2 standard deviation of the check standard is appropriate for representing the process variability. It is

standard deviation

computed with $v = K - 1$ degrees of freedom as:

$$s_{chkstd} = s_2 = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{Y}_{k\bullet} - \bar{Y}_{\bullet\bullet}\right)^2}$$

where

$$\bar{Y}_{\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{k\bullet}$$

is the grand mean of the KJ check standard measurements.

Use in quality control

The check standard data and standard deviations that are described in this section are used for controlling two aspects of a measurement process:

1. Control of short-term variability
2. Control of bias and long-term variability

Case study: Resistivity check standard

For an example, see the case study for resistivity where several check standards were measured $J = 6$ times per day over several days.

2. Measurement Process Characterization

2.2. Statistical control of a measurement process

The purpose of this section is to outline the steps that can be taken to exercise statistical control over the measurement process and demonstrate the validity of the uncertainty statement. Measurement processes can change both with respect to bias and variability. A change in instrument precision may be readily noted as measurements are being recorded, but changes in bias or long-term variability are difficult to catch when the process is looking at a multitude of artifacts over time.

What are the issues for control of a measurement process?

1. Purpose
2. Assumptions
3. Role of the check standard

How are bias and long-term variability controlled?

1. Shewhart control chart
2. Exponentially weighted moving average control chart
3. Data collection and analysis
4. Control procedure
5. Remedial actions & strategies

How is short-term variability controlled?

1. Control chart for standard deviations
2. Data collection and analysis
3. Control procedure
4. Remedial actions and strategies

2.2.1. What are the issues in controlling the measurement process?

Purpose is to guarantee the 'goodness' of measurement results

The purpose of statistical control is to guarantee the 'goodness' of measurement results within predictable limits and to validate the statement of uncertainty of the measurement result.

Statistical control methods can be used to test the measurement process for change with respect to bias and variability from its historical levels. However, if the measurement process is improperly specified or calibrated, then the control procedures can only guarantee comparability among measurements.

Assumption of normality is not stringent

The assumptions that relate to measurement processes apply to statistical control; namely that the errors of measurement are uncorrelated over time and come from a population with a single distribution. The tests for control depend on the assumption that the underlying distribution is normal (Gaussian), but the test procedures are robust to slight departures from normality. Practically speaking, all that is required is that the distribution of measurements be bell-shaped and symmetric.

Check standard is mechanism for controlling the process

Measurements on a check standard provide the mechanism for controlling the measurement process.

Measurements on the check standard should produce identical results except for the effect of random errors, and tests for control are basically tests of whether or not the random errors from the process continue to be drawn from the same statistical distribution as the historical data on the check standard.

Changes that can be monitored and tested with the check standard database are:

1. Changes in bias and long-term variability
2. Changes in instrument precision or short-term variability

2.2.2. How are bias and variability controlled?

Bias and variability are controlled by monitoring measurements on a check standard over time

Bias and long-term variability are controlled by monitoring measurements on a check standard over time. A change in the measurement on the check standard that persists at a constant level over several measurement sequences indicates possible:

1. Change or damage to the reference standards
2. Change or damage to the check standard artifact
3. Procedural change that vitiates the assumptions of the measurement process

A change in the variability of the measurements on the check standard can be due to one of many causes such as:

1. Loss of environmental controls
2. Change in handling techniques
3. Severe degradation in instrumentation.

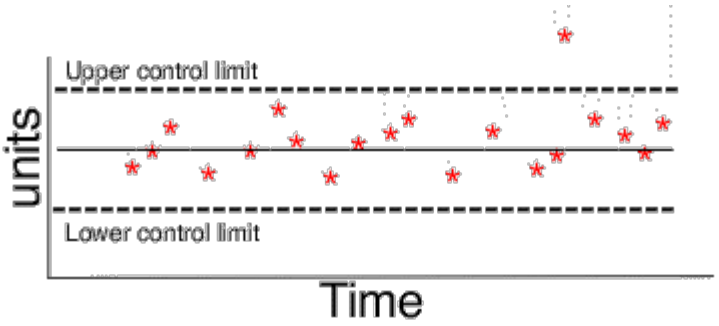
The control procedure monitors the progress of measurements on the check standard over time and signals when a significant change occurs. There are two control chart procedures that are suitable for this purpose.

Shewhart Chart is easy to implement

The Shewhart control chart has the advantage of being intuitive and easy to implement. It is characterized by a center line and symmetric upper and lower control limits. The chart is good for detecting large changes but not for quickly detecting small changes (of the order of one-half to one standard deviation) in the process.

Depiction of Shewhart control chart

In the simplistic illustration of a Shewhart control chart shown below, the measurements are within the control limits with the exception of one measurement which exceeds the upper control limit.

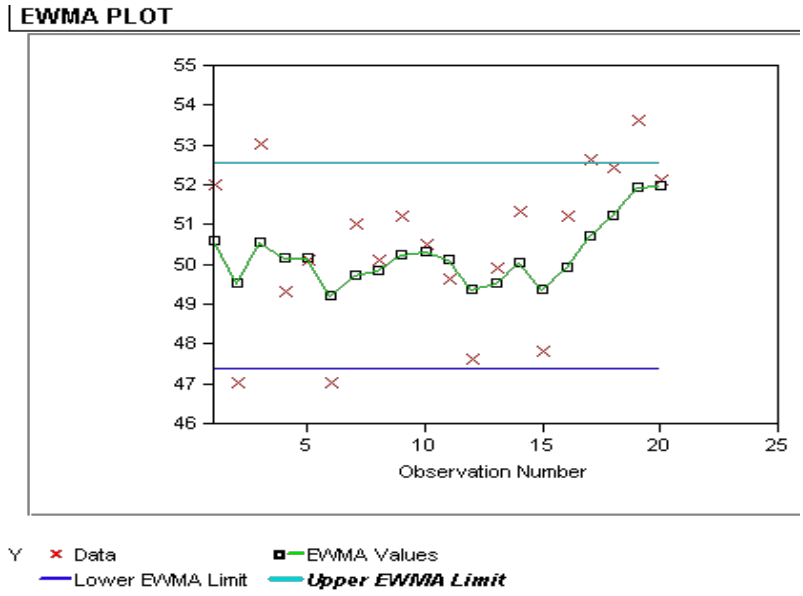


EWMA Chart is better for detecting small changes

The EWMA control chart (exponentially weighted moving average) is more difficult to implement but should be considered if the goal is quick detection of small changes. The decision process for the EWMA chart is based on an exponentially decreasing (over time) function of prior measurements on the check standard while the decision process for the Shewhart chart is based on the current measurement only.

Example of EWMA Chart

In the EWMA control chart below, the red dots represent the measurements. Control is exercised via the exponentially weighted moving average (shown as the curved line) which, in this case, is approaching its upper control limit.



Artifacts for process control must be stable and available

The check standard artifacts for controlling the bias or long-term variability of the process must be of the same type and geometry as items that are measured in the workload. The artifacts must be stable and available to the measurement process on a continuing basis. Usually, one artifact is sufficient. It can be:

Case study: Resistivity

1. An individual item drawn at random from the workload
2. A specific item reserved by the laboratory for the purpose.

Topic covered in this section>

The topics covered in this section include:

1. Shewhart control chart methodology

2. EWMA control chart methodology
3. Data collection & analysis
4. Monitoring
5. Remedies and strategies for dealing with out-of-control signals.

2.2.2.1. Shewhart control chart

Example of Shewhart control chart for mass calibrations

The Shewhart control chart has a baseline and upper and lower limits, shown as dashed lines, that are symmetric about the baseline. Measurements are plotted on the chart versus a time line. Measurements that are outside the limits are considered to be out of control.

Baseline is the average from historical data

The baseline for the control chart is the accepted value, an average of the historical check standard values. A minimum of 100 check standard values is required to establish an accepted value.

Caution - control limits are computed from the process standard deviation -- not from rational subsets

The upper (UCL) and lower (LCL) control limits are:

$$UCL = \text{Accepted value} + k * \text{process standard deviation}$$

$$LCL = \text{Accepted value} - k * \text{process standard deviation}$$

where the process standard deviation is the standard deviation computed from the check standard database.

Individual measurements cannot be assessed using the standard deviation from short-term repetitions

This procedure is an *individual observations* control chart. The previously described control charts depended on *rational subsets*, which use the standard deviations computed from the rational subsets to calculate the control limits. For a measurement process, the subgroups would consist of short-term repetitions which can characterize the precision of the instrument but not the long-term variability of the process. In measurement science, the interest is in assessing individual measurements (or averages of short-term repetitions). Thus, the standard deviation over time is the appropriate measure of variability.

Choice of k depends on number of measurements we are willing to

To achieve tight control of the measurement process, set

$$k = 2$$

in which case approximately 5% of the measurements from a process that is in control will produce out-of-control

reject signals. This assumes that there is a sufficiently large number of degrees of freedom (>100) for estimating the process standard deviation.

To flag only those measurements that are egregiously out of control, set

$$k = 3$$

in which case approximately 1% of the measurements from an in-control process will produce out-of-control signals.

2.2.2.1.1. EWMA control chart

Small changes only become obvious over time

Because it takes time for the patterns in the data to emerge, a permanent shift in the process may not immediately cause individual violations of the control limits on a Shewhart control chart. The Shewhart control chart is not powerful for detecting small changes, say of the order of 1 - 1/2 standard deviations. The EWMA (exponentially weighted moving average) control chart is better suited to this purpose.

Example of EWMA control chart for mass calibrations

The exponentially weighted moving average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time from the current measurement. The data

$$Y_1, Y_2, \ldots, Y_t$$

are the check standard measurements ordered in time. The EWMA statistic at time t is computed recursively from individual data points, with the first EWMA statistic, $EWMA_1$, being the arithmetic average of historical data.

$$EWMA_{t+1} = \lambda Y_t + (1-\lambda) EWMA_t$$

Control mechanism for EWMA

The EWMA control chart can be made sensitive to small changes or a gradual drift in the process by the choice of the weighting factor, λ. A weighting factor of 0.2 - 0.3 is usually suggested for this purpose (Hunter), and 0.15 is also a popular choice.

Limits for the control chart

The target or center line for the control chart is the average of historical data. The upper (UCL) and lower (LCL) limits are

$$UCL = EWMA_1 + ks\sqrt{\frac{\lambda}{(2-\lambda)}}$$

$$LCL = EWMA_1 - ks\sqrt{\frac{\lambda}{(2-\lambda)}}$$

where *s* times the radical expression is a good approximation to the standard deviation of the EWMA statistic and the factor *k* is chosen in the same way as for the Shewhart control chart -- generally to be 2 or 3.

Procedure for implementing the EWMA control chart

The implementation of the EWMA control chart is the same as for any other type of control procedure. The procedure is built on the assumption that the "good" historical data are representative of the in-control process, with future data from the same process tested for agreement with the historical data. To start the procedure, a target (average) and process standard deviation are computed from historical check standard data. Then the procedure enters the monitoring stage with the EWMA statistics computed and tested against the control limits. The EWMA statistics are weighted averages, and thus their standard deviations are smaller than the standard deviations of the raw data and the corresponding control limits are narrower than the control limits for the Shewhart individual observations chart.

2.2.2.2. Data collection

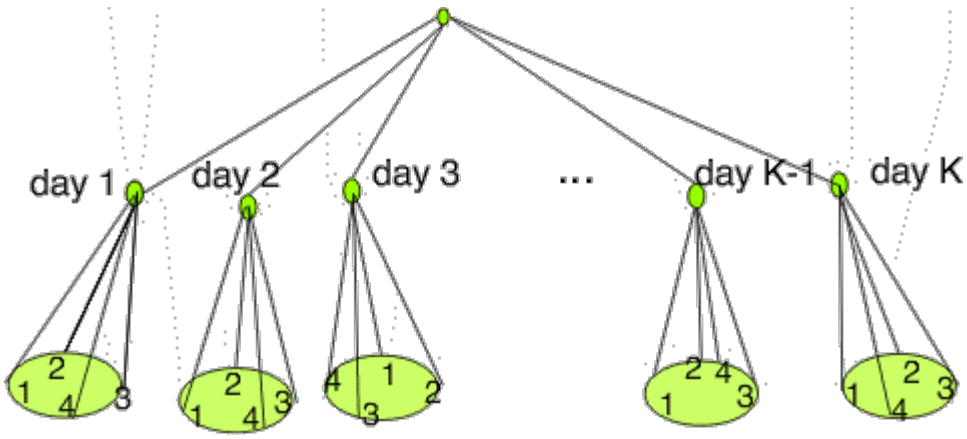
Measurements should cover a sufficiently long time period to cover all environmental conditions

A schedule should be set up for making measurements on the artifact (check standard) chosen for control purposes. The measurements are structured to sample all environmental conditions in the laboratory and all other sources of influence on the measurement result, such as operators and instruments.

For high-precision processes where the uncertainty of the result must be guaranteed, a measurement on the check standard should be included with every measurement sequence, if possible, and at least once a day.

For each occasion, *J* measurements are made on the check standard. If there is no interest in controlling the short-term variability or precision of the instrument, then one measurement is sufficient. However, a dual purpose is served by making two or three measurements that track both the bias and the short-term variability of the process with the same database.

Depiction of check standard measurements with J = 4 repetitions per day on the surface of a silicon wafer over K days where the repetitions are randomized over position on the wafer



K days - 4 repetitions

2-level design for measurements on a check standard

Notation

For *J* measurements on each of *K* days, the measurements are denoted by

$$Y_{kj}\,(k=1,\cdots,K;\ j=1,\cdots J)$$

The check standard value is

The check standard value for the *k*th day is

defined as an average of short-term repetitions

$$\bar{Y}_{k\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{kj}$$

Accepted value of check standard

The accepted value, or baseline for the control chart, is

$$\bar{Y}_{\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_{k\bullet}$$

Process standard deviation

The process standard deviation is

$$s_2 = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{Y}_{k\bullet} - \bar{Y}_{\bullet\bullet}\right)^2}$$

Caution

Check standard measurements should be structured in the same way as values reported on the test items. For example, if the reported values are averages of two measurements made within 5 minutes of each other, the check standard values should be averages of the two measurements made in the same manner.

Database

Case study: Resistivity

Averages and short-term standard deviations computed from *J* repetitions should be recorded in a file along with identifications for all significant factors. The best way to record this information is to use one file with one line (row in a spreadsheet) of information in fixed fields for each group. A list of typical entries follows:

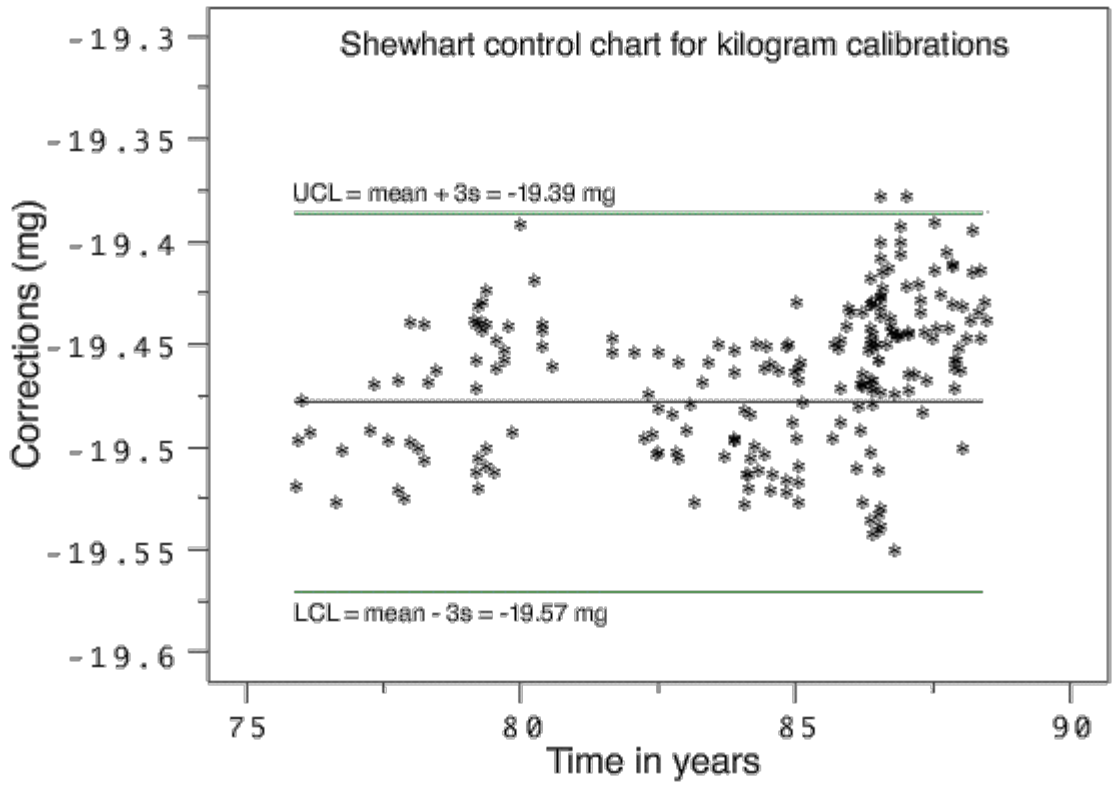
1. Month
2. Day
3. Year
4. Check standard identification
5. Identification for the measurement design (if applicable)
6. Instrument identification
7. Check standard value
8. Repeatability (short-term) standard deviation from *J* repetitions
9. Degrees of freedom
10. Operator identification
11. Environmental readings (if pertinent)

2.2.2.3. Monitoring bias and long-term variability

Monitoring stage

Once the baseline and control limits for the control chart have been determined from historical data, and any bad observations removed and the control limits recomputed, the measurement process enters the monitoring stage. A Shewhart control chart and EWMA control chart for monitoring a mass calibration process are shown below. For the purpose of comparing the two techniques, the two control charts are based on the same data where the baseline and control limits are computed from the data taken prior to 1985. The monitoring stage begins at the start of 1985. Similarly, the control limits for both charts are 3-standard deviation limits. The check standard data and analysis are explained more fully in another section.

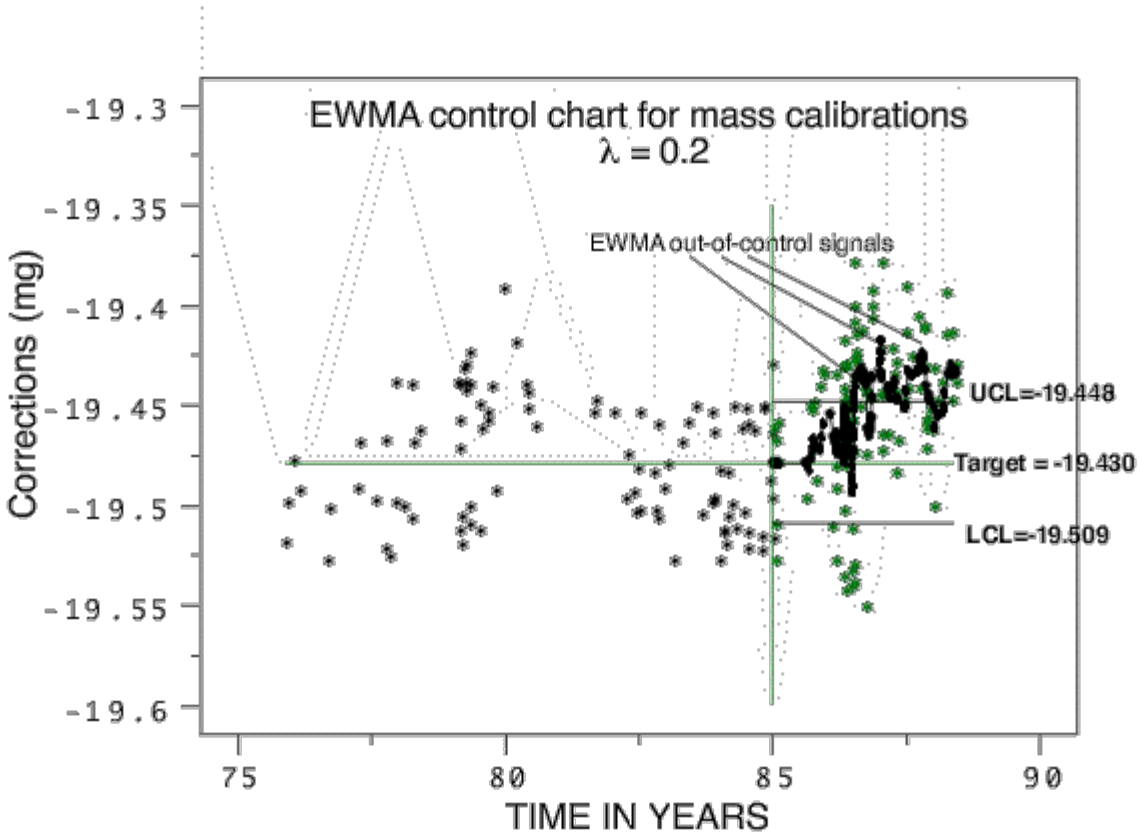
Shewhart control chart of measurements of kilogram check standard showing outliers and a shift in the process that occurred after 1985



EWMA chart for measurements on kilogram check standard

In the EWMA control chart below, the control data after 1985 are shown in green, and the EWMA statistics are shown as black dots superimposed on the raw data. The EWMA statistics, and not the raw data, are of interest in looking for out-of-control signals. Because the EWMA statistic is a weighted average, it has a smaller standard deviation than a single control measurement, and, therefore, the EWMA control limits are narrower than the limits for the Shewhart control chart shown above.

showing multiple violations of the control limits for the EWMA statistics



Measurements that exceed the control limits require action

The control strategy is based on the predictability of future measurements from historical data. Each new check standard measurement is plotted on the control chart in real time. These values are expected to fall within the control limits if the process has not changed. Measurements that exceed the control limits are probably out-of-control and require remedial action. Possible causes of out-of-control signals need to be understood when developing strategies for dealing with outliers.

Signs of significant trends or shifts

The control chart should be viewed in its entirety on a regular basis] to identify drift or shift in the process. In the Shewhart control chart shown above, only a few points exceed the control limits. The small, but significant, shift in the process that occurred after 1985 can only be identified by examining the plot of control measurements over time. A re-analysis of the kilogram check standard data shows that the control limits for the Shewhart control chart should be updated based on the the data after 1985. In the EWMA control chart, multiple violations of the control limits occur after 1986. In the calibration environment, the incidence of several violations should alert the control engineer that a shift in the process has occurred, possibly because of damage or change in the value of a reference standard, and the process requires review.

2.2.2.4. Remedial actions

Consider possible causes for out-of-control signals and take corrective long-term actions

There are many possible causes of out-of-control signals.

A. Causes that do not warrant corrective action for the process (but which do require that the current measurement be discarded) are:

1. Chance failure where the process is actually in-control
2. Glitch in setting up or operating the measurement process
3. Error in recording of data

B. Changes in bias can be due to:

1. Damage to artifacts
2. Degradation in artifacts (wear or build-up of dirt and mineral deposits)

C. Changes in long-term variability can be due to:

1. Degradation in the instrumentation
2. Changes in environmental conditions
3. Effect of a new or inexperienced operator

4-step strategy for short-term

An immediate strategy for dealing with out-of-control signals associated with high precision measurement processes should be pursued as follows:

Repeat measurements

1. Repeat the measurement sequence to establish whether or not the out-of-control signal was simply a chance occurrence, glitch, or whether it flagged a permanent change or trend in the process.

Discard measurements on test items

2. With high precision processes, for which a check standard is measured along with the test items, new values should be assigned to the test items based on new measurement data.

Check for

3. Examine the patterns of recent data. If the process is

drift

gradually drifting out of control because of degradation in instrumentation or artifacts, then:

- Instruments may need to be repaired
- Reference artifacts may need to be recalibrated.

Reevaluate

4. Reestablish the process value and control limits from more recent data if the measurement process cannot be brought back into control.

2.2.3. How is short-term variability controlled?

Emphasis on instruments

Short-term variability or instrument precision is controlled by monitoring standard deviations from repeated measurements on the instrument(s) of interest. The database can come from measurements on a single artifact or a representative set of artifacts.

Artifacts - Case study: Resistivity

The artifacts must be of the same type and geometry as items that are measured in the workload, such as:

1. Items from the workload
2. A single check standard chosen for this purpose
3. A collection of artifacts set aside for this specific purpose

Concepts covered in this section

The concepts that are covered in this section include:

1. Control chart methodology for standard deviations
2. Data collection and analysis
3. Monitoring
4. Remedies and strategies for dealing with out-of-control signals

2.2.3.1. Control chart for standard deviations

Degradation of instrument or anomalous behavior on one occasion

Changes in the precision of the instrument, particularly anomalies and degradation, must be addressed. Changes in precision can be detected by a statistical control procedure based on the *F*-distribution where the short-term standard deviations are plotted on the control chart.

The base line for this type of control chart is the pooled standard deviation, s_1, as defined in Data collection and analysis.

Example of control chart for a mass balance

Only the upper control limit, *UCL*, is of interest for detecting degradation in the instrument. As long as the short-term standard deviations fall within the upper control limit established from historical data, there is reason for confidence that the precision of the instrument has not degraded (i.e., common cause variations).

The control limit is based on the F-distribution

The control limit is

$$UCL = s_1\sqrt{F_{\alpha, J-1, K(J-1)}}$$

where the quantity under the radical is the upper α critical value from the *F* table with degrees of freedom *(J - 1)* and *K(J - 1)*. The numerator degrees of freedom, *v1 = (J -1)*, are associated with the standard deviation computed from the current measurements, and the denominator degrees of freedom, *v2 = K(J -1)*, correspond to the pooled standard deviation of the historical data. The probability α is chosen to be small, say 0.05.

The justification for this control limit, as opposed to the more conventional standard deviation control limit, is that we are essentially performing the following hypothesis test:

$H_0: \sigma_1 = \sigma_2$
$H_a: \sigma_2 > \sigma_1$

where σ_1 is the population value for the s_1 defined above and σ_2 is the population value for the standard deviation of the current values being tested. Generally, s_1 is based on

sufficient historical data that it is reasonable to make the assumption that σ_1 is a "known" value.

The upper control limit above is then derived based on the standard *F* test for equal standard deviations. Justification and details of this derivation are given in Cameron and Hailes (1974).

Sample Code

Sample code for computing the *F* value for the case where $\alpha = 0.05$, $J = 6$, and $K = 6$, is available for both Dataplot and R.

2.2.3.2. Data collection

Case study: Resistivity

A schedule should be set up for making measurements with a single instrument (once a day, twice a week, or whatever is appropriate for sampling all conditions of measurement).

Short-term standard deviations

The measurements are denoted

$$Y_{kj}\left(k=1,\cdots,K;j=1,\cdots J\right)$$

where there are J measurements on each of K occasions. The average for the kth occasion is:

$$\bar{Y}_{k\bullet}=\frac{1}{J}\sum_{j=1}^{J}Y_{kj}$$

The short-term (repeatability) standard deviation for the kth occasion is:

$$s_1=\sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left(Y_{kj}-\bar{Y}_{k\bullet}\right)^2}$$

with *(J-1)* degrees of freedom.

Pooled standard deviation

The repeatability standard deviations are pooled over the K occasions to obtain an estimate with $K(J-1)$ degrees of freedom of the level-1 standard deviation

$$s_1=\sqrt{\frac{1}{K}\sum_{k=1}^{K}s_{1k}^2}$$

Note: The same notation is used for the repeatability standard deviation whether it is based on one set of measurements or pooled over several sets.

Database

The individual short-term standard deviations along with identifications for all significant factors are recorded in a file. The best way to record this information is by using one file with one line (row in a spreadsheet) of information in fixed

fields for each group. A list of typical entries follows.

1. Identification of test item or check standard
2. Date
3. Short-term standard deviation
4. Degrees of freedom
5. Instrument
6. Operator

2.2.3.3. Monitoring short-term precision

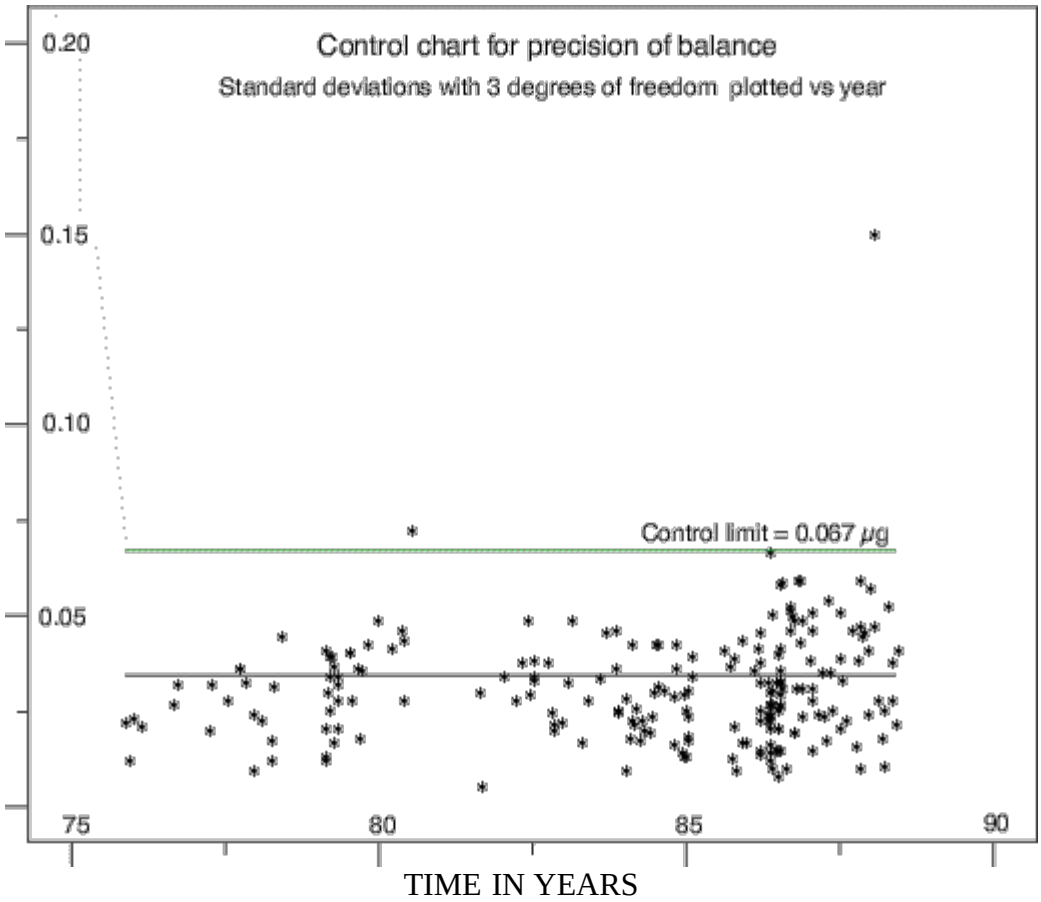
Monitoring future precision

Once the base line and control limit for the control chart have been determined from historical data, the measurement process enters the monitoring stage. In the control chart shown below, the control limit is based on the data taken prior to 1985.

Each new standard deviation is monitored on the control chart

Each new short-term standard deviation based on *J* measurements is plotted on the control chart; points that exceed the control limits probably indicate lack of statistical control. Drift over time indicates degradation of the instrument. Points out of control require remedial action, and possible causes of out of control signals need to be understood when developing strategies for dealing with outliers.

Control chart for precision for a mass balance from historical standard deviations for the balance with 3 degrees of freedom each. The control chart identifies two outliers and slight degradation over time in the precision of the balance



Monitoring where the number of

There is no requirement that future standard deviations be based on *J*, the number of measurements in the historical database. However, a change in the number of measurements leads to a change in the test for control, and it may

measurements are different from J

not be convenient to draw a control chart where the control limits are changing with each new measurement sequence.

For a new standard deviation based on *J'* measurements, the precision of the instrument is in control if

$$s_{new} < s_1 \sqrt{F_{\alpha, J'-1, K(J-1)}}$$

.

Notice that the numerator degrees of freedom, *v1 = J'- 1*, changes but the denominator degrees of freedom, *v2 = K(J - 1)*, remains the same.

2.2.3.4. Remedial actions

Examine possible causes

A. Causes that do not warrant corrective action (but which do require that the current measurement be discarded) are:

1. Chance failure where the precision is actually in control
2. Glitch in setting up or operating the measurement process
3. Error in recording of data

B. Changes in instrument performance can be due to:

1. Degradation in electronics or mechanical components
2. Changes in environmental conditions
3. Effect of a new or inexperienced operator

Repeat measurements

Repeat the measurement sequence to establish whether or not the out-of-control signal was simply a chance occurrence, glitch, or whether it flagged a permanent change or trend in the process.

Assign new value to test item

With high precision processes, for which the uncertainty must be guaranteed, new values should be assigned to the test items based on new measurement data.

Check for degradation

Examine the patterns of recent standard deviations. If the process is gradually drifting out of control because of degradation in instrumentation or artifacts, instruments may need to be repaired or replaced.

2.3. Calibration

The purpose of this section is to outline the procedures for calibrating artifacts and instruments while guaranteeing the 'goodness' of the calibration results. Calibration is a measurement process that assigns values to the property of an artifact or to the response of an instrument relative to reference standards or to a designated measurement process. The purpose of calibration is to eliminate or reduce bias in the user's measurement system relative to the reference base. The calibration procedure compares an "unknown" or test item(s) or instrument with reference standards according to a specific algorithm.

What are the issues for calibration?

1. Artifact or instrument calibration
2. Reference base
3. Reference standard(s)

What is artifact (single-point) calibration?

1. Purpose
2. Assumptions
3. Bias
4. Calibration model

What are calibration designs?

1. Purpose
2. Assumptions
3. Properties of designs
4. Restraint
5. Check standard in a design
6. Special types of bias (left-right effect & linear drift)
7. Solutions to calibration designs
8. Uncertainty of calibrated values

Catalog of calibration designs

1. Mass weights
2. Gage blocks
3. Electrical standards - saturated standard cells, zeners, resistors
4. Roundness standards
5. Angle blocks

6. Indexing tables
7. Humidity cylinders

Control of artifact calibration

1. Control of the precision of the calibrating instrument
2. Control of bias and long-term variability

What is instrument calibration over a regime?

1. Models for instrument calibration
2. Data collection
3. Assumptions
4. What can go wrong with the calibration procedure?
5. Data analysis and model validation
6. Calibration of future measurements
7. Uncertainties of calibrated values
    1. From propagation of error for a quadratic calibration
    2. From check standard measurements for a linear calibration
    3. Comparison of check standard technique and propagation of error

Control of instrument calibration

1. Control chart for linear calibration
2. Critical values of t* statistic

2.3.1. Issues in calibration

Calibration reduces bias

Calibration is a measurement process that assigns values to the property of an artifact or to the response of an instrument relative to reference standards or to a designated measurement process. The purpose of calibration is to eliminate or reduce bias in the user's measurement system relative to the reference base.

Artifact & instrument calibration

The calibration procedure compares an "unknown" or test item(s) or instrument with reference standards according to a specific algorithm. Two general types of calibration are considered in this Handbook:

- artifact calibration at a single point
- instrument calibration over a regime

Types of calibration not discussed

The procedures in this Handbook are appropriate for calibrations at secondary or lower levels of the traceability chain where reference standards for the unit already exist. Calibration from first principles of physics and reciprocity calibration are not discussed.

2.3.1.1. Reference base

Ultimate authority

The most critical element of any measurement process is the relationship between a single measurement and the reference base for the unit of measurement. The reference base is the ultimate source of authority for the measurement unit.

Base and derived units of measurement

The base units of measurement in the Le Systeme International d'Unites (SI) are (Taylor):

- kilogram - mass
- meter - length
- second - time
- ampere - electric current
- kelvin - thermodynamic temperature
- mole - amount of substance
- candela - luminous intensity

These units are maintained by the Bureau International des Poids et Mesures in Paris. Local reference bases for these units and SI derived units such as:

- pascal - pressure
- newton - force
- hertz - frequency
- ohm - resistance
- degrees Celsius - Celsius temperature, etc.

are maintained by national and regional standards laboratories.

Other sources

Consensus values from interlaboratory tests or instrumentation/standards as maintained in specific environments may serve as reference bases for other units of measurement.

2.3.1.2. Reference standards

Primary reference standards

A reference standard for a unit of measurement is an artifact that embodies the quantity of interest in a way that ties its value to the reference base.

At the highest level, a primary reference standard is assigned a value by direct comparison with the reference base. Mass is the only unit of measurement that is defined by an artifact. The kilogram is defined as the mass of a platinum-iridium kilogram that is maintained by the Bureau International des Poids et Mesures in Sevres, France.

Primary reference standards for other units come from realizations of the units embodied in artifact standards. For example, the reference base for length is the meter which is defined as the length of the path by light in vacuum during a time interval of 1/299,792,458 of a second.

Secondary reference standards

Secondary reference standards are calibrated by comparing with primary standards using a high precision comparator and making appropriate corrections for non-ideal conditions of measurement.

Secondary reference standards for mass are stainless steel kilograms, which are calibrated by comparing with a primary standard on a high precision balance and correcting for the buoyancy of air. In turn these weights become the reference standards for assigning values to test weights.

Secondary reference standards for length are gage blocks, which are calibrated by comparing with primary gage block standards on a mechanical comparator and correcting for temperature. In turn, these gage blocks become the reference standards for assigning values to test sets of gage blocks.

2.3.2. What is artifact (single-point) calibration?

Purpose

Artifact calibration is a measurement process that assigns values to the property of an artifact relative to a reference standard(s). The purpose of calibration is to eliminate or reduce bias in the user's measurement system relative to the reference base.

The calibration procedure compares an "unknown" or test item(s) with a reference standard(s) of the same nominal value (hence, the term single-point calibration) according to a specific algorithm called a calibration design.

Assumptions

The calibration procedure is based on the assumption that individual readings on test items and reference standards are subject to:

- Bias that is a function of the measuring system or instrument
- Random error that may be uncontrollable

What is bias?

The operational definition of bias is that it is the difference between values that would be assigned to an artifact by the client laboratory and the laboratory maintaining the reference standards. Values, in this sense, are understood to be the long-term averages that would be achieved in both laboratories.

Calibration model for eliminating bias requires a reference standard that is very close in value to the test item

One approach to eliminating bias is to select a reference standard that is almost identical to the test item; measure the two artifacts with a comparator type of instrument; and take the difference of the two measurements to cancel the bias. The only requirement on the instrument is that it be linear over the small range needed for the two artifacts.

The test item has value X^*, as yet to be assigned, and the reference standard has an assigned value R^*. Given a measurement, X, on the test item and a measurement, R, on the reference standard,

$$X = Bias + X^* + error_1$$
$$R = Bias + R^* + error_2,$$

the difference between the test item and the reference is estimated by

$$D = X - R,$$

and the value of the test item is reported as

$$\hat{Test} = X^* = D + R^*.$$

Need for redundancy leads to calibration designs

A deficiency in relying on a single difference to estimate D is that there is no way of assessing the effect of random errors. The obvious solution is to:

- Repeat the calibration measurements J times
- Average the results
- Compute a standard deviation from the J results

Schedules of redundant intercomparisons involving measurements on several reference standards and test items in a connected sequence are called calibration designs and are discussed in later sections.

2.3.3. What are calibration designs?

Calibration designs are redundant schemes for intercomparing reference standards and test items

Calibration designs are redundant schemes for intercomparing reference standards and test items in such a way that the values can be assigned to the test items based on known values of reference standards. Artifacts that traditionally have been calibrated using calibration designs are:

- mass weights
- resistors
- voltage standards
- length standards
- angle blocks
- indexing tables
- liquid-in-glass thermometers, etc.

Outline of section

The topics covered in this section are:

- Designs for elimination of left-right bias and linear drift
- Solutions to calibration designs
- Uncertainties of calibrated values

A catalog of calibration designs is provided in the next section.

Assumptions for calibration designs include demands on the quality of the artifacts

The assumptions that are necessary for working with calibration designs are that:

- Random errors associated with the measurements are independent.
- All measurements come from a distribution with the same standard deviation.
- Reference standards and test items respond to the measuring environment in the same manner.
- Handling procedures are consistent from item to item.
- Reference standards and test items are stable during the time of measurement.
- Bias is canceled by taking the difference between measurements on the test item and the reference standard.

Important concept - Restraint

The restraint is the known value of the reference standard or, for designs with two or more reference standards, the restraint is the summation of the values of the reference standards.

Requirements & properties of designs

Basic requirements are:

- The differences must be nominally zero.
- The design must be solvable for individual items given the restraint.

It is possible to construct designs which do not have these properties. This will happen, for example, if reference standards are only compared among themselves and test items are only compared among themselves without any intercomparisons.

Practical considerations determine a 'good' design

We do not apply 'optimality' criteria in constructing calibration designs because the construction of a 'good' design depends on many factors, such as convenience in manipulating the test items, time, expense, and the maximum load of the instrument.

- The number of measurements should be small.
- The degrees of freedom should be greater than three.
- The standard deviations of the estimates for the test items should be small enough for their intended purpose.

Check standard in a design

Designs listed in this Handbook have provision for a check standard in each series of measurements. The check standard is usually an artifact, of the same nominal size, type, and quality as the items to be calibrated. Check standards are used for:

- Controlling the calibration process
- Quantifying the uncertainty of calibrated results

Estimates that can be computed from a design

Calibration designs are solved by a restrained least-squares technique (Zelen) which gives the following estimates:

- Values for individual reference standards
- Values for individual test items
- Value for the check standard
- Repeatability standard deviation and degrees of freedom
- Standard deviations associated with values for reference standards and test items

2.3.3.1. Elimination of special types of bias

Assumptions which may be violated

Two of the usual assumptions relating to calibration measurements are not always valid and result in biases. These assumptions are:

- Bias is canceled by taking the difference between the measurement on the test item and the measurement on the reference standard
- Reference standards and test items remain stable throughout the measurement sequence

Ideal situation

In the ideal situation, bias is eliminated by taking the difference between a measurement X on the test item and a measurement R on the reference standard. However, there are situations where the ideal is not satisfied:

- Left-right (or constant instrument) bias
- Bias caused by instrument drift

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.1. Elimination of special types of bias

2.3.3.1.1. Left-right (constant instrument) bias

Left-right bias *which is not eliminated by differencing*

A situation can exist in which a bias, P, which is constant and independent of the direction of measurement, is introduced by the measurement instrument itself. This type of bias, which has been observed in measurements of standard voltage cells (Eicke & Cameron) and is not eliminated by reversing the direction of the current, is shown in the following equations.

$$Y_1 = X - R + P + error_1$$
$$Y_2 = R - X + P + error_2$$

Elimination of left-right bias requires two measurements in reverse direction

The difference between the test and the reference can be estimated without bias only by taking the difference between the two measurements shown above where P cancels in the differencing so that

$$D = Y_1 - Y_2 = 2X - 2R.$$

*The value of the test item depends on the known value of the reference standard, R**

The test item, X, can then be estimated without bias by

$$\hat{Test} = X^* = \frac{1}{2}(Y_1 - Y_2) + R^*$$

and P can be estimated by

$$\hat{P} = \frac{1}{2}(Y_1 + Y_2).$$

Calibration designs that are left-right balanced

This type of scheme is called left-right balanced and the principle is extended to create a catalog of left-right balanced designs for intercomparing reference standards among themselves. These designs are appropriate *ONLY* for comparing reference standards in the same environment, or enclosure, and are not appropriate for comparing, say, across standard voltage cells in two boxes.

1. Left-right balanced design for a group of 3 artifacts
2. Left-right balanced design for a group of 4 artifacts
3. Left-right balanced design for a group of 5 artifacts
4. Left-right balanced design for a group of 6 artifacts

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.1. Elimination of special types of bias

2.3.3.1.2. Bias caused by instrument drift

Bias caused by linear drift *over the time of measurement*

The requirement that reference standards and test items be stable during the time of measurement cannot always be met because of changes in temperature caused by body heat, handling, etc.

Representation of linear drift

Linear drift for an even number of measurements is represented by

..., -5d, -3d, -1d, +1d, +3d, +5d, ...

and for an odd number of measurements by

..., -3d, -2d, -1d, 0d, +1d, +2d, +3d, ...

Assumptions for drift elimination

The effect can be mitigated by a drift-elimination scheme (Cameron/Hailes) which assumes:

- Linear drift over time
- Equally spaced measurements in time

Example of drift-elimination scheme

An example is given by substitution weighing where scale deflections on a balance are observed for X, a test weight, and R, a reference weight.

$$Y_1 = X - 3d_1 + error_1$$
$$Y_2 = R - 1d_2 + error_2$$
$$Y_3 = R + 1d_3 + error_3$$
$$Y_4 = X + 3d_4 + error_4$$

Estimates of drift-free difference and size of drift

The drift-free difference between the test and the reference is estimated by

$$D = \frac{1}{2}\{(Y_1 - Y_2) - (Y_3 - Y_4)\}$$

and the size of the drift is estimated by

$$\hat{d} = \frac{1}{4}\{-Y_1 + Y_2 - Y_3 + Y_4\}$$

Calibration designs for eliminating linear drift

This principle is extended to create a catalog of drift-elimination designs for multiple reference standards and test items. These designs are listed under calibration designs for gauge blocks because they have traditionally been used to counteract the effect of temperature build-up in the comparator during calibration.

2.3.3.2. Solutions to calibration designs

Solutions for designs listed in the catalog

Solutions for all designs that are cataloged in this Handbook are included with the designs. Solutions for other designs can be computed from the instructions on the following page given some familiarity with matrices.

Measurements for the 1,1,1 design

The use of the tables shown in the catalog are illustrated for three artifacts; namely, a reference standard with known value R^* and a check standard and a test item with unknown values. All artifacts are of the same nominal size. The design is referred to as a 1,1,1 design for

- n = 3 difference measurements
- m = 3 artifacts

Convention for showing the measurement sequence and identifying the reference and check standards

The convention for showing the measurement sequence is shown below. Nominal values are underlined in the first line showing that this design is appropriate for comparing three items of the same nominal size such as three one-kilogram weights. The reference standard is the first artifact, the check standard is the second, and the test item is the third.

	1	1	1
Y(1) =	+	-	
Y(2) =	+		-
Y(3) =		+	-
Restraint	+		
Check standard		+	

Limitation of this design

This design has degrees of freedom

$$v = n - m + 1 = 1$$

Convention for showing least-squares estimates for individual items

The table shown below lists the coefficients for finding the estimates for the individual items. The estimates are computed by taking the cross-product of the appropriate column for the item of interest with the column of measurement data and dividing by the divisor shown at the top of the table.

```
                    SOLUTION MATRIX
                      DIVISOR = 3

OBSERVATIONS      1      1      1

   Y(1)           0     -2     -1
   Y(2)           0     -1     -2
   Y(3)           0      1     -1
   R*             3      3      3
```

Solutions for individual items from the table above

For example, the solution for the reference standard is shown under the first column; for the check standard under the second column; and for the test item under the third column. Notice that the estimate for the reference standard is guaranteed to be R*, regardless of the measurement results, because of the restraint that is imposed on the design. The estimates are as follows:

$$\hat{R^*} = \frac{1}{3}(0 \cdot Y_1 + 0 \cdot Y_2 + 0 \cdot Y_3) + R^*$$

$$\hat{Chk} = \frac{1}{3}(-2 \cdot Y_1 - 1 \cdot Y_2 + 1 \cdot Y_3) + R^*$$

$$\hat{Test} = \frac{1}{3}(-1 \cdot Y_1 - 2 \cdot Y_2 - 1 \cdot Y_3) + R^*$$

Convention for showing standard deviations for individual items and combinations of items

The standard deviations are computed from two tables of factors as shown below. The standard deviations for combinations of items include appropriate covariance terms.

```
FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT  FACTOR
      K1       1   1   1
 1  0.0000    +
 1  0.8165        +
 1  0.8165            +
 2  1.4142        +   +
 1  0.8165        +

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT  FACTOR

      K2       1   1   1
 1  0.0000    +
 1  1.4142        +
 1  1.4142            +
 2  2.4495        +   +
 1  1.4142        +
```

Unifying equation

The standard deviation for each item is computed using the unifying equation:

$$s_{test} = \sqrt{K_1^2 s_1^2 + K_2^2 s_{days}^2}$$

Standard deviations for 1,1,1 design from the

For the 1,1,1 design, the standard deviations are:

$$s_{R^*} = 0$$

tables of factors

$$s_{chk} = \sqrt{(0.8165\ s_1)^2 + (1.4142\ s_{days})^2} = \sqrt{\frac{2}{3}s_1^2 + 2s_{days}^2}$$

$$s_{test} = \sqrt{(0.8165\ s_1)^2 + (1.4142\ s_{days})^2} = \sqrt{\frac{2}{3}s_1^2 + 2s_{days}^2}$$

$$s_{chk+test} = \sqrt{(1.4142\ s_1)^2 + (2.4495\ s_{days})^2} = \sqrt{2s_1^2 + 6s_{days}^2}$$

Process standard deviations must be known from historical data

In order to apply these equations, we need an estimate of the standard deviation, s_{days}, that describes day-to-day changes in the measurement process. This standard deviation is in turn derived from the level-2 standard deviation, s_2, for the check standard. This standard deviation is estimated from historical data on the check standard; it can be negligible, in which case the calculations are simplified.

The repeatability standard deviation s_1, is estimated from historical data, usually from data of several designs.

Steps in computing standard deviations

The steps in computing the standard deviation for a test item are:

- Compute the repeatability standard deviation from the design or historical data.
- Compute the standard deviation of the check standard from historical data.
- Locate the factors, K_1 and K_2 for the **check standard**; for the 1,1,1 design the factors are 0.8165 and 1.4142, respectively, where the check standard entries are last in the tables.
- Apply the unifying equation to the check standard to estimate the standard deviation for days. Notice that the standard deviation of the check standard is the same as the level-2 standard deviation, s_2, that is referred to on some pages. The equation for the between-days standard deviation from the unifying equation is

  $$s_{days} = \frac{1}{K_2}\sqrt{s_2^2 - K_1^2 s_1^2}$$

  .

  Thus, for the example above

  $$s_{days} = \frac{1}{\sqrt{2}}\sqrt{s_2^2 - \frac{2}{3}s_1^2}$$

  .
- This is the number that is entered into the NIST mass calibration software as the between-time standard deviation. If you are using this software, this is the only computation that you need to make because the standard deviations for the test items are computed automatically by the software.

- If the computation under the radical sign gives a negative number, set $s_{days}=0$. (This is possible and indicates that there is no contribution to uncertainty from day-to-day effects.)

- For completeness, the computations of the standard deviations for the test item and for the sum of the test and the check standard using the appropriate factors are shown below.

$$s_{test} = \sqrt{\frac{2}{3}s_1^2 + 2s_{days}^2} = \sqrt{\frac{2}{3}s_1^2 + 2\frac{1}{2}\left(s_2^2 - \frac{2}{3}s_1^2\right)} = s_2$$

$$s_{test+chk} = \sqrt{2s_1^2 + 6s_{days}^2} = \sqrt{2s_1^2 + 6\frac{1}{2}\left(s_2^2 - \frac{2}{3}s_1^2\right)} = \sqrt{3}s_2$$

2.3.3.2.1. General matrix solutions to calibration designs

Requirements

Solutions for all designs that are cataloged in this Handbook are included with the designs. Solutions for other designs can be computed from the instructions below given some familiarity with matrices. The matrix manipulations that are required for the calculations are:

- transposition (indicated by ')
- multiplication
- inversion

Notation

- n = number of difference measurements
- m = number of artifacts
- (n - m + 1) = degrees of freedom
- X= (nxm) design matrix
- r'= (mx1) vector identifying the restraint
- v_i' = (mx1) vector identifying ith item of interest consisting of a 1 in the ith position and zeros elsewhere
- R*= value of the reference standard
- Y= (mx1) vector of observed difference measurements

Convention for showing the measurement sequence

The convention for showing the measurement sequence is illustrated with the three measurements that make up a 1,1,1 design for 1 reference standard, 1 check standard, and 1 test item. Nominal values are underlined in the first line .

	1	1	1
Y(1) =	+	-	
Y(2) =	+		-
Y(3) =		+	-

Matrix algebra for solving a design

The (mxn) design matrix X is constructed by replacing the pluses (+), minues (-) and blanks with the entries 1, -1, and 0 respectively.

The (mxm) matrix of normal equations, $X'X$, is formed and augmented by the restraint vector to form an (m+1)x(m+1) matrix, A:

$$A = \begin{bmatrix} X'X & r' \\ r & 0 \end{bmatrix}$$

Inverse of design matrix

The A matrix is inverted and shown in the form:

$$A^{-1} = \begin{bmatrix} Q & h' \\ h & 0 \end{bmatrix}$$

where Q is an mxm matrix that, when multiplied by s^2, yields the usual variance-covariance matrix.

Estimates of values of individual artifacts

The least-squares estimates for the values of the individual artifacts are contained in the (mx1) matrix, B, where

$$B = QX'Y + h'R^*$$

where Q is the upper left element of the A^{-1} matrix shown above. The structure of the individual estimates is contained in the QX' matrix; i.e. the estimate for the ith item can be computed from XQ and Y by

- Cross multiplying the ith column of XQ with Y
- And adding R^**(nominal test)/(nominal restraint)*

Clarify with an example

We will clarify the above discussion with an example from the mass calibration process at NIST. In this example, two NIST kilograms are compared with a customer's unknown kilogram.

The design matrix, X, is

$$X = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

The first two columns represent the two NIST kilograms while the third column represents the customers kilogram (i.e., the kilogram being calibrated).

The measurements obtained, i.e., the Y matrix, are

$$Y = \begin{bmatrix} -0.3800 \\ -1.5900 \\ -1.2150 \end{bmatrix}$$

The measurements are the differences between two measurements, as specified by the design matrix, measured in grams. That is, $Y(1)$ is the difference in measurement between NIST kilogram one and NIST kilogram two, $Y(2)$ is the difference in measurement between NIST kilogram one and the customer kilogram, and $Y(3)$ is the difference in measurement between NIST kilogram two and the customer kilogram.

The value of the reference standard, R^*, is 0.82329.

Then

$$X'X = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

If there are three weights with known values for weights one and two, then

r = [1 1 0]

Thus

$$A = \begin{bmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & -1 & 1 \\ -1 & -1 & 2 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

and so

$$A^{-1} = \frac{1}{6}\begin{bmatrix} 1 & -1 & 0 & 3 \\ -1 & 1 & 0 & 3 \\ 0 & 0 & 3 & 3 \\ 3 & 3 & 3 & 0 \end{bmatrix}$$

From A^{-1}, we have

$$Q = \frac{1}{6}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

We then compute QX'

$$QX' = \frac{1}{6}\begin{bmatrix} 2 & 1 & -1 \\ -2 & -1 & 1 \\ 0 & -3 & -3 \end{bmatrix}$$

We then compute $B = QX'Y + h'R^*$

$$B = \frac{1}{6}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}\begin{bmatrix} -0.3800 \\ -1.5900 \\ -1.2150 \end{bmatrix} + 0.82329\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

This yields the following least-squares coefficient estimates:

$$B = \begin{bmatrix} 0.2225 \\ 0.6008 \\ 1.8141 \end{bmatrix}$$

Standard deviations of estimates

The standard deviation for the ith item is:

$$s_{item_i} = \sqrt{v_i Q v_i' s_1^2 + v_i D v_i' s_{days}^2}$$

where

$$D = (QX'X)(QX'X)'$$

The process standard deviation, which is a measure of the overall precision of the (NIST) mass calibrarion process,

$$s_1 = \sqrt{\frac{1}{n-m+1} Y'(I - XQX')Y}$$

is the residual standard deviation from the design, and s_{days} is the standard deviation for days, which can only be estimated from check standard measurements.

Example

We continue the example started above. Since n = 3 and m = 3, the formula reduces to:

$$s_1 = \sqrt{Y'(I - XQX')Y}$$

Substituting the values shown above for X, Y, and Q results in

$$(I - XQX') = \begin{bmatrix} 0.3333 & -0.3333 & 0.3333 \\ -0.3333 & 0.3333 & -0.3333 \\ 0.3333 & -0.3333 & 0.3333 \end{bmatrix}$$

and

$$Y'(I - XQX')Y = 0.0000083333$$

Finally, taking the square root gives

$$s_1 = 0.002887$$

The next step is to compute the standard deviation of item 3 (the customers kilogram), that is s_{item_3}. We start by substitituting the values for X and Q and computing D

$$D = (QX'X)(QX'X)' = \begin{bmatrix} 0.5000 & -0.5000 & 0.0000 \\ -0.5000 & 0.5000 & 0.0000 \\ 0.0000 & 0.0000 & 1.5000 \end{bmatrix}$$

Next, we substitute v_i' = [0 0 1] and $s_{days}^2 = 0.02111^2$ (this value is taken from a check standard and not computed from the values given in this example).

We obtain the following computations

$$v_i Q v_i' = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \frac{1}{6} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} [0 \ 0 \ 1] = 0.5$$

and

$$v_i D v_i' = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5000 & -0.5000 & 0.0000 \\ -0.5000 & 0.5000 & 0.0000 \\ 0.0000 & 0.0000 & 1.5000 \end{bmatrix} [0 \ 0 \ 1] = 1.5$$

and

$$s_{item_i} = \sqrt{v_i Q v_i' s_1^2 + v_i D v_i' s_{days}^2}$$

$$s_{item_3} = \sqrt{0.5(0.002887)^2 + 1.5(0.0211)^2} = 0.02593$$

2.3.3.3. Uncertainties of calibrated values

Uncertainty analysis follows the ISO principles

This section discusses the calculation of uncertainties of calibrated values from calibration designs. The discussion follows the guidelines in the section on classifying and combining components of uncertainty. Two types of evaluations are covered.

1. type A evaluations of time-dependent sources of random error
2. type B evaluations of other sources of error

The latter includes, but is not limited to, uncertainties from sources that are not replicated in the calibration design such as uncertainties of values assigned to reference standards.

Uncertainties for test items

Uncertainties associated with calibrated values for test items from designs require calculations that are specific to the individual designs. The steps involved are outlined below.

Outline for the section on uncertainty analysis

- Historical perspective
- Assumptions
- Example of more realistic model
- Computation of repeatability standard deviations
- Computation of level-2 standard deviations
- Combination of repeatability and level-2 standard deviations
- Example of computations for 1,1,1,1 design
- Type B uncertainty associated with the restraint
- Expanded uncertainty of calibrated values

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.3. Uncertainties of calibrated values

2.3.3.3.1. Type A evaluations for calibration designs

Change over time

Type A evaluations for calibration processes must take into account changes in the measurement process that occur over time.

Historically, uncertainties considered only instrument imprecision

Historically, computations of uncertainties for calibrated values have treated the precision of the comparator instrument as the primary source of random uncertainty in the result. However, as the precision of instrumentation has improved, effects of other sources of variability have begun to show themselves in measurement processes. This is not universally true, but for many processes, instrument imprecision (short-term variability) cannot explain all the variation in the process.

Effects of environmental changes

Effects of humidity, temperature, and other environmental conditions which cannot be closely controlled or corrected must be considered. These tend to exhibit themselves over time, say, as between-day effects. The discussion of between-day (level-2) effects relating to gauge studies carries over to the calibration setting, but the computations are not as straightforward.

Assumptions which are specific to this section

The computations in this section depend on specific assumptions:

1. Short-term effects associated with instrument response
   - come from a single distribution
   - vary randomly from measurement to measurement within a design.
2. Day-to-day effects
   - come from a single distribution
   - vary from artifact to artifact but remain constant for a single calibration
   - vary from calibration to calibration

These

These assumptions have proved useful for characterizing

assumptions have proved useful but may need to be expanded in the future

high precision measurement processes, but more complicated models may eventually be needed which take the relative magnitudes of the test items into account. For example, in mass calibration, a 100 g weight can be compared with a summation of 50g, 30g and 20 g weights in a single measurement. A sophisticated model might consider the size of the effect as relative to the nominal masses or volumes.

Example of the two models for a design for calibrating test item using 1 reference standard

To contrast the simple model with the more complicated model, a measurement of the difference between X, the test item, with unknown and yet to be determined value, X^*, and a reference standard, R, with known value, R^*, and the reverse measurement are shown below.

Model (1) takes into account only instrument imprecision so that:

(1)

$$Y_1 = X - R + error_1$$
$$Y_2 = R - X + error_2$$

with the error terms random errors that come from the imprecision of the measuring instrument.

Model (2) allows for both instrument imprecision and level-2 effects such that:

(2)

$$Y_1 = (X + \Delta_X) - (R + \Delta_R) + error_1$$
$$Y_2 = (R + \Delta_R) - (X + \Delta_X) + error_2$$

where the delta terms explain small changes in the values of the artifacts that occur over time. For both models, the value of the test item is estimated as

$$\hat{Test} = X^* = \frac{1}{2}(Y_1 - Y_2) + R^*$$

Standard deviations from both models

For model (l), the standard deviation of the test item is

$$s_{test} = \frac{s_1}{\sqrt{2}}$$

For model (2), the standard deviation of the test item is

$$s_{test} = \sqrt{\frac{s_1^2}{2} + \frac{s_2^2}{2}}$$

.

Note on relative contributions of both components to uncertainty

In both cases, s_1 is the repeatability standard deviation that describes the precision of the instrument and s_2 is the level-2 standard deviation that describes day-to-day changes. One thing to notice in the standard deviation for the test item is the contribution of s_2 relative to the total uncertainty. If s_2 is large relative to s_1, or dominates, the uncertainty will not be appreciably reduced by adding measurements to the calibration design.

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.3. Uncertainties of calibrated values

2.3.3.3.2. Repeatability and level-2 standard deviations

Repeatability standard deviation comes from the data of a single design

The repeatability standard deviation of the instrument can be computed in two ways.

1. It can be computed as the residual standard deviation from the design and should be available as output from any software package that reduces data from calibration designs. The matrix equations for this computation are shown in the section on solutions to calibration designs. The standard deviation has degrees of freedom

$$v = n - m + 1$$

for n difference measurements and m items. Typically the degrees of freedom are very small. For two differences measurements on a reference standard and test item, the degrees of freedom is $v=1$.

A more reliable estimate comes from pooling over historical data

2. A more reliable estimate of the standard deviation can be computed by pooling variances from K calibrations (and then taking its square root) using the same instrument (assuming the instrument is in statistical control). The formula for the pooled estimate is

$$s_1 = \sqrt{\frac{1}{\sum_k \nu_k} \sum_k \nu_k s_k^2}$$

Level-2 standard deviation is estimated from check standard measurements

The level-2 standard deviation cannot be estimated from the data of the calibration design. It cannot generally be estimated from repeated designs involving the test items. The best mechanism for capturing the day-to-day effects is a check standard, which is treated as a test item and included in each calibration design. Values of the check standard, estimated over time from the calibration design,

are used to estimate the standard deviation.

Assumptions

The check standard value must be stable over time, and the measurements must be in statistical control for this procedure to be valid. For this purpose, it is necessary to keep a historical record of values for a given check standard, and these values should be kept by instrument and by design.

Computation of level-2 standard deviation

Given K historical check standard values,

$$C_1, C_2, \cdots, C_K$$

the standard deviation of the check standard values is computed as

$$s_c = s_2 = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(C_k - \overline{C_.}\right)^2}$$

where

$$\overline{C_.} = \frac{1}{K}\sum_{k=1}^{K} C_k$$

with degrees of freedom $v = K - 1$.

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.3. Uncertainties of calibrated values

2.3.3.3.3. Combination of repeatability and level-2 standard deviations

Standard deviation of test item depends on several factors

The final question is how to combine the repeatability standard deviation and the standard deviation of the check standard to estimate the standard deviation of the test item. This computation depends on:

- structure of the design
- position of the check standard in the design
- position of the reference standards in the design
- position of the test item in the design

Derivations require matrix algebra

Tables for estimating standard deviations for all test items are reported along with the solutions for all designs in the catalog. The use of the tables for estimating the standard deviations for test items is illustrated for the 1,1,1,1 design. Matrix equations can be used for deriving estimates for designs that are not in the catalog.

The check standard for each design is either an additional test item in the design, other than the test items that are submitted for calibration, or it is a construction, such as the difference between two reference standards as estimated by the design.

2.3.3.3.4. Calculation of standard deviations for 1,1,1,1 design

Design with 2 reference standards and 2 test items

An example is shown below for a 1,1,1,1 design for two reference standards, R_1 and R_2, and two test items, X_1 and X_2, and six difference measurements. The restraint, R^*, is the sum of values of the two reference standards, and the check standard, which is independent of the restraint, is the difference between the values of the reference standards. The design and its solution are reproduced below.

Check standard is the difference between the 2 reference standards

OBSERVATIONS	1	1	1	1
Y(1)	+	-		
Y(2)	+		-	
Y(3)	+			-
Y(4)		+	-	
Y(5)		+		-
Y(6)			+	-
RESTRAINT	+	+		
CHECK STANDARD	+	-		

DEGREES OF FREEDOM = 3

SOLUTION MATRIX
DIVISOR = 8

OBSERVATIONS	1	1	1	1
Y(1)	2	-2	0	0
Y(2)	1	-1	-3	-1
Y(3)	1	-1	-1	-3
Y(4)	-1	1	-3	-1
Y(5)	-1	1	-1	-3
Y(6)	0	0	2	-2
R*	4	4	4	4

Explanation of solution matrix

The solution matrix gives values for the test items of

$$X_1^* = \frac{1}{8}(-3Y_2 - Y_3 - 3Y_4 - Y_5 + 2Y_6) + \frac{1}{2}R^*$$

$$X_2^* = \frac{1}{8}(-Y_2 - 3Y_3 - Y_4 - 3Y_5 - 2Y_6) + \frac{1}{2}R^*$$

Factors for computing contributions of repeatability and level-2 standard deviations to uncertainty

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR K_1	1	1	1	1
1	0.3536	+			
1	0.3536		+		
1	0.6124			+	
1	0.6124				+
0	0.7071	+	-		

FACTORS FOR LEVEL-2 STANDARD DEVIATIONS

WT	FACTOR K_2	1	1	1	1
1	0.7071	+			
1	0.7071		+		
1	1.2247			+	
1	1.2247				+
0	1.4141	+	-		

The first table shows factors for computing the contribution of the repeatability standard deviation to the total uncertainty. The second table shows factors for computing the contribution of the between-day standard deviation to the uncertainty. Notice that the check standard is the last entry in each table.

Unifying equation

The unifying equation is:

$$s_{test} = \sqrt{K_1^2 s_1^2 + K_2^2 s_{days}^2}$$

Standard deviations are computed using the factors from the tables with the unifying equation

The steps in computing the standard deviation for a test item are:

- Compute the repeatability standard deviation from historical data.
- Compute the standard deviation of the check standard from historical data.
- Locate the factors, K_1 and K_2, for the check standard.
- Compute the between-day variance (using the unifying equation for the check standard). For this example,

$$s_2^2 = \left\{ \left(\sqrt{.5}s_1\right)^2 + \left(\sqrt{2}s_{days}\right)^2 \right\} = \left\{ \frac{1}{2}s_1^2 + 2s_{days}^2 \right\}$$

implies

$$\Rightarrow$$

$$s_{days}^2 = \frac{1}{2}\left\{ s_2^2 - \frac{1}{2}s_1^2 \right\}$$

.

- If this variance estimate is negative, set s_{days} = 0. (This is possible and indicates that there is no contribution to uncertainty from day-to-day effects.)
- Locate the factors, K_1 and K_2, for the test items, and compute the standard deviations using the unifying equation. For this example,

$$s_{x_1} = \sqrt{\frac{3}{8}s_1^2 + \frac{3}{2}s_{days}^2} = \sqrt{\frac{3}{8}s_1^2 + \frac{3}{2}\frac{1}{2}\{s_2^2 - \frac{1}{2}s_1^2\}} = \sqrt{\frac{3}{4}s_2^2}$$

and

$$s_{x_2} = s_{x_1}$$

2. Measurement Process Characterization
2.3. Calibration
2.3.3. What are calibration designs?
2.3.3.3. Uncertainties of calibrated values

2.3.3.3.5. Type B uncertainty

Type B uncertainty associated with the restraint

The reference standard is assumed to have known value, R^*, for the purpose of solving the calibration design. For the purpose of computing a standard uncertainty, it has a type B uncertainty that contributes to the uncertainty of the test item.

The value of R^* comes from a higher-level calibration laboratory or process, and its value is usually reported along with its uncertainty, U. If the laboratory also reports the k factor for computing U, then the standard deviation of the restraint is

$$s_{R^*} = \frac{U}{k}$$

If k is not reported, then a conservative way of proceeding is to assume $k = 2$.

Situation where the test is different in size from the reference

Usually, a reference standard and test item are of the same nominal size and the calibration relies on measuring the small difference between the two; for example, the intercomparison of a reference kilogram compared with a test kilogram. The calibration may also consist of an intercomparison of the reference with a summation of artifacts where the summation is of the same nominal size as the reference; for example, a reference kilogram compared with 500 g + 300 g + 200 g test weights.

Type B uncertainty for the test artifact

The type B uncertainty that accrues to the test artifact from the uncertainty of the reference standard is proportional to their nominal sizes; i.e.,

$$s_B = \frac{Nominal\ Test}{Nominal\ Restraint} s_{R^*}$$

2.3.3.3.6. Expanded uncertainties

Standard uncertainty

The standard uncertainty for the test item is

$$u = \sqrt{s_{test}^2 + \left(\frac{Nominal\ Test}{Nominal\ Restraint}\right)^2 s_{R^*}^2}$$

Expanded uncertainty

The expanded uncertainty is computed as

$$U = ku$$

where k is either the critical value from the t table for degrees of freedom v or k is set equal to 2.

Problem of the degrees of freedom

The calculation of degrees of freedom, v, can be a problem. Sometimes it can be computed using the Welch-Satterthwaite approximation and the structure of the uncertainty of the test item. Degrees of freedom for the standard deviation of the restraint is assumed to be infinite. The coefficients in the Welch-Satterthwaite formula must all be positive for the approximation to be reliable.

Standard deviation for test item from the 1,1,1,1 design

For the 1,1,1,1 design, the standard deviation of the test items can be rewritten by substituting in the equation

$$s_{X_1} = s_{X_2} = \sqrt{\frac{3}{8}s_1^2 + \frac{3}{2}s_{days}^2} = \sqrt{\frac{3}{8}s_1^2 + \frac{3}{2}\left\{\frac{1}{2}s_2^2 - \frac{1}{4}s_1^2\right\}} = \frac{\sqrt{3}}{2}s_2$$

so that the degrees of freedom depends only on the degrees of freedom in the standard deviation of the check standard. This device may not work satisfactorily for all designs.

Standard uncertainty from the 1,1,1,1 design

To complete the calculation shown in the equation at the top of the page, the nominal value of the test item (which is equal to 1) is divided by the nominal value of the restraint (which is also equal to 1), and the result is squared. Thus, the standard uncertainty is

$$u = \sqrt{\frac{3}{4}s_2^2 + s_{R^*}^2}$$

Degrees of

Therefore, the degrees of freedom is approximated as

freedom using the Welch-Satterthwaite approximation

$$\nu = \frac{u^4}{\frac{\frac{9}{16}s_2^4}{n-1}}$$

where *n - 1* is the degrees of freedom associated with the check standard uncertainty. Notice that the standard deviation of the restraint drops out of the calculation because of an infinite degrees of freedom.

2.3.4. Catalog of calibration designs

Important concept - Restraint

The designs are constructed for measuring differences among reference standards and test items, singly or in combinations. Values for individual standards and test items can be computed from the design only if the value (called the *restraint* = R^*) of one or more reference standards is known. The methodology for constructing and solving calibration designs is described briefly in matrix solutions and in more detail in a NIST publication. (Cameron et al.).

Designs listed in this catalog

Designs are listed by traditional subject area although many of the designs are appropriate generally for intercomparisons of artifact standards.

- Designs for mass weights
- Drift-eliminating designs for gage blocks
- Left-right balanced designs for electrical standards
- Designs for roundness standards
- Designs for angle blocks
- Drift-eliminating design for thermometers in a bath
- Drift-eliminating designs for humidity cylinders

Properties of designs in this catalog

Basic requirements are:

1. The differences must be nominally zero.
2. The design must be solvable for individual items given the restraint.

Other desirable properties are:

1. The number of measurements should be small.
2. The degrees of freedom should be greater than zero.
3. The standard deviations of the estimates for the test items should be small enough for their intended purpose.

Information:

Given

Design

- n = number of difference measurements
- m = number of artifacts (reference standards + test items) to be calibrated

Solution

Factors for computing standard deviations

the following information is shown for each design:

- Design matrix -- (n x m)
- Vector that identifies standards in the restraint -- (1 x m)
- Degrees of freedom = (n - m + 1)
- Solution matrix for given restraint -- (n x m)
- Table of factors for computing standard deviations

Convention for showing the measurement sequence

Nominal sizes of standards and test items are shown at the top of the design. Pluses (+) indicate items that are measured together; and minuses (-) indicate items are not measured together. The difference measurements are constructed from the design of pluses and minuses. For example, a 1,1,1 design for one reference standard and two test items of the same nominal size with three measurements is shown below:

	1	1	1
Y(1) =	+	-	
Y(2) =	+		-
Y(3) =	+	-	

Solution matrix

Example and interpretation

The cross-product of the column of difference measurements and R^* with a column from the solution matrix, divided by the named divisor, gives the value for an individual item. For example,

Solution matrix
Divisor = 3

	1	1	1
Y(1)	0	-2	-1
Y(2)	0	-1	-2
Y(3)	0	+1	-1
R*	+3	+3	+3

implies that estimates for the restraint and the two test items are:

$$\hat{R^*} = \frac{1}{3}\{0 \cdot Y_1 + 0 \cdot Y_2 + 0 \cdot Y_3 + 3R^*\} = R^*$$

$$\hat{Test_1} = \frac{1}{3}\{-2Y_1 - Y_2 + Y_3 + 3R^*\}$$

$$\hat{Test_2} = \frac{1}{3}\{-Y_1 - 2Y_2 - Y_3 + 3R^*\}$$

Interpretation of table of factors

The factors in this table provide information on precision. The repeatability standard deviation, s_1, is multiplied by the appropriate factor to obtain the standard deviation for an

individual item or combination of items. For example,

```
          Sum     Factor      1       1
1
          1       0.0000      +
          1       0.8166              +
          1       0.8166
+
          2       1.4142               +
+
```

implies that the standard deviations for the estimates are:

$$s_{R^*} = 0$$

$$s_{test_1} = 0.8661 \cdot s_1$$

$$s_{test_2} = 0.8661 \cdot s_1$$

$$s_{test_{1+2}} = 1.4142 \cdot s_1$$

2.3.4.1. Mass weights

Tie to kilogram reference standards

Near-accurate mass measurements require a sequence of designs that relate the masses of individual weights to a reference kilogram(s) standard (Jaeger & Davis). Weights generally come in sets, and an entire set may require several series to calibrate all the weights in the set.

Example of weight set

A 5,3,2,1 weight set would have the following weights:

1000 g

500g, 300g, 200g, 100g

50g, 30g 20g, 10g

5g, 3g, 2g, 1g

0.5g, 0.3g, 0.2g, 0.1g

Depiction of a design with three series for calibrating a 5,3,2,1 weight set with weights between 1 kg and 10 g

First series using

The calibrations start with a comparison of the one kilogram test weight with the reference kilograms (see the graphic above). The 1,1,1,1 design requires two kilogram reference

1,1,1,1 design

standards with known values, *R1** and *R2**. The fourth kilogram in this design is actually a summation of the 500, 300, 200 g weights which becomes the restraint in the next series.

The restraint for the first series is the known average mass of the reference kilograms,

$$R^{\bullet} = \frac{R_1^{\bullet} + R_2^{\bullet}}{2}$$

The design assigns values to all weights including the individual reference standards. For this design, the check standard is not an artifact standard but is defined as the difference between the values assigned to the reference kilograms by the design; namely,

$$C = \left(\hat{R}_1^{*} - \hat{R}_2^{*} \right)$$

2nd series using 5,3,2,1,1,1 design

The second series is a 5,3,2,1,1,1 design where the restraint over the 500g, 300g and 200g weights comes from the value assigned to the summation in the first series; i.e.,

$$R^{\bullet} = \sum 1 = \hat{X}_{500} + \hat{X}_{300} + \hat{X}_{200}$$

The weights assigned values by this series are:

- 500g, 300g, 200 g and 100g test weights
- 100 g check standard (2nd 100g weight in the design)
- Summation of the 50g, 30g, 20g weights.

Other starting points

The calibration sequence can also start with a 1,1,1 design. This design has the disadvantage that it does **not** have provision for a check standard.

Better choice of design

A better choice is a 1,1,1,1,1 design which allows for two reference kilograms and a kilogram check standard which occupies the 4th position among the weights. This is preferable to the 1,1,1,1 design but has the disadvantage of requiring the laboratory to maintain three kilogram standards.

Important detail

The solutions are only applicable for the restraints as shown.

Designs for decreasing weight sets

1. 1,1,1 design
2. 1,1,1,1 design
3. 1,1,1,1,1 design
4. 1,1,1,1,1,1 design
5. 2,1,1,1 design
6. 2,2,1,1,1 design

7. 2,2,2,1,1 design
8. 5,2,2,1,1,1 design
9. 5,2,2,1,1,1,1 design
10. 5,3,2,1,1,1 design
11. 5,3,2,1,1,1,1 design
12. 5,3,2,2,1,1,1 design
13. 5,4,4,3,2,2,1,1 design
14. 5,5,2,2,1,1,1,1 design
15. 5,5,3,2,1,1,1 design
16. 1,1,1,1,1,1,1,1 design
17. 3,2,1,1,1 design

Design for pound weights

1. 1,2,2,1,1 design

Designs for increasing weight sets

1. 1,1,1 design
2. 1,1,1,1 design
3. 5,3,2,1,1 design
4. 5,3,2,1,1,1 design
5. 5,2,2,1,1,1 design
6. 3,2,1,1,1 design

2.3.4.1.1. Design for 1,1,1

Design 1,1,1

```
OBSERVATIONS    1    1    1

     Y(1)       +    -
     Y(2)       +         -
     Y(3)            +    -


RESTRAINT       +

CHECK STANDARD      +

DEGREES OF FREEDOM   =   1

                    SOLUTION MATRIX
                      DIVISOR   =   3

OBSERVATIONS       1       1       1

   Y(1)            0      -2      -1
   Y(2)            0      -1      -2
   Y(3)            0       1      -1
   R*              3       3       3

   R* = value of reference weight


FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
WT   FACTOR
                   1    1    1
 1   0.0000        +
 1   0.8165             +
 1   0.8165                  +
 2   1.4142             +    +
 1   0.8165             +

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
WT   FACTOR
                   1    1    1
 1   0.0000        +
 1   1.4142             +
 1   1.4142                  +
 2   2.4495             +    +
 1   1.4142             +
```

Explanation of notation and interpretation of tables

2.3.4.1.2. Design for 1,1,1,1

```
Design 1,1,1,1

 OBSERVATIONS   1
1   1   1


     Y(1)       +
-
     Y(2)       +
-
     Y(3)       +
-
     Y(4)
+   -
     Y(5)
+       -
     Y(6)
+   -

 RESTRAINT      +
+

 CHECK STANDARD +
-

 DEGREES OF
FREEDOM  =  3


SOLUTION MATRIX

DIVISOR  =  8

 OBSERVATIONS
1      1      1
1


    Y(1)
2     -2      0
0
    Y(2)
1     -1     -3
-1
    Y(3)
1     -1     -1
-3
    Y(4)
-1     1     -3
-1
    Y(5)
-1     1     -1
-3
    Y(6)
0      0      2
-2
    R*
4      4      4
```

```
4

    R* = sum of
two reference
standards


 FACTORS FOR
REPEATABILITY
STANDARD
DEVIATIONS
 WT  FACTOR
        K1      1
1   1   1
  1  0.3536    +
  1  0.3536
+
  1  0.6124
+
  1  0.6124
+
  0  0.7071    +
-


 FACTORS FOR
BETWEEN-DAY
STANDARD
DEVIATIONS
 WT  FACTOR
        K2      1
1   1   1
  1  0.7071    +
  1  0.7071
+
  1  1.2247
+
  1  1.2247
+
  0  1.4141    +
-
```

Explanation of notation and interpretation of tables

2.3.4.1.3. Design for 1,1,1,1,1

CASE 1: CHECK STANDARD = DIFFERENCE BETWEEN FIRST TWO WEIGHTS

```
 OBSERVATIONS    1    1    1    1
1

      Y(1)        +    -
      Y(2)        +         -
      Y(3)        +              -
      Y(4)        +
-
      Y(5)             +    -
      Y(6)             +         -
      Y(7)             +
-
      Y(8)                  +    -
      Y(9)                  +
-
      Y(10)                       +
-

 RESTRAINT       +    +
 CHECK STANDARD  +    -
 DEGREES OF FREEDOM   =    6
```

```
SOLUTION MATRIX
DIVISOR =  10

 OBSERVATIONS       1       1
1       1       1

    Y(1)             2      -2
0       0       0
    Y(2)             1      -1
-3      -1      -1
    Y(3)             1      -1
-1      -3      -1
    Y(4)             1      -1
-1      -1      -3
    Y(5)            -1       1
-3      -1      -1
    Y(6)            -1       1
-1      -3      -1
    Y(7)            -1       1
-1      -1      -3
    Y(8)             0       0
2      -2       0
    Y(9)             0       0
```

CASE 2: CHECK STANDARD = FOURTH WEIGHT

```
 OBSERVATIONS    1    1    1    1
1

      Y(1)        +    -
      Y(2)        +         -
      Y(3)        +              -
      Y(4)        +
-
      Y(5)             +    -
      Y(6)             +         -
      Y(7)             +
-
      Y(8)                  +    -
      Y(9)                  +
-
      Y(10)                       +
-

 RESTRAINT       +    +
 CHECK STANDARD                  +
 DEGREES OF FREEDOM   =    6
```

```
SOLUTION MATRIX
DIVISOR =  10

 OBSERVATIONS       1       1
1       1       1

    Y(1)             2      -2
0       0       0
    Y(2)             1      -1
-3      -1      -1
    Y(3)             1      -1
-1      -3      -1
    Y(4)             1      -1
-1      -1      -3
    Y(5)            -1       1
-3      -1      -1
    Y(6)            -1       1
-1      -3      -1
    Y(7)            -1       1
-1      -1      -3
    Y(8)             0       0
2      -2       0
    Y(9)             0       0
2       0      -2
```

```
2       0      -2
    Y(10)          0       0
0       2      -2
    R*             5       5
5       5       5

    R* = sum of two reference
standards
```

```
 FACTORS FOR REPEATABILITY
STANDARD DEVIATIONS
 WT  FACTOR
        K1      1   1   1   1
1
  1  0.3162    +
  1  0.3162        +
  1  0.5477            +
  1  0.5477                +
  1  0.5477
+
  2  0.8944                +
+
  3  1.2247            +   +
+
  0  0.6325    +   -
```

```
 FACTORS FOR BETWEEN-DAY
STANDARD DEVIATIONS
 WT  FACTOR
        K2      1   1   1   1
1
  1  0.7071    +
  1  0.7071        +
  1  1.2247            +
  1  1.2247                +
  1  1.2247
+
  2  2.0000                +
+
  3  2.7386            +   +
+
  0  1.4142     +   -
```

```
    Y(10)          0       0
0       2      -2
    R*             5       5
5       5       5

    R* = sum of two reference
standards
```

```
 FACTORS FOR REPEATABILITY
STANDARD DEVIATIONS
 WT  FACTOR
        K1      1   1   1   1
1
  1  0.3162    +
  1  0.3162        +
  1  0.5477            +
  1  0.5477                +
  1  0.5477
+
  2  0.8944                +
+
  3  1.2247            +   +
+
  1  0.5477                +
```

```
 FACTORS FOR BETWEEN-DAY
STANDARD DEVIATIONS
 WT  FACTOR
        K2      1   1   1   1
1
  1  0.7071    +
  1  0.7071        +
  1  1.2247            +
  1  1.2247                +
  1  1.2247
+
  2  2.0000                +
+
  3  2.7386            +   +
+
  1  1.2247                +
```

Explanation of notation and interpretation of tables

2.3.4.1.4. Design for 1,1,1,1,1,1

Design 1,1,1,1,1,1

OBSERVATIONS	1	1	1	1	1	1
X(1)	+	-				
X(2)	+		-			
X(3)	+			-		
X(4)	+				-	
X(5)	+					-
X(6)		+	-			
X(7)		+		-		
X(8)		+			-	
X(9)		+				-
X(10)			+	-		
X(11)			+		-	
X(12)			+			-
X(13)				+	-	
X(14)				+		-
X(15)					+	-
RESTRAINT	+	+				
CHECK STANDARD					+	

DEGREES OF FREEDOM = 10

SOLUTION MATRIX
DIVISOR = 8

OBSERVATIONS	1	1	1	1	1	1
Y(1)	1	-1	0	0	0	0
Y(2)	1	0	-1	0	0	0
Y(3)	1	0	0	-1	0	0
Y(4)	1	0	0	0	-1	0
Y(5)	2	1	1	1	1	0
Y(6)	0	1	-1	0	0	0
Y(7)	0	1	0	-1	0	0
Y(8)	0	1	0	0	-1	0
Y(9)	1	2	1	1	1	0
Y(10)	0	0	1	-1	0	0
Y(11)	0	0	1	0	-1	0
Y(12)	1	1	2	1	1	0
Y(13)	0	0	0	1	-1	0
Y(14)	1	1	1	2	1	0
Y(15)	1	1	1	1	2	0
R*	6	6	6	6	6	6

R* = sum of two reference standards

FACTORS FOR COMPUTING REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	1	1	1	1	1	1
1	0.2887	+					

```
 1  0.2887        +
 1  0.5000            +
 1  0.5000                +
 1  0.5000                    +
 1  0.5000                        +
 2  0.8165            +   +
 3  1.1180            +   +   +
 4  1.4142            +   +   +   +
 1  0.5000                    +


FACTORS FOR COMPUTING BETWEEN-DAY STANDARD DEVIATIONS
WT  FACTOR
                1   1   1   1   1   1
 1  0.7071     +
 1  0.7071         +
 1  1.2247             +
 1  1.2247                 +
 1  1.2247                     +
 1  1.2247                         +
 2  2.0000             +   +
 3  2.7386             +   +   +
 4  3.4641             +   +   +   +
 1  1.2247                     +
```

Explanation of notation and interpretation of tables

2.3.4.1.5. Design for 2,1,1,1

```
Design 2,1,1,1

 OBSERVATIONS   2   1   1   1

     Y(1)       +   -       -
     Y(2)       +       -   -
     Y(3)       +   -   -
     Y(4)           +       -
     Y(5)           +   -
     Y(6)               +   -

 RESTRAINT      +

 CHECK STANDARD         +

DEGREES OF FREEDOM  =      3


                     SOLUTION MATRIX
                      DIVISOR  =  4

 OBSERVATIONS      2      1      1      1

    Y(1)           0     -1      0     -1
    Y(2)           0      0     -1     -1
    Y(3)           0     -1     -1      0
    Y(4)           0      1      0     -1
    Y(5)           0      1     -1      0
    Y(6)           0      0      1     -1
    R*             4      2      2      2

    R* = value of the reference standard


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 WT  FACTOR
                2   1   1   1
  2  0.0000     +
  1  0.5000         +
  1  0.5000             +
  1  0.5000                 +
  2  0.7071         +   +
  3  0.8660         +   +   +
  1  0.5000             +


 FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
 WT  FACTOR
                2   1   1   1
  2  0.0000     +
  1  1.1180         +
  1  1.1180             +
  1  1.1180                 +
  2  1.7321         +   +
```

```
  3  2.2913        +    +    +
  1  1.1180             +
```

[Explanation of notation and interpretation of tables]

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.1. Mass weights

2.3.4.1.6. Design for 2,2,1,1,1

Design 2,2,1,1,1

OBSERVATIONS	2	2	1	1	1
Y(1)	+	-	-	+	
Y(2)	+	-		-	+
Y(3)	+	-	+		-
Y(4)	+	-			
Y(5)	+		-	-	
Y(6)	+		-		-
Y(7)	+			-	-
Y(8)		+	-	-	
Y(9)		+	-		-
Y(10)		+		-	-
RESTRAINT	+	+	+		
CHECK STANDARD				+	

DEGREES OF FREEDOM = 6

SOLUTION MATRIX
DIVISOR = 275

OBSERVATIONS	2	2	1	1	1
Y(1)	47	-3	-44	66	11
Y(2)	25	-25	0	-55	55
Y(3)	3	-47	44	-11	-66
Y(4)	25	-25	0	0	0
Y(5)	29	4	-33	-33	22
Y(6)	29	4	-33	22	-33
Y(7)	7	-18	11	-44	-44
Y(8)	4	29	-33	-33	22
Y(9)	4	29	-33	22	-33
Y(10)	-18	7	11	-44	-44
R*	110	110	55	55	55

R* = sum of three reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	2	2	1	1	1
2	0.2710	+				
2	0.2710		+			
1	0.3347			+		
1	0.4382				+	
1	0.4382					+
2	0.6066				+	+
3	0.5367			+	+	+
1	0.4382				+	

```
FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
WT   FACTOR
                2   2   1   1   1
 2   0.8246     +
 2   0.8246         +
 1   0.8485             +
 1   1.0583                 +
 1   1.0583                     +
 2   1.5748                 +   +
 3   1.6971             +   +   +
 1   1.0583                 +
```

Explanation of notation and interpretation of tables

2.3.4.1.7. Design for 2,2,2,1,1

Design 2,2,2,1,1

OBSERVATIONS	2	2	2	1	1
Y(1)	+	-			
Y(2)	+		-		
Y(3)		+	-		
Y(4)	+			-	-
Y(5)		+		-	-
Y(6)			+	-	-
Y(7)				+	-
RESTRAINT	+	+			
CHECK STANDARD			+		

DEGREES OF FREEDOM = 3

SOLUTION MATRIX
DIVISOR = 16

OBSERVATIONS	2	2	2	1	1
Y(1)	4	-4	0	0	0
Y(2)	2	-2	-6	-1	-1
Y(3)	-2	2	-6	-1	-1
Y(4)	2	-2	-2	-3	-3
Y(5)	-2	2	-2	-3	-3
Y(6)	0	0	4	-2	-2
Y(7)	0	0	0	8	-8
R*	8	8	8	4	4

R* = sum of the two reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	2	2	2	1	1
2	0.3536	+				
2	0.3536		+			
2	0.6124			+		
1	0.5863				+	
1	0.5863					+
2	0.6124				+	+
4	1.0000			+	+	+
1	0.5863				+	

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT	FACTOR	2	2	2	1	1
2	0.7071	+				
2	0.7071		+			

```
2  1.2247            +
1  1.0607               +
1  1.0607                    +
2  1.5811                 +  +
4  2.2361            +    +  +
1  1.0607                 +
```

Explanation of notation and interpretation of tables

2.3.4.1.8. Design for 5,2,2,1,1,1

Design 5,2,2,1,1,1

OBSERVATIONS	5	2	2	1	1	1
Y(1)	+	-	-	-	-	+
Y(2)	+	-	-	-	+	-
Y(3)	+	-	-	+	-	-
Y(4)	+	-		-	-	-
Y(5)	+		-	-	-	-
Y(6)		+	-	+	-	
Y(7)		+	-	-		+
Y(8)		+	-		+	-
RESTRAINT	+	+	+	+		
CHECK STANDARD					+	

DEGREES OF FREEDOM = 3

SOLUTION MATRIX
DIVISOR = 70

OBSERVATIONS	5	2	2	1	1	1
Y(1)	15	-8	-8	1	1	21
Y(2)	15	-8	-8	1	21	1
Y(3)	5	-12	-12	19	-1	-1
Y(4)	0	2	12	-14	-14	-14
Y(5)	0	12	2	-14	-14	-14
Y(6)	-5	8	-12	9	-11	-1
Y(7)	5	12	-8	-9	1	11
Y(8)	0	10	-10	0	10	-10
R*	35	14	14	7	7	7

R* = sum of the four reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	2	2	1	1	1
5	0.3273	+					
2	0.3854		+				
2	0.3854			+			
1	0.4326				+		
1	0.4645					+	
1	0.4645						+
1	0.4645					+	

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT	FACTOR	5	2	2	1	1	1
5	1.0000	+					

```
2  0.8718        +
2  0.8718            +
1  0.9165                +
1  1.0198                    +
1  1.0198                        +
1  1.0198                    +
```

Explanation of notation and interpretation of tables

2.3.4.1.9. Design for 5,2,2,1,1,1,1

Design 5,2,2,1,1,1,1

OBSERVATIONS	5	2	2	1	1	1	1
Y(1)	+	-	-	-			
Y(2)	+	-	-		-		
Y(3)	+	-	-			-	
Y(4)	+	-	-				-
Y(5)		+		+	-	-	-
Y(6)			+	+	-	-	-
Y(7)		+	+	-	-	-	-
Y(8)					+	-	
Y(9)					+		-
Y(10)						+	-
RESTRAINT	+	+	+	+			
CHECK STANDARD						+	

DEGREES OF FREEDOM = 4

SOLUTION MATRIX
DIVISOR = 60

OBSERVATIONS	5	2	2	1	1	1	1
Y(1)	12	0	0	-12	0	0	0
Y(2)	6	-4	-4	2	-12	3	3
Y(3)	6	-4	-4	2	3	-12	3
Y(4)	6	-4	-4	2	3	3	-12
Y(5)	-6	28	-32	10	-6	-6	-6
Y(6)	-6	-32	28	10	-6	-6	-6
Y(7)	6	8	8	-22	-6	-6	-6
Y(8)	0	0	0	0	15	-15	0
Y(9)	0	0	0	0	15	0	-15
Y(10)	0	0	0	0	0	15	-15
R*	30	12	12	6	6	6	6

R* = sum of the four reference standards

```
FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
WT  FACTOR
                5   2   2   1   1   1   1
 5  0.3162      +
 2  0.7303          +
 2  0.7303              +
 1  0.4830                  +
 1  0.4472                      +
 1  0.4472                          +
 1  0.4472                              +
 2  0.5477                      +   +
 3  0.5477                      +   +   +
 1  0.4472                          +
```

```
FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
WT  FACTOR
                5   2   2   1   1   1   1
 5  1.0000      +
 2  0.8718          +
 2  0.8718              +
 1  0.9165                  +
 1  1.0198                      +
 1  1.0198                          +
 1  1.0198                              +
 2  1.4697                      +   +
 3  1.8330                      +   +   +
 1  1.0198                          +
```

Explanation of notation and interpretation of tables

2.3.4.1.10. Design for 5,3,2,1,1,1

OBSERVATIONS	5	3	2	1	1	1
Y(1)	+	-	-	+	-	
Y(2)	+	-	-		+	-
Y(3)	+	-	-	-		+
Y(4)	+	-	-			
Y(5)	+		-	-	-	-
Y(6)		+	-	+	-	-
Y(7)		+	-	-	+	-
Y(8)		+	-	-	-	+
Y(9)			+	-	-	
Y(10)			+	-		-
Y(11)			+		-	-
RESTRAINT	+	+	+			
CHECK STANDARD				+		

DEGREES OF FREEDOM = 6

SOLUTION MATRIX
DIVISOR = 920

OBSERVATIONS	5	3	2	1	1	1
Y(1)	100	-68	-32	119	-111	4
Y(2)	100	-68	-32	4	119	-111
Y(3)	100	-68	-32	-111	4	119
Y(4)	100	-68	-32	4	4	4
Y(5)	60	-4	-56	-108	-108	-108
Y(6)	-20	124	-104	128	-102	-102
Y(7)	-20	124	-104	-102	128	-102
Y(8)	-20	124	-104	-102	-102	128
Y(9)	-20	-60	80	-125	-125	-10
Y(10)	-20	-60	80	-125	-10	-125
Y(11)	-20	-60	80	-10	-125	-125
R*	460	276	184	92	92	92

R* = sum of the three reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	3	2	1	1	1
5	0.2331	+					
3	0.2985		+				
2	0.2638			+			
1	0.3551				+		
1	0.3551					+	
1	0.3551						+
2	0.5043				+	+	
3	0.6203				+	+	+
1	0.3551					+	

```
FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
WT   FACTOR
                5   3   2   1   1   1
 5   0.8660     +
 3   0.8185         +
 2   0.8485             +
 1   1.0149                 +
 1   1.0149                     +
 1   1.0149                         +
 2   1.4560                 +   +
 3   1.8083                 +   +   +
 1   1.0149                     +
```

Explanation of notation and interpretation of tables

2.3.4.1.11. Design for 5,3,2,1,1,1,1

Design 5,3,2,1,1,1,1

OBSERVATIONS	5	3	2	1	1	1	1
Y(1)	+	-	-				
Y(2)	+	-		-	-		
Y(3)	+	-				-	-
Y(4)	+		-	-	-	-	
Y(5)	+		-	-	-		-
Y(6)	+		-	-		-	-
Y(7)	+		-		-	-	-
Y(8)		+	-	-			
Y(9)		+	-		-		
Y(10)		+	-			-	
Y(11)		+	-				-
RESTRAINT	+	+	+				
CHECK STANDARD						+	

DEGREES OF FREEDOM = 5

SOLUTION MATRIX
DIVISOR = 40

OBSERVATIONS	5	3	2	1	1	1	1
Y(1)	20	-4	-16	12	12	12	12
Y(2)	0	-4	4	-8	-8	2	2
Y(3)	0	-4	4	2	2	-8	-8
Y(4)	0	0	0	-5	-5	-10	10
Y(5)	0	0	0	-5	-5	10	-10
Y(6)	0	0	0	-10	10	-5	-5
Y(7)	0	0	0	10	-10	-5	-5
Y(8)	0	4	-4	-12	8	3	3
Y(9)	0	4	-4	8	-12	3	3
Y(10)	0	4	-4	3	3	-12	8
Y(11)	0	4	-4	3	3	8	-12
R*	20	12	8	4	4	4	4

R* = sum of the three reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	3	2	1	1	1	1
5	0.5000	+						
3	0.2646		+					
2	0.4690			+				
1	0.6557				+			
1	0.6557					+		
1	0.6557						+	
1	0.6557							+
2	0.8485				+	+		
3	1.1705				+	+	+	
4	1.3711				+	+	+	+
1	0.6557						+	

FACTORS FOR LEVEL-2 STANDARD DEVIATIONS

WT	FACTOR	5	3	2	1	1	1	1
5	0.8660	+						
3	0.8185		+					
2	0.8485			+				
1	1.0149				+			
1	1.0149					+		
1	1.0149						+	
1	1.0149							+
2	1.4560				+	+		
3	1.8083				+	+	+	
4	2.1166				+	+	+	+
1	1.0149						+	

Explanation of notation and interpretation of tables

HOME TOOLS & AIDS SEARCH BACK NEXT

2.3.4.1.12. Design for 5,3,2,2,1,1,1

OBSERVATIONS	5	3	2	2	1	1	1
Y(1)	+	-	-				
Y(2)	+	-		-			
Y(3)	+		-	-	-		
Y(4)	+		-	-		-	
Y(5)	+		-	-			-
Y(6)		+		-	-		
Y(7)		+		-		-	
Y(8)		+		-			-
Y(9)		+			-	-	-
Y(10)					+	-	
Y(11)						+	-
Y(12)					-		+
RESTRAINT	+	+	+				
CHECK STANDARDS						+	

DEGREES OF FREEDOM = 6

SOLUTION MATRIX
DIVISOR = 10

OBSERVATIONS	5	3	2	2	1	1
Y(1)	2	0	-2	2	0	0
Y(2)	0	-6	6	-4	-2	-2
Y(3)	1	1	-2	0	-1	1
Y(4)	1	1	-2	0	1	-1
Y(5)	1	1	-2	0	1	1
Y(6)	-1	1	0	-2	-1	1
Y(7)	-1	1	0	-2	1	-1
Y(8)	-1	1	0	-2	1	1
Y(9)	0	-2	2	2	-4	-4
Y(10)	0	0	0	0	2	-2
Y(11)	0	0	0	0	0	2
Y(12)	0	0	0	0	-2	0
R*	5	3	2	2	1	1

R* = sum of the three reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	3	2	2	1	1	1
5	0.3162	+						
3	0.6782		+					
2	0.7483			+				
2	0.6000				+			
1	0.5831					+		
1	0.5831						+	
1	0.5831							+
3	0.8124				+	+		
4	1.1136				+	+	+	
1	0.5831						+	

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT	FACTOR	5	3	2	2	1	1	1
5	0.8660	+						
3	0.8185		+					
2	0.8485			+				
2	1.0583				+			
1	1.0149					+		
1	1.0149						+	
1	1.0149							+
3	1.5067				+	+		
4	1.8655				+	+	+	
1	1.0149						+	

Explanation of notation and interpretation of tables

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.1. Mass weights

2.3.4.1.13. Design for 5,4,4,3,2,2,1,1

OBSERVATIONS	5	4	4	3	2	2	1	1
Y(1)	+	+		-	-	-	-	-
Y(2)	+		+	-	-	-	-	-
Y(3)	+	-					-	
Y(4)	+		-					-
Y(5)	+		-				-	
Y(6)	+	-			-			
Y(7)	+			-		-	-	
Y(8)	+			-	-	-		
Y(9)	+	-		-				
Y(10)	+		-		-			
Y(11)	+		-	-				
Y(12)	+	-	-					
RESTRAINT							+	+
CHECK STANDARD							+	-

DEGREES OF FREEDOM = 5

SOLUTION MATRIX
DIVISOR = 916

OBSERVATIONS	5	4	4	3	2	2	1	1
Y(1)	232	325	123	8	-37	135	-1	1
Y(2)	384	151	401	108	73	105	101	-101
Y(3)	432	84	308	236	168	204	-144	144
Y(4)	608	220	196	400	440	-120	408	-408
Y(5)	280	258	30	136	58	234	-246	246
Y(6)	24	-148	68	64	-296	164	-8	8
Y(7)	-104	-122	-142	28	214	-558	-118	118
Y(8)	-512	-354	-382	-144	-250	-598	18	-18
Y(9)	76	-87	139	-408	55	443	51	-51
Y(10)	-128	26	-210	-36	-406	194	-110	110
Y(11)	-76	87	-139	-508	-55	473	-51	51
Y(12)	-300	-440	-392	116	36	-676	100	-100
R*	1224	696	720	516	476	120	508	408

R* = sum of the two reference standards (for going-up calibrations)

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	4	4	3	2	2	1	1
5	1.2095	+							
4	0.8610		+						
4	0.9246			+					
3	0.9204				+				
2	0.8456					+			
2	1.4444						+		
1	0.5975							+	
1	0.5975								+
4	1.5818					+	+		
7	1.7620				+	+	+		
11	2.5981			+	+	+	+		
15	3.3153		+	+	+	+	+		
20	4.4809	+	+	+	+	+	+		
0	1.1950							+	-

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT	FACTOR	5	4	4	3	2	2	1	1
5	2.1380	+							
4	1.4679		+						
4	1.4952			+					
3	1.2785				+				
2	1.2410					+			
2	1.0170						+		
1	0.7113							+	
1	0.7113								+
4	1.6872					+	+		
7	2.4387				+	+	+		
11	3.4641			+	+	+	+		
15	4.4981		+	+	+	+	+		
20	6.2893	+	+	+	+	+	+		
0	1.4226							+	-

Explanation of notation and interpretation of tables

2.3.4.1.14. Design for 5,5,2,2,1,1,1,1

Design 5,5,2,2,1,1,1,1

OBSERVATIONS	5	5	2	2	1	1	1	1
Y(1)	+		-	-	-			
Y(2)		+	-	-		-		
Y(3)	+		-	-			-	
Y(4)		+	-	-				-
Y(5)			+	+	-	-	-	-
Y(6)			+		-		-	
Y(7)			+			-		-
Y(8)				+	-			-
Y(9)				+		-	-	
Y(10)					+		-	
Y(11)						+		-
RESTRAINT	+	+						
CHECK STANDARD							+	

DEGREES OF FREEDOM = 4

SOLUTION MATRIX
DIVISOR = 120

OBSERVATIONS	5	5	2	2	1	1
Y(1)	30	-30	-12	-12	-22	-10
Y(2)	-30	30	-12	-12	-10	-22
Y(3)	30	-30	-12	-12	10	-2
Y(4)	-30	30	-12	-12	-2	10
Y(5)	0	0	6	6	-12	-12
Y(6)	-30	30	33	-27	-36	24
Y(7)	30	-30	33	-27	24	-36
Y(8)	0	0	-27	33	-18	6
Y(9)	0	0	-27	33	6	-18
Y(10)	0	0	0	0	32	8
Y(11)	0	0	0	0	8	32
R*	60	60	24	24	12	12

1	1
10	-2
-2	10
22	-10
10	-22
12	-12
36	-24
24	-36
6	-18
18	-6
32	-8
-8	-32
12	12

R* = sum of the two reference standards

FACTORS FOR COMPUTING REPEATABILITY STANDARD DEVIATIONS

WT	FACTOR	5	5	2	2	1	1	1	1
5	0.6124	+							
5	0.6124		+						
2	0.5431			+					
2	0.5431				+				
1	0.5370					+			
1	0.5370						+		
1	0.5370							+	
1	0.5370								+
2	0.6733					+	+		
4	0.8879				+	+	+		
6	0.8446			+	+	+	+		
11	1.0432		+	+	+	+	+		
16	0.8446	+	+	+	+	+	+		
1	0.5370							+	

FACTORS FOR COMPUTING LEVEL-2 STANDARD DEVIATIONS

WT	FACTOR	5	5	2	2	1	1	1	1
5	0.7071	+							
5	0.7071		+						
2	1.0392			+					
2	1.0392				+				
1	1.0100					+			
1	1.0100						+		
1	1.0100							+	
1	1.0100								+
2	1.4422					+	+		
4	1.8221				+	+	+		
6	2.1726			+	+	+	+		
11	2.2847		+	+	+	+	+		
16	2.1726	+	+	+	+	+	+		
1	1.0100							+	

Explanation of notation and interpretation of tables

2.3.4.1.15. Design for 5,5,3,2,1,1,1

```
 OBSERVATIONS    5    5    3
2    1    1    1

     Y(1)        +         -
-
     Y(2)             +    -
-
     Y(3)        +
-    -    -    -
     Y(4)             +
-    -    -    -
     Y(5)        +         -
-    -
     Y(6)        +         -
-         -
     Y(7)        +         -
-    -
     Y(8)             +    -
-    -
     Y(9)             +    -
-         -
     Y(10)            +    -
-    -

 RESTRAINT       +    +

 CHECK STANDARD
+

 DEGREES OF FREEDOM  =
4


SOLUTION MATRIX

DIVISOR =  10

 OBSERVATIONS       5
5       3       2       1
1       1


    Y(1)             1      -
1      -2      -3       1
1       1
    Y(2)            -1
1      -2      -3       1
1       1
    Y(3)             1      -
1       2      -2      -1
-1      -1
    Y(4)            -1
1       2      -2      -1
-1      -1
    Y(5)             1      -
1      -1       1      -2
```

```
-2      3
    Y(6)            1     -
1     -1      1     -2
3     -2
    Y(7)            1     -
1     -1      1       3
-2     -2
    Y(8)           -1
1     -1      1     -2
-2      3
    Y(9)           -1
1     -1      1     -2
3     -2
    Y(10)          -1
1     -1      1       3
-2     -2
    R*              5
5      3      2       1
1      1

    R* = sum of the two
reference standards


 FACTORS FOR REPEATABILITY
STANDARD DEVIATIONS
 WT  FACTOR
                 5   5   3
2   1   1   1
  5  0.3162    +
  5  0.3162        +
  3  0.4690            +
  2  0.5657
+
  1  0.6164
+
  1  0.6164
+
  1  0.6164
+
  3  0.7874
+   +
  6  0.8246            +
+   +
 11  0.8832        +   +
+   +
 16  0.8246    +   +   +
+   +
  1  0.6164
+


 FACTORS FOR BETWEEN-DAY
STANDARD DEVIATIONS
 WT  FACTOR
                 5   5   3
2   1   1   1
  5  0.7071    +
  5  0.7071        +
  3  1.0863            +
  2  1.0392
+
  1  1.0100
+
  1  1.0100
+
  1  1.0100
+
  3  1.4765
+   +
  6  1.9287            +
+   +
 11  2.0543        +   +
+   +
 16  1.9287    +   +   +
+   +
  1  1.0100
+
```

[Explanation of notation and interpretation of tables]

2.3.4.1.16. Design for 1,1,1,1,1,1,1,1 weights

```
 OBSERVATIONS    1   1   1   1   1   1   1   1

     Y(1)        +       -
     Y(2)        +           -
     Y(3)        +               -
     Y(4)        +                   -
     Y(5)        +                       -
     Y(6)        +                           -
     Y(7)            +   -
     Y(8)            +       -
     Y(9)            +           -
     Y(10)           +               -
     Y(11)           +                   -
     Y(12)           +                       -

 RESTRAINT       +   +

 CHECK STANDARD                          +

DEGREES OF FREEDOM  =  5
```

SOLUTION MATRIX
DIVISOR = 12

OBSERVATIONS	1	1	1	1	1	1	1	1
Y(1)	1	-1	-6	0	0	0	0	0
Y(2)	1	-1	0	-6	0	0	0	0
Y(3)	1	-1	0	0	-6	0	0	0
Y(4)	1	-1	0	0	0	-6	0	0
Y(5)	1	-1	0	0	0	0	-6	0
Y(6)	1	-1	0	0	0	0	0	-6
Y(7)	-1	1	-6	0	0	0	0	0
Y(8)	-1	1	0	-6	0	0	0	0
Y(9)	-1	1	0	0	-6	0	0	0
Y(10)	-1	1	0	0	0	-6	0	0
Y(11)	-1	1	0	0	0	0	-6	0
Y(12)	-1	1	0	0	0	0	0	-6
R*	6	6	6	6	6	6	6	6

R* = sum of the two reference standards

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT	K1	1	1	1	1	1	1	1	1
1	0.2887	+							
1	0.2887		+						
1	0.7071			+					
1	0.7071				+				
1	0.7071					+			
1	0.7071						+		
1	0.7071							+	
1	0.7071								+
2	1.0000			+	+				
3	1.2247			+	+	+			
4	1.4142			+	+	+	+		
5	1.5811			+	+	+	+	+	
6	1.7321			+	+	+	+	+	+
1	0.7071							+	

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT	K2	1	1	1	1	1	1	1	1
1	0.7071	+							
1	0.7071		+						
1	1.2247			+					
1	1.2247				+				
1	1.2247					+			
1	1.2247						+		
1	1.2247							+	
1	1.2247								+
2	2.0000			+	+				
3	2.7386			+	+	+			
4	3.4641			+	+	+	+		
5	4.1833			+	+	+	+	+	
6	4.8990			+	+	+	+	+	+
1	1.2247							+	

Explanation of notation and interpretation of tables

2.3.4.1.17. Design for 3,2,1,1,1 weights

```
OBSERVATIONS    3   2   1   1   1

    Y(1)        +   -   -
    Y(2)        +   -       -
    Y(3)        +   -           -
    Y(4)        +       -   -   -
    Y(5)            +   -   -
    Y(6)            +   -       -
    Y(7)            +       -   -
    Y(8)                +   -
    Y(9)                +       -
    Y(10)                   +   -

RESTRAINT       +   +

CHECK STANDARD              +

DEGREES OF FREEDOM = 6
```

SOLUTION MATRIX
DIVISOR = 25

OBSERVATIONS	3	2	1	1	1
Y(1)	3	-3	-4	1	1
Y(2)	3	-3	1	-4	1
Y(3)	3	-3	1	1	-4
Y(4)	1	-1	-3	-3	-3
Y(5)	-2	2	-4	-4	1
Y(6)	-2	2	-4	1	-4
Y(7)	-2	2	1	-4	-4
Y(8)	0	0	5	-5	0
Y(9)	0	0	5	0	-5
Y(10)	0	0	0	5	-5
R*	15	10	5	5	5

R* = sum of the two reference standards

```
FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

 WT     K1      3   2   1   1   1
 3   0.2530     +
 2   0.2530         +
 1   0.4195             +
 1   0.4195                 +
 1   0.4195                     +
 2   0.5514             +   +
 3   0.6197             +   +   +
 1   0.4195                 +

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS
```

WT	K2	3	2	1	1	1
3	0.7211	+				
2	0.7211		+			
1	1.0392			+		
1	1.0392				+	
1	1.0392					+
2	1.5232			+	+	
3	1.9287			+	+	+
1	1.0392				+	

[Explanation of notation and interpretation of tables]

2.3.4.1.18. Design for 10-and 20-pound weights

```
OBSERVATIONS    1   2   2   1   1

    Y(1)        +           -
    Y(2)        +               -
    Y(3)        +   -           +
    Y(4)        +       -   +
    Y(5)        +   -       +
    Y(6)        +       -       +
    Y(7)            +   -

RESTRAINT       +

CHECK STANDARD              +

DEGREES OF FREEDOM = 3


                        SOLUTION MATRIX
                          DIVISOR  =   24

OBSERVATIONS       1      2      2      1      1

   Y(1)            0    -12    -12    -16     -8
   Y(2)            0    -12    -12     -8    -16
   Y(3)            0     -9     -3     -4      4
   Y(4)            0     -3     -9      4     -4
   Y(5)            0     -9     -3      4     -4
   Y(6)            0     -3     -9     -4      4
   Y(7)            0      6     -6      0      0
   R*             24     48     48     24     24

   R* = Value of the reference standard

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

WT    K1       1   2   2   1   1
 2  0.9354        +
 2  0.9354            +
 1  0.8165                +
 1  0.8165                    +
 4  1.7321        +   +
 5  2.3805        +   +   +
 6  3.0000        +   +   +   +
 1  0.8165                +

FACTORS FOR BETWEEN-DAY STANDARD DEVIATIONS

WT    K2       1   2   2   1   1
 2  2.2361        +
 2  2.2361            +
 1  1.4142                +
 1  1.4142                    +
 4  4.2426        +   +
 5  5.2915        +   +   +
```

```
6  6.3246        +    +    +    +
1  1.4142                  +
```

[Explanation of notation and interpretation of tables]

2.3.4.2. Drift-elimination designs for gauge blocks

Tie to the defined unit of length

The unit of length in many industries is maintained and disseminated by gauge blocks. The highest accuracy calibrations of gauge blocks are done by laser intererometry which allows the transfer of the unit of length to a gauge piece. Primary standards laboratories maintain master sets of English gauge blocks and metric gauge blocks which are calibrated in this manner. Gauge blocks ranging in sizes from 0.1 to 20 inches are required to support industrial processes in the United States.

Mechanical comparison of gauge blocks

However, the majority of gauge blocks are calibrated by comparison with master gauges using a mechanical comparator specifically designed for measuring the small difference between two blocks of the same nominal length. The measurements are temperature corrected from readings taken directly on the surfaces of the blocks. Measurements on 2 to 20 inch blocks require special handling techniques to minimize thermal effects. A typical calibration involves a set of 81 gauge blocks which are compared one-by-one with master gauges of the same nominal size.

Calibration designs for gauge blocks

Calibration designs allow comparison of several gauge blocks of the same nominal size to one master gauge in a manner that promotes economy of operation and minimizes wear on the master gauge. The calibration design is repeated for each size until measurements on all the blocks in the test sets are completed.

Problem of thermal drift

Measurements on gauge blocks are subject to drift from heat build-up in the comparator. This drift must be accounted for in the calibration experiment or the lengths assigned to the blocks will be contaminated by the drift term.

Elimination of linear drift

The designs in this catalog are constructed so that the solutions are immune to linear drift if the measurements are equally spaced over time. The size of the drift is the average of the n difference measurements. Keeping track of drift from design to design is useful because a marked change from its usual range of values may indicate a problem with the measurement system.

Assumption for Doiron designs

Mechanical measurements on gauge blocks take place successively with one block being inserted into the comparator followed by a second block and so on. This scenario leads to the assumption that the individual measurements are subject to drift (Doiron). Doiron lists designs meeting this criterion which also allow for:

- two master blocks, *R1* and *R2*
- one check standard = difference between *R1* and *R2*
- one - nine test blocks

Properties of drift-elimination designs that use 1 master block

The designs are constructed to:

- Be immune to linear drift
- Minimize the standard deviations for test blocks (as much as possible)
- Spread the measurements on each block throughout the design
- Be completed in 5-10 minutes to keep the drift at the 5 nm level

Caution

Because of the large number of gauge blocks that are being intercompared and the need to eliminate drift, the Doiron designs are not completely balanced with respect to the test blocks. Therefore, the standard deviations are not equal for all blocks. If all the blocks are being calibrated for use in one facility, it is easiest to quote the largest of the standard deviations for all blocks rather than try to maintain a separate record on each block.

Definition of master block and check standard

At the National Institute of Standards and Technology (NIST), the first two blocks in the design are NIST masters which are designated *R1* and *R2*, respectively. The *R1* block is a steel block, and the *R2* block is a chrome-carbide block. If the test blocks are steel, the reference is *R1*; if the test blocks are chrome-carbide, the reference is *R2*. The check standard is always the difference between *R1* and *R2* as estimated from the design and is independent of *R1* and *R2*. The designs are listed in this section of the catalog as:

1. Doiron design for 3 gauge blocks - 6 measurements
2. Doiron design for 3 gauge blocks - 9 measurements
3. Doiron design for 4 gauge blocks - 8 measurements
4. Doiron design for 4 gauge blocks - 12 measurements
5. Doiron design for 5 gauge blocks - 10 measurements
6. Doiron design for 6 gauge blocks - 12 measurements
7. Doiron design for 7 gauge blocks - 14 measurements
8. Doiron design for 8 gauge blocks - 16 measurements
9. Doiron design for 9 gauge blocks - 18 measurements
10. Doiron design for 10 gauge blocks - 20 measurements
11. Doiron design for 11 gauge blocks - 22 measurements

Properties of designs that use 2 master blocks

Historical designs for gauge blocks (Cameron and Hailes) work on the assumption that the difference measurements are contaminated by linear drift. This assumption is more restrictive and covers the case of drift in successive measurements but produces fewer designs. The Cameron/Hailes designs meeting this criterion allow for:

- two reference (master) blocks, *R1* and *R2*
- check standard = difference between the two master blocks

and assign equal uncertainties to values of all test blocks.

The designs are listed in this section of the catalog as:

1. Cameron-Hailes design for 2 masters + 2 test blocks
2. Cameron-Hailes design for 2 masters + 3 test blocks
3. Cameron-Hailes design for 2 masters + 4 test blocks
4. Cameron-Hailes design for 2 masters + 5 test blocks

Important concept - check standard

The check standards for the designs in this section are not artifact standards but constructions from the design. The value of one master block or the average of two master blocks is the restraint for the design, and values for the masters, *R1* and *R2*, are estimated from a set of measurements taken according to the design. The check standard value is the difference between the estimates, *R1* and *R2*. Measurement control is exercised by comparing the current value of the check standard with its historical average.

2.3.4.2.1. Doiron 3-6 Design

```
Doiron 3-6 design

OBSERVATIONS   1   1   1

    Y(1)       +   -
    Y(2)       -       +
    Y(3)           +   -
    Y(4)       -   +
    Y(5)           -   +
    Y(6)       +       -

RESTRAINT      +

CHECK STANDARD     +

DEGREES OF FREEDOM  =      4

                  SOLUTION MATRIX
                   DIVISOR  =  6

OBSERVATIONS     1      1      1

   Y(1)          0      -2     -1
   Y(2)          0       1      2
   Y(3)          0       1     -1
   Y(4)          0       2      1
   Y(5)          0      -1      1
   Y(6)          0      -1     -2
   R*            6       6      6

   R* = Value of the reference standard

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 NOM FACTOR
                1   1   1
  1  0.0000     +
  1  0.5774         +
  1  0.5774             +
  1  0.5774         +
```

Explanation of notation and interpretation of tables

2.3.4.2.2. Doiron 3-9 Design

```
 Doiron 3-9 Design

 OBSERVATIONS   1   1   1

     Y(1)       +   -
     Y(2)           -   +
     Y(3)       +       -
     Y(4)       -   +
     Y(5)       -       +
     Y(6)           +   -
     Y(7)       -   +
     Y(8)           -   +
     Y(9)       +       -

 RESTRAINT      +

 CHECK STANDARD     +

 DEGREES OF FREEDOM  =      7


               SOLUTION MATRIX
                DIVISOR  =  9

 OBSERVATIONS     1       1      1

    Y(1)          0      -2     -1
    Y(2)          0      -1      1
    Y(3)          0      -1     -2
    Y(4)          0       2      1
    Y(5)          0       1      2
    Y(6)          0       1     -1
    Y(7)          0       2      1
    Y(8)          0      -1      1
    Y(9)          0      -1     -2
    R(1)          9       9      9


 FACTORS FOR COMPUTING REPEATABILITY STANDARD
DEVIATIONS
 NOM FACTOR
                1   1   1
  1  0.0000     +
  1  0.4714         +
  1  0.4714             +
  1  0.4714         +
```

Explanation of notation and interpretation of tables

2.3.4.2.3. Doiron 4-8 Design

```
Doiron 4-8 Design

OBSERVATIONS    1    1    1    1

    Y(1)        +    -
    Y(2)                  +    -
    Y(3)        -                +
    Y(4)             +    -
    Y(5)        -    +
    Y(6)                  -    +
    Y(7)        +                -
    Y(8)             -    +

RESTRAINT       +


CHECK STANDARD      +

DEGREES OF FREEDOM  =    5

                SOLUTION MATRIX
                 DIVISOR  =  8

OBSERVATIONS      1        1       1       1

   Y(1)           0       -3      -2      -1
   Y(2)           0        1       2      -1
   Y(3)           0        1       2       3
   Y(4)           0        1      -2      -1
   Y(5)           0        3       2       1
   Y(6)           0       -1      -2       1
   Y(7)           0       -1      -2      -3
   Y(8)           0       -1       2       1
   R*             8        8       8       8

   R* = Value of the reference standard


FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
NOM FACTOR
                1    1    1    1
 1  0.0000     +
 1  0.6124          +
 1  0.7071               +
 1  0.6124                    +
 1  0.6124          +
```

Explanation of notation and interpretation of tables

2.3.4.2.4. Doiron 4-12 Design

Doiron 4-12 Design

OBSERVATIONS	1	1	1	1
Y(1)	+	-		
Y(2)	+			+
Y(3)			+	-
Y(4)	-	+		
Y(5)		+	-	
Y(6)		-		+
Y(7)	+		-	
Y(8)		+		-
Y(9)	+			-
Y(10)		-	+	
Y(11)	-		+	
Y(12)			-	+
RESTRAINT	+			
CHECK STANDARD		+		

DEGREES OF FREEDOM = 9

SOLUTION MATRIX
DIVISOR = 8

OBSERVATIONS	1	1	1	1
Y(1)	0	-2	-1	-1
Y(2)	0	1	1	2
Y(3)	0	0	1	-1
Y(4)	0	2	1	1
Y(5)	0	1	-1	0
Y(6)	0	-1	0	1
Y(7)	0	-1	-2	-1
Y(8)	0	1	0	-1
Y(9)	0	-1	-1	-2
Y(10)	0	-1	1	0
Y(11)	0	1	2	1
Y(12)	0	0	-1	1
R*	6	6	6	4

R* = Value of the reference standard

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

NOM	FACTOR	1	1	1	1
1	0.0000	+			
1	0.5000		+		
1	0.5000			+	
1	0.5000				+
1	0.5000		+		

Explanation of notation and interpretation of tables

2.3.4.2.5. Doiron 5-10 Design

Doiron 5-10 Design

OBSERVATIONS	1	1	1	1	1
Y(1)	+	-			
Y(2)			-	+	
Y(3)	+		-		
Y(4)		-			+
Y(5)			-	+	
Y(6)	+			-	
Y(7)	-		+		
Y(8)			+		-
Y(9)	-			+	
Y(10)		+			-
RESTRAINT	+				
CHECK STANDARD		+			

DEGREES OF FREEDOM = 6

SOLUTION MATRIX
DIVISOR = 90

OBSERVATIONS	1	1	1	1	1
Y(1)	0	-50	-10	-10	-30
Y(2)	0	20	4	-14	30
Y(3)	0	-10	-29	-11	-15
Y(4)	0	-20	5	5	15
Y(5)	0	0	-18	18	0
Y(6)	0	-10	-11	-29	-15
Y(7)	0	10	29	11	15
Y(8)	0	-20	14	-4	-30
Y(9)	0	10	11	29	15
Y(10)	0	20	-5	-5	-15
R*	90	90	90	90	90

R* = Value of the reference standard

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

NOM	FACTOR	1	1	1	1	1
1	0.0000	+				
1	0.7454		+			
1	0.5676			+		
1	0.5676				+	
1	0.7071					+
1	0.7454		+			

Explanation of notation and interpretation of tables

2.3.4.2.6. Doiron 6-12 Design

Doiron 6-12 Design

OBSERVATIONS	1	1	1	1	1	1
Y(1)	+	-				
Y(2)			-	+		
Y(3)		-			+	
Y(4)				-	+	
Y(5)		-		+		
Y(6)			+	-		
Y(7)	+				-	
Y(8)		+		-		
Y(9)		+	-			
Y(10)	-			+		
Y(11)		+			-	
Y(12)	-		+			
RESTRAINT	+					
CHECK STANDARD		+				

DEGREES OF FREEDOM = 7

SOLUTION MATRIX
DIVISOR = 360

OBSERVATIONS	1	1	1	1	1	1
Y(1)	0	-136	-96	-76	-72	-76
Y(2)	0	-4	-24	-79	72	11
Y(3)	0	-20	-120	-35	0	55
Y(4)	0	4	24	-11	-72	79
Y(5)	0	-60	0	75	0	-15
Y(6)	0	20	120	-55	0	35
Y(7)	0	-76	-96	-61	-72	-151
Y(8)	0	64	24	4	-72	4
Y(9)	0	40	-120	-20	0	-20
Y(10)	0	72	72	72	144	72
Y(11)	0	60	0	15	0	-75
Y(12)	0	76	96	151	72	61
R*	360	360	360	360	360	360

R* = Value of the reference standard

```
FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
NOM FACTOR
                1   1   1   1   1   1
 1  0.0000      +
 1  0.6146          +
 1  0.7746              +
 1  0.6476                  +
 1  0.6325                      +
 1  0.6476                          +
 1  0.6146          +
```

Explanation of notation and interpretation of tables

2.3.4.2.7. Doiron 7-14 Design

Doiron 7-14 Design

OBSERVATIONS	1	1	1	1	1	1	1
Y(1)	+	-					
Y(2)				-		+	
Y(3)			+		-		
Y(4)		+					-
Y(5)	+					-	
Y(6)			-		+		
Y(7)		+		-			
Y(8)			+				-
Y(9)					+	-	
Y(10)	-						+
Y(11)		-				+	
Y(12)			-	+			
Y(13)	-			+			
Y(14)					-		+

RESTRAINT +

CHECK STANDARD +

DEGREES OF FREEDOM = 8

PARAMETER VALUES
DIVISOR = 1015

OBSERVATIONS	1	1	1	1	1	1	1
Y(1)	0	-406	-203	-203	-203	-203	-203
Y(2)	0	0	-35	-210	35	210	0
Y(3)	0	0	175	35	-175	-35	0
Y(4)	0	203	-116	29	-116	29	-261
Y(5)	0	-203	-229	-214	-264	-424	-174
Y(6)	0	0	-175	-35	175	35	0
Y(7)	0	203	-61	-221	-26	-11	29
Y(8)	0	0	305	90	130	55	-145
Y(9)	0	0	220	15	360	-160	145
Y(10)	0	203	319	174	319	174	464
Y(11)	0	-203	26	11	61	221	-29
Y(12)	0	0	-360	160	-220	-15	-145
Y(13)	0	203	264	424	229	214	174
Y(14)	0	0	-130	-55	-305		

```
-90    145
    R*        1015      1015   1015   1015   1015
1015   1015

    R* = Value of the reference standard


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 NOM FACTOR
                1   1   1   1   1   1   1
  1  0.0000     +
  1  0.6325         +
  1  0.7841             +
  1  0.6463                 +
  1  0.7841                     +
  1  0.6463                         +
  1  0.6761                             +
  1  0.6325         +
```

Explanation of notation and interpretation of tables

2.3.4.2.8. Doiron 8-16 Design

```
 Doiron 8-16 Design

 OBSERVATIONS   1   1   1   1   1   1   1   1

     Y(1)       +   -
     Y(2)                       +       -
     Y(3)               -               +
     Y(4)       -               +
     Y(5)           +           -
     Y(6)                               -   +
     Y(7)           -   +
     Y(8)       -                   +
     Y(9)                       -       +
     Y(10)      -           +
     Y(11)          +                       -
     Y(12)              -       +
     Y(13)      -           +
     Y(14)                  -               +
     Y(15)          +       -
     Y(16)          +                   -

 RESTRAINT      +

 CHECK STANDARD     +

 DEGREES OF FREEDOM  =      9

                          SOLUTION MATRIX
                           DIVISOR = 2852

 OBSERVATIONS  1          1      1      1      1
1      1      1

    Y(1)       0      -1392   -620   -472   -516
-976   -824   -916
    Y(2)       0         60    248    -78     96
878   -112   -526
    Y(3)       0        352    124   -315    278
255    864    289
    Y(4)       0        516    992    470   1396
706    748    610
    Y(5)       0       -356    620     35    286
-979    -96   -349
    Y(6)       0         92      0     23   -138
253   -552    667
    Y(7)       0       -148   -992    335   -522
-407   -104    -81
    Y(8)       0       -416    372    113    190
995     16    177
    Y(9)       0        308   -248    170   -648
134    756    342
    Y(10)      0        472    620    955    470
585    640    663
    Y(11)      0        476   -124   -191    -94
-117   -128   -703
    Y(12)      0       -104   -620   -150    404
-286      4   -134
```

```
    Y(13)      0         472    620    955    470
585    640    663
    Y(14)      0         444    124   -292    140
508    312    956
    Y(15)      0         104    620    150   -404
286     -4    134
    Y(16)      0         568   -124   -168   -232
136   -680    -36
    R*      2852        2852   2852   2852   2852
2852   2852   2852

    R* = value of reference block


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 WT  FACTOR
                1   1   1   1   1   1   1   1
  1  0.0000     +
  1  0.6986         +
  1  0.7518             +
  1  0.5787                 +
  1  0.6996                     +
  1  0.8313                         +
  1  0.7262                             +
  1  0.7534                                 +
  1  0.6986         +
```

[Explanation of notation and interpretation of tables]

2.3.4.2.9. Doiron 9-18 Design

```
Doiron 9-18 Design

OBSERVATIONS   1   1   1   1   1   1   1   1   1

    Y(1)       +   -
    Y(2)                                   -       +
    Y(3)           +   -
    Y(4)                   -   +
    Y(5)                           +       -
    Y(6)           -       +
    Y(7)               +                   -
    Y(8)                                   +   -
    Y(9)       -               +
    Y(10)                  +       -
    Y(11)      -                           +
    Y(12)                          -       +
    Y(13)      -           +
    Y(14)          +       -
    Y(15)                      -       +
    Y(16)      +                           -
    Y(17)              -           +
    Y(18)                              +   -

RESTRAINT      +

CHECK STANDARD     +

DEGREES OF FREEDOM  =     10

                          SOLUTION MATRIX
                           DIVISOR = 8247

 OBSERVATIONS   1         1      1      1      1
1      1      1      1

    Y(1)         0     -3680  -2305  -2084  -1175   -
1885  -1350  -1266   -654
    Y(2)         0      -696  -1422   -681  -1029
-984  -2586   -849   1203
    Y(3)         0      1375  -3139    196   -491   -
1279  -1266   -894   -540
    Y(4)         0      -909   -222  -1707   1962
-432    675    633    327
    Y(5)         0       619   1004    736   -329
2771   -378  -1674   -513
    Y(6)         0     -1596   -417   1140    342
303     42    186     57
    Y(7)         0       955   2828    496   -401
971  -1689   -411   -525
    Y(8)         0       612    966    741   1047
1434    852   2595  -1200
    Y(9)         0      1175   1666   1517   3479
1756   2067   2085   1038
    Y(10)        0       199  -1276   1036   -239   -
3226   -801  -1191   -498
    Y(11)        0       654   1194    711   1038
1209   1719   1722   2922
```

```
    Y(12)       0        91    494    -65  -1394
887    504   2232    684
    Y(13)       0      2084   1888   3224   1517
2188   1392   1452    711
    Y(14)       0      1596    417  -1140   -342
-303    -42   -186    -57
    Y(15)       0       175    950   -125  -1412
437   2238    486    681
    Y(16)       0      -654  -1194   -711  -1038  -
1209  -1719  -1722  -2922
    Y(17)       0      -420  -2280    300     90
2250   -423    483     15
    Y(18)       0        84    456    -60    -18
-450   1734  -1746     -3
    R*       8247      8247   8247   8247   8247
8247   8247   8247   8247

    R* = Value of the reference standard


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 NOM FACTOR
                1   1   1   1   1   1   1   1   1
  1  0.0000    +
  1  0.6680        +
  1  0.8125            +
  1  0.6252                +
  1  0.6495                    +
  1  0.8102                        +
  1  0.7225                            +
  1  0.7235                                +
  1  0.5952                                    +
  1  0.6680        +
```

[Explanation of notation and interpretation of tables]

2.3.4.2.10. Doiron 10-20 Design

```
 Doiron 10-20 Design

 OBSERVATIONS   1   1   1   1   1   1   1   1   1
1

     Y(1)       +   -
     Y(2)                           +   -
     Y(3)                   -
+
     Y(4)                       +   -
     Y(5)           +                       -
     Y(6)               +   -
     Y(7)                               +   -
     Y(8)       -
+
     Y(9)                           +   -
     Y(10)                                  +
-
     Y(11)                      +       -
     Y(12)          +   -
     Y(13)              +
-
     Y(14)                          -           +
     Y(15)              +                   -
     Y(16)      +               -
     Y(17)              -                   +
     Y(18)                  +   -
     Y(19)      -                           +
     Y(20)          -               +

 RESTRAINT      +

 CHECK STANDARD     +

 DEGREES OF FREEDOM  =     11

                            SOLUTION MATRIX
                          DIVISOR  =  33360

 OBSERVATIONS 1          1      1      1      1
1      1      1      1

    Y(1)       0     -15300  -9030  -6540  -5970  -
9570  -7770  -6510  -9240
    Y(2)       0       1260   1594   1716   3566
3470   9078  -5678    -24
    Y(3)       0       -960  -2856  -7344  -2664  -
1320  -1992  -1128    336
    Y(4)       0      -3600  -1536    816   5856  -
9120  -1632  -1728  -3744
    Y(5)       0       6060    306  -1596   -906  -
1050   -978  -2262  -8376
    Y(6)       0       2490   8207  -8682  -1187
1165   2769   2891    588
    Y(7)       0      -2730    809  -1494   -869  -
2885    903   6557  -8844
    Y(8)       0       5580   7218  11412   6102
```

```
6630   6366   5514   8472
    Y(9)      0      1800  -2012   -408   -148
7340  -7524  -1916   1872
    Y(10)     0      3660   1506  -3276    774
3990   2382   3258   9144
    Y(11)     0     -1800  -3548    408   5708  -
1780  -9156  -3644  -1872
    Y(12)     0      6270  -9251  -3534  -1609
455  -3357  -3023    516
    Y(13)     0       960   2856   7344   2664
1320   1992   1128   -336
    Y(14)     0      -330   -391    186  -2549  -
7925  -2457   1037   6996
    Y(15)     0      2520   8748   3432   1572
1380   1476  -5796    -48
    Y(16)     0     -5970  -7579  -8766 -15281  -
9425  -9573  -6007  -6876
    Y(17)     0     -1260  -7154  -1716   1994
2090   7602    118     24
    Y(18)     0       570   2495   9990  -6515  -
1475  -1215    635   1260
    Y(19)     0      6510   9533   6642   6007
7735   9651  15329   8772
    Y(20)     0     -5730     85   1410   3455
8975   3435   1225   1380
    R*    33360      33360  33360  33360  33360
33360  33360  33360  33360

    R* = Value of the reference standard


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 NOM FACTOR
                1   1   1   1   1   1   1   1   1
1
  1  0.0000    +
  1  0.6772        +
  1  0.7403            +
  1  0.7498                +
  1  0.6768                    +
  1  0.7456                        +
  1  0.7493                            +
  1  0.6779                                +
  1  0.7267                                    +
  1  0.6961
+
  1  0.6772        +
```

[Explanation of notation and interpretation of tables]

2.3.4.2.11. Doiron 11-22 Design

```
 Doiron 11-22 Design

 OBSERVATIONS   1   1   1   1   1   1   1   1   1
1   1

     Y(1)       +   -
     Y(2)                       +   -
     Y(3)                               +
-
     Y(4)                   +                   -
     Y(5)           +   -
     Y(6)
+   -
     Y(7)                           -       +
     Y(8)       -               +
     Y(9)               +   -
     Y(10)      +                   -
     Y(11)          +                       -
     Y(12)              -                   +
     Y(13)                          +   -
     Y(14)          -
+
     Y(15)                  +   -
     Y(16)                                  +
-
     Y(17)                              +   -
     Y(18)      -
+
     Y(19)              +                       -
     Y(20)                      -           +
     Y(21)                  -
+
     Y(22)                          +
-

 RESTRAINT      +


 CHECK STANDARD     +


 DEGREES OF FREEDOM  =     12


                            SOLUTION MATRIX
                            DIVISOR = 55858

 OBSERVATIONS      1      1      1      1      1
1      1      1      1      1

    Y(1)      0    -26752 -18392 -15532  -9944  -
8778 -14784 -15466 -16500 -10384 -17292
    Y(2)      0      1166   1119   3976  12644 -
11757  -1761   2499   1095  -2053   1046
    Y(3)      0      5082   4446   3293   4712
160   5882  15395   3527  -9954    487
    Y(4)      0      -968  -1935  10496   2246   -
635  -4143   -877 -13125   -643  -1060
    Y(5)      0      8360 -18373  -8476  -3240  -
3287  -8075  -1197  -9443  -1833  -2848
    Y(6)      0     -6908  -7923  -9807  -2668
```

```
431  -4753  -1296 -10224   9145 -18413
    Y(7)      0      1716   3084   6091    404  -
2452 -10544  -2023  15073    332   5803
    Y(8)      0      9944  13184  15896  24476
11832  13246  14318  13650   9606  12274
    Y(9)      0      2860  12757 -11853  -2712
145   3585    860    578   -293  -2177
    Y(10)     0     -8778 -12065 -11920 -11832 -
23589 -15007 -11819 -12555 -11659 -11228
    Y(11)     0     11286   1729   -271  -4374  -
3041  -3919 -14184   -180  -3871   1741
    Y(12)     0     -3608 -13906  -4734     62
2942  11102   2040  -2526    604  -2566
    Y(13)     0     -6006 -10794  -7354  -1414
8582 -18954  -6884 -10862  -1162  -6346
    Y(14)     0     -9460   1748   6785   2330
2450   2790     85   6877   4680  16185
    Y(15)     0      5588  10824  19965  -8580
88   6028   1485  11715   2904  10043
    Y(16)     0      -792   5803   3048   1376
1327   5843   1129  15113  -1911 -10100
    Y(17)     0      -682   6196   3471  -1072
3188  15258 -10947   6737  -1434   2023
    Y(18)     0     10384  12217  12510   9606
11659  12821  14255  13153  24209  15064
    Y(19)     0      1892  10822  -1357   -466   -
490   -558    -17 -12547   -936  -3237
    Y(20)     0      5522   3479    -93 -10158    -
13   5457  15332   3030   4649   3277
    Y(21)     0      1760  -3868 -13544  -3622   -
692  -1700   -252  -1988   2554  11160
    Y(22)     0     -1606   -152   -590   2226
11930   2186  -2436   -598 -12550  -3836
    R*    55858     55858  55858  55858  55858
55858  55858  55858  55858  55858  55858

    R* = Value of the reference standard

 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 NOM FACTOR
                1   1   1   1   1   1   1   1   1
1   1
  1  0.0000    +
  1  0.6920        +
  1  0.8113            +
  1  0.8013                +
  1  0.6620                    +
  1  0.6498                        +
  1  0.7797                            +
  1  0.7286                                +
  1  0.8301                                    +
  1  0.6583
+
  1  0.6920        +
```

Explanation of notation and interpretation of tables

2.3.4.3. Designs for electrical quantities

Standard cells

Banks of saturated standard cells that are nominally one volt are the basis for maintaining the unit of voltage in many laboratories.

Bias problem

It has been observed that potentiometer measurements of the difference between two saturated standard cells, connected in series opposition, are effected by a thermal emf which remains constant even when the direction of the circuit is reversed.

Designs for eliminating bias

A calibration design for comparing standard cells can be constructed to be left-right balanced so that:

- A constant bias, P, does not contaminate the estimates for the individual cells.
- P is estimated as the average of difference measurements.

Designs for electrical quantities

Designs are given for the following classes of electrical artifacts. These designs are left-right balanced and may be appropriate for artifacts other than electrical standards.

- Saturated standard reference cells
- Saturated standard test cells
- Zeners
- Resistors

Standard cells in a single box

Left-right balanced designs for comparing standard cells among themselves where the restraint is over all reference cells are listed below. These designs are **not** appropriate for assigning values to test cells.

Estimates for individual standard cells and the bias term, P, are shown under the heading, 'SOLUTION MATRIX'. These designs also have the advantage of requiring a change of connections to only one cell at a time.

1. Design for 3 standard cells
2. Design for 4 standard cells
3. Design for 5 standard cells
4. Design for 6 standard cells

Test cells Calibration designs for assigning values to test cells in a common environment on the basis of comparisons with reference cells with known values are shown below. The designs in this catalog are left-right balanced.

1. Design for 4 test cells and 4 reference cells
2. Design for 8 test cells and 8 reference cells

Zeners Increasingly, zeners are replacing saturated standard cells as artifacts for maintaining and disseminating the volt. Values are assigned to test zeners, based on a group of reference zeners, using calibration designs.

1. Design for 4 reference zeners and 2 test zeners
2. Design for 4 reference zeners and 3 test zeners

Standard resistors Designs for comparing standard resistors that are used for maintaining and disseminating the ohm are listed in this section.

1. Design for 3 reference resistors and 1 test resistor
2. Design for 4 reference resistors and 1 test resistor

2.3.4.3.1. Left-right balanced design for 3 standard cells

Design 1,1,1

```
                  CELLS
OBSERVATIONS    1   1   1
    Y(1)        +   -
    Y(2)        +       -
    Y(3)            +   -
    Y(4)        -   +
    Y(5)        -       +
    Y(6)            -   +

RESTRAINT       +   +   +

DEGREES OF FREEDOM  =  3
```

```
                      SOLUTION MATRIX
                        DIVISOR = 6
OBSERVATIONS       1      1      1      P
   Y(1)            1     -1      0      1
   Y(2)            1      0     -1      1
   Y(3)            0      1     -1      1
   Y(4)           -1      1      0      1
   Y(5)           -1      0      1      1
   Y(6)            0     -1      1      1
   R*              2      2      2      0

   R* = AVERAGE VALUE OF 3 REFERENCE CELLS

   P = LEFT-RIGHT BIAS
```

```
FACTORS FOR COMPUTING STANDARD DEVIATIONS
  V  FACTOR      CELLS
                1   1   1
  1  0.3333     +
  1  0.3333         +
  1  0.3333             +
```

Explanation of notation and interpretation of tables

2.3.4.3.2. Left-right balanced design for 4 standard cells

Design 1,1,1,1

```
OBSERVATIONS    1   1   1   1
    Y(1)        +   -
    Y(2)        +       -
    Y(3)            +   -
    Y(4)            +       -
    Y(5)                +   -
    Y(6)        -       +
    Y(7)            -   +
    Y(8)            -       +
    Y(9)        -           +
    Y(10)               -   +
    Y(11)       -   +
    Y(12)       +           -

RESTRAINT       +   +   +   +


DEGREES OF FREEDOM  =  8
```

SOLUTION MATRIX
DIVISOR = 8

OBSERVATIONS	1	1	1	1	P
Y(1)	1	-1	0	0	1
Y(2)	1	0	-1	0	1
Y(3)	0	1	-1	0	1
Y(4)	0	1	0	-1	1
Y(5)	0	0	1	-1	1
Y(6)	-1	0	1	0	1
Y(7)	0	-1	1	0	1
Y(8)	0	-1	0	1	1
Y(9)	-1	0	0	1	1
Y(10)	0	0	-1	1	1
Y(11)	-1	1	0	0	1
Y(12)	1	0	0	-1	1
R*	2	2	2	2	0

R* = AVERAGE VALUE OF 4 REFERENCE CELLS

P = LEFT-RIGHT BIAS

FACTORS FOR COMPUTING STANDARD DEVIATIONS

V	FACTOR	CELLS			
		1	1	1	1
1	0.3062	+			
1	0.3062		+		
1	0.3062			+	
1	0.3062				+

Explanation of notation and interpretation of tables

2.3.4.3.3. Left-right balanced design for 5 standard cells

Design 1,1,1,1,1

```
OBSERVATIONS     1   1   1   1   1

      Y(1)       +   -
      Y(2)       +       -
      Y(3)           +   -
      Y(4)           +       -
      Y(5)               +   -
      Y(6)               +       -
      Y(7)                   +   -
      Y(8)       -           +
      Y(9)       -                   +
      Y(10)          -               +

  RESTRAINT      +   +   +   +   +

  DEGREES OF FREEDOM  =  5
```

SOLUTION MATRIX
DIVISOR = 5

OBSERVATIONS	1	1	1	1	1	P
Y(1)	1	-1	0	0	0	1
Y(2)	1	0	-1	0	0	1
Y(3)	0	1	-1	0	0	1
Y(4)	0	1	0	-1	0	1
Y(5)	0	0	1	-1	0	1
Y(6)	0	0	1	0	-1	1
Y(7)	0	0	0	1	-1	1
Y(8)	-1	0	0	1	0	1
Y(9)	-1	0	0	0	1	1
Y(10)	0	-1	0	0	1	1
R*	1	1	1	1	1	0

R* = AVERAGE VALUE OF 5 REFERENCE CELLS

P = LEFT-RIGHT BIAS

FACTORS FOR COMPUTING REPEATABILITY STANDARD DEVIATIONS

V	FACTOR	CELLS 1	1	1	1	1
1	0.4000	+				
1	0.4000		+			
1	0.4000			+		
1	0.4000				+	
1	0.4000					+

Explanation of notation and interpretation of tables

2.3.4.3.4. Left-right balanced design for 6 standard cells

Design 1,1,1,1,1,1

```
                        CELLS
OBSERVATIONS     1   1   1   1   1   1
     Y(1)        +   -
     Y(2)        +       -
     Y(3)            +   -
     Y(4)            +       -
     Y(5)                +   -
     Y(6)                +       -
     Y(7)                    +   -
     Y(8)                    +       -
     Y(9)                        +   -
     Y(10)       -               +
     Y(11)       -                   +
     Y(12)           -               +
     Y(13)       +               -
     Y(14)           +               -
     Y(15)               +               -

 RESTRAINT       +   +   +   +   +   +

 DEGREES OF FREEDOM = 9
```

SOLUTION MATRIX
DIVISOR = 6

OBSERVATIONS	1	1	1	1	1	1
Y(1)	1	-1	0	0	0	0
Y(2)	1	0	-1	0	0	0
Y(3)	0	1	-1	0	0	0
Y(4)	0	1	0	-1	0	0
Y(5)	0	0	1	-1	0	0
Y(6)	0	0	1	0	-1	0
Y(7)	0	0	0	1	-1	0
Y(8)	0	0	0	1	0	-1
Y(9)	0	0	0	0	1	-1
Y(10)	-1	0	0	0	1	0
Y(11)	-1	0	0	0	0	1
Y(12)	0	-1	0	0	0	1

P
1
1
1
1
1
1
1
1
1
1

```
    Y(13)          1      0      0     -1      0      0
1
    Y(14)          0      1      0      0     -1      0
1
    Y(15)          0      0      1      0      0     -1
1
    R*             1      1      1      1      1      1
0

    R* = AVERAGE VALUE OF 6 REFERENCE CELLS

    P = LEFT-RIGHT BIAS


 FACTORS FOR COMPUTING STANDARD DEVIATIONS
  V  FACTOR            CELLS
                  1   1   1   1   1   1
  1  0.3727      +
  1  0.3727          +
  1  0.3727              +
  1  0.3727                  +
  1  0.3727                      +
  1  0.3727                          +
```

Explanation of notation and interpretation of tables

2.3.4.3.5. Left-right balanced design for 4 references and 4 test items

Design for 4 references and 4 test items.

```
 OBSERVATIONS    1   1   1   1   1   1   1   1

     Y(1)        +                   -
     Y(2)        +                           -
     Y(3)                +                   -
     Y(4)                +           -
     Y(5)            +                   -
     Y(6)            +                           -
     Y(7)                        +               -
     Y(8)                        +           -
     Y(9)        -                           +
     Y(10)       -                                   +
     Y(11)               -                           +
     Y(12)               -                   +
     Y(13)           -               +
     Y(14)           -                           +
     Y(15)                       -               +
     Y(16)                       -   +

 RESTRAINT       +   +   +   +

 DEGREES OF FREEDOM = 8
```

SOLUTION MATRIX
DIVISOR = 16

OBSERVATIONS	1	1	1	1	1	1	1	P
Y(1)	3	-1	-1	-1	-4	0	0	1
Y(2)	3	-1	-1	-1	0	0	-4	1
Y(3)	-1	-1	3	-1	0	0	-4	1
Y(4)	-1	-1	3	-1	-4	0	0	1
Y(5)	-1	3	-1	-1	0	-4	0	1
Y(6)	-1	3	-1	-1	0	0	0	1
Y(7)	-1	-1	-1	3	0	0	0	1
Y(8)	-1	-1	-1	3	0	-4	0	1
Y(9)	-3	1	1	1	0	4	0	1
Y(10)	-3	1	1	1	0	0	0	1
Y(11)	1	1	-3	1	0	0	0	1
Y(12)	1	1	-3	1	0	4	0	1

```
0      0     1
    Y(13)          1     -3      1      1      4      0
0      0     1
    Y(14)          1     -3      1      1      0      0
4      0     1
    Y(15)          1      1      1     -3      0      0
4      0     1
    Y(16)          1      1      1     -3      4      0
0      0     1
    R*             4      4      4      4      4      4
4      4     0

    R* = AVERAGE VALUE OF REFERENCE CELLS

    P = ESTIMATE OF LEFT-RIGHT BIAS


 FACTORS FOR COMPUTING STANDARD DEVIATIONS
  V  FACTORS              CELLS
                    1   1   1   1   1   1   1   1
  1  0.4330         +
  1  0.4330             +
  1  0.4330                 +
  1  0.4330                     +
  1  0.5000                         +
  1  0.5000                             +
  1  0.5000                                 +
  1  0.5000                                     +
```

[Explanation of notation and interpretation of tables]

2.3.4.3.6. Design for 8 references and 8 test items

Design for 8 references and 8 test items.

	TEST CELLS								REFERENCE		
CELLS											
OBSERVATIONS	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1							
Y(1)	+								-		
Y(2)	-									+	
Y(3)		-									+
Y(4)		+									
-											
Y(5)			+								
-											
Y(6)			-								
+											
Y(7)				-							
+											
Y(8)				+							
-											
Y(9)					+				-		
Y(10)						+				-	
Y(11)						-					+
Y(12)							-				
+											
Y(13)							+				
-											
Y(14)								+			
-											
Y(15)								-			
+											
Y(16)					-						
+											
RESTRAINT									+	+	+
+	+	+	+	+							

DEGREES OF FREEDOM = 0

SOLUTION MATRIX FOR TEST CELLS
DIVISOR = 16

OBSERVATIONS	1	1	1	1	1	1	1	1
Y(1)	8	4	0	-4	-6	6	2	-2
Y(2)	-8	4	0	-4	-6	6	2	-2
Y(3)	4	-8	-4	0	2	6	-6	-2
Y(4)	4	8	-4	0	2	6	-6	-2
Y(5)	0	-4	8	4	2	-2	-6	6
Y(6)	0	-4	-8	4	2	-2	-6	6
Y(7)	-4	0	4	-8	-6	-2	2	6
Y(8)	-4	0	4	8	-6	-2	2	6

```
2       6
    Y(9)           -6      -2       2       6       8      -4
0       4
    Y(10)          -6       6       2      -2      -4       8
4       0
    Y(11)          -6       6       2      -2      -4      -8
4       0
    Y(12)           2       6      -6      -2       0       4
-8      -4
    Y(13)           2       6      -6      -2       0       4
8      -4
    Y(14)           2      -2      -6       6       4       0
-4      8
    Y(15)           2      -2      -6       6       4       0
-4     -8
    Y(16)          -6      -2       2       6      -8      -4
0       4
    R               2       2       2       2       2       2
2       2

                          SOLUTION MATRIX FOR REFERENCE
CELLS
                                      DIVISOR = 16
 OBSERVATIONS       1       1       1       1       1       1
1       1      P

    Y(1)           -7       7       5       3       1      -1
-3      -5      1
    Y(2)           -7       7       5       3       1      -1
-3      -5      1
    Y(3)            3       5       7      -7      -5      -3
-1       1      1
    Y(4)            3       5       7      -7      -5      -3
-1       1      1
    Y(5)            1      -1      -3      -5      -7       7
5       3      1
    Y(6)            1      -1      -3      -5      -7       7
5       3      1
    Y(7)           -5      -3      -1       1       3       5
7      -7      1
    Y(8)           -5      -3      -1       1       3       5
7      -7      1
    Y(9)           -7      -5      -3      -1       1       3
5       7      1
    Y(10)          -5      -7       7       5       3       1
-1      -3      1
    Y(11)          -5      -7       7       5       3       1
-1      -3      1
    Y(12)           1       3       5       7      -7      -5
-3      -1      1
    Y(13)           1       3       5       7      -7      -5
-3      -1      1
    Y(14)           3       1      -1      -3      -5      -7
7       5      1
    Y(15)           3       1      -1      -3      -5      -7
7       5      1
    Y(16)          -7      -5      -3      -1       1       3
5       7      1
    R*              2       2       2       2       2       2
2       2      0

    R* = AVERAGE VALUE OF 8 REFERENCE CELLS

    P = ESTIMATE OF LEFT-RIGHT BIAS


FACTORS FOR COMPUTING STANDARD DEVIATIONS FOR TEST CELLS
  V  FACTORS           TEST CELLS
                  1   1   1   1   1   1   1   1
  1  1.1726       +
  1  1.1726           +
  1  1.1726               +
  1  1.1726                   +
  1  1.1726                       +
  1  1.1726                           +
  1  1.1726                               +
  1  1.1726                                   +
```

[Explanation of notation and interpretation of tables]

2.3.4.3.7. Design for 4 reference zeners and 2 test zeners

Design for 4 references zeners and 2 test zeners.

```
                        ZENERS
 OBSERVATIONS    1   1   1   1   1   1

      Y(1)       +                   -
      Y(2)       +                       -
      Y(3)           +               -
      Y(4)           +                   -
      Y(5)               +           -
      Y(6)               +               -
      Y(7)                   +       -
      Y(8)                   +           -
      Y(9)                   -       +
      Y(10)                  -           +
      Y(11)              -           +
      Y(12)              -               +
      Y(13)          -               +
      Y(14)          -                   +
      Y(15)      -                   +
      Y(16)      -                       +

 RESTRAINT       +   +   +   +

 CHECK STANDARD  +   -

 DEGREES OF FREEDOM  =  10
```

SOLUTION MATRIX
DIVISOR = 16

OBSERVATIONS	1	1	1	1	1	1	P
Y(1)	3	-1	-1	-1	-2	0	1
Y(2)	3	-1	-1	-1	0	-2	1
Y(3)	-1	3	-1	-1	-2	0	1
Y(4)	-1	3	-1	-1	0	-2	1
Y(5)	-1	-1	3	-1	-2	0	1
Y(6)	-1	-1	3	-1	0	-2	1
Y(7)	-1	-1	-1	3	-2	0	1
Y(8)	-1	-1	-1	3	0	-2	1
Y(9)	1	1	1	-3	2	0	1
Y(10)	1	1	1	-3	0	2	

1						
	Y(11)	1	1	-3	1	2	0
1							
	Y(12)	1	1	-3	1	0	2
1							
	Y(13)	1	-3	1	1	2	0
1							
	Y(14)	1	-3	1	1	0	2
1							
	Y(15)	-3	1	1	1	2	0
1							
	Y(16)	-3	1	1	1	0	2
1							
	R*	4	4	4	4	4	4
0							

R* = AVERAGE VALUE OF 4 REFERENCE STANDARDS

P = LEFT-RIGHT EFFECT

FACTORS FOR COMPUTING STANDARD DEVIATIONS

V	FACTORS	ZENERS					
		1	1	1	1	1	P
1	0.4330	+					
1	0.4330		+				
1	0.4330			+			
1	0.4330				+		
1	0.3536					+	
1	0.3536						+
1	0.2500						+

Explanation of notation and interpretation of tables

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

2.3.4.3.8. Design for 4 reference zeners and 3 test zeners

Design for 4 references and 3 test zeners.

```
                        ZENERS

 OBSERVATIONS    1    1    1    1    1    1    1

      Y(1)       -    +
      Y(2)       -         +
      Y(3)            +         -
      Y(4)       +                              -
      Y(5)       +                         -
      Y(6)       +                    -
      Y(7)            -               +
      Y(8)            -                    +
      Y(9)            -                         +
      Y(10)                -                     +
      Y(11)                -                +
      Y(12)                -          +
      Y(13)                     +     -
      Y(14)                     +          -
      Y(15)                     +               -
      Y(16)                +    -
      Y(17)           +         -
      Y(18)      -         +


 RESTRAINT       +    +    +    +

 CHECK STANDARD  +    -


 DEGREES OF FREEDOM = 11


                              SOLUTION MATRIX
                              DIVISOR  = 1260

 OBSERVATIONS      1       1       1       1       1       1
1     P


     Y(1)        -196     196     -56      56       0       0
0    70
     Y(2)        -160     -20     160      20       0       0
0    70
     Y(3)          20     160     -20    -160       0       0
0    70
     Y(4)         143     -53     -17     -73       0       0    -
315    70
     Y(5)         143     -53     -17     -73       0    -315
0    70
     Y(6)         143     -53     -17     -73    -315       0
0    70
     Y(7)          53    -143      73      17     315       0
0    70
     Y(8)          53    -143      73      17       0     315
0    70
```

```
    Y(9)          53   -143     73     17      0      0
315   70
    Y(10)         17     73   -143     53      0      0
315   70
    Y(11)         17     73   -143     53      0    315
0   70
    Y(12)         17     73   -143     53    315      0
0   70
    Y(13)        -73    -17    -53    143   -315      0
0   70
    Y(14)        -73    -17    -53    143      0   -315
0   70
    Y(15)        -73    -17    -53    143      0      0   -
315   70
    Y(16)         56    -56    196   -196      0      0
0   70
    Y(17)         20    160    -20   -160      0      0
0   70
    Y(18)       -160    -20    160     20      0      0
0   70
    R*           315    315    315    315    315    315
315   0
```

R* = Average value of the 4 reference zeners

P = left-right effect

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

V	K1	1	1	1	1	1	1	1
1	0.5000					+		
1	0.5000						+	
1	0.5000							+
2	0.7071					+	+	
3	0.8660					+	+	+
0	0.5578	+	-					

Explanation of notation and interpretation of tables

2.3.4.3.9. Design for 3 references and 1 test resistor

Design 1,1,1,1

```
 OBSERVATIONS    1    1    1    1

     Y(1)        +    -
     Y(2)        +         -
     Y(3)        +              -
     Y(4)        -              +
     Y(5)        -         +
     Y(6)        -    +

 RESTRAINT       +    +    +

 DEGREES OF FREEDOM = 3
```

SOLUTION MATRIX
DIVISOR = 6

OBSERVATIONS	1	1	1	1
Y(1)	1	-2	1	1
Y(2)	1	1	-2	1
Y(3)	0	0	0	-3
Y(4)	0	0	0	3
Y(5)	-1	-1	2	-1
Y(6)	-1	2	-1	-1
R	2	2	2	2

R = AVERAGE VALUE OF 3 REFERENCE RESISTORS

FACTORS FOR COMPUTING STANDARD DEVIATIONS

OHM FACTORS		RESISTORS			
		1	1	1	1
1	0.3333	+			
1	0.5270		+		
1	0.5270			+	
1	0.7817				+

Explanation of notation and interpretation of tables

2.3.4.3.10. Design for 4 references and 1 test resistor

Design 1,1,1,1,1

```
OBSERVATIONS    1    1    1    1    1

     Y(1)       +                   -
     Y(2)            +              -
     Y(3)                 +         -
     Y(4)                      +    -
     Y(5)                      -    +
     Y(6)                 -         +
     Y(7)            -              +
     Y(8)       -                   +

RESTRAINT       +    +    +    +

DEGREES OF FREEDOM   =       4


                          SOLUTION MATRIX
                            DIVISOR = 8

OBSERVATIONS       1       1       1       1       1

   Y(1)            3      -1      -1      -1      -1
   Y(2)           -1       3      -1      -1      -1
   Y(3)           -1      -1       3      -1      -1
   Y(4)           -1      -1      -1       3      -1
   Y(5)            1       1       1      -3       1
   Y(6)            1       1      -3       1       1
   Y(7)            1      -3       1       1       1
   Y(8)           -3       1       1       1       1
   R               2       2       2       2       2

   R = AVERAGE VALUE OF REFERENCE RESISTORS


FACTORS FOR COMPUTING STANDARD DEVIATIONS
OHM FACTORS
                1    1    1    1    1
 1   0.6124     +
 1   0.6124          +
 1   0.6124               +
 1   0.6124                    +
 1   0.3536                         +
```

Explanation of notation and interpretation of tables

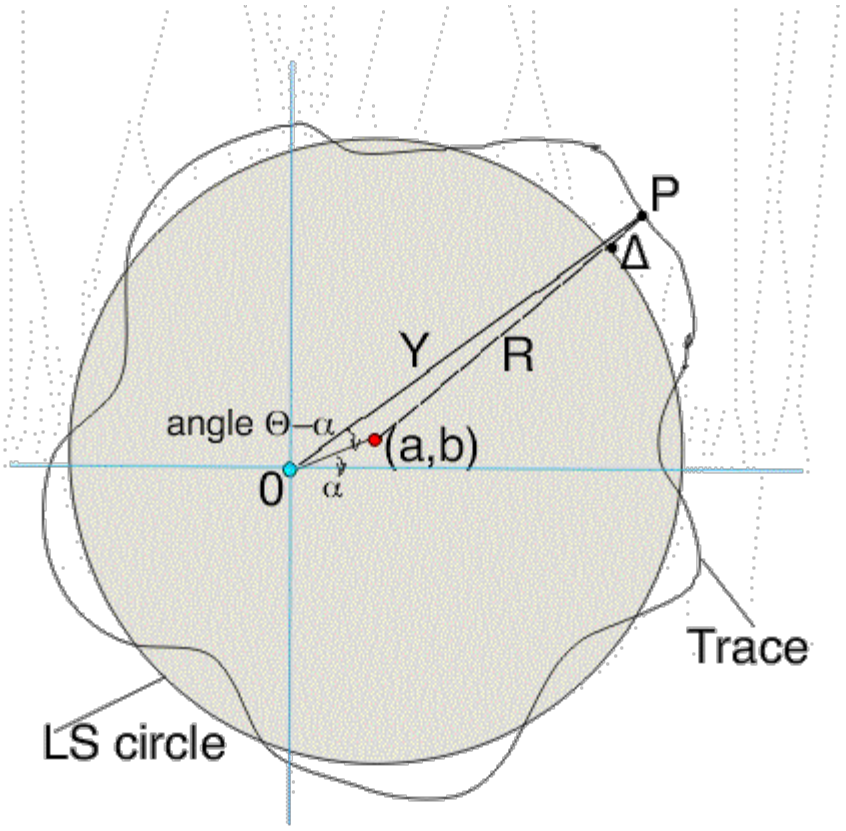
2.3.4.4. Roundness measurements

Roundness measurements

Measurements of roundness require 360° traces of the workpiece made with a turntable-type instrument or a stylus-type instrument. A least squares fit of points on the trace to a circle define the parameters of noncircularity of the workpiece. A diagram of the measurement method is shown below.

The diagram shows the trace and Y, the distance from the spindle center to the trace at the angle.

A least squares circle fit to data at equally spaced angles gives estimates of P - R, the noncircularity, where R = radius of the circle and P = distance from the center of the circle to the trace.



Low precision measurements

Some measurements of roundness do not require a high level of precision, such as measurements on cylinders, spheres, and ring gages where roundness is not of primary importance. For this purpose, a single trace is made of the workpiece.

Weakness of single trace method

The weakness of this method is that the deviations contain both the spindle error and the workpiece error, and these two errors cannot be separated with the single trace. Because the spindle error is usually small and within known limits, its effect can be ignored except when

the most precise measurements are needed.

High precision measurements

High precision measurements of roundness are appropriate where an object, such as a hemisphere, is intended to be used primarily as a roundness standard.

Measurement method

The measurement sequence involves making multiple traces of the roundness standard where the standard is rotated between traces. Least-squares analysis of the resulting measurements enables the noncircularity of the spindle to be separated from the profile of the standard.

Choice of measurement method

A synopsis of the measurement method and the estimation technique are given in this chapter for:

- Single-trace method
- Multiple-trace method

The reader is encouraged to obtain a copy of the publication on roundness (Reeve) for a more complete description of the measurement method and analysis.

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.4. Roundness measurements

2.3.4.4.1. Single-trace roundness design

Low precision measurements

Some measurements of roundness do not require a high level of precision, such as measurements on cylinders, spheres, and ring gages where roundness is not of primary importance. The diagram of the measurement method shows the trace and *Y*, the distance from the spindle center to the trace at the angle. A least-squares circle fit to data at equally spaced angles gives estimates of *P* - *R*, the noncircularity, where *R* = radius of the circle and *P* = distance from the center of the circle to the trace.

Single trace method

For this purpose, a single trace covering exactly 360° is made of the workpiece and measurements Y_i at angles θ_i of the distance between the center of the spindle and the trace, are made at

$$\theta_i \{i = 1, \cdots, N\}$$

equally spaced angles. A least-squares circle fit to the data gives the following estimators of the parameters of the circle.

$$\hat{R} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\hat{a} = \frac{2}{N}\sum_{i=1}^{N} Y_i \cos(\theta_i)$$

$$\hat{b} = \frac{2}{N}\sum_{i=1}^{N} Y_i \sin(\theta_i)$$

.

Noncircularity of workpiece

The deviation of the trace from the circle at angle θ_i, which defines the noncircularity of the workpiece, is estimated by:

$$\hat{\Delta} = Y_i - \hat{R} - \hat{a}\cos(\theta_i) - \hat{b}\sin(\theta_i)$$

Weakness of single trace method

The weakness of this method is that the deviations contain both the spindle error and the workpiece error, and these two errors cannot be separated with the single trace. Because the spindle error is usually small and within known limits, its effect can be ignored except when the most precise measurements are needed.

2.3.4.4.2. Multiple-trace roundness designs

High precision measurements

High precision roundness measurements are required when an object, such as a hemisphere, is intended to be used primarily as a roundness standard. The method outlined on this page is appropriate for either a turntable-type instrument or a spindle-type instrument.

Measurement method

The measurement sequence involves making multiple traces of the roundness standard where the standard is rotated between traces. Least-squares analysis of the resulting measurements enables the noncircularity of the spindle to be separated from the profile of the standard. The reader is referred to the publication on the subject (Reeve) for details covering measurement techniques and analysis.

Method of n traces

The number of traces that are made on the workpiece is arbitrary but should not be less than four. The workpiece is centered as well as possible under the spindle. The mark on the workpiece which denotes the zero angular position is aligned with the zero position of the spindle as shown in the graph. A trace is made with the workpiece in this position. The workpiece is then rotated clockwise by *360/n* degrees and another trace is made. This process is continued until *n* traces have been recorded.

Mathematical model for estimation

For $i = 1,...,n$, the ith angular position is denoted by

$$\theta_i = \frac{360(i-1)}{n}\text{deg}$$

Definition of terms relating to distances to the least squares circle

The deviation from the least squares circle (LSC) of the workpiece at the θ_i position is α_i.

The deviation of the spindle from its LSC at the α_i position is β_i.

Terms relating to

For the jth graph, let the three parameters that define the LSC be given by

parameters of least squares circle

$$R_j, a_j, b_j$$

defining the radius *R*, *a*, and *b* as shown in the graph. In an idealized measurement system these parameters would be constant for all *j*. In reality, each rotation of the workpiece causes it to shift a small amount vertically and horizontally. To account for this shift, separate parameters are needed for each trace.

Correction for obstruction to stylus

Let Y_{ij} be the observed distance (in polar graph units) from the center of the *j*th graph to the point on the curve that corresponds to the θ_i position of the spindle. If *K* is the magnification factor of the instrument in microinches/polar graph unit and δ is the angle between the lever arm of the stylus and the tangent to the workpiece at the point of contact (which normally can be set to zero if there is no obstruction), the transformed observations to be used in the estimation equations are:

$$Z_{ij} = K\cos(\delta)Y_{ij}$$

.

Estimates for parameters

The estimation of the individual parameters is obtained as a least-squares solution that requires six restraints which essentially guarantee that the sum of the vertical and horizontal deviations of the spindle from the center of the LSC are zero. The expressions for the estimators are as follows:

$$\hat{\alpha}_i = \sum_{j=1}^{n}\sum_{k=1}^{n} t_{k-i+j+n} Z_{kj}$$

$$\hat{\beta}_i = \sum_{j=1}^{n}\sum_{k=1}^{n} t_{k-i+1+n} Z_{kj}$$

$$\hat{R}_j = \sum_{k=1}^{n} Z_{kj}$$

$$\hat{a}_j = \frac{2}{n}\sum_{k=1}^{n} Z_{kj}\cos(\theta_k)$$

$$\hat{b}_j = \frac{2}{n}\sum_{k=1}^{n} Z_{kj}\sin(\theta_k)$$

where

$$t_m = \frac{n-3}{n^2}; m = 1$$

$$t_m = -\frac{1}{n^2}(1 + 2\cos(\theta_m)); 2 \le m \le n$$

Finally, the standard deviations of the profile estimators are given by:

$$s_{\alpha_j} = s_{\beta_j} = \frac{\sqrt{n-3}}{n} s; (i = 1, \cdots, n.)$$

Computation of standard deviation

The computation of the residual standard deviation of the fit requires, first, the computation of the predicted values,

$$\hat{Z}_{ij} = \hat{\alpha}_{i+j-1} - \hat{\beta}_i + \hat{R}_j + \hat{a}_j \cos(\theta_i) + \hat{b}_j \sin(\theta_i)$$

The residual standard deviation with $v = n*n - 5n + 6$ degrees of freedom is

$$s = \sqrt{\frac{1}{n^2 - 5n + 6} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(Z_{ij} - \hat{Z}_{ij} \right)^2}$$

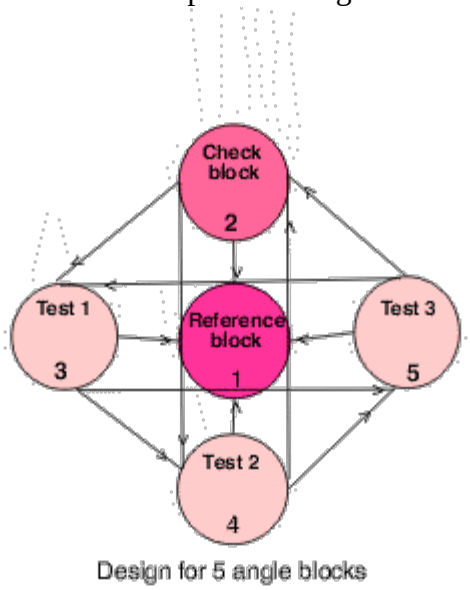
2.3.4.5. Designs for angle blocks

Purpose

The purpose of this section is to explain why calibration of angle blocks of the same size in groups is more efficient than calibration of angle blocks individually.

Calibration schematic for five angle blocks showing the reference as block 1 in the center of the diagram, the check standard as block 2 at the top; and the test blocks as blocks 3, 4, and 5.

A schematic of a calibration scheme for 1 reference block, 1 check standard, and three test blocks is shown below. The reference block, *R*, is shown in the center of the diagram and the check standard, *C*, is shown at the top of the diagram.



Design for 5 angle blocks

Block sizes

Angle blocks normally come in sets of

1, 3, 5, 20, and 30 seconds

1, 3, 5, 20, 30 minutes

1, 3, 5, 15, 30, 45 degrees

and blocks of the same nominal size from 4, 5 or 6 different sets can be calibrated simultaneously using one of the designs shown in this catalog.

- Design for 4 angle blocks

- Design for 5 angle blocks
- Design for 6 angle blocks

Restraint

The solution to the calibration design depends on the known value of a reference block, which is compared with the test blocks. The reference block is designated as block 1 for the purpose of this discussion.

Check standard

It is suggested that block 2 be reserved for a check standard that is maintained in the laboratory for quality control purposes.

Calibration scheme

A calibration scheme developed by Charles Reeve (Reeve) at the National Institute of Standards and Technology for calibrating customer angle blocks is explained on this page. The reader is encouraged to obtain a copy of the publication for details on the calibration setup and quality control checks for angle block calibrations.

Series of measurements for calibrating 4, 5, and 6 angle blocks simultaneously

For all of the designs, the measurements are made in groups of seven starting with the measurements of blocks in the following order: 2-3-2-1-2-4-2. Schematically, the calibration design is completed by counter-clockwise rotation of the test blocks about the reference block, one-at-a-time, with 7 readings for each series reduced to 3 difference measurements. For *n* angle blocks (including the reference block), this amounts to *n - 1* series of 7 readings. The series for 4, 5, and 6 angle blocks are shown below.

Measurements for 4 angle blocks

```
Series 1: 2-3-2-1-2-4-2
Series 2: 4-2-4-1-4-3-4
Series 3: 3-4-3-1-3-2-3
```

Measurements for 5 angle blocks (see diagram)

```
Series 1: 2-3-2-1-2-4-2
Series 2: 5-2-5-1-5-3-5
Series 3: 4-5-4-1-4-2-4
Series 4: 3-4-3-1-3-5-3
```

Measurements for 6 angle blocks

```
Series 1: 2-3-2-1-2-4-2
Series 2: 6-2-6-1-6-3-6
Series 3: 5-6-5-1-5-2-5
Series 4: 4-5-4-1-4-6-4
Series 5: 3-4-3-1-3-5-3
```

Equations for the measurements in the first series showing error sources

The equations explaining the seven measurements for the first series in terms of the errors in the measurement system are:

$$
\begin{aligned}
Z_{11} &= B + X_1 + \quad\; error_{11} \\
Z_{12} &= B + X_2 + \quad d + error_{12} \\
Z_{13} &= B + X_3 + 2d + error_{13} \\
Z_{14} &= B + X_4 + 3d + error_{14} \\
Z \quad &= B + X \quad + 4d + error
\end{aligned}
$$

$$Z_{16} = B + X_6^{15} + 5d + error_{16}^{15}$$
$$Z_{17} = B + X_7^{5} + 6d + error_{17}$$

with B a bias associated with the instrument, d is a linear drift factor, X is the value of the angle block to be determined; and the error terms relate to random errors of measurement.

Calibration procedure depends on difference measurements

The check block, C, is measured before and after each test block, and the difference measurements (which are not the same as the difference measurements for calibrations of mass weights, gage blocks, etc.) are constructed to take advantage of this situation. Thus, the 7 readings are reduced to 3 difference measurements for the first series as follows:

$$Y_{11} = (Z_{11} - 2Z_{12} + Z_{13})/2$$
$$Y_{12} = (Z_{13} - 2Z_{14} + Z_{15})/2$$
$$Y_{13} = (Z_{15} - 2Z_{16} + X_{17})/2$$

For all series, there are $3(n - 1)$ difference measurements, with the first subscript in the equations above referring to the series number. The difference measurements are free of drift and instrument bias.

Design matrix

As an example, the design matrix for $n = 4$ angle blocks is shown below.

1	1	1	1
0	1	-1	0
-1	1	0	0
0	1	0	-1
0	-1	0	1
-1	0	0	1
0	0	-1	1
0	0	1	-1
-1	0	1	0
0	-1	1	0

The design matrix is shown with the solution matrix for identification purposes only because the least-squares solution is weighted (Reeve) to account for the fact that test blocks are measured twice as many times as the reference block. The weight matrix is not shown.

Solutions to the calibration designs measurements

Solutions to the angle block designs are shown on the following pages. The solution matrix and factors for the repeatability standard deviation are to be interpreted as explained in solutions to calibration designs . As an example, the solution for the design for *n=4* angle blocks is as follows:

The solution for the reference standard is shown under the first column of the solution matrix; for the check standard under the second column; for the first test block under the third column; and for the second test block under the fourth column. Notice that the estimate for the reference block is guaranteed to be R^*, regardless

of the measurement results, because of the restraint that is imposed on the design. Specifically,

$$\hat{Reference} = 0 + R^*$$

$$\hat{Check} = \frac{1}{24}\begin{pmatrix} 2.272Y_{11} + 9.352Y_{12} + 2.272Y_{13} \\ -5.052Y_{21} + 7.324Y_{22} - 1.221Y_{23} \\ -1.221Y_{31} + 7.324Y_{32} - 5.052Y_{33} \end{pmatrix} + R^*$$

$$\hat{Test}_1 = \frac{1}{24}\begin{pmatrix} -5.052Y_{11} + 7.324Y_{12} - 1.221Y_{13} \\ -1.221Y_{21} + 7.324Y_{22} - 5.052Y_{23} \\ +2.272Y_{31} + 9.352Y_{32} + 2.272Y_{33} \end{pmatrix} + R^*$$

$$\hat{Test}_2 = \frac{1}{24}\begin{pmatrix} -1.221Y_{11} + 7.324Y_{12} - 5.052Y_{13} \\ +2.272Y_{21} + 9.352Y_{22} + 2.272Y_{23} \\ -5.052Y_{31} + 7.324Y_{32} - 1.221Y_{33} \end{pmatrix} + R^*$$

Solutions are correct only for the restraint as shown.

Calibrations can be run for top and bottom faces of blocks

The calibration series is run with the blocks all face "up" and is then repeated with the blocks all face "down", and the results averaged. The difference between the two series can be large compared to the repeatability standard deviation, in which case a between-series component of variability must be included in the calculation of the standard deviation of the reported average.

Calculation of standard deviations when the blocks are measured in two orientations

For *n* blocks, the differences between the values for the blocks measured in the top (denoted by "t") and bottom (denoted by "b") positions are denoted by:

$$\delta_i = \hat{X}_i^t - \hat{X}_i^t \qquad \text{for } i = I, ..., n$$

The standard deviation of the average (for each block) is calculated from these differences to be:

$$s_{avg\ test} = \sqrt{\frac{1}{4(n-1)}\sum_{i=1}^{n}\delta_i^2}$$

Standard deviations when the blocks are measured in only one

If the blocks are measured in only one orientation, there is no way to estimate the between-series component of variability and the standard deviation for the value of each block is computed as

$$s_{test} = K_1 s_1$$

orientation	where K_1 is shown under "Factors for computing repeatability standard deviations" for each design and s_1 is the repeatability standard deviation as estimated from the design. Because this standard deviation may seriously underestimate the uncertainty, a better approach is to estimate the standard deviation from the data on the check standard over time. An expanded uncertainty is computed according to the ISO guidelines.

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.5. Designs for angle blocks

2.3.4.5.1. Design for 4 angle blocks

DESIGN MATRIX

	1	1	1	1
Y(1)	0	1	-1	0
Y(2)	-1	1	0	0
Y(3)	0	1	0	-1
Y(4)	0	-1	0	1
Y(5)	-1	0	0	1
Y(6)	0	0	-1	1
Y(7)	0	0	1	-1
Y(8)	-1	0	1	0
Y(9)	0	-1	1	0

REFERENCE +

CHECK STANDARD +

DEGREES OF FREEDOM = 6

SOLUTION MATRIX
DIVISOR = 24

OBSERVATIONS	1	1	1	1
Y(11)	0	2.2723000	-5.0516438	-1.2206578
Y(12)	0	9.3521166	7.3239479	7.3239479
Y(13)	0	2.2723000	-1.2206578	-5.0516438
Y(21)	0	-5.0516438	-1.2206578	2.2723000
Y(22)	0	7.3239479	7.3239479	9.3521166
Y(23)	0	-1.2206578	-5.0516438	2.2723000
Y(31)	0	-1.2206578	2.2723000	-5.0516438
Y(32)	0	7.3239479	9.3521166	7.3239479
Y(33)	0	-5.0516438	2.2723000	-1.2206578
R*	1	1.	1.	1.

R* = VALUE OF REFERENCE ANGLE BLOCK

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

SIZE	K1	1	1	1	1
1	0.0000	+			
1	0.9749		+		
1	0.9749			+	
1	0.9749				+
1	0.9749		+		

Explanation of notation and interpretation of tables

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.5. Designs for angle blocks

2.3.4.5.2. Design for 5 angle blocks

DESIGN MATRIX

1	1	1	1	1
0	1	-1	0	0
-1	1	0	0	0
0	1	0	-1	0
0	-1	0	0	1
-1	0	0	0	1
0	0	-1	0	1
0	0	0	1	-1
-1	0	0	1	0
0	-1	0	1	0
0	0	1	-1	0
-1	0	1	0	0
0	0	1	0	-1

REFERENCE +

CHECK STANDARD +

DEGREES OF FREEDOM = 8

SOLUTION MATRIX
DIVISOR = 24

OBSERVATIONS	1	1	1	1	1
Y(11)	0.00000	3.26463	-5.48893	-0.21200	-1.56370
Y(12)	0.00000	7.95672	5.38908	5.93802	4.71618
Y(13)	0.00000	2.48697	-0.89818	-4.80276	-0.78603
Y(21)	0.00000	-5.48893	-0.21200	-1.56370	3.26463
Y(22)	0.00000	5.38908	5.93802	4.71618	7.95672
Y(23)	0.00000	-0.89818	-4.80276	-0.78603	2.48697
Y(31)	0.00000	-0.21200	-1.56370	3.26463	-5.48893
Y(32)	0.00000	5.93802	4.71618	7.95672	5.38908
Y(33)	0.00000	-4.80276	-0.78603	2.48697	-0.89818
Y(41)	0.00000	-1.56370	3.26463	-5.48893	-0.21200
Y(42)	0.00000	4.71618	7.95672	5.38908	5.93802
Y(43)	0.00000	-0.78603	2.48697	-0.89818	-4.80276
R*	1.	1.	1.	1.	1.

R* = VALUE OF REFERENCE ANGLE BLOCK

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

SIZE	K1	1	1	1	1	1
1	0.0000	+				
1	0.7465		+			
1	0.7465			+		
1	0.7456				+	
1	0.7456					+
1	0.7465		+			

Explanation of notation and interpretation of tables

2. Measurement Process Characterization
2.3. Calibration
2.3.4. Catalog of calibration designs
2.3.4.5. Designs for angle blocks

2.3.4.5.3. Design for 6 angle blocks

DESIGN MATRIX

1	1	1	1	1	1
0	1	-1	0	0	0
-1	1	0	0	0	0
0	1	0	-1	0	0
0	-1	0	0	0	1
-1	0	0	0	0	1
0	0	-1	0	0	1
0	0	0	0	1	-1
-1	0	0	0	1	0
0	-1	0	0	1	0
0	0	0	1	-1	0
-1	0	0	1	0	0
0	0	0	1	0	-1
0	0	1	-1	0	0
-1	0	1	0	0	0
0	0	1	0	-1	0

REFERENCE +

CHECK STANDARD +

DEGREES OF FREEDOM = 10

SOLUTION MATRIX
DIVISOR = 24

OBSERVATIONS	1	1	1	1	1
Y(11)	0.0000	3.2929	-5.2312	-0.7507	-0.6445
-0.6666					
Y(12)	0.0000	6.9974	4.6324	4.6495	3.8668
3.8540					
Y(13)	0.0000	3.2687	-0.7721	-5.2098	-0.6202
-0.6666					
Y(21)	0.0000	-5.2312	-0.7507	-0.6445	-0.6666
3.2929					
Y(22)	0.0000	4.6324	4.6495	3.8668	3.8540
6.9974					
Y(23)	0.0000	-0.7721	-5.2098	-0.6202	-0.6666
3.2687					
Y(31)	0.0000	-0.7507	-0.6445	-0.6666	3.2929
-5.2312					
Y(32)	0.0000	4.6495	3.8668	3.8540	6.9974
4.6324					
Y(33)	0.0000	-5.2098	-0.6202	-0.6666	3.2687
-0.7721					
Y(41)	0.0000	-0.6445	-0.6666	3.2929	-5.2312
-0.7507					
Y(42)	0.0000	3.8668	3.8540	6.9974	4.6324
4.6495					
Y(43)	0.0000	-0.6202	-0.6666	3.2687	-0.7721
-5.2098					
Y(51)	0.0000	-0.6666	3.2929	-5.2312	-0.7507
-0.6445					

```
  Y(52)      0.0000     3.8540     6.9974     4.6324     4.6495
3.8668
  Y(53)      0.0000    -0.6666     3.2687    -0.7721    -5.2098
-0.6202
  R*         1.         1.         1.         1.         1.
1.

  R* = VALUE OF REFERENCE ANGLE BLOCK


 FACTORS FOR REPEATABILITY STANDARD DEVIATIONS
 SIZE  K1
                1   1   1   1   1   1
  1  0.0000     +
  1  0.7111         +
  1  0.7111             +
  1  0.7111                 +
  1  0.7111                     +
  1  0.7111                         +
  1  0.7111         +
```

Explanation of notation and interpretation of tables

2.3.4.6. Thermometers in a bath

Measurement sequence

Calibration of liquid in glass thermometers is usually carried out in a controlled bath where the temperature in the bath is increased steadily over time to calibrate the thermometers over their entire range. One way of accounting for the temperature drift is to measure the temperature of the bath with a standard resistance thermometer at the beginning, middle and end of each run of K test thermometers. The test thermometers themselves are measured twice during the run in the following time sequence:

$$R_1, T_1, T_2, \cdots, T_K, R_2, T'_K, \cdots, T'_2, T'_1, R_3$$

where R_1, R_2, R_3 represent the measurements on the standard resistance thermometer and T_1, T_2, ... , T_K and T'_1, T'_2, ... , T'_K represent the pair of measurements on the K test thermometers.

Assumptions regarding temperature

The assumptions for the analysis are that:

- Equal time intervals are maintained between measurements on the test items.
- Temperature increases by ΔT with each interval.
- A temperature change of ϕ is allowed for the reading of the resistance thermometer in the middle of the run.

Indications for test thermometers

It can be shown (Cameron and Hailes) that the average reading for a test thermometer is its indication at the temperature implied by the average of the three resistance readings. The standard deviation associated with this indication is calculated from difference readings where

$$d_i = T_i - T'_i$$

is the difference for the ith thermometer. This difference is an estimate of $\phi + 2(K - i)\Delta$.

Estimates of drift

The estimates of the shift due to the resistance thermometer and temperature drift are given by:

$$\hat{\phi} = \frac{2}{K(K+1)}\left(3\sum_{i=1}^{K} id_i - (K+1)\sum_{i=1}^{K} d_i\right)$$

$$\hat{\Delta} = \frac{3}{K(K^2-1)}\left((K+1)\sum_{i=1}^{K} d_i - 2\sum_{i=1}^{K} id_i\right)$$

Standard deviations

The residual variance is given by

$$s^2 = \frac{1}{(K-2)}\sum_{i=1}^{K}\left(d_i - \hat{\phi} - 2(K-i)\hat{\Delta}\right)^2 .$$

The standard deviation of the indication assigned to the ith test thermometer is

$$s_{test} = \frac{s}{\sqrt{2}}$$

and the standard deviation for the estimates of shift and drift are

$$s_{\phi} = \frac{\sqrt{2}(2K-1)}{K(K+1)}s$$

$$s_{\Delta} = \frac{\sqrt{3}}{K(K^2-1)}s$$

respectively.

2.3.4.7. Humidity standards

Humidity standards

The calibration of humidity standards usually involves the comparison of reference weights with cylinders containing moisture. The designs shown in this catalog are drift-eliminating and may be suitable for artifacts other than humidity cylinders.

List of designs

- 2 reference weights and 3 cylinders

2.3.4.7.1. Drift-elimination design for 2 reference weights and 3 cylinders

```
OBSERVATIONS    1   1   1   1   1

    Y(1)        +   -
    Y(2)                    +   -
    Y(3)                +   -
    Y(4)            +   -
    Y(5)        -                   +
    Y(6)        -               +
    Y(7)                +           -
    Y(8)            +           -
    Y(9)            -               +
    Y(10)       +       -

RESTRAINT       +   +

CHECK STANDARD  +   -

DEGREES OF FREEDOM = 6
```

SOLUTION MATRIX
DIVISOR = 10

OBSERVATIONS	1	1	1	1	1
Y(1)	2	-2	0	0	0
Y(2)	0	0	0	2	-2
Y(3)	0	0	2	-2	0
Y(4)	-1	1	-3	-1	-1
Y(5)	-1	1	1	1	3
Y(6)	-1	1	1	3	1
Y(7)	0	0	2	0	-2
Y(8)	-1	1	-1	-3	-1
Y(9)	1	-1	1	1	3
Y(10)	1	-1	-3	-1	-1
R*	5	5	5	5	5

R* = average value of the two reference weights

FACTORS FOR REPEATABILITY STANDARD DEVIATIONS

```
WT    K1        1   1   1   1   1
 1  0.5477                    +
 1  0.5477                +
 1  0.5477            +
 2  0.8944            +   +
 3  1.2247            +   +   +
 0  0.6325    +   -
```

Explanation of notation and interpretation of tables

2.3.5. Control of artifact calibration

Purpose

The purpose of statistical control in the calibration process is to guarantee the 'goodness' of calibration results within predictable limits and to validate the statement of uncertainty of the result. Two types of control can be imposed on a calibration process that makes use of statistical designs:

1. Control of instrument precision or short-term variability
2. Control of bias and long-term variability
    - Example of a Shewhart control chart
    - Example of an EWMA control chart

Short-term standard deviation

The short-term standard deviation from each design is the basis for controlling instrument precision. Because the measurements for a single design are completed in a short time span, this standard deviation estimates the basic precision of the instrument. Designs should be chosen to have enough measurements so that the standard deviation from the design has at least 3 degrees of freedom where the degrees of freedom are $(n - m + 1)$ with

- n = number of difference measurements
- m = number of artifacts.

Check standard

Measurements on a check standard provide the mechanism for controlling the bias and long-term variability of the calibration process. The check standard is treated as one of the test items in the calibration design, and its value as computed from each calibration run is the basis for accepting or rejecting the calibration. All designs cataloged in this Handbook have provision for a check standard.

The check standard should be of the same type and geometry as items that are measured in the designs. These artifacts must be stable and available to the calibration process on a continuing basis. There should be a check standard at each critical level of measurement. For example, for mass calibrations there should be check standards at the 1 kg; 100 g, 10 g, 1 g, 0.1 g levels, etc. For gage blocks, there should be check standards at all nominal lengths.

A check standard can also be a mathematical construction,

such as the computed difference between the calibrated values of two reference standards in a design.

Database of check standard values

The creation and maintenance of the database of check standard values is an important aspect of the control process. The results from each calibration run are recorded in the database. The best way to record this information is in one file with one line (row in a spreadsheet) of information in fixed fields for each calibration run. A list of typical entries follows:

1. Date
2. Identification for check standard
3. Identification for the calibration design
4. Identification for the instrument
5. Check standard value
6. Repeatability standard deviation from design
7. Degrees of freedom
8. Operator identification
9. Flag for out-of-control signal
10. Environmental readings (if pertinent)

2.3.5.1. Control of precision

Control parameters from historical data

A modified control chart procedure is used for controlling instrument precision. The procedure is designed to be implemented in real time after a baseline and control limit for the instrument of interest have been established from the database of short-term standard deviations. A separate control chart is required for each instrument -- except where instruments are of the same type with the same basic precision, in which case they can be treated as one.

The baseline is the process standard deviation that is pooled from $k = 1, ..., K$ individual repeatability standard deviations, s_k , in the database, each having ν_k degrees of freedom. The pooled repeatability standard deviation is

$$s_1 = \sqrt{\frac{1}{\nu}\sum_{k=1}^{K}\nu_k s_k^2}$$

with degrees of freedom

$$\nu = \sum_{k=1}^{K}\nu_k \quad .$$

Control procedure is invoked in real-time for each calibration run

The control procedure compares each new repeatability standard deviation that is recorded for the instrument with an upper control limit, *UCL*. Usually, only the upper control limit is of interest because we are primarily interested in detecting degradation in the instrument's precision. A possible complication is that the control limit is dependent on the degrees of freedom in the new standard deviation and is computed as follows:

$$UCL = s_1\sqrt{F_{\alpha,\nu_{new},\nu}} \quad .$$

The quantity under the radical is the upper α percentage point from the *F* table where α is chosen small to be, say, 0.05. The other two terms refer to the degrees of freedom in the new standard deviation and the degrees of freedom in the process standard deviation.

Limitation of graphical method

The graphical method of plotting every new estimate of repeatability on a control chart does not work well when the *UCL* can change with each calibration design, depending on the degrees of freedom. The algebraic equivalent is to test if the new standard deviation exceeds its control limit, in which case the short-term precision is judged to be out of control and the current calibration run is rejected. For more guidance, see Remedies and strategies for dealing with out-of-control signals.

As long as the repeatability standard deviations are in control, there is reason for confidence that the precision of the instrument has not degraded.

Case study: Mass balance precision

It is recommended that the repeatability standard deviations be plotted against time on a regular basis to check for gradual degradation in the instrument. Individual failures may not trigger a suspicion that the instrument is in need of adjustment or tuning.

2.3.5.1.1. Example of control chart for precision

Example of a control chart for precision of a mass balance

Mass calibrations usually start with the comparison of kilograms standards using a high precision balance as a comparator. Many of the measurements at the kilogram level that were made at NIST between 1975 and 1989 were made on balance #12 using a 1,1,1,1 calibration design. The redundancy in the calibration design produces estimates for the individual kilograms and a repeatability standard deviation with three degrees of freedom for each calibration run. These standard deviations estimate the precision of the balance.

Need for monitoring precision

The precision of the balance is monitored to check for:

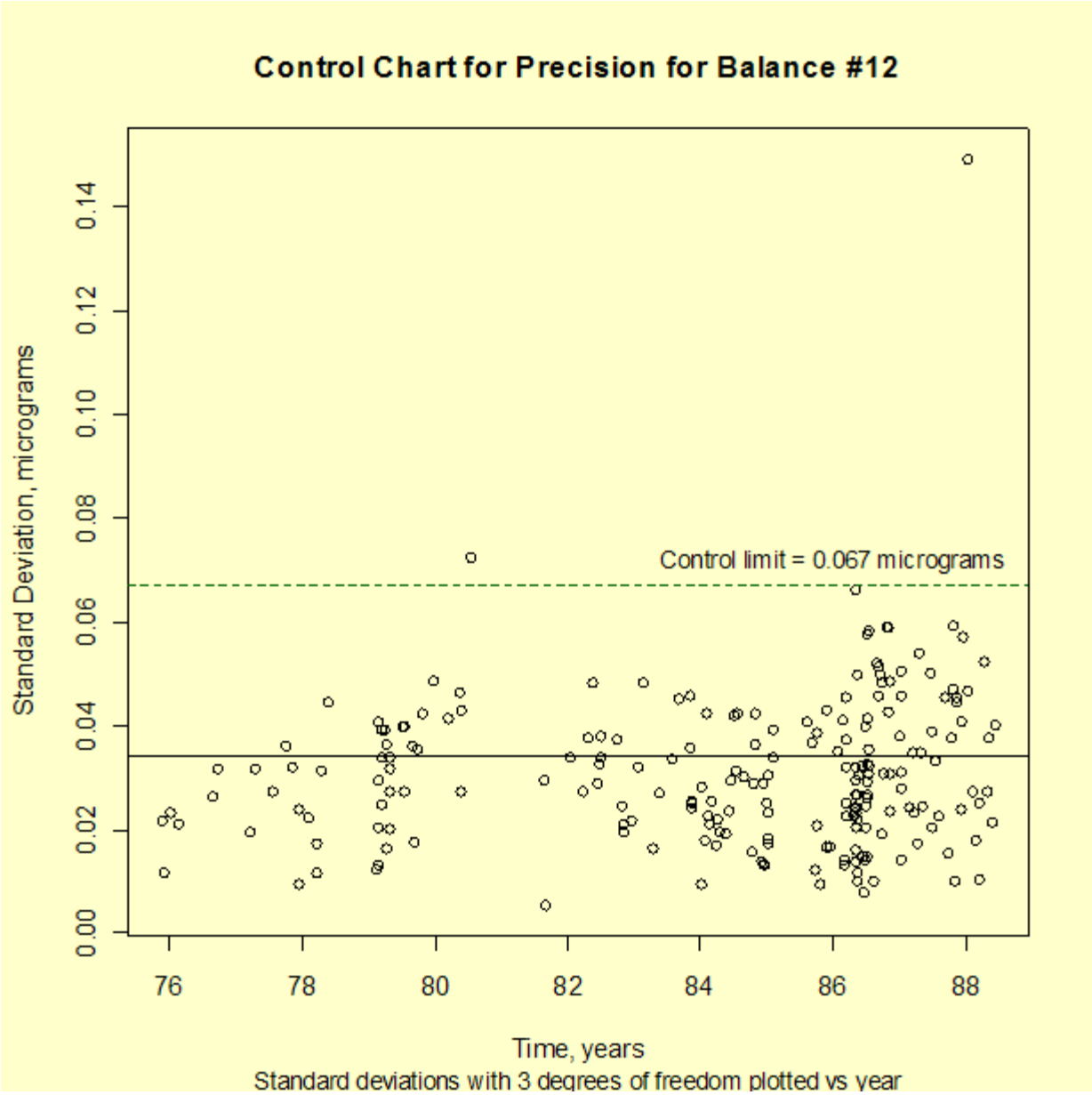
1. Slow degradation in the balance
2. Anomalous behavior at specific times

Monitoring technique for standard deviations

The standard deviations over time and many calibrations are tracked and monitored using a control chart for standard deviations. The database and control limits are updated on a yearly or bi-yearly basis and standard deviations for each calibration run in the next cycle are compared with the control limits. In this case, the standard deviations from 117 calibrations between 1975 and 1985 were pooled to obtain a repeatability standard deviation with v = 3*117 = 351 degrees of freedom, and the control limits were computed at the 1 % significance level.

Control chart for precision

The following control chart for precision for balance #12 can be generated using both Dataplot code and R code.



Interpretation of the control chart

The control chart shows that the precision of the balance remained in control through the first five months of 1988 with only two violations of the control limits. For those occasions, the calibrations were discarded and repeated. Clearly, for the second violation, something significant occurred that invalidated the calibration results.

Further interpretation of the control chart

However, it is also clear from the pattern of standard deviations over time that the precision of the balance was gradually degrading and more and more points were approaching the control limits. This finding led to a decision to replace this balance for high accuracy calibrations.

2. Measurement Process Characterization
2.3. Calibration
2.3.5. Control of artifact calibration

2.3.5.2. Control of bias and long-term variability

Control parameters are estimated using historical data

A control chart procedure is used for controlling bias and long-term variability. The procedure is designed to be implemented in real time after a baseline and control limits for the check standard of interest have been established from the database of check standard values. A separate control chart is required for each check standard. The control procedure outlined here is based on a Shewhart control chart with upper and lower control limits that are symmetric about the average. The EWMA control procedure that is sensitive to small changes in the process is discussed on another page.

For a Shewhart control procedure, the average and standard deviation of historical check standard values are the parameters of interest

The check standard values are denoted by

$$C_k \, (k=1, \cdots, K)$$

The baseline is the process average which is computed from the check standard values as

$$\bar{C} = \frac{1}{K} \sum_{k=1}^{K} C_k$$

The process standard deviation is

$$s_2 = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (C_k - \bar{C})^2}$$

with K - 1 degrees of freedom.

The control limits depend on the t distribution and the degrees of freedom in the process standard deviation

If $\bar{C}$ has been computed from historical data, the upper and lower control limits are:

$$UCL = \overline{C} + t_{1-\alpha/2,\, K-1} \cdot s_2$$

$$LCL = \overline{C} - t_{1-\alpha/2,\, K-1} \cdot s_2$$

where $t_{1\text{-}\alpha/2,\, K\text{-}1}$ denotes the 1-α/2 critical value from the *t*

table with $v = K - 1$ degrees of freedom.

Sample code

Sample code for computing the t value for a conservative case where $\alpha = 0.05$, $J = 6$, and $K = 6$, is available for both Dataplot and R.

Simplification for large degrees of freedom

It is standard practice to use a value of 3 instead of a critical value from the t table, given the process standard deviation has large degrees of freedom, say, $v > 15$.

The control procedure is invoked in real-time and a failure implies that the current calibration should be rejected

The control procedure compares the check standard value, C, from each calibration run with the upper and lower control limits. This procedure should be implemented in real time and does not necessarily require a graphical presentation. The check standard value can be compared algebraically with the control limits. The calibration run is judged to be out-of-control if either:

$$C > UCL$$

or

$$C < LCL$$

Actions to be taken

If the check standard value exceeds one of the control limits, the process is judged to be out of control and the current calibration run is rejected. The best strategy in this situation is to repeat the calibration to see if the failure was a chance occurrence. Check standard values that remain in control, especially over a period of time, provide confidence that no new biases have been introduced into the measurement process and that the long-term variability of the process has not changed.

Out-of-control signals that recur require investigation

Out-of-control signals, particularly if they recur, can be symptomatic of one of the following conditions:

- Change or damage to the reference standard(s)
- Change or damage to the check standard
- Change in the long-term variability of the calibration process

For more guidance, see Remedies and strategies for dealing with out-of-control signals.

Caution - be sure to plot the data

If the tests for control are carried out algebraically, it is recommended that, at regular intervals, the check standard values be plotted against time to check for drift or anomalies in the measurement process.

2.3.5.2.1. Example of Shewhart control chart for mass calibrations

Example of a control chart for mass calibrations at the kilogram level

Mass calibrations usually start with the comparison of four kilogram standards using a high precision balance as a comparator. Many of the measurements at the kilogram level that were made at NIST between 1975 and 1989 were made on balance #12 using a 1,1,1,1 calibration design. The restraint for this design is the known average of two kilogram reference standards. The redundancy in the calibration design produces individual estimates for the two test kilograms and the two reference standards.

Check standard

There is no slot in the 1,1,1,1 design for an artifact check standard when the first two kilograms are reference standards; the third kilogram is a test weight; and the fourth is a summation of smaller weights that act as the restraint in the next series. Therefore, the check standard is a computed difference between the values of the two reference standards as estimated from the design. The convention with mass calibrations is to report the correction to nominal, in this case the correction to 1000 g, as shown in the control charts below.

Need for monitoring

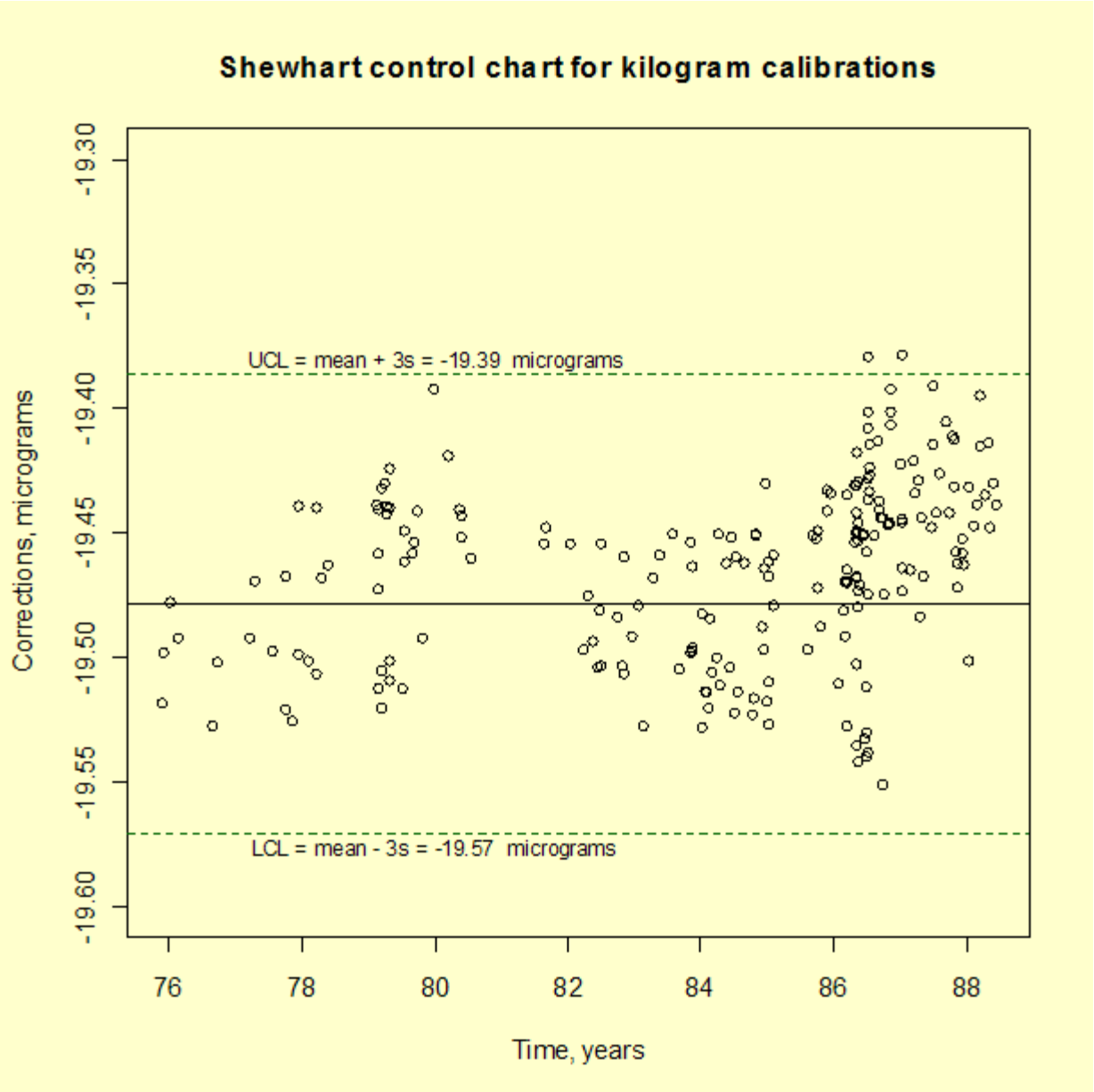
The kilogram check standard is monitored to check for:

1. Long-term degradation in the calibration process
2. Anomalous behavior at specific times

Monitoring technique for check standard values

Check standard values over time and many calibrations are tracked and monitored using a Shewhart control chart. The database and control limits are updated when needed and check standard values for each calibration run in the next cycle are compared with the control limits. In this case, the values from 117 calibrations between 1975 and 1985 were averaged to obtain a baseline and process standard deviation with $\nu = 116$ degrees of freedom. Control limits are computed with a factor of $k = 3$ to identify truly anomalous data points.

Control chart of kilogram check standard measurements showing a change in the process after 1985



Interpretation of the control chart

The control chart shows only two violations of the control limits. For those occasions, the calibrations were discarded and repeated. The configuration of points is unacceptable if many points are close to a control limit and there is an unequal distribution of data points on the two sides of the control chart -- indicating a change in either:

- process average which may be related to a change in the reference standards

or

- variability which may be caused by a change in the instrument precision or may be the result of other factors on the measurement process.

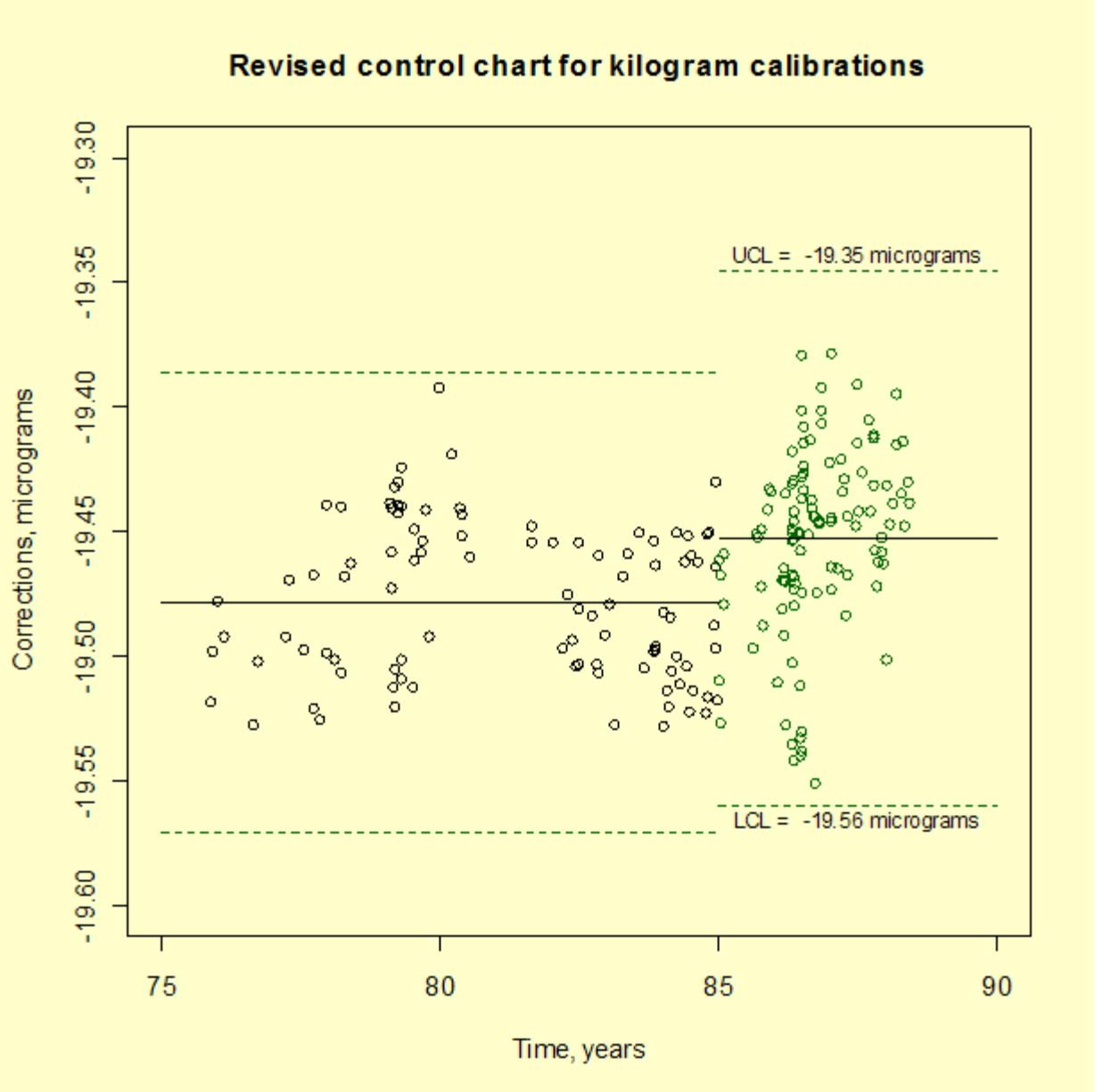
Small changes only become obvious over time

Unfortunately, it takes time for the patterns in the data to emerge because individual violations of the control limits do not necessarily point to a permanent shift in the process. The Shewhart control chart is not powerful for detecting small changes, say of the order of at most one standard deviation, which appears to be approximately the case in this application. This level of change might seem insignificant, but the calculation of uncertainties for the calibration process depends on the control limits.

Re-establishing the limits based on recent data and EWMA option

If the limits for the control chart are re-calculated based on the data after 1985, the extent of the change is obvious. Because the exponentially weighted moving average (EWMA) control chart is capable of detecting small changes, it may be a better choice for a high precision process that is producing many control values.

Revised control chart based on check standard measurements after 1985



Sample code

The original and revised Shewhart control charts can be generated using both Dataplot code and R code.

2.3.5.2.2. Example of EWMA control chart for mass calibrations

Small changes only become obvious over time

Unfortunately, it takes time for the patterns in the data to emerge because individual violations of the control limits do not necessarily point to a permanent shift in the process. The Shewhart control chart is not powerful for detecting small changes, say of the order of at most one standard deviation, which appears to be the case for the calibration data shown on the previous page. The EWMA (exponentially weighted moving average) control chart is better suited for this purpose.

Explanation of EWMA statistic at the kilogram level

The exponentially weighted moving average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time from the current measurement. The EWMA statistic at time *t* is computed recursively from individual data points which are ordered in time to be

$$Y_1, Y_2, \cdots, Y_t$$

where the first EWMA statistic is the average of historical data.

$$EWMA_{t+1} = \lambda Y_t + (1-\lambda) EWMA_t$$

Control mechanism for EWMA

The EWMA control chart can be made sensitive to small changes or a gradual drift in the process by the choice of the weighting factor, λ. A weighting factor between 0.2 - 0.3 has been suggested for this purpose (Hunter), and 0.15 is another popular choice.

Limits for the control chart

The target or center line for the control chart is the average of historical data. The upper (UCL) and lower (LCL) limits are

$$UCL = EWMA_1 + sk\sqrt{\frac{\lambda}{2-\lambda}}$$

$$LCL = EWMA_1 - sk\sqrt{\frac{\lambda}{2-\lambda}}$$

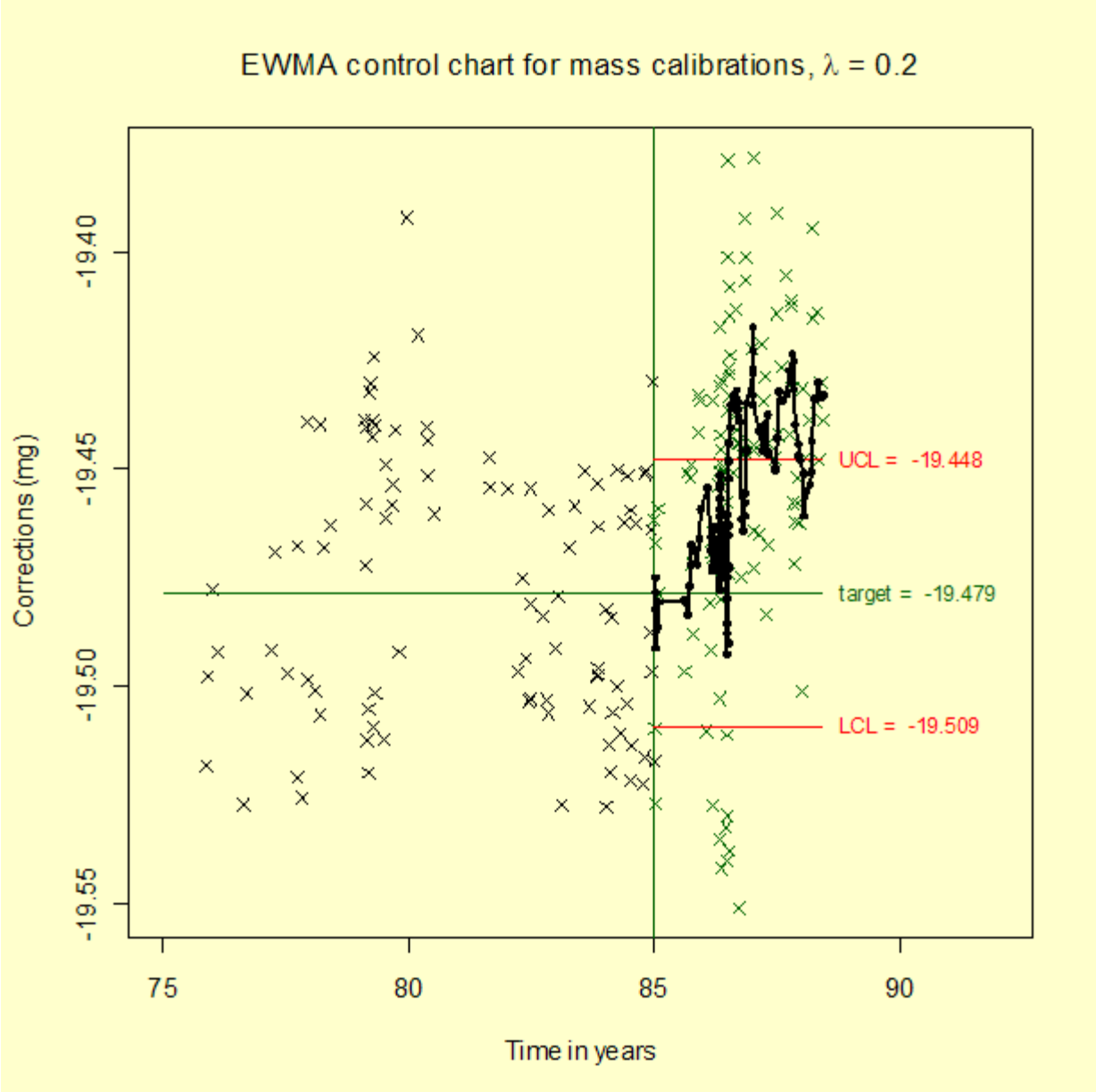
where *s* is the standard deviation of the historical data; the function under the radical is a good approximation to the component of the standard deviation of the EWMA statistic that is a function of time; and *k* is the multiplicative factor, defined in the same manner as for the Shewhart control chart, which is usually taken to be 3.

Example of EWMA chart for check

The target (average) and process standard deviation are computed from the check standard data taken prior to 1985. The computation of the EWMA statistic begins with the data taken at the start of 1985. In the control chart below, the control data after 1985 are shown in green,

standard data for kilogram calibrations showing multiple violations of the control limits for the EWMA statistics

and the EWMA statistics are shown as black dots superimposed on the raw data. The control limits are calculated according to the equation above where the process standard deviation, s = 0.03065 mg and k = 3. The EWMA statistics, and not the raw data, are of interest in looking for out-of-control signals. Because the EWMA statistic is a weighted average, it has a smaller standard deviation than a single control measurement, and, therefore, the EWMA control limits are narrower than the limits for a Shewhart control chart.



The EWMA control chart for mass calibrations can be generated using both Dataplot code and

R code.

Interpretation of the control chart

The EWMA control chart shows many violations of the control limits starting at approximately the mid-point of 1986. This pattern emerges because the process average has actually shifted about one standard deviation, and the EWMA control chart is sensitive to small changes.

2.3.6. Instrument calibration over a regime

Topics

This section discusses the creation of a calibration curve for calibrating instruments (gauges) whose responses cover a large range. Topics are:

- Models for instrument calibration
- Data collection
- Assumptions
- Conditions that can invalidate the calibration procedure
- Data analysis and model validation
- Calibration of future measurements
- Uncertainties of calibrated values

Purpose of instrument calibration

Instrument calibration is intended to eliminate or reduce bias in an instrument's readings over a range for all continuous values. For this purpose, reference standards with known values for selected points covering the range of interest are measured with the instrument in question. Then a functional relationship is established between the values of the standards and the corresponding measurements. There are two basic situations.

Instruments which require correction for bias

- The instrument reads in the same units as the reference standards. The purpose of the calibration is to identify and eliminate any bias in the instrument relative to the defined unit of measurement. For example, optical imaging systems that measure the width of lines on semiconductors read in micrometers, the unit of interest. Nonetheless, these instruments must be calibrated to values of reference standards if line width measurements across the industry are to agree with each other.

Instruments whose measurements act as surrogates for other measurements

- The instrument reads in different units than the reference standards. The purpose of the calibration is to convert the instrument readings to the units of interest. An example is densitometer measurements that act as surrogates for measurements of radiation dosage. For this purpose, reference standards are irradiated at several dosage levels and then measured by radiometry. The same reference standards are measured by densitometer. The calibrated results of future densitometer readings on medical devices are the basis for deciding if the devices have been sterilized at the proper radiation level.

Basic steps for correcting the instrument for bias

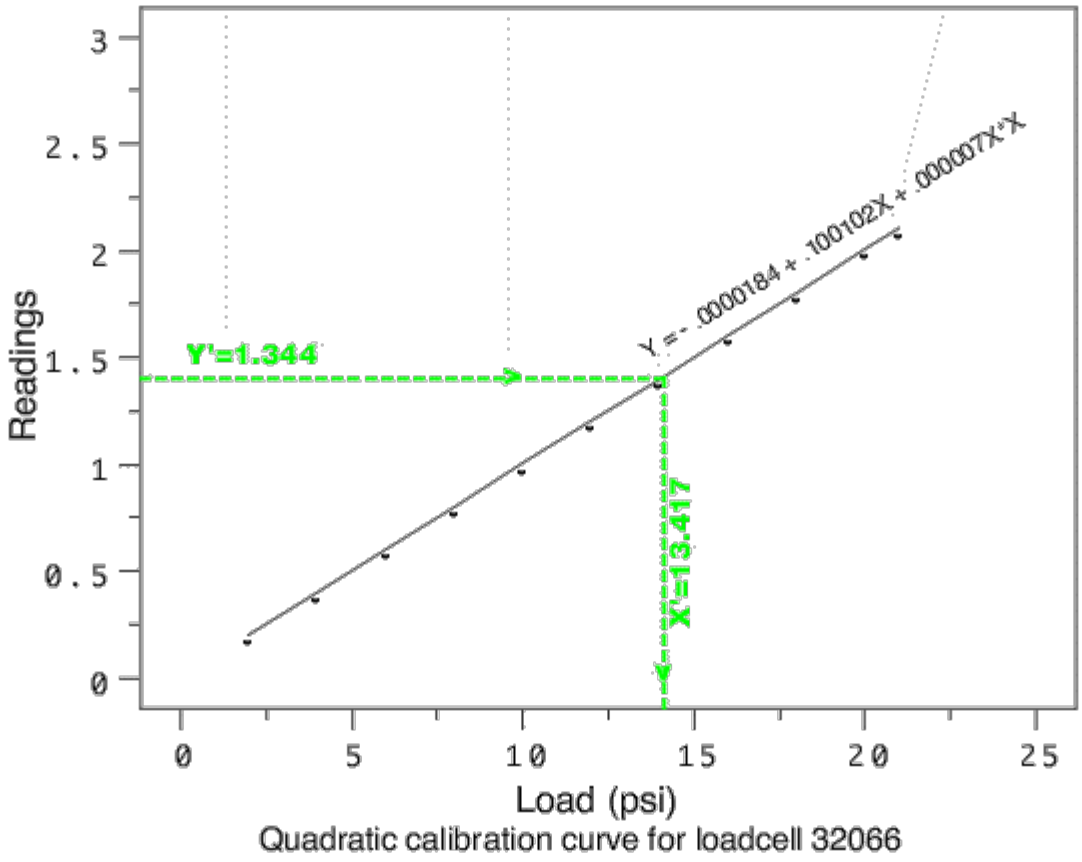
The calibration method is the same for both situations and requires the following basic steps:

- Selection of reference standards with known values to cover the range of interest.
- Measurements on the reference standards with the instrument to be calibrated.
- Functional relationship between the measured and known values of the reference standards (usually a least-squares fit to the data) called a **calibration curve**.
- Correction of all measurements by the inverse of the calibration curve.

Schematic example of a calibration curve and resulting value

A schematic explanation is provided by the figure below for load cell calibration. The loadcell measurements (shown as *) are plotted on the *y*-axis against the corresponding values of known load shown on the *x*-axis.

A quadratic fit to the loadcell data produces the calibration curve that is shown as the solid line. For a future measurement with the load cell, *Y' = 1.344* on the *y*-axis, a dotted line is drawn through *Y'* parallel to the *x*-axis. At the point where it intersects the calibration curve, another dotted line is drawn parallel to the y-axis. Its point of intersection with the *x*-axis at *X' = 13.417* is the calibrated value.



2.3.6.1. Models for instrument calibration

Notation

The following notation is used in this chapter in discussing models for calibration curves.

- Y denotes a measurement on a reference standard
- X denotes the known value of a reference standard
- ε denotes measurement error.
- a, b and c denote coefficients to be determined

Possible forms for calibration curves

There are several models for calibration curves that can be considered for instrument calibration. They fall into the following classes:

- Linear:

$$Y = a + bX + \varepsilon$$

- Quadratic:

$$Y = a + bX + cX^2 + \varepsilon$$

- Power:

$$Y = aX^b\varepsilon$$

- Non-linear:

$$Y = g(X) + \varepsilon$$

Special case of linear model - no calibration required

An instrument requires no calibration if

$$a=0 \text{ and } b=1$$

i.e., if measurements on the reference standards agree with their known values given an allowance for measurement error, the instrument is already calibrated. Guidance on collecting data, estimating and testing the coefficients is given on other pages.

Advantages of the linear

The linear model ISO 11095 is widely applied to instrument calibration because it has several advantages

model

over more complicated models.

- Computation of coefficients and standard deviations is easy.
- Correction for bias is easy.
- There is often a theoretical basis for the model.
- The analysis of uncertainty is tractable.

***Warning** on excluding the intercept term from the model*

It is often tempting to exclude the intercept, a, from the model because a zero stimulus on the x-axis should lead to a zero response on the y-axis. However, the correct procedure is to fit the full model and test for the significance of the intercept term.

Quadratic model and higher order polynomials

Responses of instruments or measurement systems which cannot be linearized, and for which no theoretical model exists, can sometimes be described by a quadratic model (or higher-order polynomial). An example is a load cell where force exerted on the cell is a non-linear function of load.

Disadvantages of quadratic models

Disadvantages of quadratic and higher-order polynomials are:

- They may require more reference standards to capture the region of curvature.
- There is rarely a theoretical justification; however, the adequacy of the model can be tested statistically.
- The correction for bias is more complicated than for the linear model.
- The uncertainty analysis is difficult.

Warning

A plot of the data, although always recommended, is not sufficient for identifying the correct model for the calibration curve. Instrument responses may not appear non-linear over a large interval. If the response and the known values are in the same units, differences from the known values should be plotted versus the known values.

Power model treated as a linear model

The power model is appropriate when the measurement error is proportional to the response rather than being additive. It is frequently used for calibrating instruments that measure dosage levels of irradiated materials.

The power model is a special case of a non-linear model that can be linearized by a natural logarithm transformation to

$$Y = \log_e(a) + b \cdot \log_e(X) + \log_e(\varepsilon)$$

so that the model to be fit to the data is of the familiar

linear form

$$W = a' + bZ + e$$

where W, Z and e are the transforms of the variables, Y, X and the measurement error, respectively, and a' is the natural logarithm of a.

Non-linear models and their limitations

Instruments whose responses are not linear in the coefficients can sometimes be described by non-linear models. In some cases, there are theoretical foundations for the models; in other cases, the models are developed by trial and error. Two classes of non-linear functions that have been shown to have practical value as calibration functions are:

1. Exponential
2. Rational

Non-linear models are an important class of calibration models, but they have several significant limitations.

- The model itself may be difficult to ascertain and verify.
- There can be severe computational difficulties in estimating the coefficients.
- Correction for bias cannot be applied algebraically and can only be approximated by interpolation.
- Uncertainty analysis is very difficult.

Example of an exponential function

An exponential function is shown in the equation below. Instruments for measuring the ultrasonic response of reference standards with various levels of defects (holes) that are submerged in a fluid are described by this function.

$$Y = \frac{e^{-aX}}{b + cX} + \varepsilon$$

Example of a rational function

A rational function is shown in the equation below. Scanning electron microscope measurements of line widths on semiconductors are described by this function (Kirby).

$$Y = \frac{a + bX + cX^2}{a_1 + b_1 X + c_1 X^2} + \varepsilon$$

2.3.6.2. Data collection

Data collection

The process of collecting data for creating the calibration curve is critical to the success of the calibration program. General rules for designing calibration experiments apply, and guidelines that are adequate for the calibration models in this chapter are given below.

Selection of reference standards

A minimum of five reference standards is required for a linear calibration curve, and ten reference standards should be adequate for more complicated calibration models.

The optimal strategy in selecting the reference standards is to space the reference standards at points corresponding to equal increments on the *y*-axis, covering the range of the instrument. Frequently, this strategy is not realistic because the person producing the reference materials is often not the same as the person who is creating the calibration curve. Spacing the reference standards at equal intervals on the *x*-axis is a good alternative.

Exception to the rule above - bracketing

If the instrument is not to be calibrated over its entire range, but only over a very short range for a specific application, then it may not be necessary to develop a complete calibration curve, and a bracketing technique (ISO 11095) will provide satisfactory results. The bracketing technique assumes that the instrument is linear over the interval of interest, and, in this case, only two reference standards are required -- one at each end of the interval.

Number of repetitions on each reference standard

A minimum of two measurements on each reference standard is required and four is recommended. The repetitions should be separated in time by days or weeks. These repetitions provide the data for determining whether a candidate model is adequate for calibrating the instrument.

2.3.6.3. Assumptions for instrument calibration

Assumption regarding reference values	The basic assumption regarding the reference values of artifacts that are measured in the calibration experiment is that they are known without error. In reality, this condition is rarely met because these values themselves usually come from a measurement process. Systematic errors in the reference values will always bias the results, and random errors in the reference values can bias the results.
Rule of thumb	It has been shown by Bruce Hoadly, in an internal NIST publication, that the best way to mitigate the effect of random fluctuations in the reference values is to plan for a large spread of values on the x-axis relative to the precision of the instrument.
Assumptions regarding measurement errors	The basic assumptions regarding measurement errors associated with the instrument are that they are: • free from outliers • independent • of equal precision • from a normal distribution.

2.3.6.4. What can go wrong with the calibration procedure

Calibration procedure may fail to eliminate bias

There are several circumstances where the calibration curve will not reduce or eliminate bias as intended. Some are discussed on this page. A critical exploratory analysis of the calibration data should expose such problems.

Lack of precision

Poor instrument precision or unsuspected day-to-day effects may result in standard deviations that are large enough to jeopardize the calibration. There is nothing intrinsic to the calibration procedure that will improve precision, and the best strategy, before committing to a particular instrument, is to estimate the instrument's precision in the environment of interest to decide if it is good enough for the precision required.

Outliers in the calibration data

Outliers in the calibration data can seriously distort the calibration curve, particularly if they lie near one of the endpoints of the calibration interval.

- Isolated outliers (single points) should be deleted from the calibration data.
- An entire day's results which are inconsistent with the other data should be examined and rectified before proceeding with the analysis.

Systematic differences among operators

It is possible for different operators to produce measurements with biases that differ in sign and magnitude. This is not usually a problem for automated instrumentation, but for instruments that depend on line of sight, results may differ significantly by operator. To diagnose this problem, measurements by different operators on the same artifacts are plotted and compared. Small differences among operators can be accepted as part of the imprecision of the measurement process, but large systematic differences among operators require resolution. Possible solutions are to retrain the operators or maintain separate calibration curves by operator.

Lack of system

The calibration procedure, once established, relies on the instrument continuing to respond in the same way over time.

control	If the system drifts or takes unpredictable excursions, the calibrated values may not be properly corrected for bias, and depending on the direction of change, the calibration may further degrade the accuracy of the measurements. To assure that future measurements are properly corrected for bias, the calibration procedure should be coupled with a statistical control procedure for the instrument.
Example of differences among repetitions in the calibration data	An important point, but one that is rarely considered, is that there can be differences in responses from repetition to repetition that will invalidate the analysis. A plot of the aggregate of the calibration data may not identify changes in the instrument response from day-to-day. What is needed is a plot of the fine structure of the data that exposes any day to day differences in the calibration data.
***Warning** - calibration can fail because of day-to-day changes*	A straight-line fit to the aggregate data will produce a 'calibration curve'. However, if straight lines fit separately to each day's measurements show very disparate responses, the instrument, at best, will require calibration on a daily basis and, at worst, may be sufficiently lacking in control to be usable.

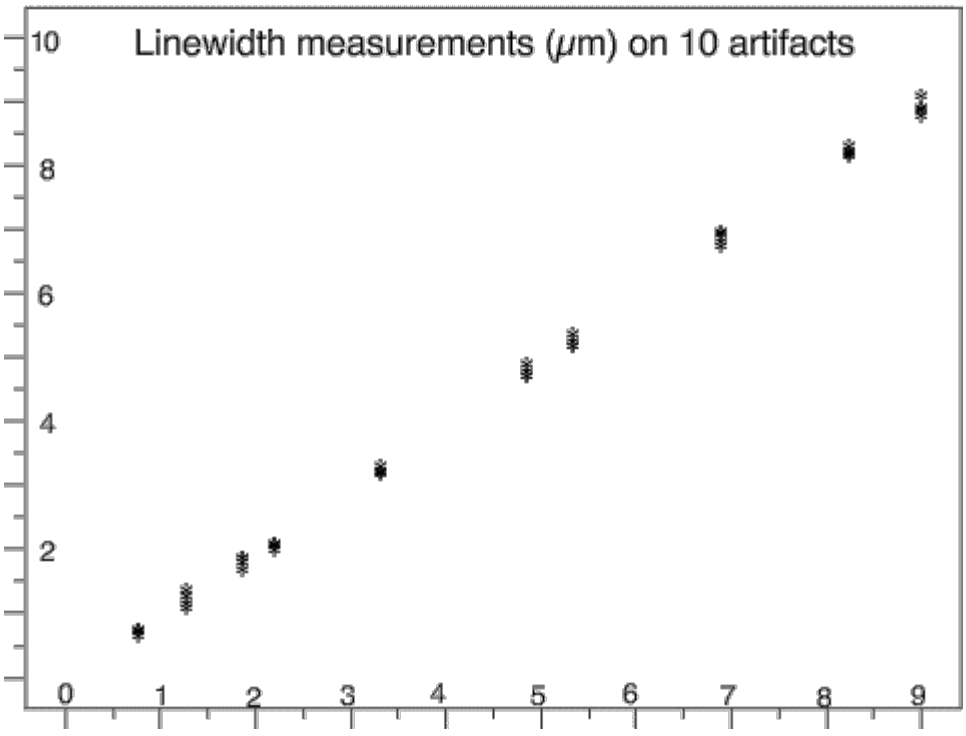
2. Measurement Process Characterization
2.3. Calibration
2.3.6. Instrument calibration over a regime
2.3.6.4. What can go wrong with the calibration procedure

2.3.6.4.1. Example of day-to-day changes in calibration

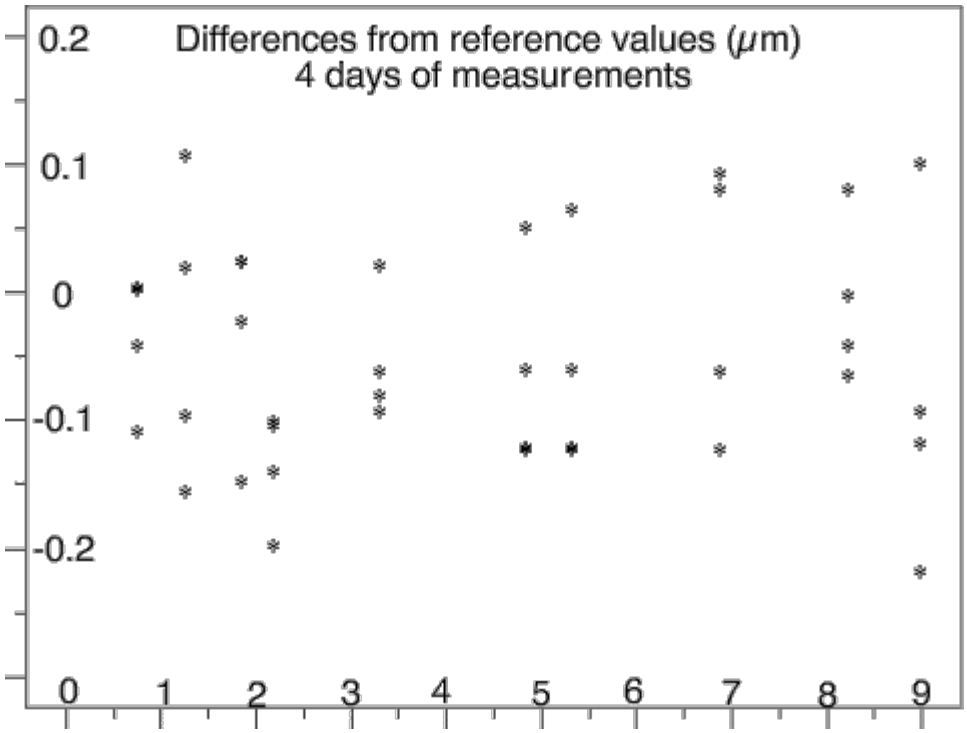
Calibration data over 4 days

Line width measurements on 10 NIST reference standards were made with an optical imaging system on each of four days. The four data points for each reference value appear to overlap in the plot because of the wide spread in reference values relative to the precision. The plot suggests that a linear calibration line is appropriate for calibrating the imaging system.

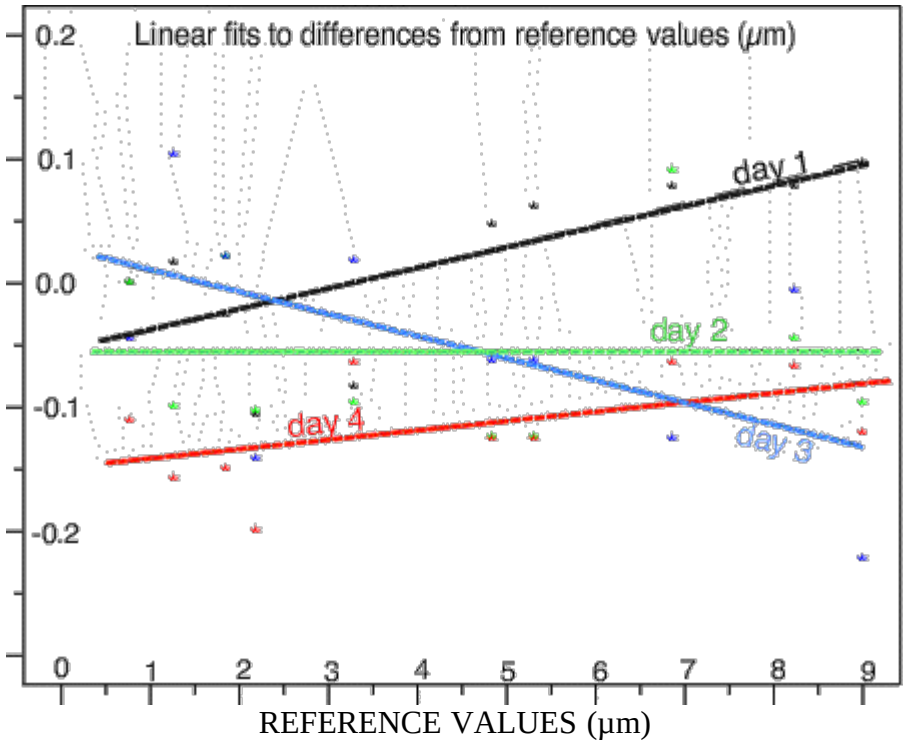
This plot shows measurements made on 10 reference materials repeated on four days with the 4 points for each day overlapping



This plot shows the differences between each measurement and the corresponding reference value. Because days are not identified, the



plot gives no indication of problems in the control of the imaging system from from day to day.



This plot, with linear calibration lines fit to each day's measurements individually, shows how the response of the imaging system changes dramatically from day to day. Notice that the slope of the calibration line goes from positive on day 1 to negative on day 3.

Interpretation of calibration findings

Given the lack of control for this measurement process, any calibration procedure built on the average of the calibration data will fail to properly correct the system on some days and invalidate resulting measurements. There is no good solution to this problem except daily calibration.

2.3.6.5. Data analysis and model validation

First step - plot the calibration data

If the model for the calibration curve is not known from theoretical considerations or experience, it is necessary to identify and validate a model for the calibration curve. To begin this process, the calibration data are plotted as a function of known values of the reference standards; this plot should suggest a candidate model for describing the data. A linear model should always be a consideration. If the responses and their known values are in the same units, a plot of differences between responses and known values is more informative than a plot of the data for exposing structure in the data.

Warning - regarding statistical software

Once an initial model has been chosen, the coefficients in the model are estimated from the data using a statistical software package. It is impossible to over-emphasize the importance of using reliable and documented software for this analysis.

Output required from a software package

The software package will use the method of least squares for estimating the coefficients. The software package should also be capable of performing a 'weighted' fit for situations where errors of measurement are non-constant over the calibration interval. The choice of weights is usually the responsibility of the user. The software package should, at the minimum, provide the following information:

- Coefficients of the calibration curve
- Standard deviations of the coefficients
- Residual standard deviation of the fit
- F-ratio for goodness of fit (if there are repetitions on the y-axis at each reference value)

Typical analysis of a quadratic fit

Load cell measurements are modeled as a quadratic function of known loads as shown below. There are three repetitions at each load level for a total of 33 measurements.

```
Parameter estimates for model y = a + b*x +
c*x*x + e:
```

```
Parameter  Estimate  Std. Error     t-value
Pr(>|t|)
a        -1.840e-05   2.451e-05      -0.751
0.459
b         1.001e-01   4.839e-06   20687.891
<2e-16
c         7.032e-06   2.014e-07      34.922
<2e-16

Residual standard error = 3.764e-05 (30 degrees
of freedom)
Multiple R-squared = 1
Adjusted R-squared = 1


Analysis of variance table:

Source of     Degrees of   Sum of         Mean
Variation      Freedom     Squares       Square
F-Ratio      Pr(>F)

Model              2         12.695       6.3475
4.48e+09    <2.2e-16
Residual          30       4.2504e-08    1.4170e-
09

(Lack of fit)      8       4.7700e-09    5.9625e-
10    0.3477       0.9368
(Pure error)      22       3.7733e-08    1.7151e-
09

Total             32         12.695
```

The analyses shown above can be reproduced using Dataplot code and R code.

Note: Dataplot reports a probability associated with the F-ratio (for example, 6.334 % for the lack-of-fit test), where a probability greater than 95 % indicates an F-ratio that is significant at the 5 % level. R reports a *p*-value that corresponds to the probability greater than the F-ratio, so a value less than 0.05 would indicate significance at the 5 % level. Other software may report in other ways; therefore, it is necessary to check the interpretation for each package.

The F-ratio is used to test the goodness of the fit to the data

The F-ratio provides information on the model as a good descriptor of the data. The F-ratio is compared with a critical value from the F-table. An F-ratio smaller than the critical value indicates that all significant structure has been captured by the model.

F-ratio < 1 always indicates a good fit

For the load cell analysis, a plot of the data suggests a linear fit. However, the linear fit gives a very large F-ratio. For the quadratic fit, the F-ratio is 0.3477 with $v_1 = 8$ and $v_2 = 22$ degrees of freedom. The critical value of F(0.05, 8, 20) = 2.45 indicates that the quadratic function is sufficient for describing the data. A fact to keep in mind is that an F-ratio < 1 does not need to be checked against a critical value; it always indicates a good fit to the data.

The t-values are used to test the significance of

The *t*-values can be compared with critical values from a *t*-table. However, for a test at the 5 % significance level, a *t*-value < 2 is a good indicator of non-significance. The *t*-value for the intercept term, a, is < 2 indicating that the

individual coefficients	intercept term is not significantly different from zero. The *t*-values for the linear and quadratic terms are significant indicating that these coefficients are needed in the model. If the intercept is dropped from the model, the analysis is repeated to obtain new estimates for the coefficients, b and c.
Residual standard deviation	The residual standard deviation estimates the standard deviation of a single measurement with the load cell.
Further considerations and tests of assumptions	The residuals (differences between the measurements and their fitted values) from the fit should also be examined for outliers and structure that might invalidate the calibration curve. They are also a good indicator of whether basic assumptions of normality and equal precision for all measurements are valid. If the initial model proves inappropriate for the data, a strategy for improving the model is followed.

2.3.6.5.1. Data on load cell #32066

Three repetitions on a load cell at eleven known loads

X	Y
2.	0.20024
2.	0.20016
2.	0.20024
4.	0.40056
4.	0.40045
4.	0.40054
6.	0.60087
6.	0.60075
6.	0.60086
8.	0.80130
8.	0.80122
8.	0.80127
10.	1.00173
10.	1.00164
10.	1.00173
12.	1.20227
12.	1.20218
12.	1.20227
14.	1.40282
14.	1.40278
14.	1.40279
16.	1.60344
16.	1.60339
16.	1.60341
18.	1.80412
18.	1.80409
18.	1.80411
20.	2.00485
20.	2.00481
20.	2.00483
21.	2.10526
21.	2.10524
21.	2.10524

2.3.6.6. Calibration of future measurements

Purpose

The purpose of creating the calibration curve is to correct future measurements made with the same instrument to the correct units of measurement. The calibration curve can be applied many, many times before it is discarded or reworked as long as the instrument remains in statistical control. Chemical measurements are an exception where frequently the calibration curve is used only for a single batch of measurements, and a new calibration curve is created for the next batch.

Notation

The notation for this section is as follows:

- Y' denotes a future measurement.
- X' denotes the associated calibrated value.
- $\hat{a}, \hat{b}, \hat{c}$ are the estimates of the coefficients, a, b, c.
- s_a, s_b, s_c are standard deviations of the coefficients, a, b, c.

Procedure

To apply a correction to a future measurement, Y^*, to obtain the calibration value X^* requires the inverse of the calibration curve.

Linear calibration line

The inverse of the calibration line for the linear model

$$Y = a + bX + \varepsilon$$

gives the calibrated value

$$X' = \frac{Y' - \hat{a}}{\hat{b}}$$

Tests for the intercept and slope of calibration

Before correcting for the calibration line by the equation above, the intercept and slope should be tested for $a=0$, and $b=1$. If both

$$\left|\frac{\hat{a}}{s_a}\right| < t_{1-\alpha/2,\nu} \quad \text{and} \quad \left|\frac{\hat{b}-1}{s_b}\right| < t_{1-\alpha/2,\nu}$$

curve -- If both conditions hold, no calibration is needed.

there is no need for calibration. If, on the other hand only the test for $a=0$ *fails*, the error is constant; if only the test for $b=1$ *fails*, the errors are related to the size of the reference standards.

Table look-up for t-factor

The factor, $t_{1-\alpha/2,\,\nu}$, is found in the *t*-table where ν is the degrees of freedom for the residual standard deviation from the calibration curve, and α is chosen to be small, say, 0.05.

Quadratic calibration curve

The inverse of the calibration curve for the quadratic model

$$Y = a + bX + cX^2 + \varepsilon$$

requires a root

$$X' = \frac{-\hat{b} \pm \sqrt{\hat{b}^2 - 4\hat{c}\left(\hat{a} - Y'\right)}}{2\hat{c}}$$

The correct root (+ or -) can usually be identified from practical considerations.

Power curve

The inverse of the calibration curve for the power model

$$Y = aX^b\varepsilon$$

gives the calibrated value

$$X' = \exp\left(\frac{\log_e(Y') - \log_e\left(\hat{a}\right)}{\hat{b}}\right)$$

where b and the natural logarithm of a are estimated from the power model transformed to a linear function.

Non-linear and other calibration curves

For more complicated models, the inverse for the calibration curve is obtained by interpolation from a graph of the function or from predicted values of the function.

2.3.6.7. Uncertainties of calibrated values

Purpose

The purpose is to quantify the uncertainty of a 'future' result that has been corrected by the calibration curve. In principle, the uncertainty quantifies any possible difference between the calibrated value and its reference base (which normally depends on reference standards).

Explanation in terms of reference artifacts

Measurements of interest are future measurements on unknown artifacts, but one way to look at the problem is to ask: If a measurement is made on one of the reference standards and the calibration curve is applied to obtain the calibrated value, how well will this value agree with the 'known' value of the reference standard?

Difficulties

The answer is not easy because of the intersection of two uncertainties associated with

1. the calibration curve itself because of limited data
2. the 'future' measurement

If the calibration experiment were to be repeated, a slightly different calibration curve would result even for a system in statistical control. An exposition of the intersection of the two uncertainties is given for the calibration of proving rings (Hockersmith and Ku).

ISO approach to uncertainty can be based on check standards or propagation of error

General procedures for computing an uncertainty based on ISO principles of uncertainty analysis are given in the chapter on modeling.

Type A uncertainties for calibrated values from calibration curves can be derived from

- check standard values
- propagation of error

An example of type A uncertainties of calibrated values from a linear calibration curve are analyzed from measurements on linewidth check standards. Comparison of the uncertainties from check standards and propagation of error for the linewidth calibration data are also illustrated.

An example of the derivation of propagation of error type A

uncertainties for calibrated values from a quadratic calibration curve for loadcells is discussed on the next page.

2. Measurement Process Characterization
2.3. Calibration
2.3.6. Instrument calibration over a regime
2.3.6.7. Uncertainties of calibrated values

2.3.6.7.1. Uncertainty for quadratic calibration using propagation of error

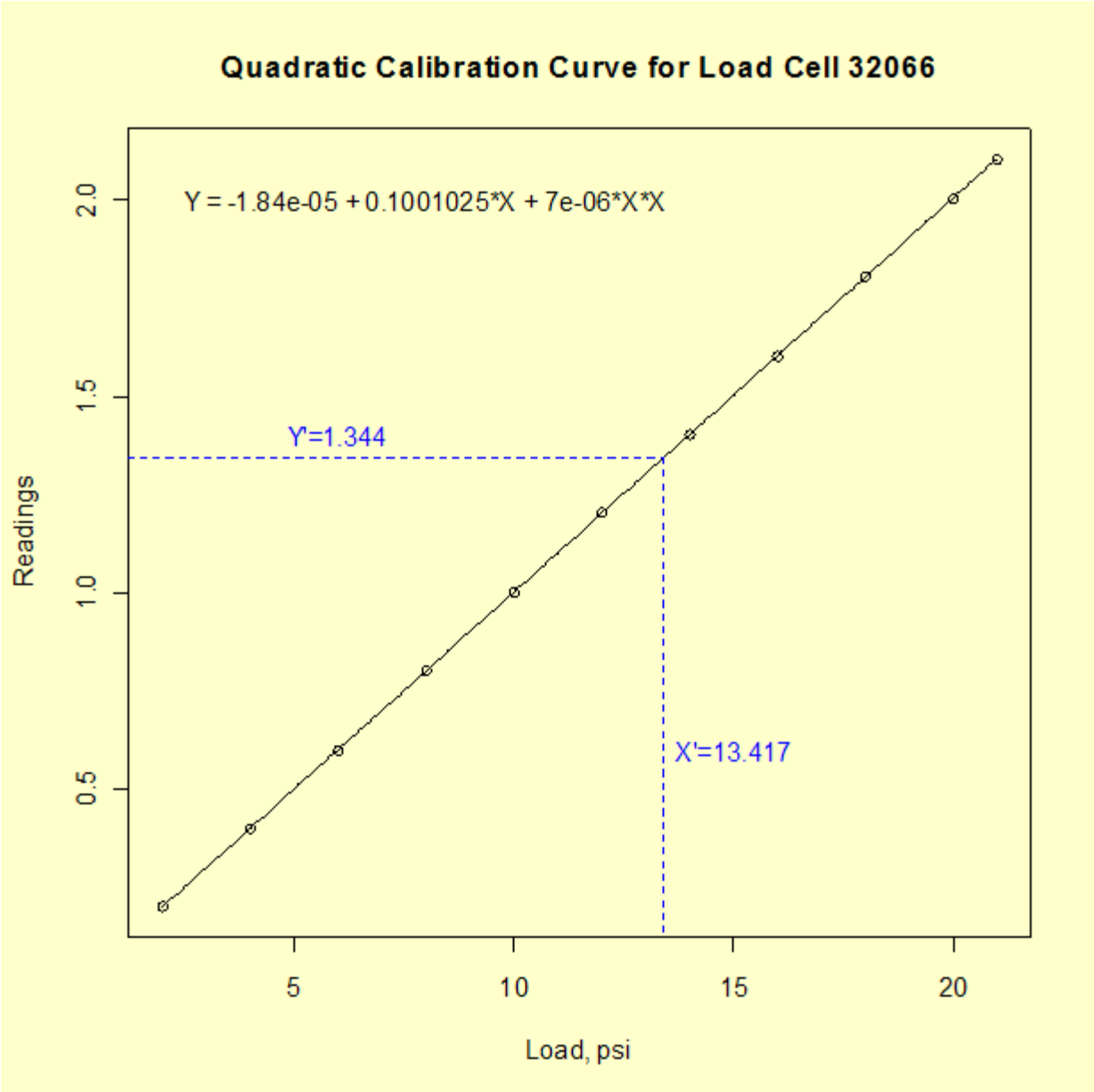
Propagation of error for uncertainty of calibrated values of loadcells

The purpose of this page is to show the propagation of error for calibrated values of a loadcell based on a quadratic calibration curve where the model for instrument response is

$$Y = a + bX + cX^2 + \varepsilon$$

The calibration data are instrument responses at known loads (psi), and estimates of the quadratic coefficients, a, b, c, and their associated standard deviations are shown with the analysis.

A graph of the calibration curve showing a measurement Y' corrected to X', the proper load (psi), is shown below.



Uncertainty of the calibrated value X'

The uncertainty to be evaluated is the uncertainty of the calibrated value, *X'*, computed for any future measurement, *Y'*, made with the calibrated instrument where

$$X' = \frac{-\hat{b} \pm \sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}}{2\hat{c}}$$

Partial derivatives

The partial derivatives are needed to compute uncertainty.

$$\frac{\partial X'}{\partial Y'} = \frac{1}{\sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}}$$

$$\frac{\partial X'}{\partial \hat{a}} = \frac{-1}{\sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}}$$

$$\frac{\partial X'}{\partial \hat{b}} = \frac{-1 + \frac{\hat{b}}{\sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}}}{2\hat{c}}$$

$$\frac{\partial X'}{\partial \hat{c}} = \frac{-\hat{a} + Y'}{\hat{c}\sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}} - \frac{-\hat{b} + \sqrt{\hat{b}^2 - 4\hat{c}(\hat{a} - Y')}}{2\hat{c}^2}$$

The variance of the calibrated value from propagation of error

The variance of X' is defined from propagation of error as follows:

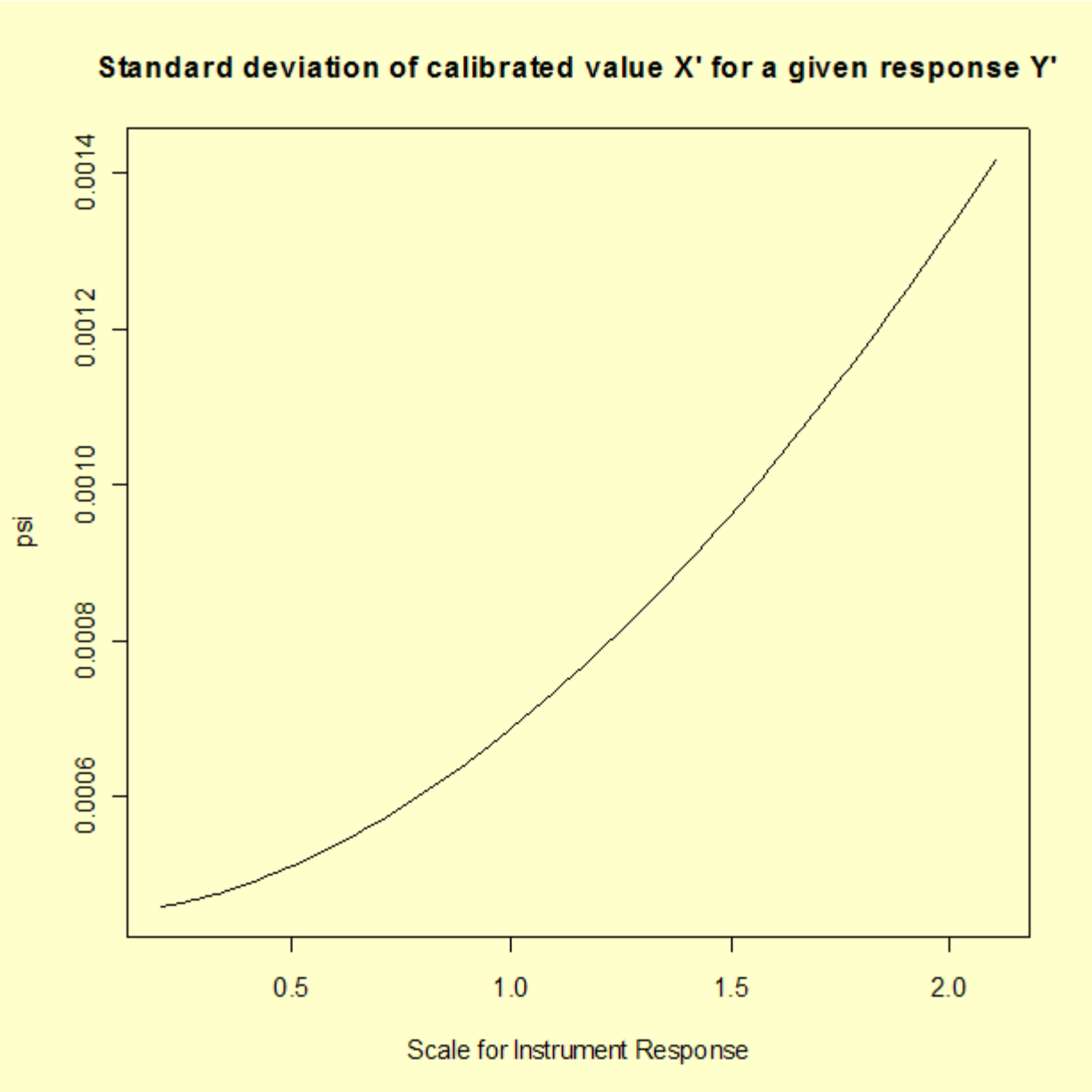
$$u^2 = \left(\frac{\partial X'}{\partial Y'}\right)^2 (s_{Y'})^2 + \left(\frac{\partial X'}{\partial \hat{a}}\right)^2 (s_{\hat{a}})^2 + \left(\frac{\partial X'}{\partial \hat{b}}\right)^2 (s_{\hat{b}})^2 + \left(\frac{\partial X'}{\partial \hat{c}}\right)^2 (s_{\hat{c}})^2$$

The values of the coefficients and their respective standard deviations from the quadratic fit to the calibration curve are substituted in the equation. The standard deviation of the measurement, *Y*, may not be the same as the standard deviation from the fit to the calibration data if the measurements to be corrected are taken with a different system; here we assume that the instrument to be calibrated has a standard deviation that is essentially the same as the instrument used for collecting the calibration data and the residual standard deviation from the quadratic fit is the appropriate estimate.

```
a  = -0.183980e-04
sa =  0.2450e-04
b  =  0.100102
sb =  0.4838e-05
c  =  0.703186e-05
sc =  0.2013e-06
sy =  0.0000376353
```

Graph showing the standard deviations of calibrated values X' for given instrument responses Y' ignoring covariance terms in the propagation of error

The standard deviation expressed above is not easily interpreted but it is easily graphed. A graph showing standard deviations of calibrated values, *X'*, as a function of instrument response, *Y'*, is shown below.



Problem with propagation of error

The propagation of errors shown above is not complete because it ignores the covariances among the coefficients, a, b, c. Unfortunately, some statistical software packages do not display these covariance terms with the other output from the analysis.

Covariance terms for loadcell data

The variance-covariance terms for the loadcell data set are shown below.

	a	b	c
a	6.0049021-10		
b	-1.0759599-10	2.3408589-11	
c	4.0191106-12	-9.5051441-13	4.0538705-14

The diagonal elements are the variances of the coefficients, a, b, c, respectively, and the off-diagonal elements are the covariance terms.

Recomputation of the standard deviation of X'

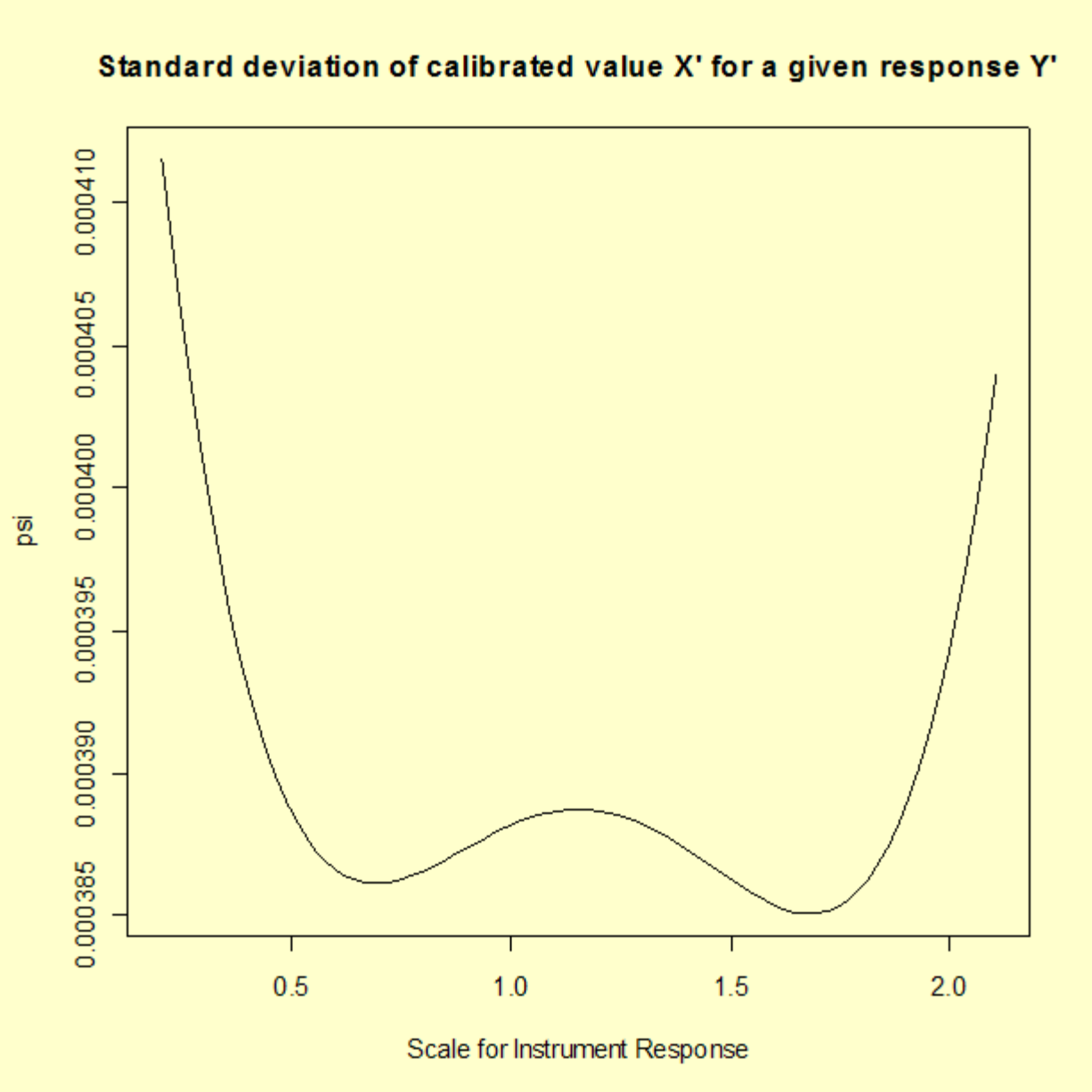
To account for the covariance terms, the variance of X' is redefined by adding the covariance terms. Appropriate substitutions are made; the standard deviations are recomputed and graphed as a function of instrument response.

$$u^2 = u^2 + 2\left(\frac{\partial X'}{\partial \hat{a}}\right)\left(\frac{\partial X'}{\partial \hat{b}}\right)s_{\hat{a}\hat{b}} + 2\left(\frac{\partial X'}{\partial \hat{a}}\right)\left(\frac{\partial X'}{\partial \hat{c}}\right)s_{\hat{a}\hat{c}} + 2\left(\frac{\partial X'}{\partial \hat{b}}\right)\left(\frac{\partial X'}{\partial \hat{c}}\right)s_{\hat{b}\hat{c}}$$

```
sab = -1.0759599e-10
sac =  4.0191106e-12
sbc = -9.5051441e-13
```

The graph below shows the correct estimates for the standard deviation of X' and gives a means for assessing the loss of accuracy that can be incurred by ignoring covariance terms. In this case, the uncertainty is reduced by including covariance terms, some of which are negative.

Graph showing the standard deviations of calibrated values, X', for given instrument responses, Y', with covariance terms included in the propagation of error



Sample code

The results in this section can be generated using R code.

2.3.6.7.2. Uncertainty for linear calibration using check standards

Check standards provide a mechanism for calculating uncertainties

The easiest method for calculating type A uncertainties for calibrated values from a calibration curve requires periodic measurements on check standards. The check standards, in this case, are artifacts at the lower, mid-point and upper ends of the calibration curve. The measurements on the check standard are made in a way that randomly samples the output of the calibration procedure.

Calculation of check standard values

The check standard values are the raw measurements on the artifacts corrected by the calibration curve. The standard deviation of these values should estimate the uncertainty associated with calibrated values. The success of this method of estimating the uncertainties depends on adequate sampling of the measurement process.

Measurements corrected by a linear calibration curve

As an example, consider measurements of linewidths on photomask standards, made with an optical imaging system and corrected by a linear calibration curve. The three control measurements were made on reference standards with values at the lower, mid-point, and upper end of the calibration interval.

Compute the calibration standard deviation

For the linewidth data, the regression equation from the calibration experiment is

$$Y = a + bX + \epsilon$$

and the estimated regression coefficients are the following.

$$\hat{a} = 0.2357$$
$$\hat{b} = 0.9870$$

Next, we calculate the difference between the "predicted" X from the regression fit and the observed X.

$$W_i = \frac{(Y_i - \hat{a})}{\hat{b}} - X_i$$

Finally, we find the calibration standard deviation by calculating the standard deviation of the computed differences.

$$S = \sqrt{\frac{\Sigma (W_i - \overline{W})^2}{n-1}}$$

The calibration standard deviation for the linewidth data is 0.119 µm.

The calculations in this section can be completed using Dataplot code and R code.

Comparison with propagation of error

The standard deviation, 0.119 µm, can be compared with a propagation of error analysis.

Other sources of uncertainty

In addition to the type A uncertainty, there may be other contributors to the uncertainty such as the uncertainties of the values of the reference materials from which the calibration curve was derived.

2. Measurement Process Characterization
2.3. Calibration
2.3.6. Instrument calibration over a regime
2.3.6.7. Uncertainties of calibrated values

2.3.6.7.3. Comparison of check standard analysis and propagation of error

Propagation of error for the linear calibration

The analysis of uncertainty for calibrated values from a linear calibration line can be addressed using propagation of error. On the previous page, the uncertainty was estimated from check standard values.

Estimates from calibration data

The calibration data consist of 40 measurements with an optical imaging system on 10 linewidth artifacts. A linear fit to the data gives a calibration curve with the following estimates for the intercept, *a*, and the slope, *b*:

Parameter	Estimate	Std. Error	*t*-value	Pr(>|*t*|)
a	0.2357623	0.02430034	9.702014	7.860745e-12
b	0.9870377	0.00344058	286.881171	5.354121e-65

with the following covariance matrix.

	a	b
a	5.905067e-04	-7.649453e-05
b	-7.649453e-05	1.183759e-05

The results shown above can be generated with R code.

Propagation of error

The propagation of error is performed for the equation

$$X' = \frac{Y' - \hat{a}}{\hat{b}}$$

so that the squared uncertainty of a calibrated value, X', is

$$u^2 = \left[\frac{\partial X'}{\partial Y'}\right]^2 s^2_{Y'} + \left[\frac{\partial X'}{\partial \hat{a}}\right]^2 s^2_{\hat{a}} + \left[\frac{\partial X'}{\partial \hat{b}}\right]^2 s^2_{\hat{b}} + 2\left[\frac{\partial X'}{\partial \hat{a}}\right]\left[\frac{\partial X'}{\partial \hat{b}}\right] s_{\hat{a}\hat{b}}$$

where

$$\frac{\partial X'}{\partial Y'} = \frac{1}{\hat{b}}$$

$$\frac{\partial X'}{\partial \hat{a}} = \frac{-1}{\hat{b}}$$

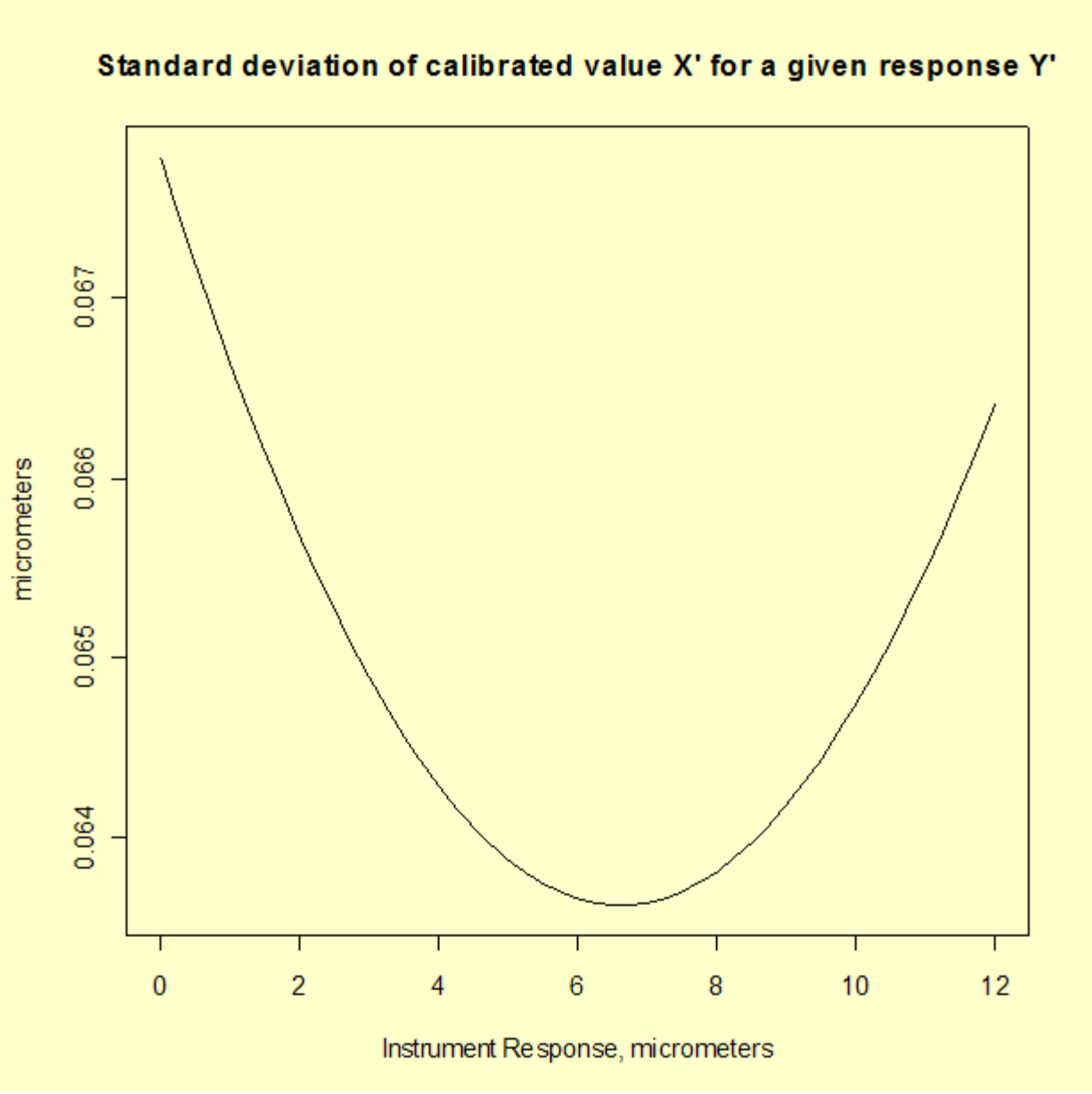
$$\frac{\partial X'}{\partial \hat{b}} = \frac{-(Y' - \hat{a})}{\hat{b}^2}$$

The uncertainty of the calibrated value, X',

$$u^2 = \left(\frac{1}{\hat{b}}\right)^2 s^2_{Y'} + \left(\frac{-1}{\hat{b}}\right)^2 s^2_{\hat{a}} + \left(\frac{-(Y' - \hat{a})}{\hat{b}^2}\right)^2 s^2_{\hat{b}} + 2\left(\frac{-1}{\hat{b}}\right)\left(\frac{-(Y' - \hat{a})}{\hat{b}^2}\right) s_{\hat{a}\hat{b}}$$

is dependent on the value of the instrument reponse Y'.

Graph showing standard deviation of calibrated value X' plotted as a function of instrument response Y' for a linear calibration



Comparison of check standard analysis and propagation of error

Comparison of the analysis of check standard data, which gives a standard deviation of 0.119 µm, and propagation of error, which gives a maximum standard deviation of 0.068 µm, suggests that the propagation of error may underestimate the type A uncertainty. The check standard measurements are undoubtedly sampling some sources of variability that do not appear in the formal propagation of error formula.

2.3.7. Instrument control for linear calibration

Purpose

The purpose of the control program is to guarantee that the calibration of an instrument does not degrade over time.

Approach

This is accomplished by exercising quality control on the instrument's output in much the same way that quality control is exercised on components in a process using a modification of the Shewhart control chart.

Check standards needed for the control program

For linear calibration, it is sufficient to control the end-points and the middle of the calibration interval to ensure that the instrument does not drift out of calibration. Therefore, check standards are required at three points; namely,

- at the lower-end of the regime
- at the mid-range of the regime
- at the upper-end of the regime

Data collection

One measurement is needed on each check standard for each checking period. It is advisable to start by making control measurements at the start of each day or as often as experience dictates. The time between checks can be lengthened if the instrument continues to stay in control.

Definition of control value

To conform to the notation in the section on instrument corrections, X^* denotes the known value of a standard, and X denotes the measurement on the standard.

A control value is defined as the difference

$$W = X^* - X$$

If the calibration is perfect, control values will be randomly distributed about zero and fall within appropriate upper and lower limits on a control chart.

Calculation of control limits

The upper and lower control limits (Croarkin and Varner)) are, respectively,

$$l_{upper} = +\frac{s}{\hat{b}} t^{\bullet}_{\alpha/2}(\nu)$$

$$l_{lower} = -\frac{s}{\hat{b}} t^{\bullet}_{\alpha/2}(\nu)$$

where s is the residual standard deviation of the fit from the calibration experiment, and $\hat{b}$ is the estimated slope of the linear calibration curve.

*Values t**

The critical value, $t^{\bullet}_{\alpha/2}$, can be found in the t* table; ν is the degrees of freedom for the residual standard deviation; and α is equal to 0.05.

*Determining t**

For the case where α = 0.05 and ν = 38, the critical value of the t* statistic is 2.497575.

R code and Dataplot code can be used to determine t* critical values using a standard t-table for the ζ quantile and ν degrees of freedom where ζ is computed as

$$\zeta = \frac{1}{2}\left[1 - e^{\left(\frac{\ln(1-\alpha)}{m}\right)}\right]$$

where m is the number of check standards.

Sensitivity to departure from linearity

If

$$l_{lower} \le W \le l_{upper}$$

the instrument is in statistical control. Statistical control in this context implies not only that measurements are repeatable within certain limits but also that instrument response remains linear. The test is sensitive to departures from linearity.

Control chart for a system corrected by a linear calibration curve

An example of measurements of line widths on photomask standards, made with an optical imaging system and corrected by a linear calibration curve, are shown as an example. The three control measurements were made on reference standards with values at the lower, mid-point, and upper end of the calibration interval.

2.3.7.1. Control chart for a linear calibration line

Purpose

Line widths of three photomask reference standards (at the low, middle and high end of the calibration line) were measured on six days with an optical imaging system that had been calibrated from similar measurements on 10 reference artifacts. The control values and limits for the control chart , which depend on the intercept and slope of the linear calibration line, monitor the calibration and linearity of the optical imaging system.

Initial calibration experiment

The initial calibration experiment consisted of 40 measurements (not shown here) on 10 artifacts and produced a linear calibration line with:

- Intercept = 0.2357
- Slope = 0.9870
- Residual standard deviation = 0.06203 micrometers
- Degrees of freedom = 38

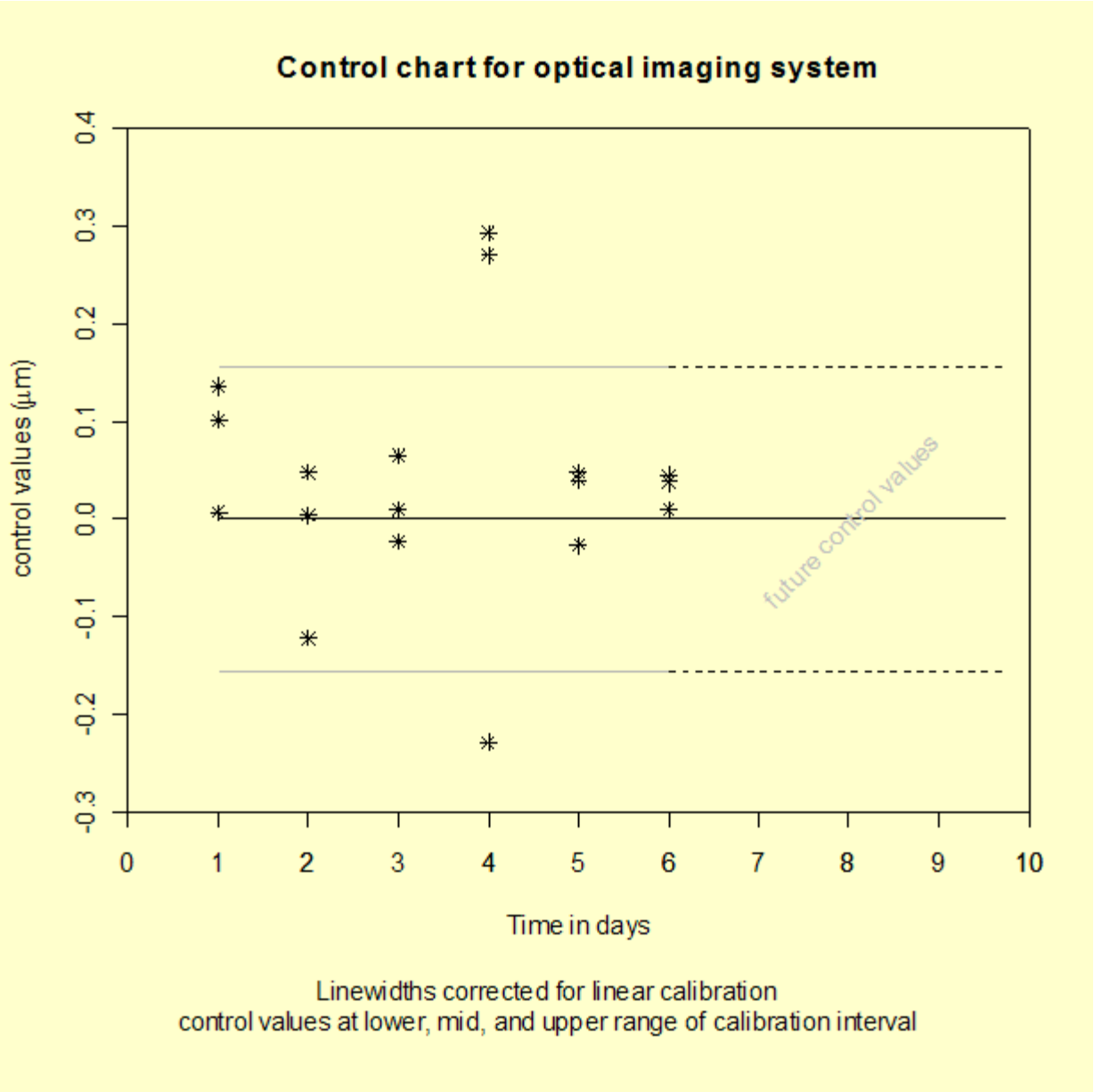
Line width measurements made with an optical imaging system

The control measurements, *Y*, and known values, *X*, for the three artifacts at the upper, mid-range, and lower end (U, M, L) of the calibration line are shown in the following table:

DAY	POSITION	X	Y
1	L	0.76	1.12
1	M	3.29	3.49
1	U	8.89	9.11
2	L	0.76	0.99
2	M	3.29	3.53
2	U	8.89	8.89
3	L	0.76	1.05
3	M	3.29	3.46
3	U	8.89	9.02
4	L	0.76	0.76
4	M	3.29	3.75
4	U	8.89	9.30
5	L	0.76	0.96
5	M	3.29	3.53
5	U	8.89	9.05
6	L	0.76	1.03
6	M	3.29	3.52
6	U	8.89	9.02

Control chart

The control chart shown below can be generated using both Dataplot code and R code.



Interpretation of control chart

The control measurements show no evidence of drift and are within the control limits except on the fourth day when all three control values are outside the limits. The cause of the problem on that day cannot be diagnosed from the data at hand, but all measurements made on that day, including workload items, should be rejected and remeasured.

2.4. Gauge R & R studies

The purpose of this section is to outline the steps that can be taken to characterize the performance of gauges and instruments used in a production setting in terms of errors that affect the measurements.

What are the issues for a gauge R & R study?

What are the design considerations for the study?

1. Artifacts
2. Operators
3. Gauges, parameter levels, configurations

How do we collect data for the study?

How do we quantify variability of measurements?

1. Repeatability
2. Reproducibility
3. Stability

How do we identify and analyze bias?

1. Resolution
2. Linearity
3. Hysteresis
4. Drift
5. Differences among gauges
6. Differences among geometries, configurations

Remedies and strategies

How do we quantify uncertainties of measurements made with the gauges?

2.4.1. What are the important issues?

Basic issues

The basic issue for the study is the behavior of gauges in a particular environment with respect to:

- Repeatability
- Reproducibility
- Stability
- Bias

Strategy

The strategy is to conduct and analyze a study that examines the behavior of similar gauges to see if:

- They exhibit different levels of precision;
- Instruments in the same environment produce equivalent results;
- Operators in the same environment produce equivalent results;
- Responses of individual gauges are affected by configuration or geometry changes or changes in setup procedures.

Other goals

Other goals are to:

- Test the resolution of instruments
- Test the gauges for linearity
- Estimate differences among gauges (bias)
- Estimate differences caused by geometries, configurations
- Estimate operator biases
- Incorporate the findings in an uncertainty budget

2.4.2. Design considerations

Design considerations

Design considerations for a gauge study are choices of:

- Artifacts (check standards)
- Operators
- Gauges
- Parameter levels
- Configurations, etc.

Selection of artifacts or check standards

The artifacts for the study are check standards or test items of a type that are typically measured with the gauges under study. It may be necessary to include check standards for different parameter levels if the gauge is a multi-response instrument. The discussion of check standards should be reviewed to determine the suitability of available artifacts.

Number of artifacts

The number of artifacts for the study should be Q $(Q > 2)$. Check standards for a gauge study are needed only for the limited time period (two or three months) of the study.

Selection of operators

Only those operators who are trained and experienced with the gauges should be enlisted in the study, with the following constraints:

- If there is a small number of operators who are familiar with the gauges, they should all be included in the study.
- If the study is intended to be representative of a large pool of operators, then a random sample of L $(L > 2)$ operators should be chosen from the pool.
- If there is only one operator for the gauge type, that operator should make measurements on K $(K > 2)$ days.

Selection of gauges

If there is only a small number of gauges in the facility, then all gauges should be included in the study.

If the study is intended to represent a larger pool of gauges, then a random sample of I $(I > 3)$ gauges should be chosen for the study.

Limit the

If the gauges operate at several parameter levels (for

initial study

example; frequencies), an initial study should be carried out at 1 or 2 levels before a larger study is undertaken.

If there are differences in the way that the gauge can be operated, an initial study should be carried out for one or two configurations before a larger study is undertaken.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

2.4.3. Data collection for time-related sources of variability

Time-related analysis

The purpose of this page is to present several options for collecting data for estimating time-dependent effects in a measurement process.

Time intervals

The following levels of time-dependent errors are considered in this section based on the characteristics of many measurement systems and should be adapted to a specific measurement situation as needed.

1. Level-1 Measurements taken over a short time to capture the precision of the gauge
2. Level-2 Measurements taken over days (of other appropriate time increment)
3. Level-3 Measurements taken over runs separated by months

Time intervals

- Simple design for 2 levels of random error
- Nested design for 2 levels of random error
- Nested design for 3 levels of random error

In all cases, data collection and analysis are straightforward, and there is no reason to estimate interaction terms when dealing with time-dependent errors. Two levels should be sufficient for characterizing most measurement systems. Three levels are recommended for measurement systems where sources of error are not well understood and have not previously been studied.

2.4.3.1. Simple design

Constraints on time and resources

In planning a gauge study, particularly for the first time, it is advisable to start with a simple design and progress to more complicated and/or labor intensive designs after acquiring some experience with data collection and analysis. The design recommended here is appropriate as a preliminary study of variability in the measurement process that occurs over time. It requires about two days of measurements separated by about a month with two repetitions per day.

Relationship to 2-level and 3-level nested designs

The disadvantage of this design is that there is minimal data for estimating variability over time. A 2-level nested design and a 3-level nested design, both of which require measurments over time, are discussed on other pages.

Plan of action

Choose at least $Q = 10$ work pieces or check standards, which are essentially identical insofar as their expected responses to the measurement method. Measure each of the check standards twice with the same gauge, being careful to randomize the order of the check standards.

After about a month, repeat the measurement sequence, randomizing anew the order in which the check standards are measured.

Notation

Measurements on the check standards are designated:

$$Y_{11}, Y_{12}$$
$$Y_{21}, Y_{22}$$

with the first index identifying the month of measurement and the second index identifying the repetition number.

Analysis of data

The level-1 standard deviation, which describes the basic precision of the gauge, is

$$s_1 = \sqrt{\frac{1}{4Q}\sum_{i=1}^{Q}\{(Y_{11} - Y_{12})^2 + (Y_{21} - Y_{22})^2\}}$$

with $v_1 = 2Q$ degrees of freedom.

The level-2 standard deviation, which describes the variability of the measurement process over time, is

$$s_2 = \sqrt{\frac{1}{Q}\sum_{i=1}^{Q}\left\{\frac{(Y_{11}+Y_{12})-(Y_{21}+Y_{22})}{2}\right\}^2}$$

with $v_2 = Q$ degrees of freedom.

Relationship to uncertainty for a test item

The standard deviation that defines the uncertainty for a single measurement on a test item, often referred to as the reproducibility standard deviation (ASTM), is given by

$$s_R = \sqrt{s_{days}^2 + s_1^2} = \frac{s_2}{\sqrt{2}}$$

The time-dependent component is

$$s_{days} = \sqrt{\frac{1}{2}s_2^2 - s_1^2}$$

There may be other sources of uncertainty in the measurement process that must be accounted for in a formal analysis of uncertainty.

2.4.3.2. 2-level nested design

Check standard measurements for estimating time-dependent sources of variability

Measurements on a check standard are recommended for studying the effect of sources of variability that manifest themselves over time. Data collection and analysis are straightforward, and there is no reason to estimate interaction terms when dealing with time-dependent errors. The measurements can be made at one of two levels. Two levels should be sufficient for characterizing most measurement systems. Three levels are recommended for measurement systems for which sources of error are not well understood and have not previously been studied.

Time intervals in a nested design

The following levels are based on the characteristics of many measurement systems and should be adapted to a specific measurement situation as needed.

- Level-1 Measurements taken over a short term to estimate gauge precision
- Level-2 Measurements taken over days (of other appropriate time increment)

Definition of number of measurements at each level

The following symbols are defined for this chapter:

- Level-1 J $(J > 1)$ repetitions
- Level-2 K $(K > 2)$ days

Schedule for making measurements

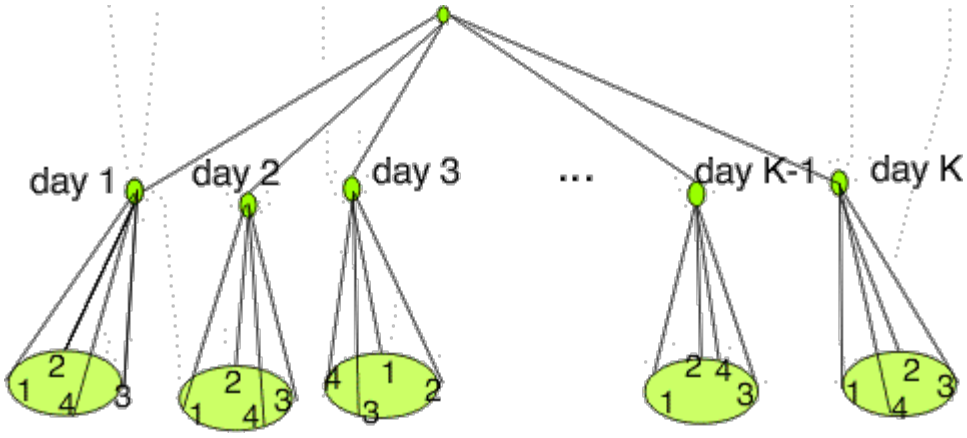
A schedule for making check standard measurements over time (once a day, twice a week, or whatever is appropriate for sampling all conditions of measurement) should be set up and adhered to. The check standard measurements should be structured in the same way as values reported on the test items. For example, if the reported values are averages of two repetitions made within 5 minutes of each other, the check standard values should be averages of the two measurements made in the same manner.

Exception

One exception to this rule is that there should be at least $J = 2$ repetitions per day, etc. Without this redundancy, there is no way to check on the short-term precision of the measurement system.

Depiction of schedule for making check standard

measurements with 4 repetitions per day over K days on the surface of a silicon wafer



K days - 4 repetitions

2-level design for check standard measurements

Operator considerations

The measurements should be taken with *ONE* operator. Operator is not usually a consideration with automated systems. However, systems that require decisions regarding line edge or other feature delineations may be operator dependent.

Case Study: Resistivity check standard

Results should be recorded along with pertinent environmental readings and identifications for significant factors. The best way to record this information is in one file with one line or row (on a spreadsheet) of information in fixed fields for each check standard measurement.

Data analysis of gauge precision

The check standard measurements are represented by

$$Y_{kj} \left(k = 1, \cdots K; j = 1, \cdots J\right)$$

for the *j*th repetition on the *k*th day. The mean for the *k*th day is

$$\bar{Y}_{k\bullet} = \frac{1}{J} \sum_{j=1}^{J} Y_{kj}$$

and the (level-1) standard deviation for gauge precision with $v = J - 1$ degrees of freedom is

$$s_k = \sqrt{\frac{1}{J-1} \sum_{j=1}^{J} \left(Y_{kj} - \bar{Y}_{k\bullet} \right)^2} .$$

Pooling increases the reliability of the estimate of the standard deviation

The pooled level-1 standard deviation with $v = K(J - 1)$ degrees of freedom is

$$s_1 = \sqrt{\frac{1}{K} \sum_{k=1}^{K} s_k^2} .$$

Data analysis of process

The level-2 standard deviation of the check standard represents the process variability. It is computed with $v = K - 1$ degrees of freedom as:

(level-2) standard deviation

$$s_{chkstd} = s_2 = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{Y}_{k\bullet} - \bar{Y}_{\bullet\bullet}\right)^2}$$

where

$$\bar{Y}_{\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{k\bullet}$$

Relationship to uncertainty for a test item

The standard deviation that defines the uncertainty for a single measurement on a test item, often referred to as the reproducibility standard deviation (ASTM), is given by

$$s_R = \sqrt{s_{day}^2 + s_1^2} = \sqrt{s_2^2 + \frac{J-1}{J}s_1^2}$$

The time-dependent component is

$$s_{day} = \sqrt{s_2^2 - \frac{1}{J}s_1^2}$$

There may be other sources of uncertainty in the measurement process that must be accounted for in a formal analysis of uncertainty.

2.4.3.3. 3-level nested design

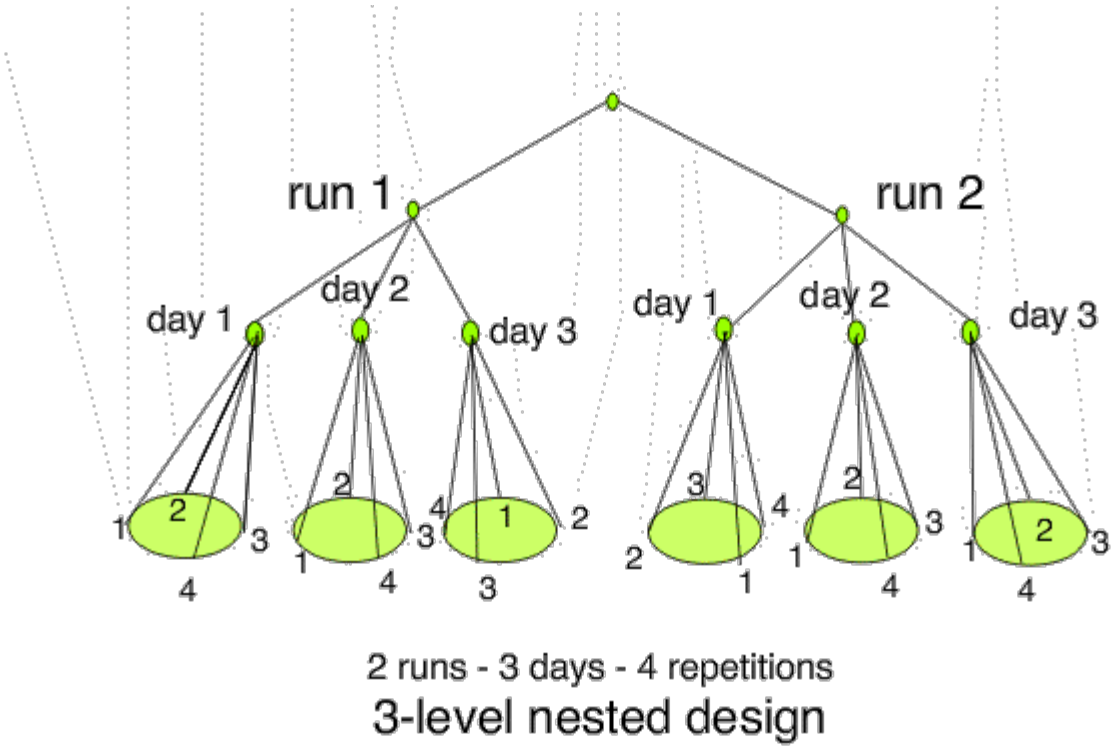
Advantages of nested designs

A nested design is recommended for studying the effect of sources of variability that manifest themselves over time. Data collection and analysis are straightforward, and there is no reason to estimate interaction terms when dealing with time-dependent errors. Nested designs can be run at several levels. Three levels are recommended for measurement systems where sources of error are not well understood and have not previously been studied.

Time intervals in a nested design

The following levels are based on the characteristics of many measurement systems and should be adapted to a specific measurement situation as need be. A typical design is shown below.

- Level-1 Measurements taken over a short-time to capture the precision of the gauge
- Level-2 Measurements taken over days (or other appropriate time increment)
- Level-3 Measurements taken over runs separated by months



Definition of number of measurements at each level

The following symbols are defined for this chapter:

- Level-1 *J (J > 1)* repetitions
- Level-2 *K (K > 2)* days
- Level-3 *L (L > 2)* runs

For the design shown above, *J = 4; K = 3 and L = 2*. The design can be repeated for:

- *Q (Q > 2)* check standards
- *I (I > 3)* gauges if the intent is to characterize several similar gauges

2-level nested design

The design can be truncated at two levels to estimate repeatability and day-to-day variability if there is no reason to estimate longer-term effects. The analysis remains the same through the first two levels.

Advantages

This design has advantages in ease of use and computation. The number of repetitions at each level need not be large because information is being gathered on several check standards.

Operator considerations

The measurements should be made with *ONE* operator. Operator is not usually a consideration with automated systems. However, systems that require decisions regarding

line edge or other feature delineations may be operator dependent. If there is reason to believe that results might differ significantly by operator, 'operators' can be substituted for 'runs' in the design. Choose *L (L > 2)* operators at random from the pool of operators who are capable of making measurements at the same level of precision. (Conduct a small experiment with operators making repeatability measurements, if necessary, to verify comparability of precision among operators.) Then complete the data collection and analysis as outlined. In this case, the level-3 standard deviation estimates operator effect.

Caution

Be sure that the design is truly nested; i.e., that each operator reports results for the same set of circumstances, particularly with regard to day of measurement so that each operator measures every day, or every other day, and so forth.

Randomize on gauges

Randomize with respect to gauges for each check standard; i.e., choose the first check standard and randomize the gauges; choose the second check standard and randomize gauges; and so forth.

Record results in a file

Record the average and standard deviation from each group of *J* repetitions by:

- check standard
- gauge

Case Study: Resistivity Gauges

Results should be recorded along with pertinent environmental readings and identifications for significant factors. The best way to record this information is in one file with one line or row (on a spreadsheet) of information in fixed fields for each check standard measurement. A list of typical entries follows.

1. Month
2. Day
3. Year
4. Operator identification
5. Check standard identification
6. Gauge identification
7. Average of J repetitions
8. Short-term standard deviation from *J* repetitions
9. Degrees of freedom
10. Environmental readings (if pertinent)

2.4.4. Analysis of variability

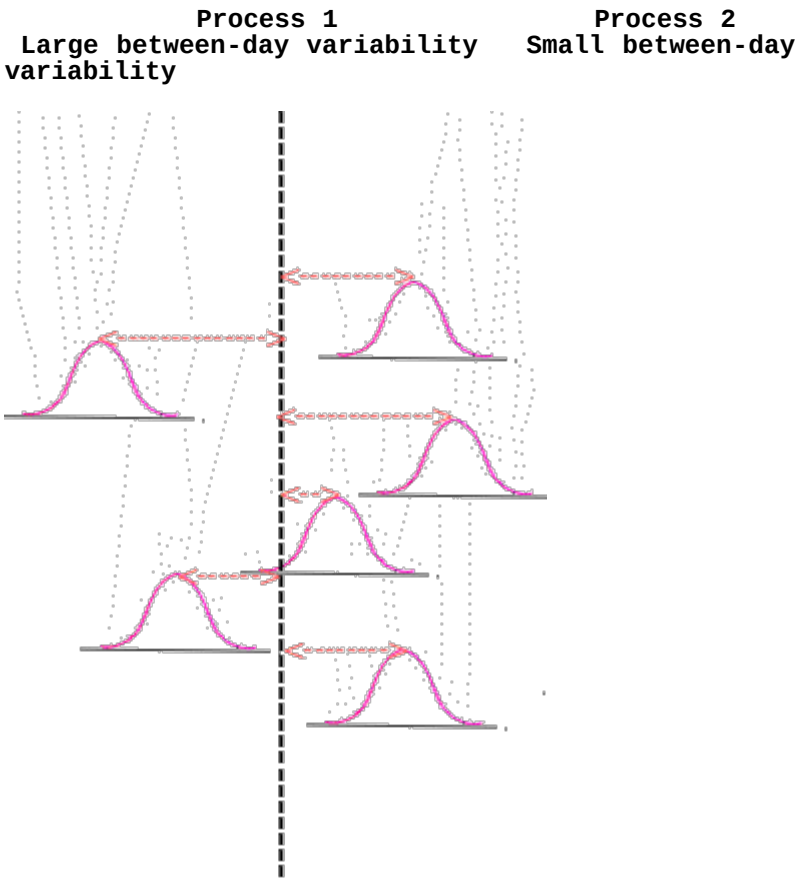
Analysis of variability from a nested design

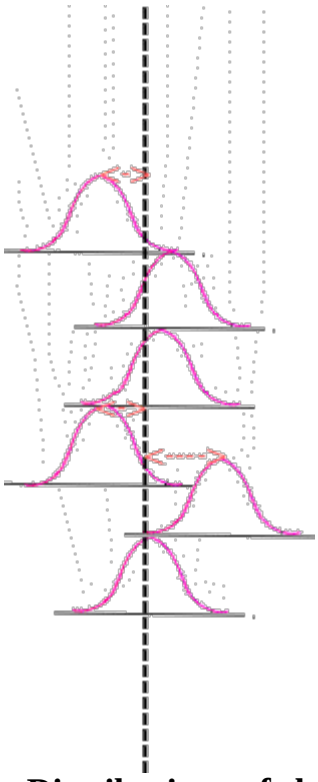
The purpose of this section is to show the effect of various levels of time-dependent effects on the variability of the measurement process with standard deviations for each level of a 3-level nested design.

- Level 1 - repeatability/short-term precision
- Level 2 - reproducibility/day-to-day
- Level 3 - stability/run-to-run

The graph below depicts possible scenarios for a 2-level design (short-term repetitions and days) to illustrate the concepts.

Depiction of 2 measurement processes with the same short-term variability over 6 days where process 1 has large between-day variability and process 2 has negligible between-day variability





Distributions of short-term measurements over 6 days where distances from centerlines illustrate between-day variability

Hint on using tabular method of analysis

An easy way to begin is with a 2-level table with *J* columns and *K* rows for the repeatability/reproducibility measurements and proceed as follows:

1. Compute an average for each row and put it in the *J+1* column.
2. Compute the level-1 (repeatability) standard deviation for each row and put it in the *J+2* column.
3. Compute the grand average and the level-2 standard deviation from data in the *J+1* column.
4. Repeat the table for each of the *L* runs.
5. Compute the level-3 standard deviation from the *L* grand averages.

Level-1: LK repeatability standard deviations can be computed from the data

The measurements from the nested design are denoted by

$$Y_{lkj}\,(l=1,\cdots,L;k=1,\cdots,K;j=1,\cdots,J)$$

Equations corresponding to the tabular analysis are shown below. Level-1 repeatability standard deviations, s_{1lk}, are pooled over the *K* days and *L* runs. Individual standard deviations with *(J - 1)* degrees of freedom each are computed from *J* repetitions as

$$s_{1lk} = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(Y_{lkj} - \bar{Y}_{lk.})^2}$$

where

$$\bar{Y}_{lk\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{lkj}$$

Level-2: L reproducibility standard deviations can be computed from the data

The level-2 standard deviation, s_{2l}, is pooled over the L runs. Individual standard deviations with *(K - 1)* degrees of freedom each are computed from K daily averages as

$$s_{2l} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(Y_{lk.} - \bar{Y}_{l..})^2}$$

where

$$\bar{Y}_{l\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{lk\bullet}$$

Level-3: A single global standard deviation can be computed from the L-run averages

A level-3 standard deviation with *(L - 1)* degrees of freedom is computed from the L-run averages as

$$s_3 = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}\left(\bar{Y}_{l\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}\right)^2}$$

where

$$\bar{Y}_{..} = \frac{1}{L}\sum_{l=1}^{L}\bar{Y}_{l..}$$

Relationship to uncertainty for a test item

The standard deviation that defines the uncertainty for a single measurement on a test item is given by

$$s_R = \sqrt{s_{runs}^2 + s_{days}^2 + s_1^2} = \sqrt{s_3^2 + \frac{K-1}{K}s_2^2 + \frac{J-1}{J}s_1^2}$$

where the pooled values, s_1 and s_2, are the usual

$$s_1 = \sqrt{\frac{\sum_{l=1}^{L}\sum_{k=1}^{K} s_{1lk}^2}{LK}}$$

and

$$s_2 = \sqrt{\frac{1}{L}\sum_{l=1}^{L} s_{2l}^2}$$

There may be other sources of uncertainty in the measurement process that must be accounted for in a formal analysis of uncertainty.

2.4.4.1. Analysis of repeatability

Case study: Resistivity probes

The repeatability quantifies the basic precision for the gauge. A level-1 repeatability standard deviation is computed for each group of J repetitions, and a graphical analysis is recommended for deciding if repeatability is dependent on the check standard, the operator, or the gauge. Two graphs are recommended. These should show:

- Plot of repeatability standard deviations versus check standard with day coded
- Plot of repeatability standard deviations versus check standard with gauge coded

Typically, we expect the standard deviation to be gauge dependent -- in which case there should be a separate standard deviation for each gauge. If the gauges are all at the same level of precision, the values can be combined over all gauges.

Repeatability standard deviations can be pooled over operators, runs, and check standards

A repeatability standard deviation from J repetitions is not a reliable estimate of the precision of the gauge. Fortunately, these standard deviations can be pooled over days; runs; and check standards, if appropriate, to produce a more reliable precision measure. The table below shows a mechanism for pooling. The pooled repeatability standard deviation, s_1, has $LK(J - 1)$ degrees of freedom for measurements taken over:

- J repetitions
- K days
- L runs

Basic pooling rules

The table below gives the mechanism for pooling repeatability standard deviations over days and runs. The pooled value is an average of weighted variances and is shown as the last entry in the right-hand column of the table. The pooling can also cover check standards, if appropriate.

View of entire dataset from the nested design

To illustrate the calculations, a subset of data collected in a nested design for one check standard (#140) and one probe (#2362) are shown below. The measurements are resistivity (ohm.cm) readings with six repetitions per day. The individual level-1 standard deviations from the six repetitions and degrees of freedom are recorded in the last two columns of the database.

Run	Wafer	Probe	Month	Day	Op	Temp	Average	Stddev	df
1	140	2362	3	15	1	23.08	96.0771	0.1024	5
1	140	2362	3	17	1	23.00	95.9976	0.0943	5
1	140	2362	3	18	1	23.01	96.0148	0.0622	5
1	140	2362	3	22	1	23.27	96.0397	0.0702	5
1	140	2362	3	23	2	23.24	96.0407	0.0627	5
1	140	2362	3	24	2	23.13	96.0445	0.0622	5
2	140	2362	4	12	1	22.88	96.0793	0.0996	5
2	140	2362	4	18	2	22.76	96.1115	0.0533	5
2	140	2362	4	19	2	22.79	96.0803	0.0364	5
2	140	2362	4	19	1	22.71	96.0411	0.0768	5
2	140	2362	4	20	2	22.84	96.0988	0.1042	5
2	140	2362	4	21	1	22.94	96.0482	0.0868	5

Pooled repeatability standard deviations over days, runs

Source of Variability	Degrees of Freedom	Standard Deviations	Sum of Squares (SS)
Probe 2362	ν_i	s_{1i}	$SS_i = \nu_i \cdot s_{1i}^2$
run 1 - day 1	5	0.1024	0.05243
run 1 - day 2	5	0.0943	0.04446
run 1 - day 3	5	0.0622	0.01934
run 1 - day 4	5	0.0702	0.02464
run 1 - day 5	5	0.0627	0.01966
run 1 - day 6	5	0.0622	0.01934
run 2 - day 1	5	0.0996	0.04960
run 2 - day 2	5	0.0533	0.01420
run 2 - day 3	5	0.0364	0.00662
run 2 - day 4	5	0.0768	0.02949
run 2 - day 5	5	0.1042	0.05429
run 2 - day 6	5	0.0868	0.03767
$\nu = \sum \nu_i \rightarrow$ gives the total degrees of freedom for s_1	**60**	$SS = \sum SS_i \rightarrow$ gives the total sum of squares for s_1	**0.37176**
The pooled value of s_1 is given by			**0.07871**

$s_1 = \sqrt{SS/\nu} \rightarrow$	

The calculations displayed in the table above can be generated using both Dataplot code and R code.

2.4.4.2. Analysis of reproducibility

Case study: Resistivity gauges

Day-to-day variability can be assessed by a graph of check standard values (averaged over *J* repetitions) versus day with a separate graph for each check standard. Graphs for all check standards should be plotted on the same page to obtain an overall view of the measurement situation.

Pooling results in more reliable estimates

The level-2 standard deviations with (*K* - 1) degrees of freedom are computed from the check standard values for days and pooled over runs as shown in the table below. The pooled level-2 standard deviation has degrees of freedom
L(*K* - 1) for measurements made over:

- *K* days
- *L* runs

Mechanism for pooling

The table below gives the mechanism for pooling level-2 standard deviations over runs. The pooled value is an average of weighted variances and is the last entry in the right-hand column of the table. The pooling can be extended in the same manner to cover check standards, if appropriate.

The table was generated using a subset of data (shown on previous page) collected in a nested design on one check standard (#140) with probe (#2362) over six days. The data are analyzed for between-day effects. The level-2 standard deviations and pooled level-2 standard deviations over runs 1 and 2 are:

Level-2 standard deviations for a single gauge pooled over runs

Source of variability	Standard deviations	Degrees of freedom	Sum of squares
Days	s_{2i}	v_i	$SS_i = v_i \cdot {s_{2i}}^2$
Run 1 Run 2	0.027280 0.027560	5 5 ------ -	0.003721 0.003798 ----------

Sum		10	0.007519
Pooled value $s_2 = \sqrt{\sum SS_i / \sum v_i}$			0.02742

Relationship to day effect

The level-2 standard deviation is related to the standard deviation for between-day precision and gauge precision by

$$s_{days} = \sqrt{s_2^2 - \frac{1}{j} s_1^2}$$

The size of the day effect can be calculated by subtraction using the formula above once the other two standard deviations have been estimated reliably.

Computation of variance component for days

For our example, the variance component for between days is -0.00028072. The negative number for the variance is interpreted as meaning that the variance component for days is zero. However, with only 10 degrees of freedom for the level-2 standard deviation, this estimate is not necessarily reliable. The standard deviation for days over the entire database shows a significant component for days.

Sample code

The calculations included in this section can be implemented using both
Dataplot code and R code.

2.4.4.3. Analysis of stability

Case study: Resistivity probes

Run-to-run variability can be assessed graphically by a plot of check standard values (averaged over J repetitions) versus time with a separate graph for each check standard. Data on all check standards should be plotted on one page to obtain an overall view of the measurement situation.

Advantage of pooling

A level-3 standard deviation with (L - 1) degrees of freedom is computed from the run averages. Because there will rarely be more than two runs per check standard, resulting in one degree of freedom per check standard, it is prudent to have three or more check standards in the design to take advantage of pooling. The mechanism for pooling over check standards is shown in the table below. The pooled standard deviation has $Q(L - 1)$ degrees and is shown as the last entry in the right-hand column of the table.

Example of pooling

The following table shows how the level-3 standard deviations for a single gauge (probe #2362) are pooled over check standards. The table can be reproduced using R code.

Level-3 standard deviations for a single gauge pooled over check standards

Source of variability	Standard deviation	Degrees of freedom	Sum of squares
Level-3	s_{3i}	v_i	$SS_i = v_i \cdot {s_{3i}}^2$
Chk std 138	0.0223	1	0.0004973
Chk std 139	0.0027	1	0.0000073
Chk std 140	0.0289	1	0.0008352
Chk std 141	0.0133	1	0.0001769
Chk std 142	0.0205	1	0.0004203
Sum		5	0.0019370
Pooled value $s_3 = \sqrt{\sum SS_i / \sum v_i}$			**0.0197**

Level-3 standard deviations

A subset of data collected in a nested design on one check standard (#140) with probe (#2362) for six days and two runs is analyzed for between-run effects. The level-3 standard deviation, computed from the averages of two runs, is 0.02885 with one degree of freedom. Dataplot code and R code can be used to perform the calculations for this data.

Relationship to long-term changes, days and gauge precision

The size of the between-run effect can be calculated by subtraction using the standard deviations for days and gauge precision as

$$s_{runs} = \sqrt{s_3^2 - \frac{1}{K}s_2^2} = \sqrt{s_3^2 - \frac{1}{K}s_{days}^2 - \frac{1}{KJ}s_1^2}$$

2.4.4.4.4. Example of calculations

Example of repeatability calculations

Short-term standard deviations based on

- $J = 6$ repetitions with 5 degrees of freedom
- $K = 6$ days
- $L = 2$ runs

were recorded with a probing instrument on $Q = 5$ wafers. The standard deviations were pooled over $K = 6$ days and $L = 2$ runs to give 60 degrees of freedom for each wafer. The pooling of repeatability standard deviations over the 5 wafers is demonstrated in the table below.

Pooled repeatability standard deviation for a single gauge

Source of variability	Sum of Squares (SS)	Degrees of freedom (DF)	Std Devs
Repeatability	$SS_i = \nu_i \cdot s_{Ii}^2$	ν_i	$s_I = \sqrt{SS/\nu} \rightarrow$
Wafer #138	0.48115	60	
Wafer #139	0.69209	60	
Wafer #140	0.48483	60	
Wafer #141	1.21752	60	
Wafer #142	0.30076	60	
SUM	**3.17635**	**300**	**0.10290**

2.4.5. Analysis of bias

Definition of bias

The terms 'bias' and 'systematic error' have the same meaning in this handbook. Bias is defined (VIM) as the difference between the measurement result and its unknown 'true value'. It can often be estimated and/or eliminated by calibration to a reference standard.

Potential problem

Calibration relates output to 'true value' in an ideal environment. However, it may not assure that the gauge reacts properly in its working environment. Temperature, humidity, operator, wear, and other factors can introduce bias into the measurements. There is no single method for dealing with this problem, but the gauge study is intended to uncover biases in the measurement process.

Sources of bias

Sources of bias that are discussed in this Handbook include:

- Lack of gauge resolution
- Lack of linearity
- Drift
- Hysteresis
- Differences among gauges
- Differences among geometries
- Differences among operators
- Remedial actions and strategies

2.4.5.1. Resolution

Resolution	Resolution (MSA) is the ability of the measurement system to detect and faithfully indicate small changes in the characteristic of the measurement result.
Definition from (MSA) manual	The resolution of the instrument is δ if there is an equal probability that the indicated value of any artifact, which differs from a reference standard by less than δ, will be the same as the indicated value of the reference.
Good versus poor	A small δ implies good resolution -- the measurement system can discriminate between artifacts that are close together in value. A large δ implies poor resolution -- the measurement system can only discriminate between artifacts that are far apart in value.
Warning	The number of digits displayed does not indicate the resolution of the instrument.
Manufacturer's statement of resolution	Resolution as stated in the manufacturer's specifications is usually a function of the least-significant digit (LSD) of the instrument and other factors such as timing mechanisms. This value should be checked in the laboratory under actual conditions of measurement.
Experimental determination of resolution	To make a determination in the laboratory, select several artifacts with known values over a range from close in value to far apart. Start with the two artifacts that are farthest apart and make measurements on each artifact. Then, measure the two artifacts with the second largest difference, and so forth, until two artifacts are found which repeatedly give the same result. The difference between the values of these two artifacts estimates the resolution.
Consequence of poor resolution	No useful information can be gained from a study on a gauge with poor resolution relative to measurement needs.

2.4.5.2. Linearity of the gauge

Definition of linearity for gauge studies

Linearity is given a narrow interpretation in this Handbook to indicate that gauge response increases in equal increments to equal increments of stimulus, or, if the gauge is biased, that the bias remains constant throughout the course of the measurement process.

Data collection and repetitions

A determination of linearity requires Q $(Q > 4)$ reference standards that cover the range of interest in fairly equal increments and J $(J > 1)$ measurements on each reference standard. One measurement is made on each of the reference standards, and the process is repeated J times.

Plot of the data

A test of linearity starts with a plot of the measured values versus corresponding values of the reference standards to obtain an indication of whether or not the points fall on a straight line with slope equal to 1 -- indicating linearity.

Least-squares estimates of bias and slope

A least-squares fit of the data to the model

$$Y = a + bX + \text{measurement error}$$

where Y is the measurement result and X is the value of the reference standard, produces an estimate of the intercept, a, and the slope, b.

Output from software package

The intercept and bias are estimated using a statistical software package that should provide the following information:

- Estimates of the intercept and slope, $\hat{a}, \hat{b}$
- Standard deviations of the intercept and slope
- Residual standard deviation of the fit
- F-test for goodness of fit

Test for linearity

Tests for the slope and bias are described in the section on instrument calibration. If the slope is different from one, the gauge is non-linear and requires calibration or repair. If the intercept is different from zero, the gauge has a bias.

Causes of non-linearity

The reference manual on Measurement Systems Analysis (MSA) lists possible causes of gauge non-linearity that should be investigated if the gauge shows symptoms of non-linearity.

1. Gauge not properly calibrated at the lower and upper ends of the operating range
2. Error in the value of X at the maximum or minimum range
3. Worn gauge
4. Internal design problems (electronics)

Note - on artifact calibration

The requirement of linearity for artifact calibration is not so stringent. Where the gauge is used as a comparator for measuring small differences among test items and reference standards of the same nominal size, as with calibration designs, the only requirement is that the gauge be linear over the small on-scale range needed to measure both the reference standard and the test item.

Situation where the calibration of the gauge is neglected

Sometimes it is not economically feasible to correct for the calibration of the gauge (Turgel and Vecchia). In this case, the bias that is incurred by neglecting the calibration is estimated as a component of uncertainty.

2.4.5.3. Drift

Definition	Drift can be defined (VIM) as a slow change in the response of a gauge.
Instruments used as comparators for calibration	Short-term drift can be a problem for comparator measurements. The cause is frequently heat build-up in the instrument during the time of measurement. It would be difficult, and probably unproductive, to try to pinpoint the extent of such drift with a gauge study. The simplest solution is to use drift-free designs for collecting calibration data. These designs mitigate the effect of linear drift on the results. Long-term drift should not be a problem for comparator measurements because such drift would be constant during a calibration design and would cancel in the difference measurements.
Instruments corrected by linear calibration	For instruments whose readings are corrected by a linear calibration line, drift can be detected using a control chart technique and measurements on three or more check standards.
Drift in direct reading instruments and uncertainty analysis	For other instruments, measurements can be made on a daily basis on two or more check standards over a preset time period, say, one month. These measurements are plotted on a time scale to determine the extent and nature of any drift. Drift rarely continues unabated at the same rate and in the same direction for a long time period. Thus, the expectation from such an experiment is to document the maximum change that is likely to occur during a set time period and plan adjustments to the instrument accordingly. A further impact of the findings is that uncorrected drift is treated as a type A component in the uncertainty analysis.

2.4.5.4. Differences among gauges

Purpose

A gauge study should address whether gauges agree with one another and whether the agreement (or disagreement) is consistent over artifacts and time.

Data collection

For each gauge in the study, the analysis requires measurements on

- *Q (Q > 2)* check standards
- *K (K > 2)* days

The measurements should be made by a single operator.

Data reduction

The steps in the analysis are:

1. Measurements are averaged over days by artifact/gauge configuration.
2. For each artifact, an average is computed over gauges.
3. Differences from this average are then computed for each gauge.
4. If the design is run as a 3-level design, the statistics are computed separately for each run.

Data from a gauge study

The data in the table below come from resistivity (ohm.cm) measurements on *Q = 5* artifacts on *K = 6* days. Two runs were made which were separated by about a month's time. The artifacts are silicon wafers and the gauges are four-point probes specifically designed for measuring resistivity of silicon wafers. Differences from the wafer means are shown in the table.

Biases for 5 probes from a gauge study with 5 artifacts on 6 days

Table of biases for probes and silicon wafers (ohm.cm)

	Wafers				
Probe	138	139	140	141	142
1	0.02476	-0.00356	0.04002	0.03938	0.00620
181	0.01076	0.03944	0.01871	-0.01072	0.03761
182	0.01926	0.00574	-0.02008	0.02458	

```
-0.00439

  2062   -0.01754   -0.03226   -0.01258   -0.02802
-0.00110

  2362   -0.03725   -0.00936   -0.02608   -0.02522
-0.03830
```

Plot of differences among probes

A graphical analysis can be more effective for detecting differences among gauges than a table of differences. The differences are plotted versus artifact identification with each gauge identified by a separate plotting symbol. For ease of interpretation, the symbols for any one gauge can be connected by dotted lines.

Interpretation

Because the plots show differences from the average by artifact, the center line is the zero-line, and the differences are estimates of bias. Gauges that are *consistently* above or below the other gauges are biased high or low, respectively, relative to the average. The best estimate of bias for a particular gauge is its average bias over the *Q* artifacts. For this data set, notice that probe #2362 is consistently biased low relative to the other probes.

Strategies for dealing with differences among gauges

Given that the gauges are a random sample of like-kind gauges, the best estimate in any situation is an average over all gauges. In the usual production or metrology setting, however, it may only be feasible to make the measurements on a particular piece with one gauge. Then, there are two methods of dealing with the differences among gauges.

1. Correct each measurement made with a particular gauge for the bias of that gauge and report the standard deviation of the correction as a type A uncertainty.
2. Report each measurement as it occurs and assess a type A uncertainty for the differences among the gauges.

2.4.5.5. Geometry/configuration differences

How to deal with configuration differences

The mechanism for identifying and/or dealing with differences among geometries or configurations in an instrument is basically the same as dealing with differences among the gauges themselves.

Example of differences among wiring configurations

An example is given of a study of configuration differences for a single gauge. The gauge, a 4-point probe for measuring resistivity of silicon wafers, can be wired in several ways. Because it was not possible to test all wiring configurations during the gauge study, measurements were made in only two configurations as a way of identifying possible problems.

Data on wiring configurations and a plot of differences between the 2 wiring configurations

Measurements were made on six wafers over six days (except for 5 measurements on wafer 39) with probe #2062 wired in two configurations. This sequence of measurements was repeated after about a month resulting in two runs. Differences between measurements in the two configurations on the same day are shown in the following table.

Differences between wiring configurations

Wafer	Day	Probe	Run 1	Run 2
17.	1	2062.	-0.0108	0.0088
17.	2	2062.	-0.0111	0.0062
17.	3	2062.	-0.0062	0.0074
17.	4	2062.	0.0020	0.0047
17.	5	2062.	0.0018	0.0049
17.	6	2062.	0.0002	0.0000
39.	1	2062.	-0.0089	0.0075
39.	3	2062.	-0.0040	-0.0016
39.	4	2062.	-0.0022	0.0052

```
   39.    5       2062.    -0.0012
0.0085
   39.    6       2062.    -0.0034    -
0.0018

   63.    1       2062.    -0.0016
0.0092
   63.    2       2062.    -0.0111
0.0040
   63.    3       2062.    -0.0059
0.0067
   63.    4       2062.    -0.0078
0.0016
   63.    5       2062.    -0.0007
0.0020
   63.    6       2062.     0.0006
0.0017

  103.    1       2062.    -0.0050
0.0076
  103.    2       2062.    -0.0140
0.0002
  103.    3       2062.    -0.0048
0.0025
  103.    4       2062.     0.0018
0.0045
  103.    5       2062.     0.0016    -
0.0025
  103.    6       2062.     0.0044
0.0035

  125.    1       2062.    -0.0056
0.0099
  125.    2       2062.    -0.0155
0.0123
  125.    3       2062.    -0.0010
0.0042
  125.    4       2062.    -0.0014
0.0098
  125.    5       2062.     0.0003
0.0032
  125.    6       2062.    -0.0017
0.0115
```

Test of difference between configurations

Because there are only two configurations, a *t*-test is used to decide if there is a difference. If

$$t = \left| \frac{\sqrt{N}}{s_{diff}} Avg_{diff} \right| > 2$$

the difference between the two configurations is statistically significant.

The average and standard deviation computed from the 29 differences in each run are shown in the table below along with the *t*-values which confirm that the differences are significant for both runs.

Average differences between wiring configurations

Run	Probe		Average	Std dev	N
	t				
1	2062	-	0.00383	0.00514	29
	-4.0				
2	2062	+	0.00489	0.00400	29
	+6.6				

Unexpected result

The data reveal a wiring bias for both runs that changes direction between runs. This is a somewhat disturbing finding, and further study of the gauges is needed. Because neither wiring configuration is preferred or known to give the 'correct' result, the differences are treated as a component of the measurement uncertainty.

2.4.5.6. Remedial actions and strategies

Variability	The variability of the gauge in its normal operating mode needs to be examined in light of measurement requirements. If the standard deviation is too large, relative to requirements, the uncertainty can be reduced by making repeated measurements and taking advantage of the standard deviation of the average (which is reduced by a factor of $1/\sqrt{n}$ when *n* measurements are averaged).
Causes of excess variability	If multiple measurements are not economically feasible in the workload, then the performance of the gauge must be improved. Causes of variability which should be examined are: • Wear • Environmental effects such as humidity • Temperature excursions • Operator technique
Resolution	There is no remedy for a gauge with insufficient resolution. The gauge will need to be replaced with a better gauge.
Lack of linearity	Lack of linearity can be dealt with by correcting the output of the gauge to account for bias that is dependent on the level of the stimulus. Lack of linearity can be tolerated (left uncorrected) if it does not increase the uncertainty of the measurement result beyond its requirement.
Drift	It would be very difficult to correct a gauge for drift unless there is sufficient history to document the direction and size of the drift. Drift can be tolerated if it does not increase the uncertainty of the measurement result beyond its requirement.
Differences among gauges or configurations	Significant differences among gauges/configurations can be treated in one of two ways: 1. By correcting each measurement for the bias of the specific gauge/configuration.

2. By accepting the difference as part of the uncertainty of the measurement process.

Differences among operators

Differences among operators can be viewed in the same way as differences among gauges. However, an operator who is incapable of making measurements to the required precision because of an untreatable condition, such as a vision problem, should be re-assigned to other tasks.

2.4.6. Quantifying uncertainties from a gauge study

Gauge studies can be used as the basis for uncertainty assessment

One reason for conducting a gauge study is to quantify uncertainties in the measurement process that would be difficult to quantify under conditions of actual measurement.

This is a reasonable approach to take if the results are truly representative of the measurement process in its working environment. Consideration should be given to all sources of error, particularly those sources of error which do not exhibit themselves in the short-term run.

Potential problem with this approach

The potential problem with this approach is that the calculation of uncertainty depends totally on the gauge study. If the measurement process changes its characteristics over time, the standard deviation from the gauge study will not be the correct standard deviation for the uncertainty analysis. One way to try to avoid such a problem is to carry out a gauge study both before and after the measurements that are being characterized for uncertainty. The 'before' and 'after' results should indicate whether or not the measurement process changed in the interim.

Uncertainty analysis requires information about the specific measurement

The computation of uncertainty depends on the particular measurement that is of interest. The gauge study gathers the data and estimates standard deviations for sources that contribute to the uncertainty of the measurement result. However, specific formulas are needed to relate these standard deviations to the standard deviation of a measurement result.

General guidance

The following sections outline the general approach to uncertainty analysis and give methods for combining the standard deviations into a final uncertainty:

1. Approach
2. Methods for type A evaluations
3. Methods for type B evaluations
4. Propagation of error
5. Error budgets and sensitivity coefficients
6. Standard and expanded uncertainties
7. Treatment of uncorrected biases

Type A evaluations of random error

Data collection methods and analyses of random sources of uncertainty are given for the following:

1. Repeatability of the gauge
2. Reproducibility of the measurement process
3. Stability (very long-term) of the measurement process

Biases - Rule of thumb

The approach for biases is to estimate the maximum bias from a gauge study and compute a standard uncertainty from the maximum bias assuming a suitable distribution. The formulas shown below assume a uniform distribution for each bias.

Determining resolution

If the resolution of the gauge is δ, the standard uncertainty for resolution is

$$s_{resolution} = \delta / \sqrt{3}$$

Determining non-linearity

If the maximum departure from linearity for the gauge has been determined from a gauge study, and it is reasonable to assume that the gauge is equally likely to be engaged at any point within the range tested, the standard uncertainty for linearity is

$$s_{linearity} = Max\left|Y_{observed} - Y_{fitted}\right| / \sqrt{3}$$

Hysteresis

Hysteresis, as a performance specification, is defined (NCSL RP-12) as the maximum difference between the upscale and downscale readings on the same artifact during a full range traverse in each direction. The standard uncertainty for hysteresis is

$$s_{hysteresis} = Max\left|Y_{upscale} - Y_{downscale}\right| / \sqrt{3}$$

Determining drift

Drift in direct reading instruments is defined for a specific time interval of interest. The standard uncertainty for drift is

$$s_{drift} = Max\left|Y_0 - Y_t\right| / \sqrt{3}$$

where Y_0 and Y_t are measurements at time zero and t, respectively.

Other biases

Other sources of bias are discussed as follows:

1. Differences among gauges
2. Differences among configurations

Case study:

A case study on type A uncertainty analysis from a gauge

Type A uncertainties from a gauge study	study is recommended as a guide for bringing together the principles and elements discussed in this section. The study in question characterizes the uncertainty of resistivity measurements made on silicon wafers.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

2.5. Uncertainty analysis

Uncertainty measures 'goodness' of a test result

This section discusses the uncertainty of measurement results. Uncertainty is a measure of the 'goodness' of a result. Without such a measure, it is impossible to judge the fitness of the value as a basis for making decisions relating to health, safety, commerce or scientific excellence.

Contents

1. What are the issues for uncertainty analysis?
2. Approach to uncertainty analysis
    1. Steps
3. Type A evaluations
    1. Type A evaluations of random error
        1. Time-dependent components
        2. Measurement configurations
    2. Type A evaluations of material inhomogeneities
        1. Data collection and analysis
    3. Type A evaluations of bias
        1. Treatment of inconsistent bias
        2. Treatment of consistent bias
        3. Treatment of bias with sparse data
4. Type B evaluations
    1. Assumed distributions
5. Propagation of error considerations
    1. Functions of a single variable
    2. Functions of two variables
    3. Functions of several variables
6. Error budgets and sensitivity coefficients
    1. Sensitivity coefficients for measurements on the test item
    2. Sensitivity coefficients for measurements on a check standard
    3. Sensitivity coefficients for measurements with a 2-level design
    4. Sensitivity coefficients for measurements with a 3-level design
    5. Example of error budget
7. Standard and expanded uncertainties
    1. Degrees of freedom
8. Treatment of uncorrected bias

1. Computation of revised uncertainty

2.5.1. Issues

Issues for uncertainty analysis

Evaluation of uncertainty is an ongoing process that can consume time and resources. It can also require the services of someone who is familiar with data analysis techniques, particularly statistical analysis. Therefore, it is important for laboratory personnel who are approaching uncertainty analysis for the first time to be aware of the resources required and to carefully lay out a plan for data collection and analysis.

Problem areas

Some laboratories, such as test laboratories, may not have the resources to undertake detailed uncertainty analyses even though, increasingly, quality management standards such as the ISO 9000 series are requiring that all measurement results be accompanied by statements of uncertainty.

Other situations where uncertainty analyses are problematical are:

- One-of-a-kind measurements
- Dynamic measurements that depend strongly on the application for the measurement

Directions being pursued

What can be done in these situations? There is no definitive answer at this time. Several organizations, such as the National Conference of Standards Laboratories (NCSL) and the International Standards Organization (ISO) are investigating methods for dealing with this problem, and there is a document in draft that will recommend a simplified approach to uncertainty analysis based on results of interlaboratory tests.

Relationship to interlaboratory test results

Many laboratories or industries participate in interlaboratory studies where the test method itself is evaluated for:

- repeatability within laboratories
- reproducibility across laboratories

These evaluations do not lead to uncertainty statements because the purpose of the interlaboratory test is to evaluate, and then improve, the test method as it is

applied across the industry. The purpose of uncertainty analysis is to evaluate the result of a particular measurement, in a particular laboratory, at a particular time. However, the two purposes are related.

Default recommendation for test laboratories

If a test laboratory has been party to an interlaboratory test that follows the recommendations and analyses of an American Society for Testing Materials standard (ASTM E691) or an ISO standard (ISO 5725), the laboratory can, as a default, represent its standard uncertainty for a single measurement as the reproducibility standard deviation as defined in ASTM E691 and ISO 5725. This standard deviation includes components for within-laboratory repeatability common to all laboratories and between-laboratory variation.

Drawbacks of this procedure

The standard deviation computed in this manner describes a future single measurement made at a laboratory randomly drawn from the group and leads to a prediction interval (Hahn & Meeker) rather than a confidence interval. It is not an ideal solution and may produce either an unrealistically small or unacceptably large uncertainty for a particular laboratory. The procedure can reward laboratories with poor performance or those that do not follow the test procedures to the letter and punish laboratories with good performance. Further, the procedure does not take into account sources of uncertainty other than those captured in the interlaboratory test. Because the interlaboratory test is a snapshot at one point in time, characteristics of the measurement process over time cannot be accurately evaluated. Therefore, it is a strategy to be used only where there is no possibility of conducting a realistic uncertainty investigation.

2.5.2. Approach

Procedures in this chapter

The procedures in this chapter are intended for test laboratories, calibration laboratories, and scientific laboratories that report results of measurements from ongoing or well-documented processes.

Pertinent sections

The following pages outline methods for estimating the individual uncertainty components, which are consistent with materials presented in other sections of this Handbook, and rules and equations for combining them into a final expanded uncertainty. The general framework is:

1. ISO Approach
2. Outline of steps to uncertainty analysis
3. Methods for type A evaluations
4. Methods for type B evaluations
5. Propagation of error considerations
6. Uncertainty budgets and sensitivity coefficients
7. Standard and expanded uncertainties
8. Treatment of uncorrected bias

Specific situations are outlined in other places in this chapter

Methods for calculating uncertainties for specific results are explained in the following sections:

- Calibrated values of artifacts
- Calibrated values from calibration curves
  - From propagation of error
  - From check standard measurements
  - Comparison of check standards and propagation of error
- Gauge R & R studies
- Type A components for resistivity measurements
- Type B components for resistivity measurements

ISO definition of uncertainty

Uncertainty, as defined in the ISO Guide to the Expression of Uncertainty in Measurement (GUM) and the International Vocabulary of Basic and General Terms in Metrology (VIM), is a

> "parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be

attributed to the measurand."

Consistent with historical view of uncertainty

This definition is consistent with the well-established concept that an uncertainty statement assigns credible limits to the accuracy of a reported value, stating to what extent that value may differ from its reference value (Eisenhart). In some cases, reference values will be traceable to a national standard, and in certain other cases, reference values will be consensus values based on measurements made according to a specific protocol by a group of laboratories.

Accounts for both random error and bias

The estimation of a possible discrepancy takes into account both random error and bias in the measurement process. The distinction to keep in mind with regard to random error and bias is that random errors cannot be corrected, and biases can, theoretically at least, be corrected or eliminated from the measurement result.

Relationship to precision and bias statements

Precision and bias are properties of a measurement method. Uncertainty is a property of a *specific result* for a single test item that depends on a specific measurement configuration (laboratory/instrument/operator, etc.). It depends on the repeatability of the instrument; the reproducibility of the result over time; the number of measurements in the test result; and all sources of random and systematic error that could contribute to disagreement between the result and its reference value.

Handbook follows the ISO approach

This Handbook follows the ISO approach (GUM) to stating and combining components of uncertainty. To this basic structure, it adds a statistical framework for estimating individual components, particularly those that are classified as type A uncertainties.

Basic ISO tenets

The ISO approach is based on the following rules:

- Each uncertainty component is quantified by a standard deviation.
- All biases are assumed to be corrected and any uncertainty is the uncertainty of the correction.
- Zero corrections are allowed if the bias cannot be corrected and an uncertainty is assessed.
- All uncertainty intervals are symmetric.

ISO approach to classifying sources of error

Components are grouped into two major categories, depending on the source of the data and not on the type of error, and each component is quantified by a standard deviation. The categories are:

- Type A - components evaluated by statistical

methods
- Type B - components evaluated by other means (or in other laboratories)

Interpretation of this classification

One way of interpreting this classification is that it distinguishes between information that comes from sources local to the measurement process and information from other sources -- although this interpretation does not always hold. In the computation of the final uncertainty it makes no difference how the components are classified because the ISO guidelines treat type A and type B evaluations in the same manner.

Rule of quadrature

All uncertainty components (standard deviations) are combined by root-sum-squares (quadrature) to arrive at a 'standard uncertainty', u, which is the standard deviation of the reported value, taking into account all sources of error, both random and systematic, that affect the measurement result.

Expanded uncertainty for a high degree of confidence

If the purpose of the uncertainty statement is to provide coverage with a high level of confidence, an expanded uncertainty is computed as

$$U = k\,u$$

where k is chosen to be the $t_{1-\alpha/2,\nu}$ critical value from the *t*-table with ν degrees of freedom.

For large degrees of freedom, it is suggested to use $k = 2$ to approximate 95% coverage. Details for these calculations are found under degrees of freedom.

Type B evaluations

Type B evaluations apply to random errors and biases for which there is little or no data from the local process, and to random errors and biases from other measurement processes.

2.5.2.1. Steps

Steps in uncertainty analysis - define the result to be reported

The first step in the uncertainty evaluation is the definition of the result to be reported for the test item for which an uncertainty is required. The computation of the standard deviation depends on the number of repetitions on the test item and the range of environmental and operational conditions over which the repetitions were made, in addition to other sources of error, such as calibration uncertainties for reference standards, which influence the final result. If the value for the test item cannot be measured directly, but must be calculated from measurements on secondary quantities, the equation for combining the various quantities must be defined. The steps to be followed in an uncertainty analysis are outlined for two situations:

Outline of steps to be followed in the evaluation of uncertainty for a single quantity

A. **Reported value involves measurements on one quantity.**

1. Compute a type A standard deviation for random sources of error from:
   - Replicated results for the test item.
   - Measurements on a check standard.
   - Measurements made according to a 2-level designed experiment
   - Measurements made according to a 3-level designed experiment
2. Make sure that the collected data and analysis cover all sources of random error such as:
   - instrument imprecision
   - day-to-day variation
   - long-term variation

   and bias such as:
   - differences among instruments
   - operator differences.
3. Compute a standard deviation for each type B component of uncertainty.
4. Combine type A and type B standard deviations into a

standard uncertainty for the reported result using sensitivity factors.

5. Compute an expanded uncertainty.

Outline of steps to be followed in the evaluation of uncertainty involving several secondary quantities

B. - **Reported value involves more than one quantity.**

1. Write down the equation showing the relationship between the quantities.
    - Write-out the propagation of error equation and do a preliminary evaluation, if possible, based on propagation of error.
2. If the measurement result **can be replicated directly**, regardless of the number of secondary quantities in the individual repetitions, treat the uncertainty evaluation as in (A.1) to (A.5) above, being sure to evaluate all sources of random error in the process.
3. If the measurement result **cannot be replicated directly**, treat *each measurement quantity* as in (A.1) and (A.2) and:
    - Compute a standard deviation for each measurement quantity.
    - Combine the standard deviations for the individual quantities into a standard deviation for the reported result via propagation of error.
4. Compute a standard deviation for each type B component of uncertainty.
5. Combine type A and type B standard deviations into a standard uncertainty for the reported result.
6. Compute an expanded uncertainty.
7. Compare the uncerainty derived by propagation of error with the uncertainty derived by data analysis techniques.

2.5.3. Type A evaluations

Type A evaluations apply to both error and bias

Type A evaluations can apply to both random error and bias. The only requirement is that the calculation of the uncertainty component be based on a statistical analysis of data. The distinction to keep in mind with regard to random error and bias is that:

- random errors cannot be corrected
- biases can, theoretically at least, be corrected or eliminated from the result.

Caveat for biases

The ISO guidelines are based on the assumption that all biases are corrected and that the only uncertainty from this source is the uncertainty of the correction. The section on type A evaluations of bias gives guidance on how to assess, correct and calculate uncertainties related to bias.

Random error and bias require different types of analyses

How the source of error affects the reported value and the context for the uncertainty determines whether an analysis of random error or bias is appropriate.

Consider a laboratory with several instruments that can reasonably be assumed to be representative of all similar instruments. Then the differences among these instruments can be considered to be a random effect if the uncertainty statement is intended to apply to the result of any instrument, selected at random, from this batch.

If, on the other hand, the uncertainty statement is intended to apply to one specific instrument, then the bias of this instrument relative to the group is the component of interest.

The following pages outline methods for type A evaluations of:

1. Random errors
2. Bias

2.5.3.1. Type A evaluations of random components

Type A evaluations of random components

Type A sources of uncertainty fall into three main categories:

1. Uncertainties that reveal themselves over time
2. Uncertainties caused by specific conditions of measurement
3. Uncertainties caused by material inhomogeneities

Time-dependent changes are a primary source of random errors

One of the most important indicators of random error is time, with the root cause perhaps being environmental changes over time. Three levels of time-dependent effects are discussed in this section.

Many possible configurations may exist in a laboratory for making measurements

Other sources of uncertainty are related to measurement configurations within the laboratory. Measurements on test items are usually made on a single day, with a single operator, on a single instrument, etc. If the intent of the uncertainty is to characterize all measurements made in the laboratory, the uncertainty should account for any differences due to:

1. instruments
2. operators
3. geometries
4. other

Examples of causes of differences within a laboratory

Examples of causes of differences within a well-maintained laboratory are:

1. Differences among instruments for measurements of derived units, such as sheet resistance of silicon, where the instruments cannot be directly calibrated to a reference base
2. Differences among operators for optical measurements that are not automated and depend strongly on operator sightings
3. Differences among geometrical or electrical

configurations of the instrumentation

Calibrated instruments do not fall in this class	Calibrated instruments do not normally fall in this class because uncertainties associated with the instrument's calibration are reported as type B evaluations, and the instruments in the laboratory should agree within the calibration uncertainties. Instruments whose responses are not directly calibrated to the defined unit are candidates for type A evaluations. This covers situations in which the measurement is defined by a test procedure or standard practice using a specific instrument type.
Evaluation depends on the context for the uncertainty	How these differences are treated depends primarily on the context for the uncertainty statement. The differences, depending on the context, will be treated either as random differences, or as bias differences.
Uncertainties due to inhomogeneities	Artifacts, electrical devices, and chemical substances, etc. can be inhomogeneous relative to the quantity that is being characterized by the measurement process. If this fact is known beforehand, it may be possible to measure the artifact very carefully at a specific site and then direct the user to also measure at this site. In this case, there is no contribution to measurement uncertainty from inhomogeneity. However, this is not always possible, and measurements may be destructive. As an example, compositions of chemical compounds may vary from bottle to bottle. If the reported value for the lot is established from measurements on a few bottles drawn at random from the lot, this variability must be taken into account in the uncertainty statement. Methods for testing for inhomogeneity and assessing the appropriate uncertainty are discussed on another page.

2.5.3.1.1. Type A evaluations of time-dependent effects

Time-dependent changes are a primary source of random errors

One of the most important indicators of random error is time. Effects not specifically studied, such as environmental changes, exhibit themselves over time. Three levels of time-dependent errors are discussed in this section. These can be usefully characterized as:

1. Level-1 or short-term errors (repeatability, imprecision)
2. Level-2 or day-to-day errors (reproducibility)
3. Level-3 or long-term errors (stability - which may not be a concern for all processes)

Day-to-day errors can be the dominant source of uncertainty

With instrumentation that is exceedingly precise in the short run, changes over time, often caused by small environmental effects, are frequently the dominant source of uncertainty in the measurement process. The uncertainty statement is not 'true' to its purpose if it describes a situation that cannot be reproduced over time. The customer for the uncertainty is entitled to know the range of possible results for the measurement result, independent of the day or time of year when the measurement was made.

Two levels may be sufficient

Two levels of time-dependent errors are probably sufficient for describing the majority of measurement processes. Three levels may be needed for new measurement processes or processes whose characteristics are not well understood.

Measurements on test item are used to assess uncertainty only when no other data are available

Repeated measurements on the test item generally do not cover a sufficient time period to capture day-to-day changes in the measurement process. The standard deviation of these measurements is quoted as the estimate of uncertainty only if no other data are available for the assessment. For J short-term measurements, this standard deviation has $\nu = J - 1$ degrees of freedom.

A check standard is the best device for capturing all sources of random error

The best approach for capturing information on time-dependent sources of uncertainties is to intersperse the workload with measurements on a check standard taken at set intervals over the life of the process. The standard deviation of the check standard measurements estimates the overall temporal component of uncertainty directly -- thereby obviating the estimation of individual components.

Nested design for estimating type A uncertainties

Case study: Temporal uncertainty from a 3-level nested design

A less-efficient method for estimating time-dependent sources of uncertainty is a designed experiment. Measurements can be made specifically for estimating two or three levels of errors. There are many ways to do this, but the easiest method is a nested design where J short-term measurements are replicated on K days and the entire operation is then replicated over L runs (months, etc.). The analysis of these data leads to:

- s_1 = standard deviation with $(J-1)$ degrees of freedom for short-term errors
- s_2 = standard deviation with $(K-1)$ degrees of freedom for day-to-day errors
- s_3 = standard deviation with $(L-1)$ degrees of freedom for very long-term errors

Approaches given in this chapter

The computation of the uncertainty of the reported value for a test item is outlined for situations where temporal sources of uncertainty are estimated from:

1. measurements on the test item itself
2. measurements on a check standard
3. measurements from a 2-level nested design (gauge study)
4. measurements from a 3-level nested design (gauge study)

2. Measurement Process Characterization
2.5. Uncertainty analysis
2.5.3. Type A evaluations
2.5.3.1. Type A evaluations of random components

2.5.3.1.2. Measurement configuration within the laboratory

Purpose of this page

The purpose of this page is to outline options for estimating uncertainties related to the specific measurement configuration under which the test item is measured, given other possible measurement configurations. Some of these may be controllable and some of them may not, such as:

- instrument
- operator
- temperature
- humidity

The effect of uncontrollable environmental conditions in the laboratory can often be estimated from check standard data taken over a period of time, and methods for calculating components of uncertainty are discussed on other pages. Uncertainties resulting from controllable factors, such as operators or instruments chosen for a specific measurement, are discussed on this page.

First, decide on context for uncertainty

The approach depends primarily on the context for the uncertainty statement. For example, if instrument effect is the question, one approach is to regard, say, the instruments in the laboratory as a random sample of instruments of the same type and to compute an uncertainty that applies to all results regardless of the particular instrument on which the measurements are made. The other approach is to compute an uncertainty that applies to results using a specific instrument.

Next, evaluate whether or not there are differences

To treat instruments as a random source of uncertainty requires that we first determine if differences due to instruments are significant. The same can be said for operators, etc.

Plan for collecting data

To evaluate the measurement process for instruments, select a random sample of I $(I > 4)$ instruments from those available. Make measurements on Q $(Q > 2)$ artifacts with each instrument.

Graph showing differences among instruments

For a graphical analysis, differences from the average for each artifact can be plotted versus artifact, with instruments individually identified by a special plotting symbol. The plot is examined to determine if some instruments always read high or low relative to the other instruments and if this behavior is consistent across artifacts. If there are systematic and significant differences among instruments, a type A uncertainty for instruments is computed. Notice that in the graph for resistivity probes, there are differences among the probes with probes #4 and #5, for example, consistently reading low relative to the other probes. A standard deviation that describes the differences among the probes is included as a component of the uncertainty.

Standard deviation for instruments

Given the measurements,

$$Y_{11}, Y_{12}, \cdots Y_{1I}; \cdots; Y_{Q1}, Y_{Q2}, \cdots Y_{QI}$$

for each of Q artifacts and I instruments, the pooled standard deviation that describes the differences among instruments is:

$$s_{inst} = \sqrt{\frac{1}{Q}\frac{1}{I-1}\sum_{q=1}^{Q}\sum_{i=1}^{I}(Y_{qi} - \bar{Y}_{q\cdot})^2}$$

where

$$\bar{Y}_{q\cdot} = \frac{1}{I}\sum_{i=1}^{I} Y_{qi}$$

Example of resistivity measurements on silicon wafers

A two-way table of resistivity measurements (ohm.cm) with 5 probes on 5 wafers (identified as: 138, 139, 140, 141, 142) is shown below. Standard deviations for probes with 4 degrees of freedom each are shown for each wafer. The pooled standard deviation over all wafers, with 20 degrees of freedom, is the type A standard deviation for instruments.

	Wafers				
Probe	138	139	140	141	142
1	95.1548	99.3118	96.1018	101.1248	94.2593
281	95.1408	99.3548	96.0805	101.0747	94.2907
. 283	95.1493	99.3211	96.0417	101.1100	94.2487
2062	95.1125	99.2831	96.0492	101.0574	94.2520
2362	95.0928	99.3060	96.0357	101.0602	94.2148

```
Std dev     0.02643   0.02612   0.02826    0.03038
0.02711
DF                4         4         4          4
4


Pooled standard deviation  =  0.02770      DF =
20
```

2.5.3.2. Material inhomogeneity

Purpose of this page

The purpose of this page is to outline methods for assessing uncertainties related to material inhomogeneities. Artifacts, electrical devices, and chemical substances, etc. can be inhomogeneous relative to the quantity that is being characterized by the measurement process.

Effect of inhomogeneity on the uncertainty

Inhomogeneity can be a factor in the uncertainty analysis where

1. an artifact is characterized by a single value and the artifact is inhomogeneous over its surface, etc.
2. a lot of items is assigned a single value from a few samples from the lot and the lot is inhomogeneous from sample to sample.

An unfortunate aspect of this situation is that the uncertainty from inhomogeneity may dominate the uncertainty. If the measurement process itself is very precise and in statistical control, the total uncertainty may still be unacceptable for practical purposes because of material inhomogeneities.

Targeted measurements can eliminate the effect of inhomogeneity

It may be possible to measure an artifact very carefully at a specific site and direct the user to also measure at this site. In this case there is no contribution to measurement uncertainty from inhomogeneity.

Example

Silicon wafers are doped with boron to produce desired levels of resistivity (ohm.cm). Manufacturing processes for semiconductors are not yet capable (at least at the time this was originally written) of producing 2" diameter wafers with constant resistivity over the surfaces. However, because measurements made at the center of a wafer by a certification laboratory can be reproduced in the industrial setting, the inhomogeneity is not a factor in the uncertainty analysis -- as long as only the center-point of the wafer is used for future measurements.

Random

Random inhomogeneities are assessed using statistical

inhomogeneities

methods for quantifying random errors. An example of inhomogeneity is a chemical compound which cannot be sufficiently homogenized with respect to isotopes of interest. Isotopic ratio determinations, which are destructive, must be determined from measurements on a few bottles drawn at random from the lot.

Best strategy

The best strategy is to draw a sample of bottles from the lot for the purpose of identifying and quantifying between-bottle variability. These measurements can be made with a method that lacks the accuracy required to certify isotopic ratios, but is precise enough to allow between-bottle comparisons. A second sample is drawn from the lot and measured with an accurate method for determining isotopic ratios, and the reported value for the lot is taken to be the average of these determinations. There are therefore two components of uncertainty assessed:

1. component that quantifies the imprecision of the average
2. component that quantifies how much an individual bottle can deviate from the average.

Systematic inhomogeneities

Systematic inhomogeneities require a somewhat different approach. Roughness can vary systematically over the surface of a 2" square metal piece lathed to have a specific roughness profile. The certification laboratory can measure the piece at several sites, but unless it is possible to characterize roughness as a mathematical function of position on the piece, inhomogeneity must be assessed as a source of uncertainty.

Best strategy

In this situation, the best strategy is to compute the reported value as the average of measurements made over the surface of the piece and assess an uncertainty for departures from the average. The component of uncertainty can be assessed by one of several methods for evaluating bias -- depending on the type of inhomogeneity.

Standard method

The simplest approach to the computation of uncertainty for systematic inhomogeneity is to compute the maximum deviation from the reported value and, assuming a uniform, normal or triangular distribution for the distribution of inhomogeneity, compute the appropriate standard deviation. Sometimes the approximate shape of the distribution can be inferred from the inhomogeneity measurements. The standard deviation for inhomogeneity assuming a uniform distribution is:

$$s_{inh} = \frac{1}{\sqrt{3}} MaxDeviation$$

2.5.3.2.1. Data collection and analysis

Purpose of this page

The purpose of this page is to outline methods for:

- collecting data
- testing for inhomogeneity
- quantifying the component of uncertainty

Balanced measurements at 2-levels

The simplest scheme for identifying and quantifying the effect of inhomogeneity of a measurement result is a balanced (equal number of measurements per cell) 2-level nested design. For example, *K* bottles of a chemical compound are drawn at random from a lot and *J (J > 1)* measurements are made per bottle. The measurements are denoted by

$$Y_{11}, Y_{12}, \cdots Y_{1j}; \cdots; Y_{k1}, Y_{k2}, \cdots Y_{kj}$$

where the *k* index runs over bottles and the *j* index runs over repetitions within a bottle.

Analysis of measurements

The between (bottle) variance is calculated using an analysis of variance technique that is repeated here for convenience.

$$s^2_{inh} = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{Y}_{k\cdot} - \bar{Y}_{\cdot\cdot})^2 - \frac{1}{KJ(J-1)}\sum_{k=1}^{K}\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\cdot})^2$$

where

$$\bar{Y}_{k\cdot} = \frac{1}{J}\sum_{j=1}^{J} Y_{kj}$$

and

$$\bar{Y}_{\cdot\cdot} = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_{k\cdot}$$

Between bottle variance may be negative

If this variance is negative, there is no contribution to uncertainty, and the bottles are equivalent with regard to their chemical compositions. Even if the variance is positive, inhomogeneity still may not be statistically significant, in which case it is not required to be included

as a component of the uncertainty.

If the between-bottle variance is statistically significantly (i.e., judged to be greater than zero), then inhomogeneity contributes to the uncertainty of the reported value.

Certification, reported value and associated uncertainty

The purpose of assessing inhomogeneity is to be able to assign a value to the entire batch based on the average of a few bottles, and the determination of inhomogeneity is usually made by a less accurate method than the certification method. The reported value for the batch would be the **average of *N* repetitions on *Q* bottles** using the certification method.

The uncertainty calculation is summarized below for the case where the only contribution to uncertainty from the measurement method itself is the repeatability standard deviation, s_1 associated with the certification method. For more complicated scenarios, see the pages on uncertainty budgets.

If $s_{inh}^2 \leq 0 \Rightarrow s_{reported\ value} = \frac{1}{\sqrt{QN}} s_1$

If $s_{inh}^2 > 0$, we need to distinguish two cases and their interpretations:

1. The standard deviation

$$s_{reported\ value} = \sqrt{\frac{Q+1}{Q} s_{inh}^2 + \frac{1}{QN} s_1^2}$$

   leads to an interval that covers the difference between the reported value and the average for a bottle selected at random from the batch.

2. The standard deviation

$$s_{reported\ value} = \sqrt{\frac{Q+1}{Q} s_{inh}^2 + \frac{QN+1}{QN} s_1^2}$$

   allows one to test the instrument using a single measurement. The prediction interval for the difference between the reported value and a single measurement, made with the same precision as the certification measurements, on a bottle selected at random from the batch. This is appropriate when the instrument under test is similar to the certification instrument. If the difference is not within the interval, the user's instrument is in need of calibration.

Relationship to prediction intervals

When the standard deviation for inhomogeneity is included in the calculation, as in the last two cases above, the uncertainty interval becomes a *prediction interval* (Hahn & Meeker) and is interpreted as characterizing a future measurement on a bottle drawn at random from the lot.

2.5.3.3. Type A evaluations of bias

Sources of bias relate to the specific measurement environment

The sources of bias discussed on this page cover specific measurement configurations. Measurements on test items are usually made on a single day, with a single operator, with a single instrument, etc. Even if the intent of the uncertainty is to characterize only those measurements made in one specific configuration, the uncertainty must account for any significant differences due to:

1. instruments
2. operators
3. geometries
4. other

Calibrated instruments do not fall in this class

Calibrated instruments do not normally fall in this class because uncertainties associated with the instrument's calibration are reported as type B evaluations, and the instruments in the laboratory should agree within the calibration uncertainties. Instruments whose responses are not directly calibrated to the defined unit are candidates for type A evaluations. This covers situations where the measurement is defined by a test procedure or standard practice using a specific instrument type.

The best strategy is to correct for bias and compute the uncertainty of the correction

This problem was treated on the foregoing page as an analysis of random error for the case where the uncertainty was intended to apply to all measurements for all configurations. If measurements for only *one configuration* are of interest, such as measurements made with a specific instrument, or if a smaller uncertainty is required, the differences among, say, instruments are treated as biases. The *best strategy* in this situation is to correct all measurements made with a specific instrument to the average for the instruments in the laboratory and compute a type A uncertainty for the correction. This strategy, of course, relies on the assumption that the instruments in the laboratory represent a random sample of all instruments of a specific type.

Only limited comparisons can be made

However, suppose that it is possible to make comparisons among, say, only two instruments and neither is known to be 'unbiased'. This scenario requires a different strategy

among sources of possible bias

because the average will not necessarily be an unbiased result. The best strategy if there is a significant difference between the instruments, and this should be tested, is to apply a 'zero' correction and assess a type A uncertainty of the correction.

Guidelines for treatment of biases

The discussion above is intended to point out that there are many possible scenarios for biases and that they should be treated on a case-by-case basis. A plan is needed for:

- gathering data
- testing for bias (graphically and/or statistically)
- estimating biases
- assessing uncertainties associated with significant biases.

caused by:

- instruments
- operators
- configurations, geometries, etc.
- inhomogeneities

Plan for testing for assessing bias

Measurements needed for assessing biases among instruments, say, requires a random sample of I $(I > 1)$ instruments from those available and measurements on Q $(Q > 2)$ artifacts with each instrument. The same can be said for the other sources of possible bias. General strategies for dealing with significant biases are given in the table below.

Data collection and analysis for assessing biases related to:

- lack of resolution of instrument
- non-linearity of instrument
- drift

are addressed in the section on gauge studies.

Sources of data for evaluating this type of bias

Databases for evaluating bias may be available from:

- check standards
- gauge R and R studies
- control measurements

Strategies for assessing corrections and uncertainties associated with significant biases

Type of bias	Examples	Type of correction	Uncertainty
1. Inconsistent	Sign change (+ to -) Varying magnitude	Zero	Based on maximum bias

2. Consistent	Instrument bias ~ same magnitude over many artifacts	Bias (for a single instrument) = difference from average over several instruments	Standard deviation of correction
3. Not correctable because of sparse data - consistent or inconsistent	Limited testing; e.g., only 2 instruments, operators, configurations, etc.	Zero	Standard deviation of correction
4. Not correctable - consistent	Lack of resolution, non-linearity, drift, material inhomogeneity	Zero	Based on maximum bias

Strategy for no significant bias If there is no significant bias over time, there is no correction and no contribution to uncertainty.

2.5.3.3.1. Inconsistent bias

Strategy for inconsistent bias -- apply a zero correction

If there is significant bias but it changes direction over time, a zero correction is assumed and the standard deviation of the correction is reported as a type A uncertainty; namely,

$$s_{correction} = \frac{1}{\sqrt{3}} MaxBias$$

Computations based on uniform or normal distribution

The equation for estimating the standard deviation of the correction assumes that biases are uniformly distributed between {-max |bias|, + max |bias|}. This assumption is quite conservative. It gives a larger uncertainty than the assumption that the biases are normally distributed. If normality is a more reasonable assumption, substitute the number '3' for the 'square root of 3' in the equation above.

Example of change in bias over time

The results of resistivity measurements with five probes on five silicon wafers are shown below for probe #283, which is the probe of interest at this level with the artifacts being 1 ohm.cm wafers. The bias for probe #283 is negative for run 1 and positive for run 2 with the runs separated by a two-month time period. The correction is taken to be zero.

Table of biases (ohm.cm) for probe 283

Wafer	Probe	Run 1	Run 2
11	283	0.0000340	-0.0001841
26	283	-0.0001000	0.0000861
42	283	0.0000181	0.0000781
131	283	-0.0000701	0.0001580
208	283	-0.0000240	0.0001879
Average	283	-0.0000284	0.0000652

A conservative assumption is that the bias could fall somewhere within the limits ± a, with a = maximum bias or 0.0000652 ohm.cm. The standard deviation of the correction is included as a type A systematic component of the uncertainty.

$$s_{correction} = \frac{1}{\sqrt{3}} MaxBias = 0.000038\ ohm.cm$$

2.5.3.3.2. Consistent bias

Consistent bias

Bias that is significant and persists consistently over time for a specific instrument, operator, or configuration should be corrected if it can be reliably estimated from repeated measurements. Results with the instrument of interest are then corrected to:

$$\textit{Corrected result} = \textit{Measurement} - \textit{Estimate of bias}$$

The example below shows how bias can be identified graphically from measurements on five artifacts with five instruments and estimated from the differences among the instruments.

Graph showing consistent bias for probe #5

An analysis of bias for five instruments based on measurements on five artifacts shows differences from the average for each artifact plotted versus artifact with instruments individually identified by a special plotting symbol. The plot is examined to determine if some instruments always read high or low relative to the other instruments, and if this behavior is consistent across artifacts. Notice that on the graph for resistivity probes, probe #2362, (#5 on the graph), which is the instrument of interest for this measurement process, consistently reads low relative to the other probes. This behavior is consistent over 2 runs that are separated by a two-month time period.

Strategy - correct for bias

Because there is significant and consistent bias for the instrument of interest, the measurements made with that instrument should be corrected for its average bias relative to the other instruments.

Computation of bias

Given the measurements,

$$Y_{qi}\,(q=1,\cdots,Q;i=1,\cdots,I)$$

on Q artifacts with I instruments, the average bias for instrument, I' say, is

$$\bar{B}_{I'} = \frac{1}{Q}\sum_{q=1}^{Q}\left(Y_{qI'} - \bar{Y}_{q\cdot}\right)$$

where

$$\bar{Y}_{q\cdot} = \frac{1}{I}\sum_{i=1}^{I} Y_{qi}$$

Computation of correction

The correction that should be made to measurements made with instrument I' is

$$Y_{corrected} = Y_{measured} - \bar{B}_{I'}$$

Type A uncertainty of the correction

The type A uncertainty of the correction is the standard deviation of the average bias or

$$s_{correction} = \frac{1}{\sqrt{Q}} s_{bias} = \frac{1}{\sqrt{Q}} \sqrt{\frac{1}{(Q-1)} \sum_{q=1}^{Q} \left(Y_{qI'} - \bar{Y}_{q\cdot} - \bar{B}_{I'} \right)^2}$$

Example of consistent bias for probe #2362 used to measure resistivity of silicon wafers

The table below comes from the table of resistivity measurements from a type A analysis of random effects with the average for each wafer subtracted from each measurement. The differences, as shown, represent the biases for each probe with respect to the other probes. Probe #2362 has an average bias, over the five wafers, of -0.02724 ohm.cm. If measurements made with this probe are corrected for this bias, the standard deviation of the correction is a type A uncertainty.

Table of biases for probes and silicon wafers (ohm.cm)

	Wafers				
Probe	138	139	140	141	142
1	0.02476	-0.00356	0.04002	0.03938	0.00620
181	0.01076	0.03944	0.01871	-0.01072	0.03761
182	0.01926	0.00574	-0.02008	0.02458	-0.00439
2062	-0.01754	-0.03226	-0.01258	-0.02802	-0.00110
2362	-0.03725	-0.00936	-0.02608	-0.02522	-0.03830

Average bias for probe #2362 = - 0.02724

Standard deviation of bias = 0.01171 with 4 degrees of freedom

Standard deviation of correction = 0.01171/sqrt(5) = 0.00523

Note on different approaches to instrument bias

The analysis on this page considers the case where only one instrument is used to make the certification measurements; namely probe #2362, and the certified values are corrected for bias due to this probe. The analysis in the section on type A analysis of random effects considers the case where any one of the probes could be used to make the certification measurements.

2.5.3.3.3. Bias with sparse data

Strategy for dealing with limited data

The purpose of this discussion is to outline methods for dealing with biases that may be real but which cannot be estimated reliably because of the sparsity of the data. For example, a test between two, of many possible, configurations of the measurement process cannot produce a reliable enough estimate of bias to permit a correction, but it can reveal problems with the measurement process. The strategy for a significant bias is to apply a 'zero' correction. The type A uncertainty component is the standard deviation of the correction, and the calculation depends on whether the bias is

- inconsistent
- consistent

The analyses in this section can be produced using both Dataplot code and R code.

Example of differences among wiring settings

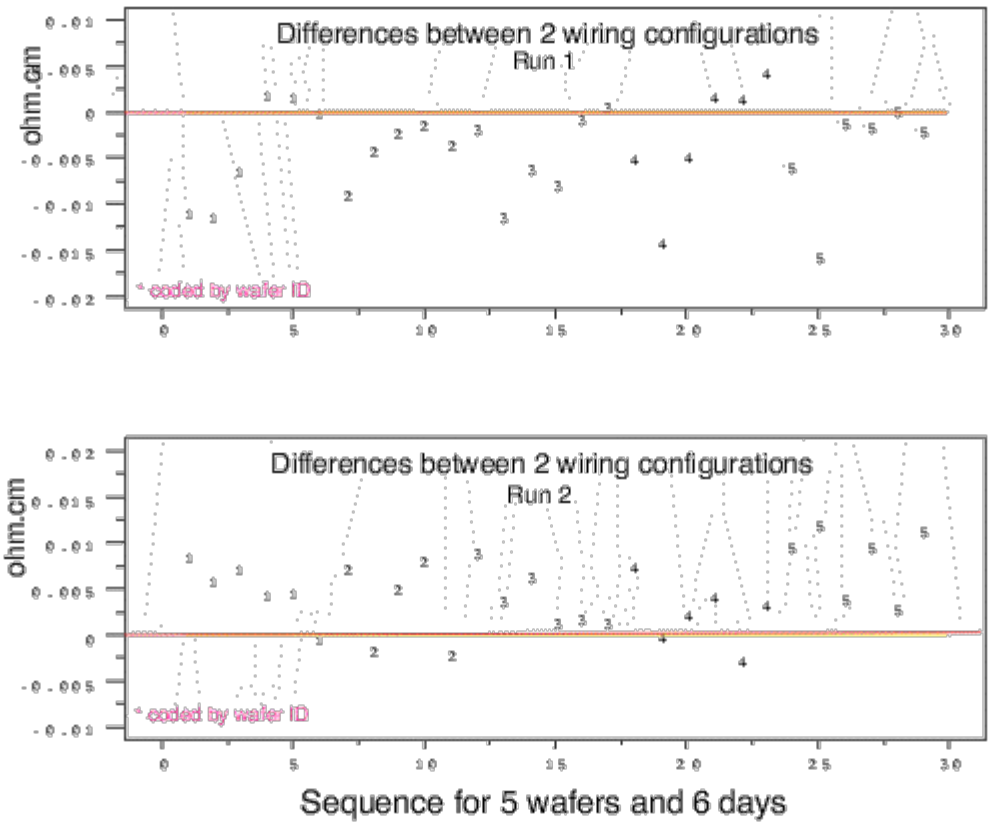
An example is given of a study of wiring settings for a single gauge. The gauge, a 4-point probe for measuring resistivity of silicon wafers, can be wired in several ways. Because it was not possible to test all wiring configurations during the gauge study, measurements were made in only two configurations as a way of identifying possible problems.

Data on wiring configurations

Measurements were made on six wafers over six days (except for 5 measurements on wafer 39) with probe #2062 wired in two configurations. This sequence of measurements was repeated after about a month resulting in two runs. A database of differences between measurements in the two configurations on the same day are analyzed for significance.

Plot the differences between the two wiring configurations

A plot of the differences between the two configurations shows that the differences for run 1 are, for the most part, less than zero, and the differences for run 2 are greater than zero.



Statistical test for difference between two configurations

A *t*-statistic is used as an approximate test where we are assuming the differences are approximately normal. The average difference and standard deviation of the difference are required for this test. If

$$t = \left| \frac{\sqrt{N}}{s_{diff}} Avg_{diff} \right| > 2$$

the difference between the two configurations is statistically significant.

The average and standard deviation computed from the N = 29 differences in each run from the table above are shown along with corresponding *t*-values which confirm that the differences are significant, but in opposite directions, for both runs.

Average differences between wiring configurations

Run	Probe	Average	Std dev	N	t
1	2062	- 0.00383	0.00514	29	- 4.0
2	2062	+ 0.00489	0.00400	29	+ 6.6

Case of inconsistent bias

The data reveal a significant wiring bias for both runs that changes direction between runs. Because of this inconsistency, a 'zero' correction is applied to the results, and the type A uncertainty is taken to be

$$s_{correction} = \frac{1}{\sqrt{3}} MaxBias$$

For this study, the type A uncertainty for wiring bias is

$$s_{correction} = \frac{1}{\sqrt{3}} 0.00489\, ohm.cm = 0.0028\, ohm.cm$$

Case of consistent bias

Even if the bias is consistent over time, a 'zero' correction is applied to the results, and for a single run, the estimated standard deviation of the correction is

$$s_{correction} = \frac{1}{\sqrt{N}} s_{diff}$$

For two runs (1 and 2), the estimated standard deviation of the correction is

$$s_{correction} = \frac{1}{\sqrt{2N}} \sqrt{s_{diff_1}^2 + s_{diff_2}^2}$$

2.5.4. Type B evaluations

Type B evaluations apply to both error and bias

Type B evaluations can apply to both random error and bias. The distinguishing feature is that the calculation of the uncertainty component is not based on a statistical analysis of data. The distinction to keep in mind with regard to random error and bias is that:

- random errors cannot be corrected
- biases can, theoretically at least, be corrected or eliminated from the result.

Sources of type B evaluations

Some examples of sources of uncertainty that lead to type B evaluations are:

- Reference standards calibrated by another laboratory
- Physical constants used in the calculation of the reported value
- Environmental effects that cannot be sampled
- Possible configuration/geometry misalignment in the instrument
- Lack of resolution of the instrument

Documented sources of uncertainty from other processes

Documented sources of uncertainty, such as calibration reports for reference standards or published reports of uncertainties for physical constants, pose no difficulties in the analysis. The uncertainty will usually be reported as an expanded uncertainty, U, which is converted to the standard uncertainty,

$$u = U/k$$

If the k factor is not known or documented, it is probably conservative to assume that $k = 2$.

Sources of uncertainty that are local to the measurement process

Sources of uncertainty that are local to the measurement process but which cannot be adequately sampled to allow a statistical analysis require type B evaluations. One technique, which is widely used, is to estimate the worst-case effect, a, for the source of interest, from

- experience
- scientific judgment

- scant data

A standard deviation, assuming that the effect is two-sided, can then be computed based on a uniform, triangular, or normal distribution of possible effects.

Following the Guide to the Expression of Uncertainty of Measurement (GUM), the convention is to assign **infinite degrees of freedom** to standard deviations derived in this manner.

2.5.4.1. Standard deviations from assumed distributions

Difficulty of obtaining reliable uncertainty estimates

The methods described on this page attempt to avoid the difficulty of allowing for sources of error for which reliable estimates of uncertainty do not exist. The methods are based on assumptions that may, or may not, be valid and require the experimenter to consider the effect of the assumptions on the final uncertainty.

Difficulty of obtaining reliable uncertainty estimates

The ISO guidelines do not allow worst-case estimates of bias to be added to the other components, but require they in some way be converted to equivalent standard deviations. The approach is to consider that any error or bias, for the situation at hand, is a random draw from a known statistical distribution. Then the standard deviation is calculated from known (or assumed) characteristics of the distribution. Distributions that can be considered are:

- Uniform
- Triangular
- Normal (Gaussian)

Standard deviation for a uniform distribution

The uniform distribution leads to the most conservative estimate of uncertainty; i.e., it gives the largest standard deviation. The calculation of the standard deviation is based on the assumption that the end-points, ± a, of the distribution are known. It also embodies the assumption that all effects on the reported value, between -a and +a, are equally likely for the particular source of uncertainty.

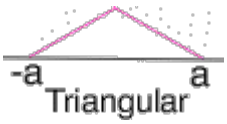
$$s_{source} = \frac{1}{\sqrt{3}} a$$

Standard deviation for a triangular

The triangular distribution leads to a less conservative estimate of uncertainty; i.e., it gives a smaller standard deviation than the uniform distribution. The calculation of the standard deviation is based on the assumption that the end-

distribution

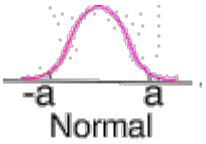
points, ± a, of the distribution are known and the mode of the triangular distribution occurs at zero.



$$s_{source} = \frac{1}{\sqrt{6}} a$$

Standard deviation for a normal distribution

The normal distribution leads to the least conservative estimate of uncertainty; i.e., it gives the smallest standard deviation. The calculation of the standard deviation is based on the assumption that the end-points, ± a, encompass 99.7 percent of the distribution.



$$s_{source} = \frac{1}{3} a$$

Degrees of freedom

In the context of using the Welch-Saitterthwaite formula with the above distributions, the degrees of freedom is assumed to be infinite.

2.5.5. Propagation of error considerations

Top-down approach consists of estimating the uncertainty from direct repetitions of the measurement result

The approach to uncertainty analysis that has been followed up to this point in the discussion has been what is called a top-down approach. Uncertainty components are estimated from direct repetitions of the measurement result. To contrast this with a propagation of error approach, consider the simple example where we estimate the area of a rectangle from replicate measurements of length and width. The area

$$area = length \times width$$

can be computed from each replicate. The standard deviation of the reported area is estimated directly from the replicates of area.

Advantages of top-down approach

This approach has the following advantages:

- proper treatment of covariances between measurements of length and width
- proper treatment of unsuspected sources of error that would emerge if measurements covered a range of operating conditions and a sufficiently long time period
- independence from propagation of error model

Propagation of error approach combines estimates from individual auxiliary measurements

The formal propagation of error approach is to compute:

1. standard deviation from the length measurements
2. standard deviation from the width measurements

and combine the two into a standard deviation for area using the approximation for products of two variables (ignoring a possible covariance between length and width),

$$s_{area} = \sqrt{width^2 s^2_{length} + length^2 s^2_{width}}$$

Exact formula

Goodman (1960) derived an exact formula for the variance between two products. Given two random variables, x and y (correspond to width and length in the above approximate formula), the exact formula for the variance is:

$$V(\bar{x}\bar{y}) = [X^2V(y) + Y^2V(x) + 2XYE_{11} + 2X\frac{E_{12}}{n} + 2Y\frac{E_{21}}{n} + \frac{V(x)V(y)}{n} + \frac{Cov((\Delta x)^2,(\Delta y)^2) - E_{11}^2}{n^2}]/n$$

with

- X = E(x) and Y = E(y) (corresponds to width and length, respectively, in the approximate formula)
- V(x) = variance of x and V(y) = variance Y (corresponds to s^2 for width and length, respectively, in the approximate formula)
- $E_{ij} = \{(\Delta x)^i, (\Delta y)^j\}$ where $\Delta x = x$ - X and $\Delta y = y$ - Y
- $Cov((\Delta x)^2, (\Delta y)^2) = E_{22} - V(x)V(y)$

To obtain the standard deviation, simply take the square root of the above formula. Also, an estimate of the statistic is obtained by substituting sample estimates for the corresponding population values on the right hand side of the equation.

Approximate formula assumes indpendence

The approximate formula assumes that length and width are independent. The exact formula assumes that length and width are not independent.

Disadvantages of propagation of error approach

In the ideal case, the propagation of error estimate above will not differ from the estimate made directly from the area measurements. However, in complicated scenarios, they may differ because of:

- unsuspected covariances
- disturbances that affect the reported value and not the elementary measurements (usually a result of mis-specification of the model)
- mistakes in propagating the error through the defining formulas

Propagation of error formula

Sometimes the measurement of interest cannot be replicated directly and it is necessary to estimate its uncertainty via propagation of error formulas (Ku). The propagation of error formula for

$$Y = f(X, Z, \ldots)$$

a function of one or more variables with measurements, X, Z, ... gives the following estimate for the standard deviation of Y:

$$s_y = \sqrt{\left(\frac{\partial Y}{\partial X}\right)^2 s_x^2 + \left(\frac{\partial Y}{\partial Z}\right)^2 s_z^2 + \cdots + \left(\frac{\partial Y}{\partial X}\right)\left(\frac{\partial Y}{\partial Z}\right) s_{xz}^2 + \cdots}$$

where

- s_x is the standard deviation of the X measurements
- s_z is the standard deviation of Z measurements
- s_y is the standard deviation of Y measurements
- $\partial Y / \partial X$ is the partial derivative of the function Y with respect to X, etc.

s_{xz} is the estimated covariance between the *X*,*Z* measurements

Treatment of covariance terms

Covariance terms can be difficult to estimate if measurements are not made in pairs. Sometimes, these terms are omitted from the formula. Guidance on when this is acceptable practice is given below:

1. If the measurements of *X, Z* are independent, the associated covariance term is *zero*.
2. Generally, reported values of test items from calibration designs have *non-zero* covariances that must be taken into account if *Y* is a summation such as the mass of two weights, or the length of two gage blocks end-to-end, etc.
3. Practically speaking, covariance terms should be included in the computation only if they have been estimated from sufficient data. See Ku (1966) for guidance on what constitutes sufficient data.

Sensitivity coefficients

The partial derivatives are the sensitivity coefficients for the associated components.

Examples of propagation of error analyses

Examples of propagation of error that are shown in this chapter are:

- Case study of propagation of error for resistivity measurements
- Comparison of check standard analysis and propagation of error for linear calibration
- Propagation of error for quadratic calibration showing effect of covariance terms

Specific formulas

Formulas for specific functions can be found in the following sections:

- functions of a single variable
- functions of two variables
- functions of many variables

2.5.5.1. Formulas for functions of one variable

Case: Y=f(X,Z)

Standard deviations of reported values that are functions of a single variable are reproduced from a paper by H. Ku (Ku).

The reported value, *Y*, is a function of the average of *N* measurements on a single variable.

Notes	***Function*** Y ***of*** $\bar{X}$ $\bar{X}$ *is an average of N measurements*	***Standard deviation of*** Y s_x *= standard deviation of X.*
	$Y = \bar{X}$	$\frac{1}{\sqrt{N}} s_x$
	$Y = \frac{\bar{X}}{1+\bar{X}}$	$\frac{s_x}{\sqrt{N}\left(1+\bar{X}\right)^2}$
	$Y = \left(\bar{X}\right)^2$	$\frac{2\bar{X}}{\sqrt{N}} s_x$
	$Y = \sqrt{\bar{X}}$	$\frac{s_x}{2\sqrt{N\bar{X}}}$
	$Y = \ln \bar{X}$	$\frac{s_x}{\sqrt{N}\,\bar{X}}$
Approximation could be seriously in error if n is small--	$Y = e^{\bar{X}}$	$\frac{e^{\bar{X}}}{\sqrt{N}} s_x$
Not directly derived from		$\frac{Y}{\sqrt{2(N-1)}}$

the formulas	$Y = \frac{100 s_x}{\bar{X}}$	Note: we need to assume that the original data follow an approximately normal distribution.

2.5.5.2. Formulas for functions of two variables

Case: Y=f(X,Z)

Standard deviations of reported values that are functions of measurements on two variables are reproduced from a paper by H. Ku (Ku).

The reported value, Y is a function of averages of N measurements on two variables.

Function Y ***of*** $\bar{X}, \bar{Z}$ $\bar{X}$ *and* $\bar{Z}$ *are averages of N measurements*	***Standard deviation of*** Y s_X = *standard dev of X;* s_Z = *standard dev of Z;* s_{XZ}^2 = *covariance of X,Z* ***Note: Covariance term is to be included only if there is a reliable estimate***
$Y = A\bar{X} + B\bar{Z}$	$\frac{1}{\sqrt{N}}\sqrt{A^2 s_X^2 + B^2 s_Z^2 + 2AB s_{XZ}^2}$
$Y = \frac{\bar{X}}{\bar{Z}}$	$\frac{1}{\sqrt{N}}\frac{\bar{X}}{\bar{Z}}\sqrt{\frac{s_X^2}{\left(\bar{X}\right)^2} + \frac{s_Z^2}{\left(\bar{Z}\right)^2} - 2\frac{s_{XZ}^2}{\bar{X}\,\bar{Z}}}$
$Y = \frac{\bar{X}}{\bar{X} + \bar{Z}}$	$\left(\frac{Y}{\bar{X}}\right)^2 \frac{1}{\sqrt{N}}\sqrt{\left(\bar{X}\right)^2 s_Z^2 + \left(\bar{Z}\right)^2 s_X^2 - 2\bar{X}\,\bar{Z}\,s_{XZ}^2}$
$Y = \bar{X}\,\bar{Z}$	$\frac{XZ}{\sqrt{N}}\sqrt{\frac{s_X^2}{\bar{X}^2} + \frac{s_Z^2}{Z^2} + 2\frac{s_{XZ}^2}{\bar{X}\bar{Z}}}$
$Y = c\left(\bar{X}\right)^a\left(\bar{Z}\right)^b$	$\frac{Y}{\sqrt{N}}\sqrt{a^2\frac{s_X^2}{\bar{X}^2} + b^2\frac{s_Z^2}{Z^2} + 2ab\frac{s_{XZ}^2}{\bar{X}\bar{Z}}}$ Note: this is an approximation. The exact result could be obtained starting from the exact formula for the standard deviation of a product derived by

	Goodman (1960).

2. Measurement Process Characterization
2.5. Uncertainty analysis
2.5.5. Propagation of error considerations

2.5.5.3. Propagation of error for many variables

Example from fluid flow with a nonlinear function

Computing uncertainty for measurands based on more complicated functions can be done using basic propagation of errors principles. For example, suppose we want to compute the uncertainty of the discharge coefficient for fluid flow (Whetstone et al.). The measurement equation is

$$C_d = \frac{\dot{m}\sqrt{1-\left(\frac{d}{D}\right)^4}}{Kd^2F\sqrt{\rho}\sqrt{\Delta P}}$$

where

$$\begin{aligned} C_d &= \text{discharge coefficient} \\ \dot{m} &= \text{mass flow rate} \\ d &= \text{orifice diameter} \\ D &= \text{pipe diameter} \\ \rho &= \text{fluid density} \\ \Delta P &= \text{differential pressure} \\ K &= \text{constant} \\ F &= \text{thermal expansion factor (constant)} \end{aligned}$$

Assuming the variables in the equation are uncorrelated, the squared uncertainty of the discharge coefficient is

$$s^2_{Cd} = \left[\frac{\partial C_d}{\partial \dot{m}}\right]^2 s^2_{\dot{m}} + \left[\frac{\partial C_d}{\partial d}\right]^2 s^2_d + \left[\frac{\partial C_d}{\partial D}\right]^2 s^2_D + \left[\frac{\partial C_d}{\partial \rho}\right]^2 s^2_\rho + \left[\frac{\partial C_d}{\partial \Delta P}\right]^2 s^2_{\Delta P}$$

and the partial derivatives are the following.

$$\frac{\partial C_d}{\partial \dot{m}} = \frac{\sqrt{1-\left(\frac{d}{D}\right)^4}}{Kd^2F\sqrt{\rho}\sqrt{\Delta P}}$$

$$\frac{\partial C_d}{\partial d} = \frac{-2\dot{m}d}{KFD^4\sqrt{\rho}\sqrt{\Delta P}\sqrt{1-\left(\frac{d}{D}\right)^4}} - \frac{2\dot{m}\sqrt{1-\left(\frac{d}{D}\right)^4}}{Kd^3F\sqrt{\rho}\sqrt{\Delta P}}$$

$$\frac{\partial C_d}{\partial D} = \frac{2\dot{m}d^2}{KFD^5\sqrt{\rho}\sqrt{\Delta P}\sqrt{1-\left(\frac{d}{D}\right)^4}}$$

$$\frac{\partial C_d}{\partial \rho} = \frac{-\dot{m}\sqrt{1-\left(\frac{d}{D}\right)^4}}{2Kd^2F\rho^{\frac{3}{2}}\sqrt{\Delta P}}$$

$$\frac{\partial C_d}{\partial \Delta P} = \frac{-\dot{m}\sqrt{1-\left(\frac{d}{D}\right)^4}}{2Kd^2F\sqrt{\rho}(\Delta P)^{\frac{3}{2}}}$$

Software can simplify propagation of error

Propagation of error for more complicated functions can be done reliably with software capable of symbolic computations or algebraic representations.

Symbolic computation software can also be used to combine the partial derivatives with the appropriate standard deviations, and then the standard deviation for the discharge coefficient can be evaluated and plotted for specific values of the secondary variables, as shown in the comparison of check standard analysis and propagation of error.

Simplification for dealing with multiplicative variables

Propagation of error for several variables can be simplified considerably for the special case where:

- the function, *Y*, is a simple multiplicative function of secondary variables, and
- uncertainty is evaluated as a percentage.

For three variables, *X*, *Z*, *W*, the function

$$Y = X \cdot Z \cdot W$$

has a standard deviation in absolute units of

$$s_Y = \sqrt{(ZW)^2 s_x^2 + (XW)^2 s_z^2 + (XZ)^2 s_w^2}$$
$$= Y\sqrt{\frac{s_x^2}{X^2} + \frac{s_z^2}{Z^2} + \frac{s_w^2}{W^2}}$$

In percent units, the standard deviation can be written as

$$\frac{s_Y}{Y} = \sqrt{\frac{s_x^2}{X^2} + \frac{s_z^2}{Z^2} + \frac{s_w^2}{W^2}}$$

if all covariances are negligible. These formulas are easily extended to more than three variables.

2.5.6. Uncertainty budgets and sensitivity coefficients

Case study showing uncertainty budget

Uncertainty components are listed in a table along with their corresponding sensitivity coefficients, standard deviations and degrees of freedom. A table of typical entries illustrates the concept.

Typical budget of type A and type B uncertainty components

Type A components	Sensitivity coefficient	Standard deviation	Degrees freedom
1. Time (repeatability)	a_1	s_1	*v1*
2. Time (reproducibility)	a_2	s_2	*v2*
3. Time (long-term)	a_3	s_3	*v3*
Type B components			
5. Reference standard	(nominal test / nominal ref)	s_4	*v4*

Sensitivity coefficients show how components are related to result

The sensitivity coefficient shows the relationship of the individual uncertainty component to the standard deviation of the reported value for a test item. The sensitivity coefficient relates to the result that is being reported and not to the method of estimating uncertainty components where the uncertainty, *u*, is

$$u = \sqrt{\sum_{i=1}^{R} a_i^2 s_i^2}$$

Sensitivity coefficients for type A components of uncertainty

This section defines sensitivity coefficients that are appropriate for type A components estimated from repeated measurements. The pages on type A evaluations, particularly the pages related to estimation of repeatability and reproducibility components, should be reviewed before continuing on this page. The convention for the notation for sensitivity coefficients for this section is that:

1. a_1 refers to the sensitivity coefficient for the

repeatability standard deviation, s_1
2. a_2 refers to the sensitivity coefficient for the reproducibility standard deviation, s_2
3. a_3 refers to the sensitivity coefficient for the stability standard deviation, s_3

with some of the coefficients possibly equal to zero.

Note on long-term errors

Even if no day-to-day nor run-to-run measurements were made in determining the reported value, the sensitivity coefficient is non-zero if that standard deviation proved to be significant in the analysis of data.

Sensitivity coefficients for other type A components of random error

Procedures for estimating differences among instruments, operators, etc., which are treated as random components of uncertainty in the laboratory, show how to estimate the standard deviations so that the sensitivity coefficients = 1.

Sensitivity coefficients for type A components for bias

This Handbook follows the ISO guidelines in that biases are corrected (correction may be zero), and the uncertainty component is the standard deviation of the correction. Procedures for dealing with biases show how to estimate the standard deviation of the correction so that the sensitivity coefficients are equal to one.

Sensitivity coefficients for specific applications

The following pages outline methods for computing sensitivity coefficients where the components of uncertainty are derived in the following manner:

1. From measurements on the test item itself
2. From measurements on a check standard
3. From measurements in a 2-level design
4. From measurements in a 3-level design

and give an example of an uncertainty budget with sensitivity coefficients from a 3-level design.

Sensitivity coefficients for type B evaluations

The majority of sensitivity coefficients for type B evaluations will be one with a few exceptions. The sensitivity coefficient for the uncertainty of a reference standard is the nominal value of the test item divided by the nominal value of the reference standard.

Case study- sensitivity coefficients for propagation of error

If the uncertainty of the reported value is calculated from propagation of error, the sensitivity coefficients are the multipliers of the individual variance terms in the propagation of error formula. Formulas are given for selected functions of:

1. functions of a single variable
2. functions of two variables
3. several variables

2.5.6.1. Sensitivity coefficients for measurements on the test item

From data on the test item itself

If the temporal component is estimated from N short-term readings on the test item itself

$$Y_1, Y_2, ..., Y_N$$

and

$$s_1 = \frac{1}{\sqrt{N-1}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

and the reported value is the average, the standard deviation of the reported value is

$$s_{reported\ value} = \frac{1}{\sqrt{N}} s_1$$

with degrees of freedom $\nu_1 = N - 1$.

Sensitivity coefficients

The sensitivity coefficient is $a_1 = \sqrt{1/N}$. The risk in using this method is that it may seriously underestimate the uncertainty.

To improve the reliability of the uncertainty calculation

If possible, the measurements on the test item should be repeated over M days and averaged to estimate the reported value. The standard deviation for the reported value is computed from the daily averages>, and the standard deviation for the temporal component is:

$$s_{reported\ value} = \frac{1}{\sqrt{M}}\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{Y}_{m\bullet} - \bar{Y}_{\bullet\bullet}\right)^2}$$

with degrees of freedom $\nu_2 = M - 1$ where $\bar{Y}_{m\bullet}$ are the daily averages and $\bar{Y}_{\bullet\bullet}$ is the grand average.

The sensitivity coefficients are: $a_1 = 0$; $a_2 = \sqrt{1/M}$.

Note on

Even if no day-to-day nor run-to-run measurements were

long-term errors	made in determining the reported value, the sensitivity coefficient is non-zero if that standard deviation proved to be significant in the analysis of data.

2.5.6.2. Sensitivity coefficients for measurements on a check standard

From measurements on check standards

If the temporal component of the measurement process is evaluated from measurements on a check standard and there are M days ($M = 1$ is permissible) of measurements on the test item that are **structured in the same manner as the measurements on the check standard**, the standard deviation for the reported value is

$$s_{reported\ value} = \frac{1}{\sqrt{M}} s_2$$

with degrees of freedom $\nu_2 = K - 1$ from the K entries in the check standard database.

Standard deviation from check standard measurements

The computation of the standard deviation from the check standard values and its relationship to components of instrument precision and day-to-day variability of the process are explained in the section on two-level nested designs using check standards.

Sensitivity coefficients

The sensitivity coefficients are: a_1; $a_2 = \sqrt{1/M}$.

2.5.6.3. Sensitivity coefficients for measurements from a 2-level design

Sensitivity coefficients from a 2-level design

If the temporal components are estimated from a 2-level nested design, and the reported value for a test item is an average over

- N short-term repetitions
- M ($M = 1$ is permissible) days

of measurements on the test item, the standard deviation for the reported value is:

$$s_{reported\ value} = \sqrt{\frac{1}{M}s_{days}^2 + \frac{1}{MN}s_1^2}$$

See the relationships in the section on 2-level nested design for definitions of the standard deviations and their respective degrees of freedom.

Problem with estimating degrees of freedom

If degrees of freedom are required for the uncertainty of the reported value, the formula above cannot be used directly and must be rewritten in terms of the standard deviations, s_1 and s_2.

$$s_{reported\ value} = \sqrt{\frac{1}{M}s_2^2 + \frac{J-N}{MNJ}s_1^2}$$

Sensitivity coefficients

The sensitivity coefficients are: $a_1 = \sqrt{(J-N)/PMNJ}$; $a_2 = \sqrt{1/M}$.

Specific sensitivity coefficients are shown in the table below for selections of N, M.

Sensitivity coefficients for two components of uncertainty

Number	Number	Short-term	Day-to-

short-term N	day-to-day M	sensitivity coefficient a_1	day sensitivity coefficient a_2
1	1	$\sqrt{(J-1)/J}$	1
N	1	$\sqrt{(J-N)/NJ}$	1
N	M	$\sqrt{(J-N)/MNJ}$	$\sqrt{1/M}$

2.5.6.4. Sensitivity coefficients for measurements from a 3-level design

Sensitivity coefficients from a 3-level design

If the temporal components are estimated from a 3-level nested design and the reported value is an average over

- *N* short-term repetitions
- *M* days
- *P* runs

of measurements on the test item, the standard deviation for the reported value is:

$$s_{reported\ value} = \sqrt{\frac{1}{P}s_{runs}^2 + \frac{1}{PM}s_{days}^2 + \frac{1}{PMN}s_1^2}$$

See the section on analysis of variability for definitions and relationships among the standard deviations shown in the equation above.

Problem with estimating degrees of freedom

If degrees of freedom are required for the uncertainty, the formula above cannot be used directly and must be rewritten in terms of the standard deviations s_1, s_2, and s_3.

$$s_{reported\ value} = \sqrt{\frac{1}{P}s_3^2 + \frac{K-M}{PMK}s_2^2 + \frac{J-N}{PMNJ}s_1^2}$$

Sensitivity coefficients

The sensitivity coefficients are:

$a_1 = \sqrt{(J-N)/PMNJ}$; $a_2 = \sqrt{(K-M)/PMK}$;
$a_3 = \sqrt{1/P}$.

Specific sensitivity coefficients are shown in the table below for selections of *N*, *M*, *P*. In addition, the following constraints must be observed:

J must be > or = N and K must be > or = M

Sensitivity coefficients for three components of uncertainty

Number short-term N	Number day-to-day M	Number run-to-run P	Short-term sensitivity coefficient a_1	Day-to-day sensitivity coefficient a_2	Run-to-run sensitivity coefficient a_3
1	1	1	$\sqrt{(J-1)/J}$	$\sqrt{(K-1)/K}$	1
N	1	1	$\sqrt{(J-N)/NJ}$	$\sqrt{(K-1)/K}$	1
N	M	1	$\sqrt{(J-N)/MNJ}$	$\sqrt{(K-M)/MK}$	1
N	M	P	$\sqrt{(J-N)/PMNJ}$	$\sqrt{(K-M)/MPK}$	$\sqrt{1/P}$

2.5.6.5. Example of uncertainty budget

Example of uncertainty budget for three components of temporal uncertainty

An uncertainty budget that illustrates several principles of uncertainty analysis is shown below. The reported value for a test item is the average of N short-term measurements where the temporal components of uncertainty were estimated from a 3-level nested design with J short-term repetitions over K days.

The number of measurements made on the test item is the same as the number of short-term measurements in the design; i.e., $N = J$. Because there were no repetitions over days or runs on the test item, $M = 1$; $P = 1$. The sensitivity coefficients for this design are shown on the foregoing page.

Example of instrument bias

This example also illustrates the case where the measuring instrument is biased relative to the other instruments in the laboratory, with a bias correction applied accordingly. The sensitivity coefficient, given that the bias correction is based on measurements on Q artifacts, is defined as $a_4 = 1$, and the standard deviation, s_4, is the standard deviation of the correction.

Example of error budget for type A and type B uncertainties

Type A components	Sensitivity coefficient	Standard deviation	Degrees freedom
1. Repeatability	$a_1 = 0$	s_1	$J - 1$
2. Reproducibility	$a_2 = \sqrt{K - 1 / K}$	s_2	$K - 1$
2. Stability	$a_3 = 1$	s_3	$L - 1$
3. Instrument bias	$a_4 = 1$	s_4	$Q - 1$

2.5.7. Standard and expanded uncertainties

Definition of standard uncertainty

The sensitivity coefficients and standard deviations are combined by root sum of squares to obtain a 'standard uncertainty'. Given R components, the standard uncertainty is:

$$u = \sqrt{\sum_{i=1}^{R} a_i^2 s_i^2}$$

Expanded uncertainty assures a high level of confidence

If the purpose of the uncertainty statement is to provide coverage with a high level of confidence, an expanded uncertainty is computed as

$$U = ku$$

where k is chosen to be the $t_{1-\alpha/2,\nu}$ critical value from the *t*-table with ν degrees of freedom. For large degrees of freedom, $k = 2$ approximates 95 % coverage.

Interpretation of uncertainty statement

The expanded uncertainty defined above is assumed to provide a high level of coverage for the unknown true value of the measurement of interest so that for any measurement result, Y,

$$Y - U \le TrueValue \le Y + U$$

2.5.7.1. Degrees of freedom

Degrees of freedom for individual components of uncertainty

Degrees of freedom for type A uncertainties are the degrees of freedom for the respective standard deviations. Degrees of freedom for Type B evaluations may be available from published reports or calibration certificates. Special cases where the standard deviation must be estimated from fragmentary data or scientific judgment are assumed to have infinite degrees of freedom; for example,

- Worst-case estimate based on a robustness study or other evidence
- Estimate based on an assumed distribution of possible errors
- Type B uncertainty component for which degrees of freedom are not documented

Degrees of freedom for the standard uncertainty

Degrees of freedom for the standard uncertainty, u, which may be a combination of many standard deviations, is not generally known. This is particularly troublesome if there are large components of uncertainty with small degrees of freedom. In this case, the degrees of freedom is approximated by the Welch-Satterthwaite formula (Brownlee).

$$\nu = \frac{u^4}{\sum_{i=1}^{R} \frac{a_i^4 s_i^4}{\nu_i}}$$

Case study: Uncertainty and degrees of freedom

A case study of type A uncertainty analysis shows the computations of temporal components of uncertainty; instrument bias; geometrical bias; standard uncertainty; degrees of freedom; and expanded uncertainty.

2.5.8. Treatment of uncorrected bias

Background

The ISO Guide (ISO) for expressing measurement uncertainties assumes that all biases are corrected and that the uncertainty applies to the corrected result. For measurements at the factory floor level, this approach has several disadvantages. It may not be practical, may be expensive and may not be economically sound to correct for biases that do not impact the commercial value of the product (Turgel and Vecchia).

Reasons for not correcting for bias

Corrections may be expensive to implement if they require modifications to existing software and "paper and pencil" corrections can be both time consuming and prone to error. In the scientific or metrology laboratory, biases may be documented in certain situations, but the mechanism that causes the bias may not be fully understood, or repeatable, which makes it difficult to argue for correction. In these cases, the best course of action is to report the measurement as taken and adjust the uncertainty to account for the "bias".

The question is how to adjust the uncertainty

A method needs to be developed which assures that the resulting uncertainty has the following properties (Phillips and Eberhardt):

1. The final uncertainty must be greater than or equal to the uncertainty that would be quoted if the bias were corrected.
2. The final uncertainty must reduce to the same uncertainty given that the bias correction is applied.
3. The level of coverage that is achieved by the final uncertainty statement should be at least the level obtained for the case of corrected bias.
4. The method should be transferable so that both the uncertainty and the bias can be used as components of uncertainty in another uncertainty statement.
5. The method should be easy to implement.

2.5.8.1. Computation of revised uncertainty

Definition of the bias and corrected measurement

If the bias is δ and the corrected measurement is defined by

$$Y_{cor} = Y - \delta ,$$

the corrected value of Y has the usual expanded uncertainty interval which is symmetric around the unknown true value for the measurement process and is of the following type:

$$Y_{cor} - U \le TrueValue \le Y_{cor} + U$$

Definition of asymmetric uncertainty interval to account for uncorrected measurement

If no correction is made for the bias, the uncertainty interval is contaminated by the effect of the bias term as follows:

$$Y - (U + \delta) \le TrueValue \le Y + (U - \delta)$$

and can be rewritten in terms of upper and lower endpoints that are asymmetric around the true value; namely,

$$Y - U_{-} \le TrueValue \le Y + U_{+}$$

Conditions on the relationship between the bias and U

The definition above can lead to a negative uncertainty limit; e.g., if the bias is positive and greater than U, the upper endpoint becomes negative. The requirement that the uncertainty limits be greater than or equal to zero for all values of the bias guarantees non-negative uncertainty limits and is accepted at the cost of somewhat wider uncertainty intervals. This leads to the following set of restrictions on the uncertainty limits:

$$U_{-} = \begin{cases} U + \delta & \text{if} \quad U + \delta > 0 \\ 0 & \text{if} \quad U + \delta \le 0 \end{cases}$$

$$U_{+} = \begin{cases} U - \delta & \text{if} \quad U - \delta > 0 \\ 0 & \text{if} \quad U - \delta \le 0 \end{cases}$$

Situation where bias is not known exactly but must be

If the bias is not known exactly, its magnitude is estimated from repeated measurements, from sparse data or from theoretical considerations, and the standard deviation is estimated from repeated measurements or from an assumed distribution. The standard deviation of the bias becomes a

estimated component in the uncertainty analysis with the standard uncertainty restructured to be:

$$u_C = \sqrt{u^2 + u_{bias}^2}$$

and the expanded uncertainty limits become:

$$Limits = \begin{cases} U_- = ku_C + \delta \\ U_+ = ku_C - \delta \end{cases}$$

.

Interpretation The uncertainty intervals described above have the desirable properties outlined on a previous page. For more information on theory and industrial examples, the reader should consult the paper by the authors of this technique (Phillips and Eberhardt).

2. Measurement Process Characterization

2.6. Case studies

Contents The purpose of this section is to illustrate the planning, procedures, and analyses outlined in the various sections of this chapter with data taken from measurement processes at the National Institute of Standards and Technology.

1. **Gauge study of resistivity probes**
2. **Check standard study for resistivity measurements**
3. **Type A uncertainty analysis**
4. **Type B uncertainty analysis and propagation of error**

2.6.1. Gauge study of resistivity probes

Purpose The purpose of this case study is to outline the analysis of a gauge study that was undertaken to identify the sources of uncertainty in resistivity measurements of silicon wafers.

Outline

1. Background and data
2. Analysis and interpretation
3. Graphs showing repeatability standard deviations
4. Graphs showing day-to-day variability
5. Graphs showing differences among gauges
6. Run this example yourself with Dataplot
7. Dataplot macros

2.6.1.1. Background and data

Description of measurements

Measurements of resistivity on 100 ohm.cm wafers were made according to an ASTM Standard Test Method (ASTM F84) to assess the sources of uncertainty in the measurement system. Resistivity measurements have been studied over the years, and it is clear from those data that there are sources of variability affecting the process beyond the basic imprecision of the gauges. Changes in measurement results have been noted over days and over months and the data in this study are structured to quantify these time-dependent changes in the measurement process.

Gauges

The gauges for the study were five probes used to measure resistivity of silicon wafers. The five gauges are assumed to represent a random sample of typical 4-point gauges for making resistivity measurements. There is a question of whether or not the gauges are essentially equivalent or whether biases among them are possible.

Check standards

The check standards for the study were five wafers selected at random from the batch of 100 ohm.cm wafers.

Operators

The effect of operator was not considered to be significant for this study.

Database of measurements

The 3-level nested design consisted of:

- $J = 6$ measurements at the center of each wafer per day
- $K = 6$ days
- $L = 2$ runs

To characterize the probes and the influence of wafers on the measurements, the design was repeated over:

- $Q = 5$ wafers (check standards 138, 139, 140, 141, 142)
- $I = 5$ probes (1, 281, 283, 2062, 2362)

The runs were separated by about one month in time. The $J = 6$ measurements at the center of each wafer are reduced to an average and repeatability standard deviation and recorded in a database with identifications for wafer, probe,

and day.

Software The analyses used in this case study can be generated using both Dataplot code and R code.

2.6.1.1.1. Database of resistivity measurements

The check standards are five wafers chosen at random from a batch of wafers

Measurements of resistivity (ohm.cm) were made according to an ASTM Standard Test Method (F4) at NIST to assess the sources of uncertainty in the measurement system. The gauges for the study were five probes owned by NIST; the check standards for the study were five wafers selected at random from a batch of wafers cut from one silicon crystal doped with phosphorous to give a nominal resistivity of 100 ohm.cm.

Measurements on the check standards are used to estimate repeatability, day effect, and run effect

The effect of operator was not considered to be significant for this study; therefore, 'day' replaces 'operator' as a factor in the nested design. Averages and standard deviations from $J = 6$ measurements at the center of each wafer are shown in the table.

- $J = 6$ measurements at the center of the wafer per day
- $K = 6$ days (one operator) per repetition
- $L = 2$ runs (complete)
- $Q = 5$ wafers (check standards 138, 139, 140, 141, 142)
- $R = 5$ probes (1, 281, 283, 2062, 2362)

Run	Wafer	Probe	Month	Day	Op	Temp	Average	Std Dev
1	138.	1.	3.	15.	1.	22.98	95.1772	0.1191
1	138.	1.	3.	17.	1.	23.02	95.1567	0.0183
1	138.	1.	3.	18.	1.	22.79	95.1937	0.1282
1	138.	1.	3.	21.	1.	23.17	95.1959	0.0398
1	138.	1.	3.	23.	2.	23.25	95.1442	0.0346
1	138.	1.	3.	23.	1.	23.20	95.0610	0.1539
1	138.	281.	3.	16.	1.	22.99	95.1591	0.0963
1	138.	281.	3.	17.	1.	22.97	95.1195	0.0606
1	138.	281.	3.	18.	1.	22.83	95.1065	0.0842
1	138.	281.	3.	21.	1.	23.28	95.0925	0.0973
1	138.	281.	3.	23.	2.	23.14	95.1990	0.1062

```
1  138.    281.      3.     23.      1.   23.16   95.1682
0.1090
1  138.    283.      3.     16.      1.   22.95   95.1252
0.0531
1  138.    283.      3.     17.      1.   23.08   95.1600
0.0998
1  138.    283.      3.     18.      1.   23.13   95.0818
0.1108
1  138.    283.      3.     21.      1.   23.28   95.1620
0.0408
1  138.    283.      3.     22.      1.   23.36   95.1735
0.0501
1  138.    283.      3.     24.      2.   22.97   95.1932
0.0287
1  138.   2062.      3.     16.      1.   22.97   95.1311
0.1066
1  138.   2062.      3.     17.      1.   22.98   95.1132
0.0415
1  138.   2062.      3.     18.      1.   23.16   95.0432
0.0491
1  138.   2062.      3.     21.      1.   23.16   95.1254
0.0603
1  138.   2062.      3.     22.      1.   23.28   95.1322
0.0561
1  138.   2062.      3.     24.      2.   23.19   95.1299
0.0349
1  138.   2362.      3.     15.      1.   23.08   95.1162
0.0480
1  138.   2362.      3.     17.      1.   23.01   95.0569
0.0577
1  138.   2362.      3.     18.      1.   22.97   95.0598
0.0516
1  138.   2362.      3.     22.      1.   23.23   95.1487
0.0386
1  138.   2362.      3.     23.      2.   23.28   95.0743
0.0256
1  138.   2362.      3.     24.      2.   23.10   95.1010
0.0420
1  139.      1.      3.     15.      1.   23.01   99.3528
0.1424
1  139.      1.      3.     17.      1.   23.00   99.2940
0.0660
1  139.      1.      3.     17.      1.   23.01   99.2340
0.1179
1  139.      1.      3.     21.      1.   23.20   99.3489
0.0506
1  139.      1.      3.     23.      2.   23.22   99.2625
0.1111
1  139.      1.      3.     23.      1.   23.22   99.3787
0.1103
1  139.    281.      3.     16.      1.   22.95   99.3244
0.1134
1  139.    281.      3.     17.      1.   22.98   99.3378
0.0949
1  139.    281.      3.     18.      1.   22.86   99.3424
0.0847
1  139.    281.      3.     22.      1.   23.17   99.4033
0.0801
1  139.    281.      3.     23.      2.   23.10   99.3717
0.0630
1  139.    281.      3.     23.      1.   23.14   99.3493
0.1157
1  139.    283.      3.     16.      1.   22.94   99.3065
0.0381
1  139.    283.      3.     17.      1.   23.09   99.3280
0.1153
1  139.    283.      3.     18.      1.   23.11   99.3000
0.0818
1  139.    283.      3.     21.      1.   23.25   99.3347
0.0972
1  139.    283.      3.     22.      1.   23.36   99.3929
0.1189
1  139.    283.      3.     23.      1.   23.18   99.2644
0.0622
1  139.   2062.      3.     16.      1.   22.94   99.3324
0.1531
1  139.   2062.      3.     17.      1.   23.08   99.3254
0.0543
1  139.   2062.      3.     18.      1.   23.15   99.2555
0.1024
1  139.   2062.      3.     18.      1.   23.18   99.1946
```

```
0.0851
1  139.   2062.      3.     22.      1.   23.27   99.3542
0.1227
1  139.   2062.      3.     24.      2.   23.23   99.2365
0.1218
1  139.   2362.      3.     15.      1.   23.08   99.2939
0.0818
1  139.   2362.      3.     17.      1.   23.02   99.3234
0.0723
1  139.   2362.      3.     18.      1.   22.93   99.2748
0.0756
1  139.   2362.      3.     22.      1.   23.29   99.3512
0.0475
1  139.   2362.      3.     23.      2.   23.25   99.2350
0.0517
1  139.   2362.      3.     24.      2.   23.05   99.3574
0.0485
1  140.      1.      3.     15.      1.   23.07   96.1334
0.1052
1  140.      1.      3.     17.      1.   23.08   96.1250
0.0916
1  140.      1.      3.     18.      1.   22.77   96.0665
0.0836
1  140.      1.      3.     21.      1.   23.18   96.0725
0.0620
1  140.      1.      3.     23.      2.   23.20   96.1006
0.0582
1  140.      1.      3.     23.      1.   23.21   96.1131
0.1757
1  140.    281.      3.     16.      1.   22.94   96.0467
0.0565
1  140.    281.      3.     17.      1.   22.99   96.1081
0.1293
1  140.    281.      3.     18.      1.   22.91   96.0578
0.1148
1  140.    281.      3.     22.      1.   23.15   96.0700
0.0495
1  140.    281.      3.     22.      1.   23.33   96.1052
0.1722
1  140.    281.      3.     23.      1.   23.19   96.0952
0.1786
1  140.    283.      3.     16.      1.   22.89   96.0650
0.1301
1  140.    283.      3.     17.      1.   23.07   96.0870
0.0881
1  140.    283.      3.     18.      1.   23.07   95.8906
0.1842
1  140.    283.      3.     21.      1.   23.24   96.0842
0.1008
1  140.    283.      3.     22.      1.   23.34   96.0189
0.0865
1  140.    283.      3.     23.      1.   23.19   96.1047
0.0923
1  140.   2062.      3.     16.      1.   22.95   96.0379
0.2190
1  140.   2062.      3.     17.      1.   22.97   96.0671
0.0991
1  140.   2062.      3.     18.      1.   23.15   96.0206
0.0648
1  140.   2062.      3.     21.      1.   23.14   96.0207
0.1410
1  140.   2062.      3.     22.      1.   23.32   96.0587
0.1634
1  140.   2062.      3.     24.      2.   23.17   96.0903
0.0406
1  140.   2362.      3.     15.      1.   23.08   96.0771
0.1024
1  140.   2362.      3.     17.      1.   23.00   95.9976
0.0943
1  140.   2362.      3.     18.      1.   23.01   96.0148
0.0622
1  140.   2362.      3.     22.      1.   23.27   96.0397
0.0702
1  140.   2362.      3.     23.      2.   23.24   96.0407
0.0627
1  140.   2362.      3.     24.      2.   23.13   96.0445
0.0622
1  141.      1.      3.     15.      1.   23.01  101.2124
0.0900
1  141.      1.      3.     17.      1.   23.08  101.1018
0.0820
```

```
1  141.      1.      3.     18.      1.   22.75  101.1119
0.0500
1  141.      1.      3.     21.      1.   23.21  101.1072
0.0641
1  141.      1.      3.     23.      2.   23.25  101.0802
0.0704
1  141.      1.      3.     23.      1.   23.19  101.1350
0.0699
1  141.    281.      3.     16.      1.   22.93  101.0287
0.0520
1  141.    281.      3.     17.      1.   23.00  101.0131
0.0710
1  141.    281.      3.     18.      1.   22.90  101.1329
0.0800
1  141.    281.      3.     22.      1.   23.19  101.0562
0.1594
1  141.    281.      3.     23.      2.   23.18  101.0891
0.1252
1  141.    281.      3.     23.      1.   23.17  101.1283
0.1151
1  141.    283.      3.     16.      1.   22.85  101.1597
0.0990
1  141.    283.      3.     17.      1.   23.09  101.0784
0.0810
1  141.    283.      3.     18.      1.   23.08  101.0715
0.0460
1  141.    283.      3.     21.      1.   23.27  101.0910
0.0880
1  141.    283.      3.     22.      1.   23.34  101.0967
0.0901
1  141.    283.      3.     24.      2.   23.00  101.1627
0.0888
1  141.   2062.      3.     16.      1.   22.97  101.1077
0.0970
1  141.   2062.      3.     17.      1.   22.96  101.0245
0.1210
1  141.   2062.      3.     18.      1.   23.19  100.9650
0.0700
1  141.   2062.      3.     18.      1.   23.18  101.0319
0.1070
1  141.   2062.      3.     22.      1.   23.34  101.0849
0.0960
1  141.   2062.      3.     24.      2.   23.21  101.1302
0.0505
1  141.   2362.      3.     15.      1.   23.08  101.0471
0.0320
1  141.   2362.      3.     17.      1.   23.01  101.0224
0.1020
1  141.   2362.      3.     18.      1.   23.05  101.0702
0.0580
1  141.   2362.      3.     22.      1.   23.22  101.0904
0.1049
1  141.   2362.      3.     23.      2.   23.29  101.0626
0.0702
1  141.   2362.      3.     24.      2.   23.15  101.0686
0.0661
1  142.      1.      3.     15.      1.   23.02   94.3160
0.1372
1  142.      1.      3.     17.      1.   23.04   94.2808
0.0999
1  142.      1.      3.     18.      1.   22.73   94.2478
0.0803
1  142.      1.      3.     21.      1.   23.19   94.2862
0.0700
1  142.      1.      3.     23.      2.   23.25   94.1859
0.0899
1  142.      1.      3.     23.      1.   23.21   94.2389
0.0686
1  142.    281.      3.     16.      1.   22.98   94.2640
0.0862
1  142.    281.      3.     17.      1.   23.00   94.3333
0.1330
1  142.    281.      3.     18.      1.   22.88   94.2994
0.0908
1  142.    281.      3.     21.      1.   23.28   94.2873
0.0846
1  142.    281.      3.     23.      2.   23.07   94.2576
0.0795
1  142.    281.      3.     23.      1.   23.12   94.3027
0.0389
1  142.    283.      3.     16.      1.   22.92   94.2846
```

```
0.1021
1  142.    283.      3.     17.      1.   23.08   94.2197
0.0627
1  142.    283.      3.     18.      1.   23.09   94.2119
0.0785
1  142.    283.      3.     21.      1.   23.29   94.2536
0.0712
1  142.    283.      3.     22.      1.   23.34   94.2280
0.0692
1  142.    283.      3.     24.      2.   22.92   94.2944
0.0958
1  142.   2062.      3.     16.      1.   22.96   94.2238
0.0492
1  142.   2062.      3.     17.      1.   22.95   94.3061
0.2194
1  142.   2062.      3.     18.      1.   23.16   94.1868
0.0474
1  142.   2062.      3.     21.      1.   23.11   94.2645
0.0697
1  142.   2062.      3.     22.      1.   23.31   94.3101
0.0532
1  142.   2062.      3.     24.      2.   23.24   94.2204
0.1023
1  142.   2362.      3.     15.      1.   23.08   94.2437
0.0503
1  142.   2362.      3.     17.      1.   23.00   94.2115
0.0919
1  142.   2362.      3.     18.      1.   22.99   94.2348
0.0282
1  142.   2362.      3.     22.      1.   23.26   94.2124
0.0513
1  142.   2362.      3.     23.      2.   23.27   94.2214
0.0627
1  142.   2362.      3.     24.      2.   23.08   94.1651
0.1010
2  138.      1.      4.     13.      1.   23.12   95.1996
0.0645
2  138.      1.      4.     15.      1.   22.73   95.1315
0.1192
2  138.      1.      4.     18.      2.   22.76   95.1845
0.0452
2  138.      1.      4.     19.      1.   22.73   95.1359
0.1498
2  138.      1.      4.     20.      2.   22.73   95.1435
0.0629
2  138.      1.      4.     21.      2.   22.93   95.1839
0.0563
2  138.    281.      4.     14.      2.   22.46   95.2106
0.1049
2  138.    281.      4.     18.      2.   22.80   95.2505
0.0771
2  138.    281.      4.     18.      2.   22.77   95.2648
0.1046
2  138.    281.      4.     20.      2.   22.80   95.2197
0.1779
2  138.    281.      4.     20.      2.   22.87   95.2003
0.1376
2  138.    281.      4.     21.      2.   22.95   95.0982
0.1611
2  138.    283.      4.     18.      2.   22.83   95.1211
0.0794
2  138.    283.      4.     13.      1.   23.17   95.1327
0.0409
2  138.    283.      4.     18.      1.   22.67   95.2053
0.1525
2  138.    283.      4.     19.      2.   23.00   95.1292
0.0655
2  138.    283.      4.     21.      2.   22.91   95.1669
0.0619
2  138.    283.      4.     21.      2.   22.96   95.1401
0.0831
2  138.   2062.      4.     15.      1.   22.64   95.2479
0.2867
2  138.   2062.      4.     15.      1.   22.67   95.2224
0.1945
2  138.   2062.      4.     19.      2.   22.99   95.2810
0.1960
2  138.   2062.      4.     19.      1.   22.75   95.1869
0.1571
2  138.   2062.      4.     21.      2.   22.84   95.3053
0.2012
```

```
2  138.   2062.      4.     21.      2.   22.92   95.1432
0.1532
2  138.   2362.      4.     12.      1.   22.74   95.1687
0.0785
2  138.   2362.      4.     18.      2.   22.75   95.1564
0.0430
2  138.   2362.      4.     19.      2.   22.88   95.1354
0.0983
2  138.   2362.      4.     19.      1.   22.73   95.0422
0.0773
2  138.   2362.      4.     20.      2.   22.86   95.1354
0.0587
2  138.   2362.      4.     21.      2.   22.94   95.1075
0.0776
2  139.      1.      4.     13.      2.   23.14   99.3274
0.0220
2  139.      1.      4.     15.      2.   22.77   99.5020
0.0997
2  139.      1.      4.     18.      2.   22.80   99.4016
0.0704
2  139.      1.      4.     19.      1.   22.68   99.3181
0.1245
2  139.      1.      4.     20.      2.   22.78   99.3858
0.0903
2  139.      1.      4.     21.      2.   22.93   99.3141
0.0255
2  139.    281.      4.     14.      2.   23.05   99.2915
0.0859
2  139.    281.      4.     15.      2.   22.71   99.4032
0.1322
2  139.    281.      4.     18.      2.   22.79   99.4612
0.1765
2  139.    281.      4.     20.      2.   22.74   99.4001
0.0889
2  139.    281.      4.     20.      2.   22.91   99.3765
0.1041
2  139.    281.      4.     21.      2.   22.92   99.3507
0.0717
2  139.    283.      4.     13.      2.   23.11   99.3848
0.0792
2  139.    283.      4.     18.      2.   22.84   99.4952
0.1122
2  139.    283.      4.     18.      2.   22.76   99.3220
0.0915
2  139.    283.      4.     19.      2.   23.03   99.4165
0.0503
2  139.    283.      4.     21.      2.   22.87   99.3791
0.1138
2  139.    283.      4.     21.      2.   22.98   99.3985
0.0661
2  139.   2062.      4.     14.      2.   22.43   99.4283
0.0891
2  139.   2062.      4.     15.      2.   22.70   99.4139
0.2147
2  139.   2062.      4.     19.      2.   22.97   99.3813
0.1143
2  139.   2062.      4.     19.      1.   22.77   99.4314
0.1685
2  139.   2062.      4.     21.      2.   22.79   99.4166
0.2080
2  139.   2062.      4.     21.      2.   22.94   99.4052
0.2400
2  139.   2362.      4.     12.      1.   22.82   99.3408
0.1279
2  139.   2362.      4.     18.      2.   22.77   99.3116
0.1131
2  139.   2362.      4.     19.      2.   22.82   99.3241
0.0519
2  139.   2362.      4.     19.      1.   22.74   99.2991
0.0903
2  139.   2362.      4.     20.      2.   22.88   99.3049
0.0783
2  139.   2362.      4.     21.      2.   22.94   99.2782
0.0718
2  140.      1.      4.     13.      1.   23.10   96.0811
0.0463
2  140.      1.      4.     15.      2.   22.75   96.1460
0.0725
2  140.      1.      4.     18.      2.   22.78   96.1582
0.1428
2  140.      1.      4.     19.      1.   22.70   96.1039
```

```
0.1056
2  140.      1.      4.     20.      2.   22.75   96.1262
0.0672
2  140.      1.      4.     21.      2.   22.93   96.1478
0.0562
2  140.    281.      4.     15.      2.   22.71   96.1153
0.1097
2  140.    281.      4.     14.      2.   22.49   96.1297
0.1202
2  140.    281.      4.     18.      2.   22.81   96.1233
0.1331
2  140.    281.      4.     20.      2.   22.78   96.1731
0.1484
2  140.    281.      4.     20.      2.   22.89   96.0872
0.0857
2  140.    281.      4.     21.      2.   22.91   96.1331
0.0944
2  140.    283.      4.     13.      2.   23.22   96.1135
0.0983
2  140.    283.      4.     18.      2.   22.85   96.1111
0.1210
2  140.    283.      4.     18.      2.   22.78   96.1221
0.0644
2  140.    283.      4.     19.      2.   23.01   96.1063
0.0921
2  140.    283.      4.     21.      2.   22.91   96.1155
0.0704
2  140.    283.      4.     21.      2.   22.94   96.1308
0.0258
2  140.   2062.      4.     15.      2.   22.60   95.9767
0.2225
2  140.   2062.      4.     15.      2.   22.66   96.1277
0.1792
2  140.   2062.      4.     19.      2.   22.96   96.1858
0.1312
2  140.   2062.      4.     19.      1.   22.75   96.1912
0.1936
2  140.   2062.      4.     21.      2.   22.82   96.1650
0.1902
2  140.   2062.      4.     21.      2.   22.92   96.1603
0.1777
2  140.   2362.      4.     12.      1.   22.88   96.0793
0.0996
2  140.   2362.      4.     18.      2.   22.76   96.1115
0.0533
2  140.   2362.      4.     19.      2.   22.79   96.0803
0.0364
2  140.   2362.      4.     19.      1.   22.71   96.0411
0.0768
2  140.   2362.      4.     20.      2.   22.84   96.0988
0.1042
2  140.   2362.      4.     21.      1.   22.94   96.0482
0.0868
2  141.      1.      4.     13.      1.   23.07  101.1984
0.0803
2  141.      1.      4.     15.      2.   22.72  101.1645
0.0914
2  141.      1.      4.     18.      2.   22.75  101.2454
0.1109
2  141.      1.      4.     19.      1.   22.69  101.1096
0.1376
2  141.      1.      4.     20.      2.   22.83  101.2066
0.0717
2  141.      1.      4.     21.      2.   22.93  101.0645
0.1205
2  141.    281.      4.     15.      2.   22.72  101.1615
0.1272
2  141.    281.      4.     14.      2.   22.40  101.1650
0.0595
2  141.    281.      4.     18.      2.   22.78  101.1815
0.1393
2  141.    281.      4.     20.      2.   22.73  101.1106
0.1189
2  141.    281.      4.     20.      2.   22.86  101.1420
0.0713
2  141.    281.      4.     21.      2.   22.94  101.0116
0.1088
2  141.    283.      4.     13.      2.   23.26  101.1554
0.0429
2  141.    283.      4.     18.      2.   22.85  101.1267
0.0751
```

```
2  141.    283.      4.     18.      2.   22.76  101.1227
0.0826
2  141.    283.      4.     19.      2.   22.82  101.0635
0.1715
2  141.    283.      4.     21.      2.   22.89  101.1264
0.1447
2  141.    283.      4.     21.      2.   22.96  101.0853
0.1189
2  141.   2062.      4.     15.      2.   22.65  101.1332
0.2532
2  141.   2062.      4.     15.      1.   22.68  101.1487
0.1413
2  141.   2062.      4.     19.      2.   22.95  101.1778
0.1772
2  141.   2062.      4.     19.      1.   22.77  101.0988
0.0884
2  141.   2062.      4.     21.      2.   22.87  101.1686
0.2940
2  141.   2062.      4.     21.      2.   22.94  101.3289
0.2072
2  141.   2362.      4.     12.      1.   22.83  101.1353
0.0585
2  141.   2362.      4.     18.      2.   22.83  101.1201
0.0868
2  141.   2362.      4.     19.      2.   22.91  101.0946
0.0855
2  141.   2362.      4.     19.      1.   22.71  100.9977
0.0645
2  141.   2362.      4.     20.      2.   22.87  101.0963
0.0638
2  141.   2362.      4.     21.      2.   22.94  101.0300
0.0549
2  142.      1.      4.     13.      1.   23.07   94.3049
0.1197
2  142.      1.      4.     15.      2.   22.73   94.3153
0.0566
2  142.      1.      4.     18.      2.   22.77   94.3073
0.0875
2  142.      1.      4.     19.      1.   22.67   94.2803
0.0376
2  142.      1.      4.     20.      2.   22.80   94.3008
0.0703
2  142.      1.      4.     21.      2.   22.93   94.2916
0.0604
2  142.    281.      4.     14.      2.   22.90   94.2557
0.0619
2  142.    281.      4.     18.      2.   22.83   94.3542
0.1027
2  142.    281.      4.     18.      2.   22.80   94.3007
0.1492
2  142.    281.      4.     20.      2.   22.76   94.3351
0.1059
2  142.    281.      4.     20.      2.   22.88   94.3406
0.1508
2  142.    281.      4.     21.      2.   22.92   94.2621
0.0946
2  142.    283.      4.     13.      2.   23.25   94.3124
0.0534
2  142.    283.      4.     18.      2.   22.85   94.3680
0.1643
2  142.    283.      4.     18.      1.   22.67   94.3442
0.0346
2  142.    283.      4.     19.      2.   22.80   94.3391
0.0616
2  142.    283.      4.     21.      2.   22.91   94.2238
0.0721
2  142.    283.      4.     21.      2.   22.95   94.2721
0.0998
2  142.   2062.      4.     14.      2.   22.49   94.2915
0.2189
2  142.   2062.      4.     15.      2.   22.69   94.2803
0.0690
2  142.   2062.      4.     19.      2.   22.94   94.2818
0.0987
2  142.   2062.      4.     19.      1.   22.76   94.2227
0.2628
2  142.   2062.      4.     21.      2.   22.74   94.4109
0.1230
2  142.   2062.      4.     21.      2.   22.94   94.2616
0.0929
2  142.   2362.      4.     12.      1.   22.86   94.2052
```

```
0.0813
2  142.   2362.      4.     18.      2.   22.83   94.2824
0.0605
2  142.   2362.      4.     19.      2.   22.85   94.2396
0.0882
2  142.   2362.      4.     19.      1.   22.75   94.2087
0.0702
2  142.   2362.      4.     20.      2.   22.86   94.2937
0.0591
2  142.   2362.      4.     21.      1.   22.93   94.2330
0.0556
```

2.6.1.2. Analysis and interpretation

Graphs of probe effect on repeatability

A graphical analysis shows repeatability standard deviations plotted by wafer and probe. Probes are coded by numbers with probe #2362 coded as #5. The plots show that for both runs the precision of this probe is better than for the other probes.

Probe #2362, because of its superior precision, was chosen as the tool for measuring all 100 ohm.cm resistivity wafers at NIST. Therefore, the remainder of the analysis focuses on this probe.

Plot of repeatability standard deviations for probe #2362 from the nested design over days, wafers, runs

The precision of probe #2362 is first checked for consistency by plotting the repeatability standard deviations over days, wafers and runs. Days are coded by letter. The plots verify that, for both runs, probe repeatability is not dependent on wafers or days although the standard deviations on days D, E, and F of run 2 are larger in some instances than for the other days. This is not surprising because repeated probing on the wafer surfaces can cause slight degradation. Then the repeatability standard deviations are pooled over:

- $K = 6$ days for $K(J - 1) = 30$ degrees of freedom
- $L = 2$ runs for $LK(J - 1) = 60$ degrees of freedom
- $Q = 5$ wafers for $QLK(J - 1) = 300$ degrees of freedom

The results of pooling are shown below. Intermediate steps are not shown, but the section on repeatability standard deviations shows an example of pooling over wafers.

Pooled level-1 standard deviations (ohm.cm)

Probe	Run 1	DF	Run 2	DF	Pooled DF
2362.	0.0658	150	0.0758	150	0.0710 300

Graphs of reproducibility and stability for

Averages of the 6 center measurements on each wafer are plotted on a single graph for each wafer. The points (connected by lines) on the left side of each graph are

probe #2362

averages at the wafer center plotted over 5 days; the points on the right are the same measurements repeated after one month as a check on the stability of the measurement process. The plots show day-to-day variability as well as slight variability from run-to-run.

Earlier work discounts long-term drift in the gauge as the cause of these changes. A reasonable conclusion is that day-to-day and run-to-run variations come from random fluctuations in the measurement process.

Level-2 (reproducibility) standard deviations computed from day averages and pooled over wafers and runs

Level-2 standard deviations (with $K - 1 = 5$ degrees of freedom each) are computed from the daily averages that are recorded in the database. Then the level-2 standard deviations are pooled over:

- $L = 2$ runs for $L(K - 1) = 10$ degrees of freedom
- $Q = 5$ wafers for $QL(K - 1) = 50$ degrees of freedom

as shown in the table below. The table shows that the level-2 standard deviations are consistent over wafers and runs.

Level-2 standard deviations (ohm.cm) for 5 wafers

		Run 1			Run 2		
Wafer	Probe	Average	Stddev	DF	Average	Stddev	DF
138.	2362.	95.0928	0.0359	5	95.1243	0.0453	5
139.	2362.	99.3060	0.0472	5	99.3098	0.0215	5
140.	2362.	96.0357	0.0273	5	96.0765	0.0276	5
141.	2362.	101.0602	0.0232	5	101.0790	0.0537	5
142.	2362.	94.2148	0.0274	5	94.2438	0.0370	5
	2362.	Pooled	0.0333	25		0.0388	25
		(over 2 runs)				0.0362	50

Level-3 (stability) standard deviations computed from run averages and pooled

Level-3 standard deviations are computed from the averages of the two runs. Then the level-3 standard deviations are pooled over the five wafers to obtain a standard deviation with 5 degrees of freedom as shown in the table below.

over
wafers

Level-3 standard deviations (ohm.cm) for 5 wafers

Wafer	Probe	Run 1 Average	Run 2 Average	Diff	Stddev	DF
138.	2362.	95.0928	95.1243	-0.0315	0.0223	1
139.	2362.	99.3060	99.3098	-0.0038	0.0027	1
140.	2362.	96.0357	96.0765	-0.0408	0.0289	1
141.	2362.	101.0602	101.0790	-0.0188	0.0133	1
142.	2362.	94.2148	94.2438	-0.0290	0.0205	1
	2362.			Pooled	0.0197	5

[*Graphs of probe biases*](#)

A graphical analysis shows the relative biases among the 5 probes. For each wafer, differences from the wafer average by probe are plotted versus wafer number. The graphs verify that probe #2362 (coded as 5) is biased low relative to the other probes. The bias shows up more strongly after the probes have been in use (run 2).

[*Formulas for computation of biases for probe #2362*](#)

Biases by probe are shown in the following table.

Differences from the mean for each wafer

Wafer	Probe	Run 1	Run 2
138.	1.	0.0248	-0.0119
138.	281.	0.0108	0.0323
138.	283.	0.0193	-0.0258
138.	2062.	-0.0175	0.0561
138.	2362.	-0.0372	-0.0507
139.	1.	-0.0036	-0.0007
139.	281.	0.0394	0.0050
139.	283.	0.0057	0.0239
139.	2062.	-0.0323	0.0373
139.	2362.	-0.0094	-0.0657
140.	1.	0.0400	0.0109
140.	281.	0.0187	0.0106
140.	283.	-0.0201	0.0003
140.	2062.	-0.0126	0.0182
140.	2362.	-0.0261	-0.0398
141.	1.	0.0394	0.0324
141.	281.	-0.0107	-0.0037
141.	283.	0.0246	-0.0191
141.	2062.	-0.0280	0.0436
141.	2362.	-0.0252	-0.0534

```
142.    1.    0.0062    0.0093
142.  281.    0.0376    0.0174
142.  283.   -0.0044    0.0192
142. 2062.   -0.0011    0.0008
142. 2362.   -0.0383   -0.0469
```

How to deal with bias due to the probe

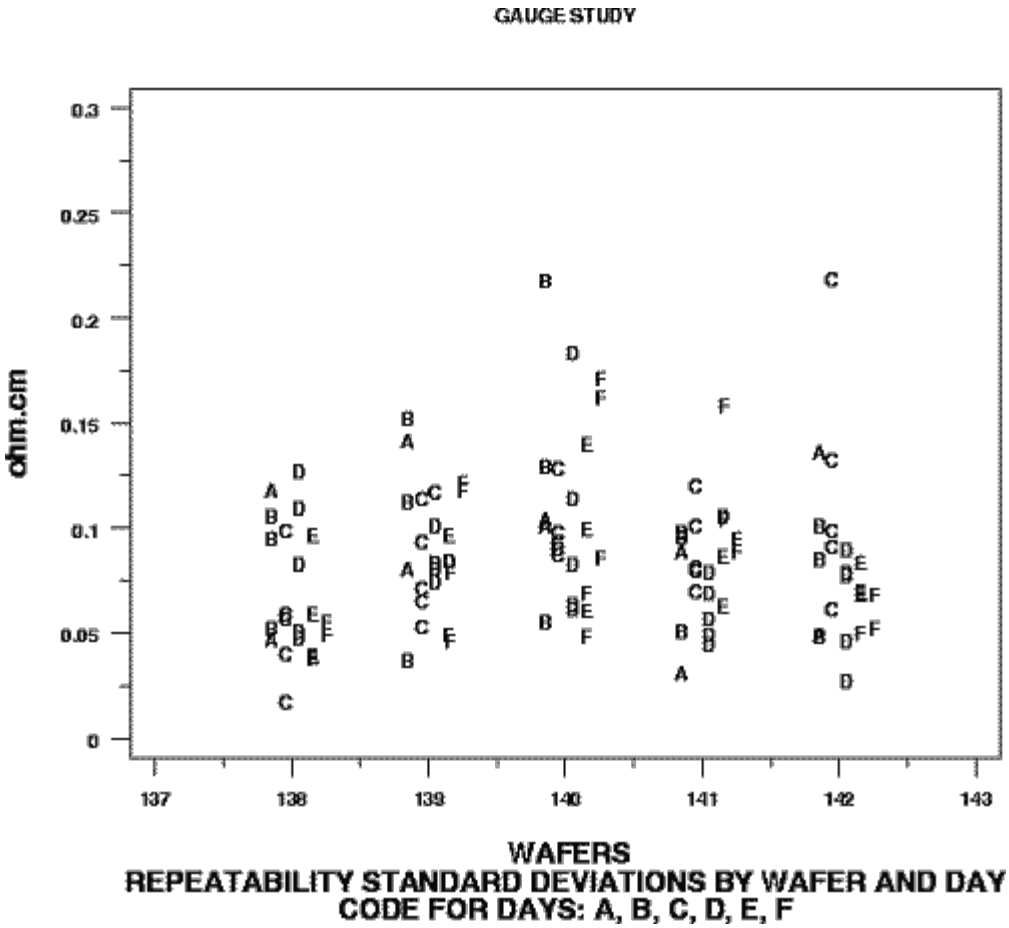
Probe #2362 was chosen for the certification process because of its superior precision, but its bias relative to the other probes creates a problem. There are two possibilities for handling this problem:

1. Correct all measurements made with probe #2362 to the average of the probes.
2. Include the standard deviation for the difference among probes in the uncertainty budget.

The better choice is (1) if we can assume that the probes in the study represent a random sample of probes of this type. This is particularly true when the unit (resistivity) is defined by a test method.

2. Measurement Process Characterization
2.6. Case studies
2.6.1. Gauge study of resistivity probes

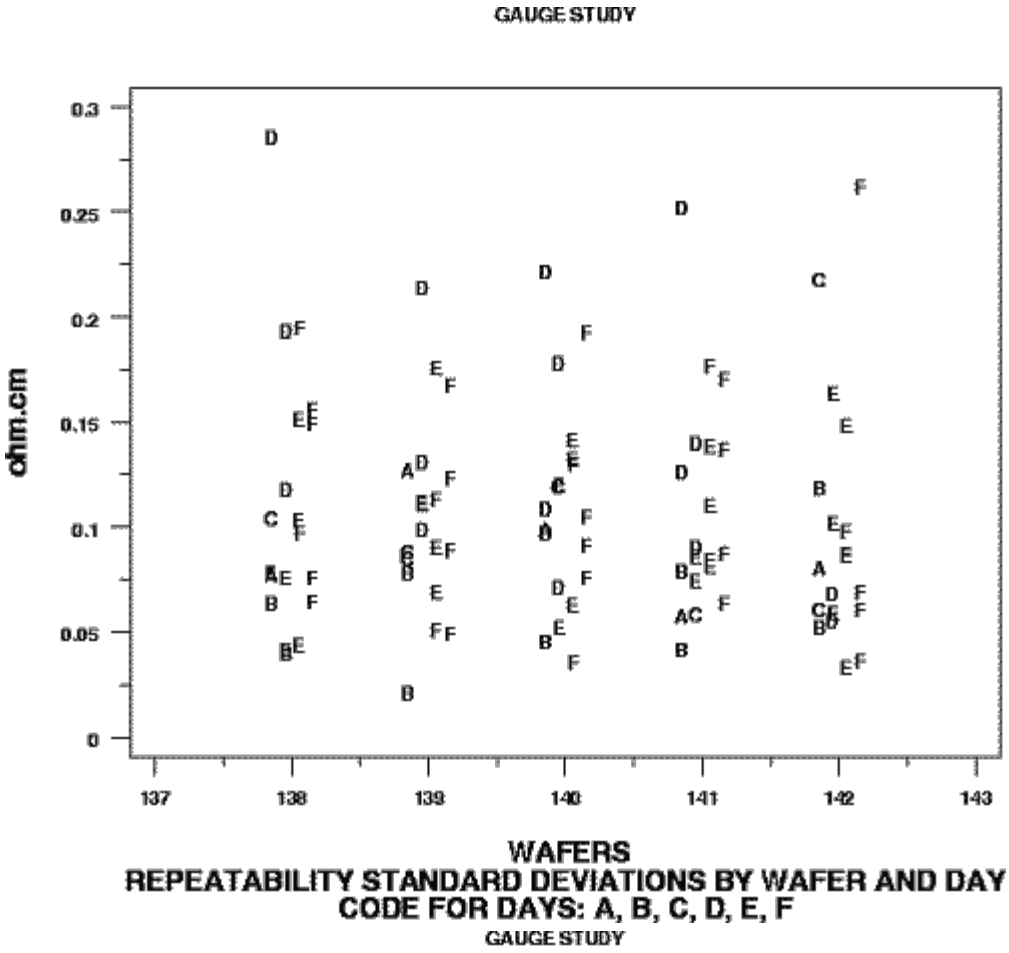
2.6.1.3. Repeatability standard deviations



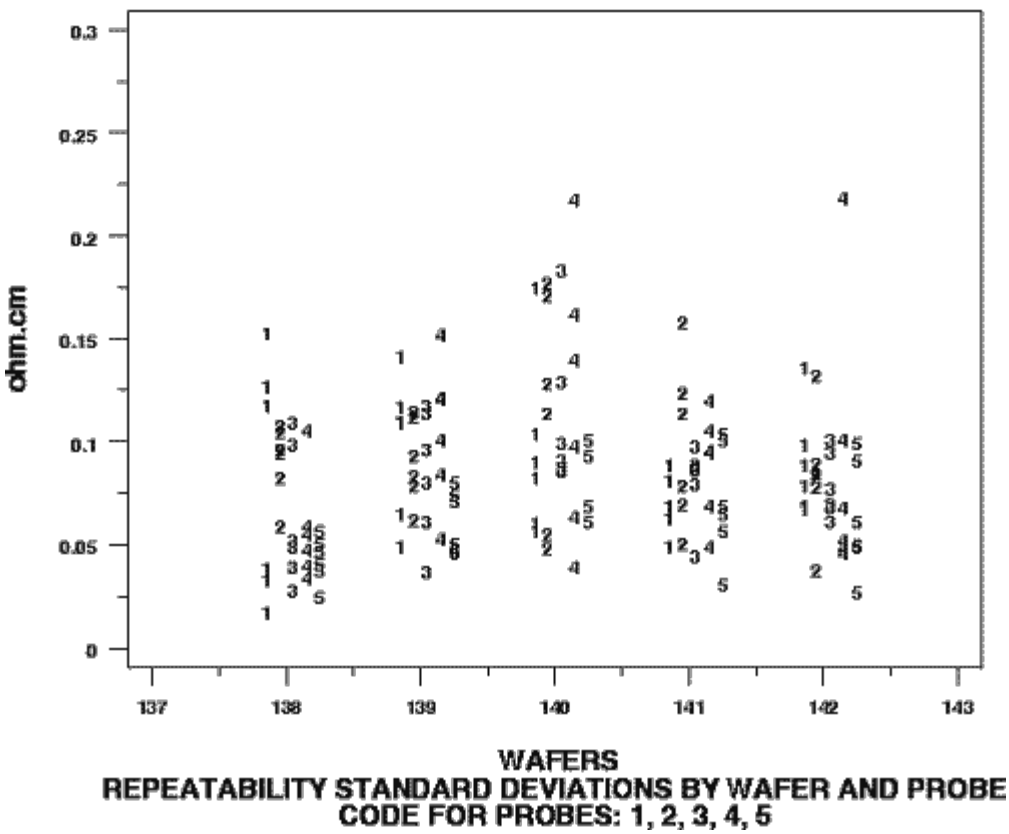
Run 1 - *Graph of repeatability standard deviations for probe #2362 -- 6 days and 5 wafers showing that repeatability is constant across wafers and days*

Run 2 - *Graph of repeatability standard deviations for probe #2362 -- 6 days and 5 wafers showing that*

repeatability is constant across wafers and days

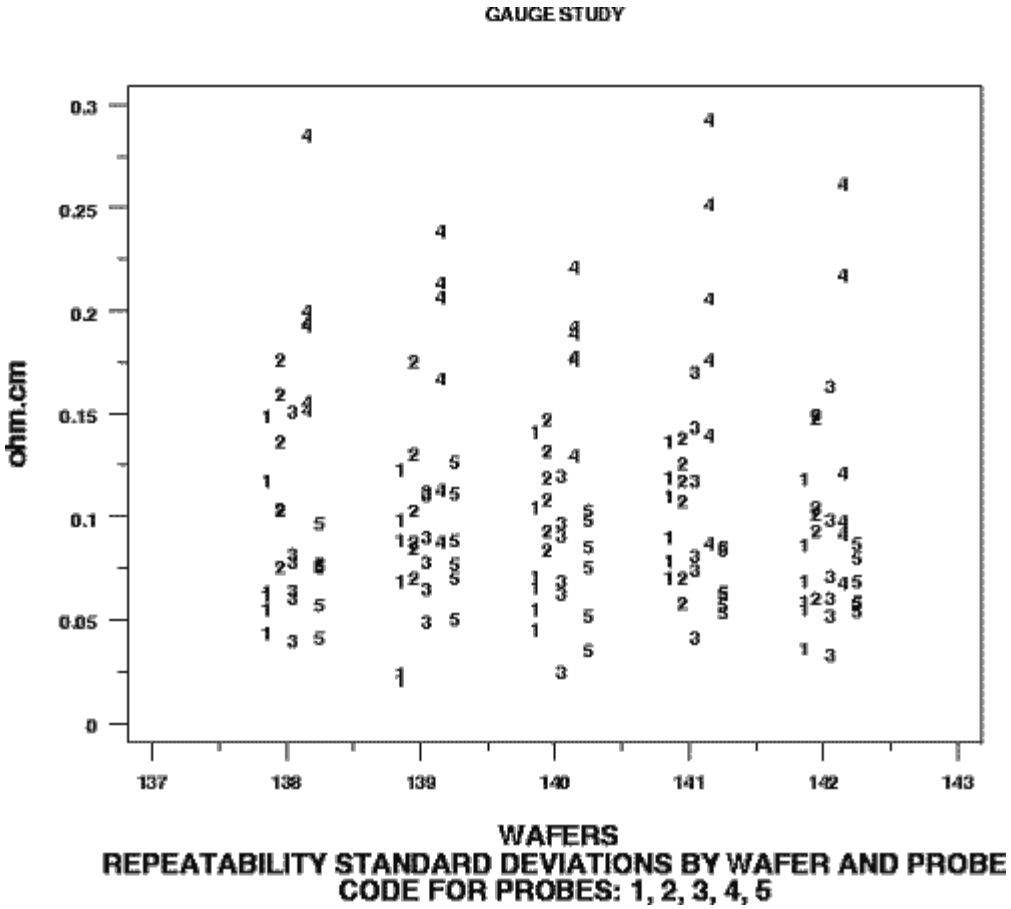


Run 1 *- Graph showing repeatability standard deviations for five probes as a function of wafers and probes*



Symbols for codes: 1 = #1; 2 = #281; 3 = #283; 4 = #2062; 5 = #2362

Run 2 - *Graph showing repeatability standard deviations for 5 probes as a function of wafers and probes*



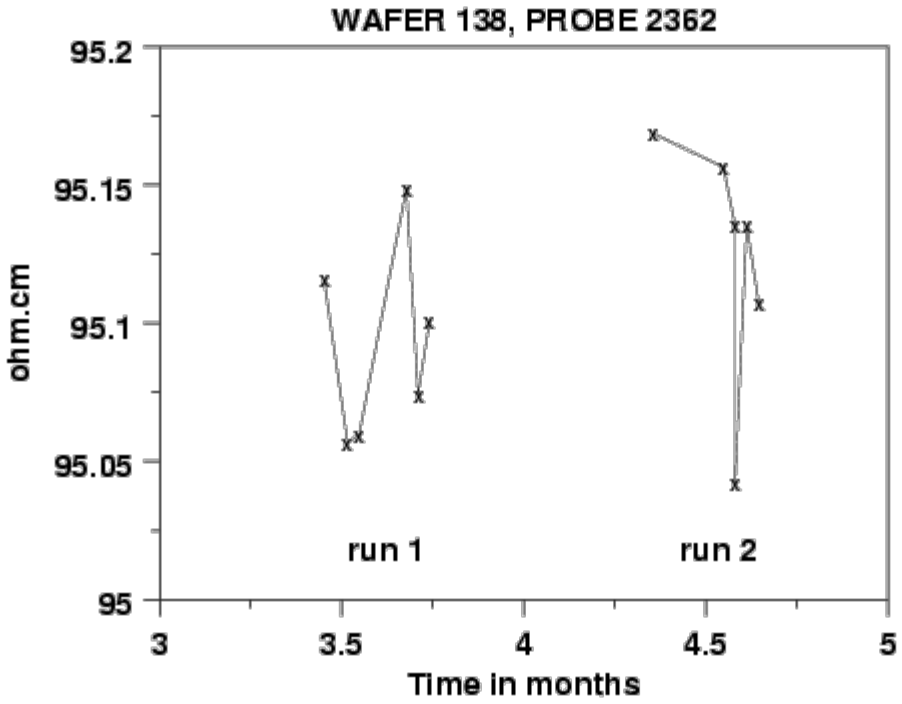
Symbols for probes: 1 = #1; 2 = #281; 3 = #283; 4 = #2062; 5 = #2362

2.6.1.4. Effects of days and long-term stability

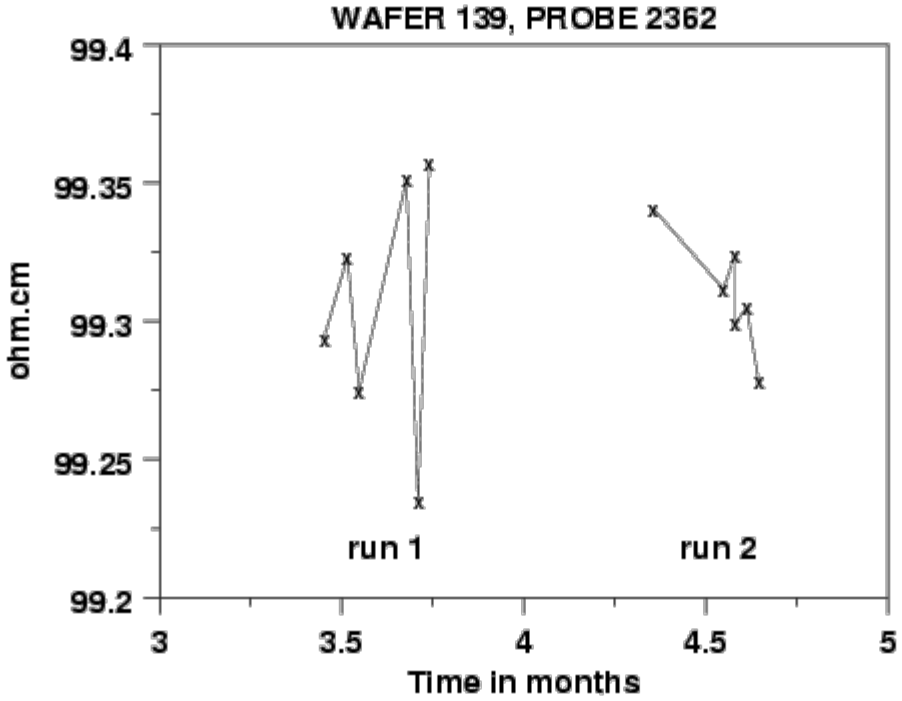
Effects of days and long-term stability on the measurements

The data points that are plotted in the five graphs shown below are averages of resistivity measurements at the center of each wafer for wafers #138, 139, 140, 141, 142. Data for each of two runs are shown on each graph. The six days of measurements for each run are separated by approximately one month and show, with the exception of wafer #139, that there is a very slight shift upwards between run 1 and run 2. The size of the effect is estimated as a level-3 standard deviation in the analysis of the data.

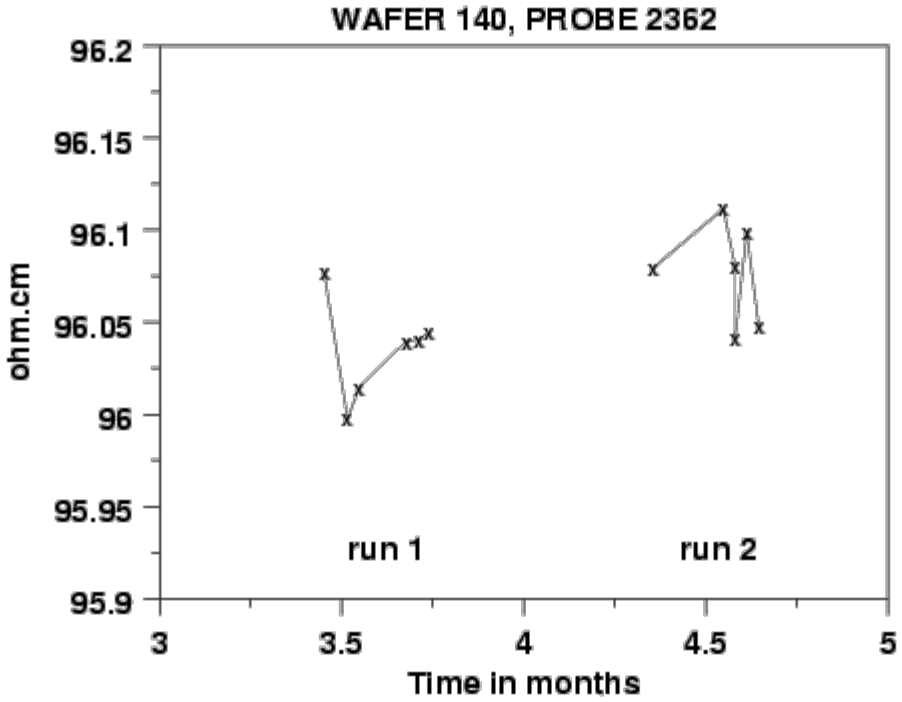
Wafer 138



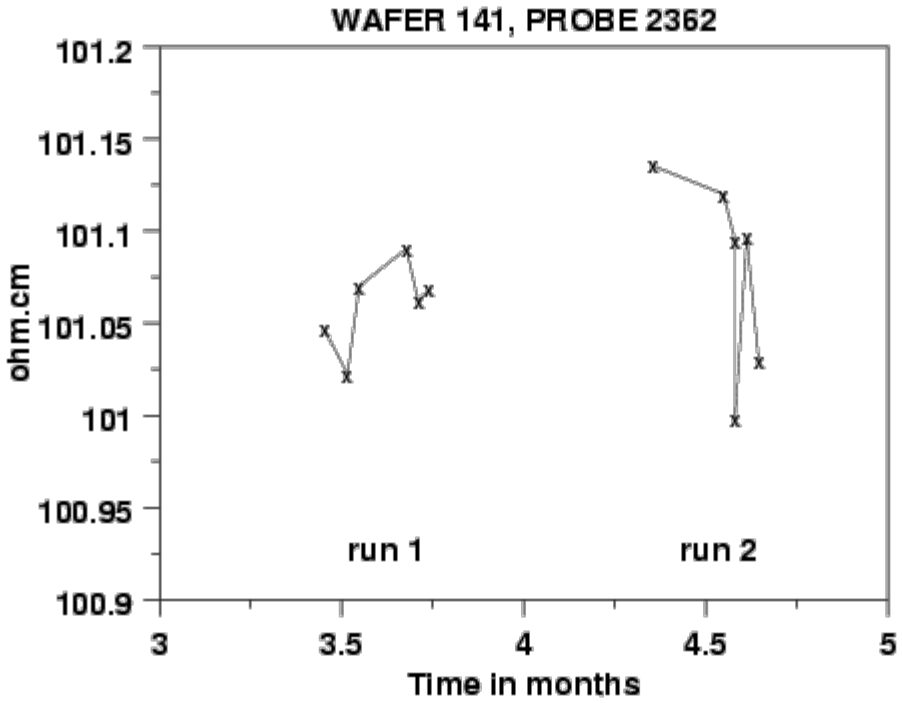
Wafer 139



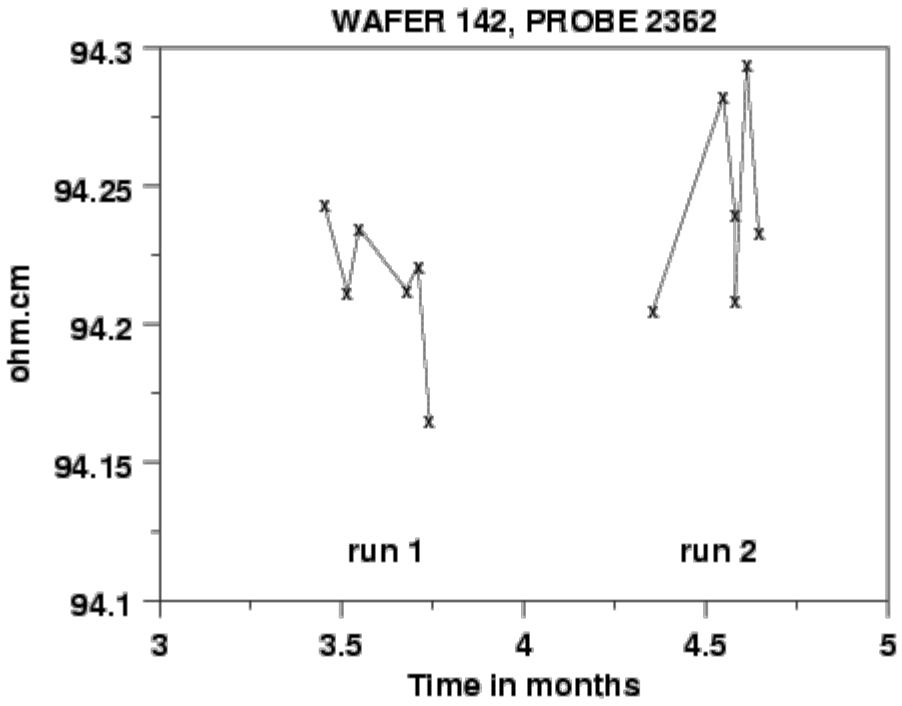
Wafer 140



Wafer 141

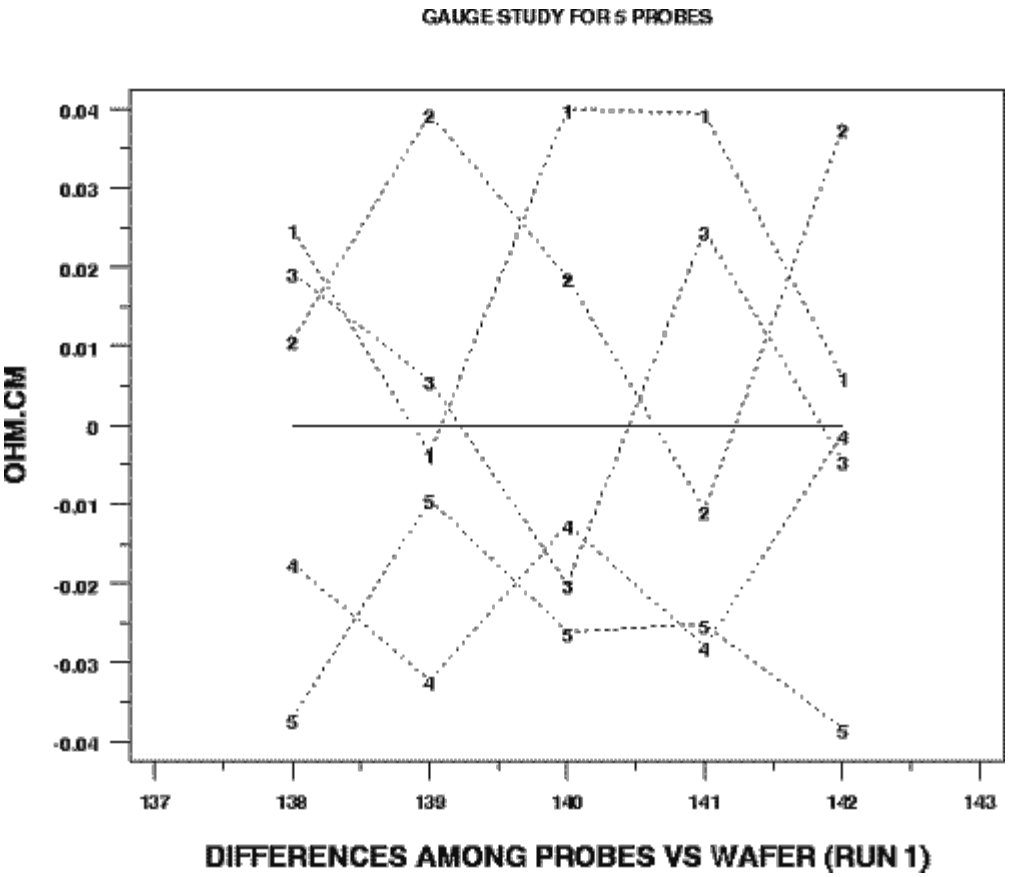

WAFER 141, PROBE 2362
101.2
101.15
101.1
101.05
101
100.95
100.9
ohm.cm
run 1
run 2
3
3.5
4
4.5
5
Time in months

Wafer 142


WAFER 142, PROBE 2362
94.3
94.25
94.2
94.15
94.1
ohm.cm
run 1
run 2
3
3.5
4
4.5
5
Time in months

2.6.1.5. Differences among 5 probes

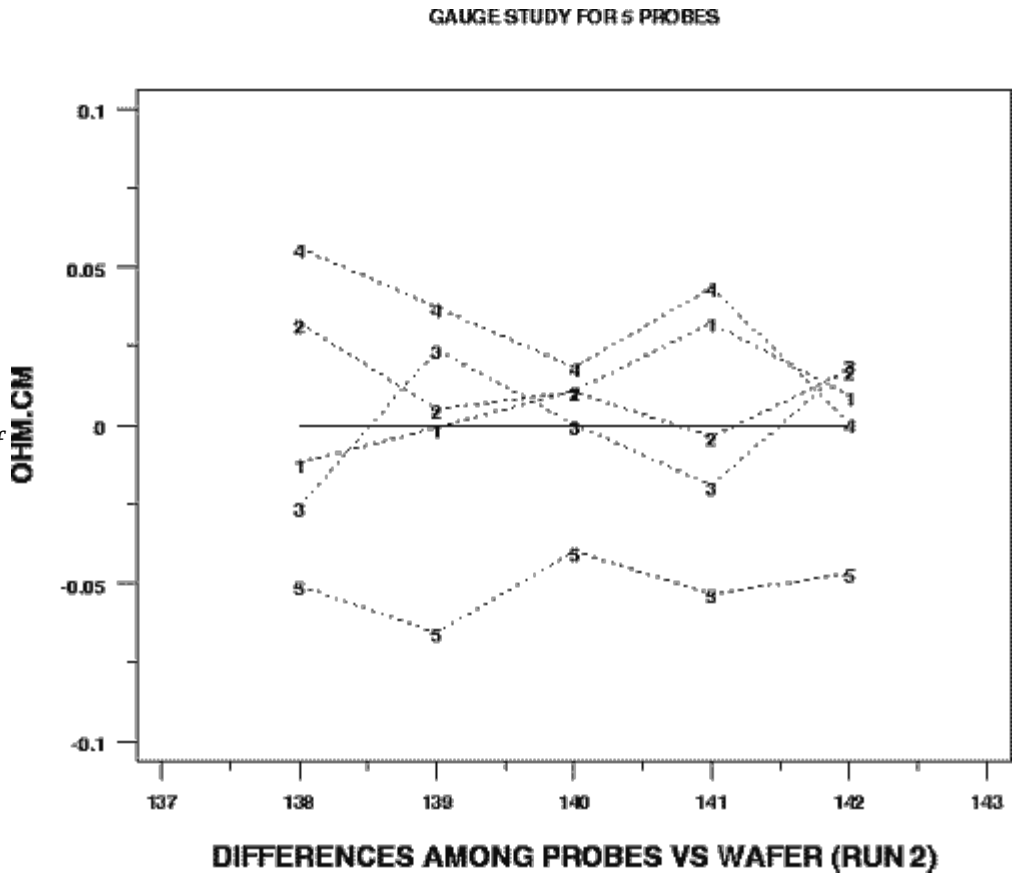
Run 1 - *Graph of differences from wafer averages for each of 5 probes showing that probes #2062 and #2362 are biased low relative to the other probes*



Symbols for probes: 1 = #1; 2 = #281; 3 = #283; 4 = #2062; 5 = #2362

Run 2 - *Graph of differences from wafer averages*

for each of 5 probes showing that probe #2362 continues to be biased low relative to the other probes



Symbols for probes: 1 = #1; 2 = #281; 3 = #283; 4 = #2062; 5 = #2362

2.6.1.6. Run gauge study example using Dataplot

View of Dataplot macros for this case study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output Window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
Graphical analyses of variability *Graphs to test for:* 1. Wafer/day effect on repeatability (run 1) 2. Wafer/day effect on repeatability (run 2) 3. Probe effect on repeatability (run 1) 4. Probe effect on repeatability (run 2) 5. Reproducibility and stability	1. and 2. Interpretation: The plots verify that, for both runs, the repeatability of probe #2362 is not dependent on wafers or days, although the standard deviations on days D, E, and F of run 2 are larger in some instances than for the other days. 3. and 4. Interpretation: Probe #2362 appears as #5 in the plots which show that, for both runs, the precision of this probe is better than for the other probes. 5. Interpretation: There is a separate plot for each wafer. The points on the left side of each plot are averages at the wafer center plotted over 5 days; the points on

	the right are the same measurements repeated after one month to check on the stability of the measurement process. The plots show day-to-day variability as well as slight variability from run-to-run.
Table of estimates for probe #2362 1. Level-1 (repeatability) 2. Level-2 (reproducibility) 3. Level-3 (stability)	1., 2. and 3.: Interpretation: The repeatability of the gauge (level-1 standard deviation) dominates the imprecision associated with measurements and days and runs are less important contributors. Of course, even if the gauge has high precision, biases may contribute substantially to the uncertainty of measurement.
Bias estimates 1. Differences among probes - run 1 2. Differences among probes - run 2	1. and 2. Interpretation: The graphs show the relative biases among the 5 probes. For each wafer, differences from the wafer average by probe are plotted versus wafer number. The graphs verify that probe #2362 (coded as 5) is biased low relative to the other probes. The bias shows up more strongly after the probes have been in use (run 2).

2.6.1.7. Dataplot macros

Plot of wafer and day effect on repeatability standard deviations for run 1

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
read mpc61.dat run wafer probe mo day op hum y
sw
y1label ohm.cm
title GAUGE STUDY
lines blank all
let z = pattern 1 2 3 4 5 6 for I = 1 1 300
let z2 = wafer + z/10 -0.25
characters a b c d e f
X1LABEL WAFERS
X2LABEL REPEATABILITY STANDARD DEVIATIONS BY
WAFER AND DAY
X3LABEL CODE FOR DAYS: A, B, C, D, E, F
TITLE RUN 1
plot sw z2 day subset run 1
```

Plot of wafer and day effect on repeatability standard deviations for run 2

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
read mpc61.dat run wafer probe mo day op hum y
sw
y1label ohm.cm
title GAUGE STUDY
lines blank all
let z = pattern 1 2 3 4 5 6 for I = 1 1 300
let z2 = wafer + z/10 -0.25
characters a b c d e f
X1LABEL WAFERS
X2LABEL REPEATABILITY STANDARD DEVIATIONS BY
WAFER AND DAY
X3LABEL CODE FOR DAYS: A, B, C, D, E, F
TITLE RUN 2
plot sw z2 day subset run 2
```

Plot of repeatability standard deviations for 5 probes - run 1

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
read mpc61.dat run wafer probe mo day op hum y
sw
y1label ohm.cm
title GAUGE STUDY
lines blank all
let z = pattern 1 2 3 4 5 6 for I = 1 1 300
let z2 = wafer + z/10 -0.25
characters 1 2 3 4 5
X1LABEL WAFERS
X2LABEL REPEATABILITY STANDARD DEVIATIONS BY
WAFER AND PROBE
X3LABEL CODE FOR PROBES: 1= SRM1; 2= 281; 3=283;
4=2062; 5=2362
TITLE RUN 1
plot sw z2 probe subset run 1
```

Plot of repeatability standard deviations for 5 probes - run 2

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
read mpc61.dat run wafer probe mo day op hum y
sw
y1label ohm.cm
title GAUGE STUDY
lines blank all
let z = pattern 1 2 3 4 5 6 for I = 1 1 300
let z2 = wafer + z/10 -0.25
characters 1 2 3 4 5
X1LABEL WAFERS
X2LABEL REPEATABILITY STANDARD DEVIATIONS BY
WAFER AND PROBE
X3LABEL CODE FOR PROBES: 1= SRM1; 2= 281; 3=283;
4=2062; 5=2362
TITLE RUN 2
plot sw z2 probe subset run 2
```

Plot of differences from the wafer mean for 5 probes - run 1

```
reset data
reset plot control
reset i/o
dimension 500 30
read mpc61a.dat wafer probe d1 d2
let biasrun1 = mean d1 subset probe 2362
print biasrun1
title GAUGE STUDY FOR 5 PROBES
Y1LABEL OHM.CM
lines dotted dotted dotted dotted dotted solid
characters 1 2 3 4 5 blank
xlimits 137 143
let zero = pattern 0 for I = 1 1 30
x1label DIFFERENCES AMONG PROBES VS WAFER (RUN
1)
plot d1 wafer probe and
plot zero wafer
```

Plot of differences from the wafer mean for 5 probes - run 2

```
reset data
reset plot control
reset i/o
dimension 500 30
read mpc61a.dat wafer probe d1 d2
let biasrun2 = mean d2 subset probe 2362
print biasrun2
title GAUGE STUDY FOR 5 PROBES
Y1LABEL OHM.CM
lines dotted dotted dotted dotted dotted solid
characters 1 2 3 4 5 blank
xlimits 137 143
let zero = pattern 0 for I = 1 1 30
x1label DIFFERENCES AMONG PROBES VS WAFER (RUN
2)
plot d2 wafer probe and
plot zero wafer
```

Plot of averages by day showing reproducibility and stability for measurements made with probe #2362 on 5 wafers

```
reset data
reset plot control
reset i/o
dimension 300 50
label size 3
read mcp61b.dat  wafer probe mo1 day1 y1 mo2
day2 y2 diff
let t = mo1+(day1-1)/31.
let t2= mo2+(day2-1)/31.
x3label WAFER 138
multiplot 3 2
plot y1 t subset wafer 138 and
plot y2 t2 subset wafer 138
x3label wafer 139
plot y1 t subset wafer 139 and
plot y2 t2 subset wafer 139
x3label WAFER 140
plot y1 t subset wafer 140 and
plot y2 t2 subset wafer 140
x3label WAFER 140
plot y1 t subset wafer 141 and
plot y2 t2 subset wafer 141
```

```
x3label WAFER 142
plot y1 t subset wafer 142 and
plot y2 t2 subset wafer 142
```

2.6.2. Check standard for resistivity measurements

Purpose The purpose of this page is to outline the analysis of check standard data with respect to controlling the precision and long-term variability of the process.

Outline

1. Background and data
2. Analysis and interpretation
3. Run this example yourself using Dataplot

2.6.2.1. Background and data

Explanation of check standard measurements	The process involves the measurement of resistivity (ohm.cm) of individual silicon wafers cut from a single crystal (# 51939). The wafers were doped with phosphorous to give a nominal resistivity of 100 ohm.cm. A single wafer (#137), chosen at random from a batch of 130 wafers, was designated as the check standard for this process.
Design of data collection and Database	The measurements were carried out according to an ASTM Test Method (F84) with NIST probe #2362. The measurements on the check standard duplicate certification measurements that were being made, during the same time period, on individual wafers from crystal #51939. For the check standard there were: • $J = 6$ repetitions at the center of the wafer on each day • $K = 25$ days The $K = 25$ days cover the time during which the individual wafers were being certified at the National Institute of Standards and Technology.
Software	The analyses used in this case study can be generated using both Dataplot code and R code.

2.6.2.1.1. Database for resistivity check standard

Description of check standard

A single wafer (#137), chosen at random from a batch of 130 wafers, is the check standard for resistivity measurements at the 100 ohm.cm level at the National Institute of Standards and Technology. The average of six measurements at the center of the wafer is the check standard value for one occasion, and the standard deviation of the six measurements is the short-term standard deviation. The columns of the database contain the following:

1. Crystal ID
2. Check standard ID
3. Month
4. Day
5. Hour
6. Minute
7. Operator
8. Humidity
9. Probe ID
10. Temperature
11. Check standard value
12. Short-term standard deviation
13. Degrees of freedom

Database of measurements on check standard

Crystal	Waf	Mo	Da	Hr	Mn	Op	Hum	Probe	Temp	Avg	Stddev	DF
51939	137	03	24	18	01	drr	42	2362	23.003	97.070	0.085	5
51939	137	03	25	12	41	drr	35	2362	23.115	97.049	0.052	5
51939	137	03	25	15	57	drr	33	2362	23.196	97.048	0.038	5
51939	137	03	28	10	10	JMT	47	2362	23.383	97.084	0.036	5
51939	137	03	28	13	31	JMT	44	2362	23.491	97.106	0.049	5
51939	137	03	28	17	33	drr	43	2362	23.352	97.014	0.036	5

```
51939 137 03 29 14 40 drr 36 2362 23.202 97.047
                     0.052 5

51939 137 03 29 16 33 drr 35 2362 23.222 97.078
                     0.117 5

51939 137 03 30 05 45 JMT 32 2362 23.337 97.065
                     0.085 5

51939 137 03 30 09 26 JMT 33 2362 23.321 97.061
                     0.052 5

51939 137 03 25 14 59 drr 34 2362 22.993 97.060
                     0.060 5

51939 137 03 31 10 10 JMT 37 2362 23.164 97.102
                     0.048 5

51939 137 03 31 13 00 JMT 37 2362 23.169 97.096
                     0.026 5

51939 137 03 31 15 32 JMT 35 2362 23.156 97.035
                     0.088 5

51939 137 04 01 13 05 JMT 34 2362 23.097 97.114
                     0.031 5

51939 137 04 01 15 32 JMT 34 2362 23.127 97.069
                     0.037 5

51939 137 04 01 10 32 JMT 48 2362 22.963 97.095
                     0.032 5

51939 137 04 06 14 38 JMT 49 2362 23.454 97.088
                     0.056 5

51939 137 04 07 10 50 JMT 34 2362 23.285 97.079
                     0.067 5

51939 137 04 07 15 46 JMT 33 2362 23.123 97.016
                     0.116 5

51939 137 04 08 09 37 JMT 33 2362 23.373 97.051
                     0.046 5

51939 137 04 08 12 53 JMT 33 2362 23.296 97.070
                     0.078 5

51939 137 04 08 15 03 JMT 33 2362 23.218 97.065
                     0.040 5

51939 137 04 11 09 30 JMT 36 2362 23.415 97.111
                     0.038 5

51939 137 04 11 11 34 JMT 35 2362 23.395 97.073
                     0.039 5
```

2.6.2.2. Analysis and interpretation

Estimates of the repeatability standard deviation and level-2 standard deviation

The level-1 standard deviations (with $J - 1 = 5$ degrees of freedom each) from the database are pooled over the $K = 25$ days to obtain a reliable estimate of repeatability. This pooled value is

$$s_1 = 0.06139 \text{ ohm.cm}$$

with $K(J - 1) = 125$ degrees of freedom. The level-2 standard deviation is computed from the daily averages to be

$$s_2 = 0.02680 \text{ ohm.cm}$$

with $K - 1 = 24$ degrees of freedom.

Relationship to uncertainty calculations

These standard deviations are appropriate for estimating the uncertainty of the average of six measurements on a wafer that is of the same material and construction as the check standard. The computations are explained in the section on sensitivity coefficients for check standard measurements. For other numbers of measurements on the test wafer, the computations are explained in the section on sensitivity coefficients for level-2 designs.

Illustrative table showing computations of repeatability and level-2 standard deviations

A tabular presentation of a subset of check standard data ($J = 6$ repetitions and $K = 6$ days) illustrates the computations. The pooled repeatability standard deviation with $K(J - 1) = 30$ degrees of freedom from this limited database is shown in the next to last row of the table. A level-2 standard deviation with $K - 1 = 5$ degrees of freedom is computed from the center averages and is shown in the last row of the table.

Control chart for probe #2362

The control chart for monitoring the precision of probe #2362 is constructed as discussed in the section on control charts for standard deviations. The upper control limit (UCL) for testing for degradation of the probe is computed using the critical value from the F table with numerator degrees of freedom $J - 1 = 5$ and denominator degrees of freedom $K(J - 1) = 125$. For a 0.05 significance level,

$$F_{0.05,5,125} = 2.29$$

$$UCL = s_1\sqrt{F_{0.05,5,125}} = 0.09238\ ohm \cdot cm$$

Interpretation of control chart for probe #2362

The control chart shows two points exceeding the upper control limit. We expect 5 % of the standard deviations to exceed the UCL for a measurement process that is in-control. Two outliers are not indicative of significant problems with the repeatability for the probe, but the probe should be monitored closely in the future.

Control chart for bias and variability

The control limits for monitoring the bias and long-term variability of resistivity with a Shewhart control chart are given by

$$UCL = Average + 2*s_2 = 97.1234\ ohm.cm$$
$$Centerline = Average = 97.0698\ ohm.cm$$
$$LCL = Average - 2*s_2 = 97.0162\ ohm.cm$$

Interpretation of control chart for bias

The control chart shows that the points scatter randomly about the center line with no serious problems, although one point exceeds the upper control limit and one point exceeds the lower control limit by a small amount. The conclusion is that there is:

- No evidence of bias, change or drift in the measurement process.
- No evidence of long-term lack of control.

Future measurements that exceed the control limits must be evaluated for long-term changes in bias and/or variability.

2.6.2.2.1. Repeatability and level-2 standard deviations

Example

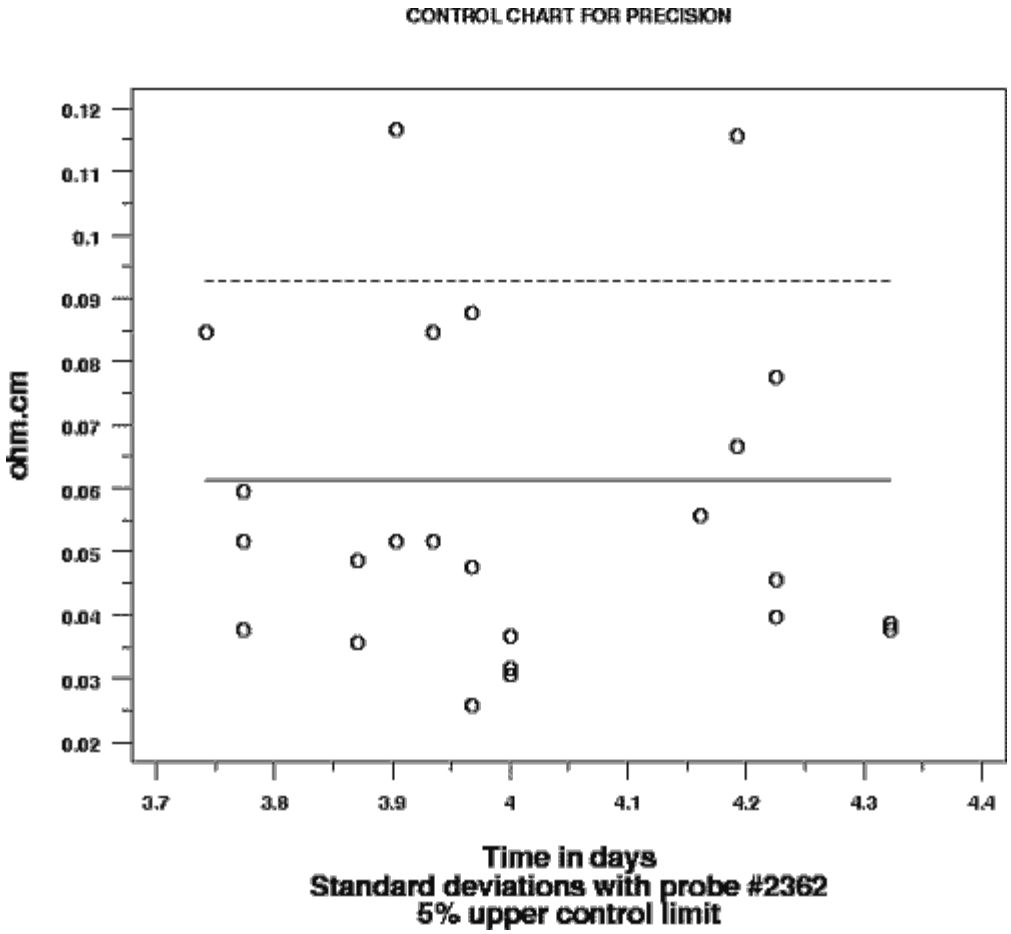
The table below illustrates the computation of repeatability and level-2 standard deviations from measurements on a check standard. The check standard measurements are resistivities at the center of a 100 ohm.cm wafer. There are $J = 6$ repetitions per day and $K = 5$ days for this example.

Table of data, averages, and repeatability standard deviations

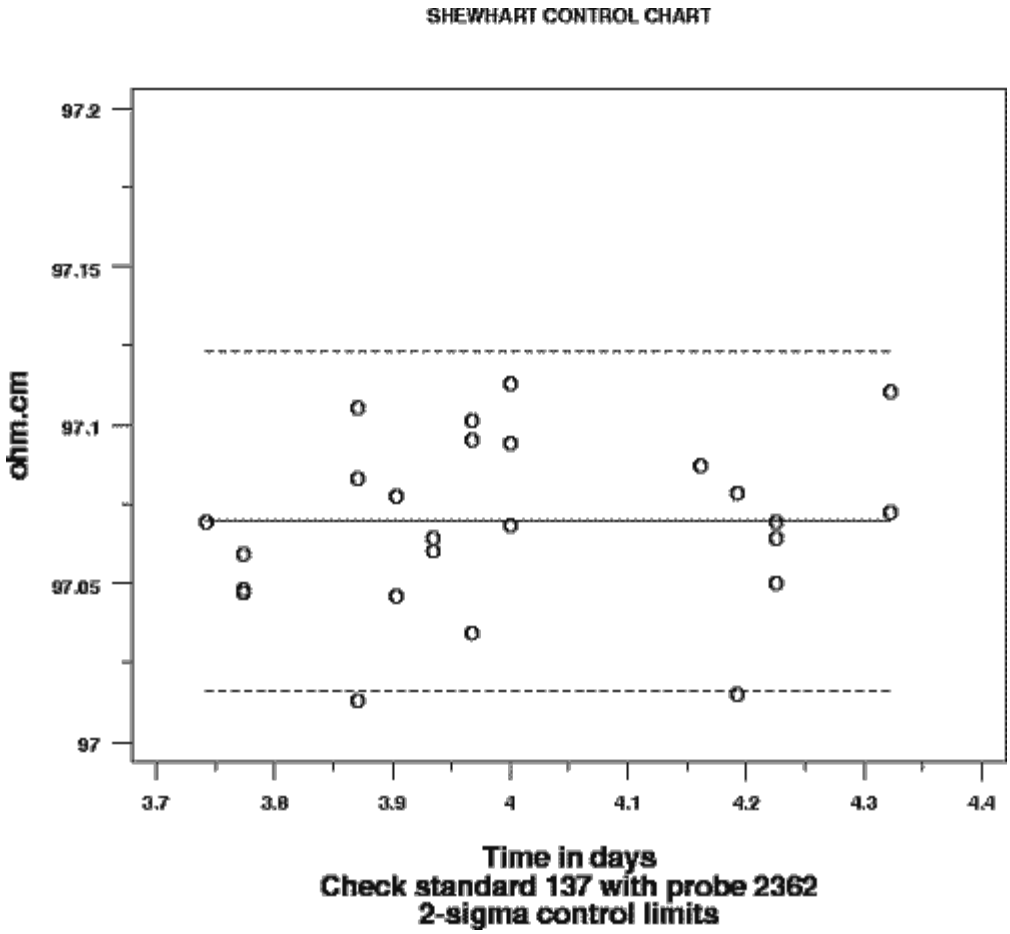
Measurements on check standard #137 Repetitions per day						
Days	**1**	**2**	**3**	**4**	**5**	**6**
1	96.920	97.054	97.057	97.035	97.189	96.965
2	97.118	96.947	97.110	97.047	96.945	97.013
3	97.034	97.084	97.023	97.045	97.061	97.074
4	97.047	97.099	97.087	97.076	97.117	97.070
5	97.127	97.067	97.106	96.995	97.052	97.121
6	96.995	96.984	97.053	97.065	96.976	96.997
Averages	97.040	97.039	97.073	97.044	97.057	97.037
Repeatability Standard Deviations	0.0777	0.0602	0.0341	0.0281	0.0896	0.0614
Pooled Repeatability Standard Deviation	**0.0625 30 df**					
Level-2 Standard Deviation	**0.0139 5 df**					

2.6.2.3. Control chart for probe precision

Control chart for probe #2362 showing violations of the control limits -- all standard deviations are based on 6 repetitions and the control limits are 95% limits



2.6.2.4. Control chart for bias and long-term variability



Shewhart control chart for measurements on a resistivity check standard showing that the process is in-control -- all measurements are averages of 6 repetitions

2.6.2.5. Run check standard example yourself

View of Dataplot macros for this case study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot. It is required that you have already downloaded and installed Dataplot and configured your browser to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output Window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
Graphical tests of assumptions Histogram Normal probability plot	The histogram and normal probability plots show no evidence of non-normality.
Control chart for precision Control chart for probe #2362 Computations: 1. Pooled repeatability standard deviation 2. Control limit	The precision control chart shows two points exceeding the upper control limit. We expect 5% of the standard deviations to exceed the UCL even when the measurement process is in-control.
Control chart for check standard Control chart for check standard #137	The Shewhart control chart shows that the points scatter randomly about the center

Computations: 1. Average check standard value 2. Process standard deviation 3. Upper and lower control limits	line with no serious problems, although one point exceeds the upper control limit and one point exceeds the lower control limit by a small amount. The conclusion is that there is no evidence of bias or lack of long-term control.

2.6.2.6. Dataplot macros

Histogram for check standard #137 to test assumption of normality

```
reset data
reset plot control
reset i/o
dimension 500 30
skip 14
read mpc62.dat crystal wafer mo day hour min op
hum probe temp y sw df
histogram y
```

Normal probability plot for check standard #137 to test assumption of normality

```
reset data
reset plot control
reset i/o
dimension 500 30
skip 14
read mpc62.dat crystal wafer mo day hour min op
hum probe temp y sw df
normal probabilty plot y
```

Control chart for precision of probe #2372 and computation of control parameter estimates

```
reset data
reset plot control
reset i/o
dimension 500 30
skip 14
read mpc62.dat crystal wafer mo day hour min op
hum probe temp y sw df
let time = mo +(day-1)/31.
let s = sw*sw
let spool = mean s
let spool = spool**.5
print spool
let f = fppf(.95, 5, 125)
let ucl = spool*(f)**.5
print ucl
title Control chart for precision
characters blank blank O
lines solid dashed blank
y1label ohm.cm
x1label Time in days
x2label Standard deviations with probe #2362
x3label 5% upper control limit
let center = sw - sw + spool
let cl = sw - sw + ucl
plot center cl sw vs time
```

Shewhart control chart for check standard #137 with computation of control

```
reset data
reset plot control
reset i/o
dimension 500 30
skip 14
read mpc62.dat crystal wafer mo day hour min op
hum probe temp y sw df
let time = mo +(day-1)/31.
let avg = mean y
let sprocess = standard deviation y
let ucl = avg + 2*sprocess
let lcl = avg - 2*sprocess
```

chart parameters

```
print avg
print sprocess
print ucl lcl
title Shewhart control chart
characters O blank blank blank
lines blank dashed solid dashed
y1label ohm.cm
x1label Time in days
x2label Check standard 137 with probe 2362
x3label 2-sigma control limits
let ybar = y - y + avg
let lc1 = y - y + lcl
let lc2 = y - y + ucl
plot y lc1 ybar lc2 vs time
```

2.6.3. Evaluation of type A uncertainty

Purpose The purpose of this case study is to demonstrate the computation of uncertainty for a measurement process with several sources of uncertainty from data taken during a gauge study.

Outline

1. Background and data for the study
2. Graphical and quantitative analyses and interpretations
3. Run this example yourself with Dataplot

2.6.3.1. Background and data

Description of measurements	The measurements in question are resistivities (ohm.cm) of silicon wafers. The intent is to calculate an uncertainty associated with the resistivity measurements of approximately 100 silicon wafers that were certified with probe #2362 in wiring configuration A, according to ASTM Method F84 (ASTM F84) which is the defined reference for this measurement. The reported value for each wafer is the average of six measurements made at the center of the wafer on a single day. Probe #2362 is one of five probes owned by the National Institute of Standards and Technology that is capable of making the measurements.
Sources of uncertainty in NIST measurements	The uncertainty analysis takes into account the following sources of variability: • Repeatability of measurements at the center of the wafer • Day-to-day effects • Run-to-run effects • Bias due to probe #2362 • Bias due to wiring configuration
Database of 3-level nested design -- for estimating time-dependent sources of uncertainty	The certification measurements themselves are not the primary source for estimating uncertainty components because they do not yield information on day-to-day effects and long-term effects. The standard deviations for the three time-dependent sources of uncertainty are estimated from a 3-level nested design. The design was replicated on each of $Q = 5$ wafers which were chosen at random, for this purpose, from the lot of wafers. The certification measurements were made between the two runs in order to check on the long-term stability of the process. The data consist of repeatability standard deviations (with $J - 1 = 5$ degrees of freedom each) from measurements at the wafer center.
Software	The analyses used in this case study can be generated using both Dataplot code and R code.

2.6.3.1.1. Database of resistivity measurements

Check standards are five wafers chosen at random from a batch of wafers

Measurements of resistivity (ohm.cm) were made according to an ASTM Standard Test Method (F4) at the National Institute of Standards and Technology to assess the sources of uncertainty in the measurement system. The gauges for the study were five probes owned by NIST; the check standards for the study were five wafers selected at random from a batch of wafers cut from one silicon crystal doped with phosphorous to give a nominal resistivity of 100 ohm.cm.

Measurements on the check standards are used to estimate repeatability, day effect, run effect

The effect of operator was not considered to be significant for this study. Averages and standard deviations from $J = 6$ measurements at the center of each wafer are shown in the table.

- $J = 6$ measurements at the center of the wafer per day
- $K = 6$ days (one operator) per repetition
- $L = 2$ runs (complete)
- $Q = 5$ wafers (check standards 138, 139, 140, 141, 142)
- $I = 5$ probes (1, 281, 283, 2062, 2362)

Run	Wafer	Probe	Month	Day	Operator	Temp	Average	Standard Deviation
1	138.	1.	3.	15.	1.	22.98	95.1772	0.1191
1	138.	1.	3.	17.	1.	23.02	95.1567	0.0183
1	138.	1.	3.	18.	1.	22.79	95.1937	0.1282
1	138.	1.	3.	21.	1.	23.17	95.1959	0.0398
1	138.	1.	3.	23.	2.	23.25	95.1442	0.0346
1	138.	1.	3.	23.	1.	23.20	95.0610	0.1539
1	138.	281.	3.	16.	1.	22.99	95.1591	0.0963
1	138.	281.	3.	17.	1.	22.97	95.1195	0.0606
1	138.	281.	3.	18.	1.	22.83	95.1065	0.0842
1	138.	281.	3.	21.	1.	23.28	95.0925	0.0973

```
1  138.    281.      3.     23.      2.   23.14   95.1990
0.1062
1  138.    281.      3.     23.      1.   23.16   95.1682
0.1090
1  138.    283.      3.     16.      1.   22.95   95.1252
0.0531
1  138.    283.      3.     17.      1.   23.08   95.1600
0.0998
1  138.    283.      3.     18.      1.   23.13   95.0818
0.1108
1  138.    283.      3.     21.      1.   23.28   95.1620
0.0408
1  138.    283.      3.     22.      1.   23.36   95.1735
0.0501
1  138.    283.      3.     24.      2.   22.97   95.1932
0.0287
1  138.   2062.      3.     16.      1.   22.97   95.1311
0.1066
1  138.   2062.      3.     17.      1.   22.98   95.1132
0.0415
1  138.   2062.      3.     18.      1.   23.16   95.0432
0.0491
1  138.   2062.      3.     21.      1.   23.16   95.1254
0.0603
1  138.   2062.      3.     22.      1.   23.28   95.1322
0.0561
1  138.   2062.      3.     24.      2.   23.19   95.1299
0.0349
1  138.   2362.      3.     15.      1.   23.08   95.1162
0.0480
1  138.   2362.      3.     17.      1.   23.01   95.0569
0.0577
1  138.   2362.      3.     18.      1.   22.97   95.0598
0.0516
1  138.   2362.      3.     22.      1.   23.23   95.1487
0.0386
1  138.   2362.      3.     23.      2.   23.28   95.0743
0.0256
1  138.   2362.      3.     24.      2.   23.10   95.1010
0.0420
1  139.      1.      3.     15.      1.   23.01   99.3528
0.1424
1  139.      1.      3.     17.      1.   23.00   99.2940
0.0660
1  139.      1.      3.     17.      1.   23.01   99.2340
0.1179
1  139.      1.      3.     21.      1.   23.20   99.3489
0.0506
1  139.      1.      3.     23.      2.   23.22   99.2625
0.1111
1  139.      1.      3.     23.      1.   23.22   99.3787
0.1103
1  139.    281.      3.     16.      1.   22.95   99.3244
0.1134
1  139.    281.      3.     17.      1.   22.98   99.3378
0.0949
1  139.    281.      3.     18.      1.   22.86   99.3424
0.0847
1  139.    281.      3.     22.      1.   23.17   99.4033
0.0801
1  139.    281.      3.     23.      2.   23.10   99.3717
0.0630
1  139.    281.      3.     23.      1.   23.14   99.3493
0.1157
1  139.    283.      3.     16.      1.   22.94   99.3065
0.0381
1  139.    283.      3.     17.      1.   23.09   99.3280
0.1153
1  139.    283.      3.     18.      1.   23.11   99.3000
0.0818
1  139.    283.      3.     21.      1.   23.25   99.3347
0.0972
1  139.    283.      3.     22.      1.   23.36   99.3929
0.1189
1  139.    283.      3.     23.      1.   23.18   99.2644
0.0622
1  139.   2062.      3.     16.      1.   22.94   99.3324
0.1531
1  139.   2062.      3.     17.      1.   23.08   99.3254
0.0543
1  139.   2062.      3.     18.      1.   23.15   99.2555
```

```
0.1024
1  139.   2062.      3.     18.      1.   23.18   99.1946
0.0851
1  139.   2062.      3.     22.      1.   23.27   99.3542
0.1227
1  139.   2062.      3.     24.      2.   23.23   99.2365
0.1218
1  139.   2362.      3.     15.      1.   23.08   99.2939
0.0818
1  139.   2362.      3.     17.      1.   23.02   99.3234
0.0723
1  139.   2362.      3.     18.      1.   22.93   99.2748
0.0756
1  139.   2362.      3.     22.      1.   23.29   99.3512
0.0475
1  139.   2362.      3.     23.      2.   23.25   99.2350
0.0517
1  139.   2362.      3.     24.      2.   23.05   99.3574
0.0485
1  140.      1.      3.     15.      1.   23.07   96.1334
0.1052
1  140.      1.      3.     17.      1.   23.08   96.1250
0.0916
1  140.      1.      3.     18.      1.   22.77   96.0665
0.0836
1  140.      1.      3.     21.      1.   23.18   96.0725
0.0620
1  140.      1.      3.     23.      2.   23.20   96.1006
0.0582
1  140.      1.      3.     23.      1.   23.21   96.1131
0.1757
1  140.    281.      3.     16.      1.   22.94   96.0467
0.0565
1  140.    281.      3.     17.      1.   22.99   96.1081
0.1293
1  140.    281.      3.     18.      1.   22.91   96.0578
0.1148
1  140.    281.      3.     22.      1.   23.15   96.0700
0.0495
1  140.    281.      3.     22.      1.   23.33   96.1052
0.1722
1  140.    281.      3.     23.      1.   23.19   96.0952
0.1786
1  140.    283.      3.     16.      1.   22.89   96.0650
0.1301
1  140.    283.      3.     17.      1.   23.07   96.0870
0.0881
1  140.    283.      3.     18.      1.   23.07   95.8906
0.1842
1  140.    283.      3.     21.      1.   23.24   96.0842
0.1008
1  140.    283.      3.     22.      1.   23.34   96.0189
0.0865
1  140.    283.      3.     23.      1.   23.19   96.1047
0.0923
1  140.   2062.      3.     16.      1.   22.95   96.0379
0.2190
1  140.   2062.      3.     17.      1.   22.97   96.0671
0.0991
1  140.   2062.      3.     18.      1.   23.15   96.0206
0.0648
1  140.   2062.      3.     21.      1.   23.14   96.0207
0.1410
1  140.   2062.      3.     22.      1.   23.32   96.0587
0.1634
1  140.   2062.      3.     24.      2.   23.17   96.0903
0.0406
1  140.   2362.      3.     15.      1.   23.08   96.0771
0.1024
1  140.   2362.      3.     17.      1.   23.00   95.9976
0.0943
1  140.   2362.      3.     18.      1.   23.01   96.0148
0.0622
1  140.   2362.      3.     22.      1.   23.27   96.0397
0.0702
1  140.   2362.      3.     23.      2.   23.24   96.0407
0.0627
1  140.   2362.      3.     24.      2.   23.13   96.0445
0.0622
1  141.      1.      3.     15.      1.   23.01  101.2124
0.0900
```

```
1  141.      1.      3.     17.      1.   23.08  101.1018
0.0820
1  141.      1.      3.     18.      1.   22.75  101.1119
0.0500
1  141.      1.      3.     21.      1.   23.21  101.1072
0.0641
1  141.      1.      3.     23.      2.   23.25  101.0802
0.0704
1  141.      1.      3.     23.      1.   23.19  101.1350
0.0699
1  141.    281.      3.     16.      1.   22.93  101.0287
0.0520
1  141.    281.      3.     17.      1.   23.00  101.0131
0.0710
1  141.    281.      3.     18.      1.   22.90  101.1329
0.0800
1  141.    281.      3.     22.      1.   23.19  101.0562
0.1594
1  141.    281.      3.     23.      2.   23.18  101.0891
0.1252
1  141.    281.      3.     23.      1.   23.17  101.1283
0.1151
1  141.    283.      3.     16.      1.   22.85  101.1597
0.0990
1  141.    283.      3.     17.      1.   23.09  101.0784
0.0810
1  141.    283.      3.     18.      1.   23.08  101.0715
0.0460
1  141.    283.      3.     21.      1.   23.27  101.0910
0.0880
1  141.    283.      3.     22.      1.   23.34  101.0967
0.0901
1  141.    283.      3.     24.      2.   23.00  101.1627
0.0888
1  141.   2062.      3.     16.      1.   22.97  101.1077
0.0970
1  141.   2062.      3.     17.      1.   22.96  101.0245
0.1210
1  141.   2062.      3.     18.      1.   23.19  100.9650
0.0700
1  141.   2062.      3.     18.      1.   23.18  101.0319
0.1070
1  141.   2062.      3.     22.      1.   23.34  101.0849
0.0960
1  141.   2062.      3.     24.      2.   23.21  101.1302
0.0505
1  141.   2362.      3.     15.      1.   23.08  101.0471
0.0320
1  141.   2362.      3.     17.      1.   23.01  101.0224
0.1020
1  141.   2362.      3.     18.      1.   23.05  101.0702
0.0580
1  141.   2362.      3.     22.      1.   23.22  101.0904
0.1049
1  141.   2362.      3.     23.      2.   23.29  101.0626
0.0702
1  141.   2362.      3.     24.      2.   23.15  101.0686
0.0661
1  142.      1.      3.     15.      1.   23.02   94.3160
0.1372
1  142.      1.      3.     17.      1.   23.04   94.2808
0.0999
1  142.      1.      3.     18.      1.   22.73   94.2478
0.0803
1  142.      1.      3.     21.      1.   23.19   94.2862
0.0700
1  142.      1.      3.     23.      2.   23.25   94.1859
0.0899
1  142.      1.      3.     23.      1.   23.21   94.2389
0.0686
1  142.    281.      3.     16.      1.   22.98   94.2640
0.0862
1  142.    281.      3.     17.      1.   23.00   94.3333
0.1330
1  142.    281.      3.     18.      1.   22.88   94.2994
0.0908
1  142.    281.      3.     21.      1.   23.28   94.2873
0.0846
1  142.    281.      3.     23.      2.   23.07   94.2576
0.0795
1  142.    281.      3.     23.      1.   23.12   94.3027
```

```
0.0389
1  142.    283.      3.     16.      1.   22.92   94.2846
0.1021
1  142.    283.      3.     17.      1.   23.08   94.2197
0.0627
1  142.    283.      3.     18.      1.   23.09   94.2119
0.0785
1  142.    283.      3.     21.      1.   23.29   94.2536
0.0712
1  142.    283.      3.     22.      1.   23.34   94.2280
0.0692
1  142.    283.      3.     24.      2.   22.92   94.2944
0.0958
1  142.   2062.      3.     16.      1.   22.96   94.2238
0.0492
1  142.   2062.      3.     17.      1.   22.95   94.3061
0.2194
1  142.   2062.      3.     18.      1.   23.16   94.1868
0.0474
1  142.   2062.      3.     21.      1.   23.11   94.2645
0.0697
1  142.   2062.      3.     22.      1.   23.31   94.3101
0.0532
1  142.   2062.      3.     24.      2.   23.24   94.2204
0.1023
1  142.   2362.      3.     15.      1.   23.08   94.2437
0.0503
1  142.   2362.      3.     17.      1.   23.00   94.2115
0.0919
1  142.   2362.      3.     18.      1.   22.99   94.2348
0.0282
1  142.   2362.      3.     22.      1.   23.26   94.2124
0.0513
1  142.   2362.      3.     23.      2.   23.27   94.2214
0.0627
1  142.   2362.      3.     24.      2.   23.08   94.1651
0.1010
2  138.      1.      4.     13.      1.   23.12   95.1996
0.0645
2  138.      1.      4.     15.      1.   22.73   95.1315
0.1192
2  138.      1.      4.     18.      2.   22.76   95.1845
0.0452
2  138.      1.      4.     19.      1.   22.73   95.1359
0.1498
2  138.      1.      4.     20.      2.   22.73   95.1435
0.0629
2  138.      1.      4.     21.      2.   22.93   95.1839
0.0563
2  138.    281.      4.     14.      2.   22.46   95.2106
0.1049
2  138.    281.      4.     18.      2.   22.80   95.2505
0.0771
2  138.    281.      4.     18.      2.   22.77   95.2648
0.1046
2  138.    281.      4.     20.      2.   22.80   95.2197
0.1779
2  138.    281.      4.     20.      2.   22.87   95.2003
0.1376
2  138.    281.      4.     21.      2.   22.95   95.0982
0.1611
2  138.    283.      4.     18.      2.   22.83   95.1211
0.0794
2  138.    283.      4.     13.      1.   23.17   95.1327
0.0409
2  138.    283.      4.     18.      1.   22.67   95.2053
0.1525
2  138.    283.      4.     19.      2.   23.00   95.1292
0.0655
2  138.    283.      4.     21.      2.   22.91   95.1669
0.0619
2  138.    283.      4.     21.      2.   22.96   95.1401
0.0831
2  138.   2062.      4.     15.      1.   22.64   95.2479
0.2867
2  138.   2062.      4.     15.      1.   22.67   95.2224
0.1945
2  138.   2062.      4.     19.      2.   22.99   95.2810
0.1960
2  138.   2062.      4.     19.      1.   22.75   95.1869
0.1571
```

```
2  138.   2062.      4.     21.      2.   22.84   95.3053
0.2012
2  138.   2062.      4.     21.      2.   22.92   95.1432
0.1532
2  138.   2362.      4.     12.      1.   22.74   95.1687
0.0785
2  138.   2362.      4.     18.      2.   22.75   95.1564
0.0430
2  138.   2362.      4.     19.      2.   22.88   95.1354
0.0983
2  138.   2362.      4.     19.      1.   22.73   95.0422
0.0773
2  138.   2362.      4.     20.      2.   22.86   95.1354
0.0587
2  138.   2362.      4.     21.      2.   22.94   95.1075
0.0776
2  139.      1.      4.     13.      2.   23.14   99.3274
0.0220
2  139.      1.      4.     15.      2.   22.77   99.5020
0.0997
2  139.      1.      4.     18.      2.   22.80   99.4016
0.0704
2  139.      1.      4.     19.      1.   22.68   99.3181
0.1245
2  139.      1.      4.     20.      2.   22.78   99.3858
0.0903
2  139.      1.      4.     21.      2.   22.93   99.3141
0.0255
2  139.    281.      4.     14.      2.   23.05   99.2915
0.0859
2  139.    281.      4.     15.      2.   22.71   99.4032
0.1322
2  139.    281.      4.     18.      2.   22.79   99.4612
0.1765
2  139.    281.      4.     20.      2.   22.74   99.4001
0.0889
2  139.    281.      4.     20.      2.   22.91   99.3765
0.1041
2  139.    281.      4.     21.      2.   22.92   99.3507
0.0717
2  139.    283.      4.     13.      2.   23.11   99.3848
0.0792
2  139.    283.      4.     18.      2.   22.84   99.4952
0.1122
2  139.    283.      4.     18.      2.   22.76   99.3220
0.0915
2  139.    283.      4.     19.      2.   23.03   99.4165
0.0503
2  139.    283.      4.     21.      2.   22.87   99.3791
0.1138
2  139.    283.      4.     21.      2.   22.98   99.3985
0.0661
2  139.   2062.      4.     14.      2.   22.43   99.4283
0.0891
2  139.   2062.      4.     15.      2.   22.70   99.4139
0.2147
2  139.   2062.      4.     19.      2.   22.97   99.3813
0.1143
2  139.   2062.      4.     19.      1.   22.77   99.4314
0.1685
2  139.   2062.      4.     21.      2.   22.79   99.4166
0.2080
2  139.   2062.      4.     21.      2.   22.94   99.4052
0.2400
2  139.   2362.      4.     12.      1.   22.82   99.3408
0.1279
2  139.   2362.      4.     18.      2.   22.77   99.3116
0.1131
2  139.   2362.      4.     19.      2.   22.82   99.3241
0.0519
2  139.   2362.      4.     19.      1.   22.74   99.2991
0.0903
2  139.   2362.      4.     20.      2.   22.88   99.3049
0.0783
2  139.   2362.      4.     21.      2.   22.94   99.2782
0.0718
2  140.      1.      4.     13.      1.   23.10   96.0811
0.0463
2  140.      1.      4.     15.      2.   22.75   96.1460
0.0725
2  140.      1.      4.     18.      2.   22.78   96.1582
```

```
0.1428
2  140.      1.      4.     19.      1.   22.70   96.1039
0.1056
2  140.      1.      4.     20.      2.   22.75   96.1262
0.0672
2  140.      1.      4.     21.      2.   22.93   96.1478
0.0562
2  140.    281.      4.     15.      2.   22.71   96.1153
0.1097
2  140.    281.      4.     14.      2.   22.49   96.1297
0.1202
2  140.    281.      4.     18.      2.   22.81   96.1233
0.1331
2  140.    281.      4.     20.      2.   22.78   96.1731
0.1484
2  140.    281.      4.     20.      2.   22.89   96.0872
0.0857
2  140.    281.      4.     21.      2.   22.91   96.1331
0.0944
2  140.    283.      4.     13.      2.   23.22   96.1135
0.0983
2  140.    283.      4.     18.      2.   22.85   96.1111
0.1210
2  140.    283.      4.     18.      2.   22.78   96.1221
0.0644
2  140.    283.      4.     19.      2.   23.01   96.1063
0.0921
2  140.    283.      4.     21.      2.   22.91   96.1155
0.0704
2  140.    283.      4.     21.      2.   22.94   96.1308
0.0258
2  140.   2062.      4.     15.      2.   22.60   95.9767
0.2225
2  140.   2062.      4.     15.      2.   22.66   96.1277
0.1792
2  140.   2062.      4.     19.      2.   22.96   96.1858
0.1312
2  140.   2062.      4.     19.      1.   22.75   96.1912
0.1936
2  140.   2062.      4.     21.      2.   22.82   96.1650
0.1902
2  140.   2062.      4.     21.      2.   22.92   96.1603
0.1777
2  140.   2362.      4.     12.      1.   22.88   96.0793
0.0996
2  140.   2362.      4.     18.      2.   22.76   96.1115
0.0533
2  140.   2362.      4.     19.      2.   22.79   96.0803
0.0364
2  140.   2362.      4.     19.      1.   22.71   96.0411
0.0768
2  140.   2362.      4.     20.      2.   22.84   96.0988
0.1042
2  140.   2362.      4.     21.      1.   22.94   96.0482
0.0868
2  141.      1.      4.     13.      1.   23.07  101.1984
0.0803
2  141.      1.      4.     15.      2.   22.72  101.1645
0.0914
2  141.      1.      4.     18.      2.   22.75  101.2454
0.1109
2  141.      1.      4.     19.      1.   22.69  101.1096
0.1376
2  141.      1.      4.     20.      2.   22.83  101.2066
0.0717
2  141.      1.      4.     21.      2.   22.93  101.0645
0.1205
2  141.    281.      4.     15.      2.   22.72  101.1615
0.1272
2  141.    281.      4.     14.      2.   22.40  101.1650
0.0595
2  141.    281.      4.     18.      2.   22.78  101.1815
0.1393
2  141.    281.      4.     20.      2.   22.73  101.1106
0.1189
2  141.    281.      4.     20.      2.   22.86  101.1420
0.0713
2  141.    281.      4.     21.      2.   22.94  101.0116
0.1088
2  141.    283.      4.     13.      2.   23.26  101.1554
0.0429
```

```
2  141.    283.      4.     18.      2.   22.85  101.1267
0.0751
2  141.    283.      4.     18.      2.   22.76  101.1227
0.0826
2  141.    283.      4.     19.      2.   22.82  101.0635
0.1715
2  141.    283.      4.     21.      2.   22.89  101.1264
0.1447
2  141.    283.      4.     21.      2.   22.96  101.0853
0.1189
2  141.   2062.      4.     15.      2.   22.65  101.1332
0.2532
2  141.   2062.      4.     15.      1.   22.68  101.1487
0.1413
2  141.   2062.      4.     19.      2.   22.95  101.1778
0.1772
2  141.   2062.      4.     19.      1.   22.77  101.0988
0.0884
2  141.   2062.      4.     21.      2.   22.87  101.1686
0.2940
2  141.   2062.      4.     21.      2.   22.94  101.3289
0.2072
2  141.   2362.      4.     12.      1.   22.83  101.1353
0.0585
2  141.   2362.      4.     18.      2.   22.83  101.1201
0.0868
2  141.   2362.      4.     19.      2.   22.91  101.0946
0.0855
2  141.   2362.      4.     19.      1.   22.71  100.9977
0.0645
2  141.   2362.      4.     20.      2.   22.87  101.0963
0.0638
2  141.   2362.      4.     21.      2.   22.94  101.0300
0.0549
2  142.      1.      4.     13.      1.   23.07   94.3049
0.1197
2  142.      1.      4.     15.      2.   22.73   94.3153
0.0566
2  142.      1.      4.     18.      2.   22.77   94.3073
0.0875
2  142.      1.      4.     19.      1.   22.67   94.2803
0.0376
2  142.      1.      4.     20.      2.   22.80   94.3008
0.0703
2  142.      1.      4.     21.      2.   22.93   94.2916
0.0604
2  142.    281.      4.     14.      2.   22.90   94.2557
0.0619
2  142.    281.      4.     18.      2.   22.83   94.3542
0.1027
2  142.    281.      4.     18.      2.   22.80   94.3007
0.1492
2  142.    281.      4.     20.      2.   22.76   94.3351
0.1059
2  142.    281.      4.     20.      2.   22.88   94.3406
0.1508
2  142.    281.      4.     21.      2.   22.92   94.2621
0.0946
2  142.    283.      4.     13.      2.   23.25   94.3124
0.0534
2  142.    283.      4.     18.      2.   22.85   94.3680
0.1643
2  142.    283.      4.     18.      1.   22.67   94.3442
0.0346
2  142.    283.      4.     19.      2.   22.80   94.3391
0.0616
2  142.    283.      4.     21.      2.   22.91   94.2238
0.0721
2  142.    283.      4.     21.      2.   22.95   94.2721
0.0998
2  142.   2062.      4.     14.      2.   22.49   94.2915
0.2189
2  142.   2062.      4.     15.      2.   22.69   94.2803
0.0690
2  142.   2062.      4.     19.      2.   22.94   94.2818
0.0987
2  142.   2062.      4.     19.      1.   22.76   94.2227
0.2628
2  142.   2062.      4.     21.      2.   22.74   94.4109
0.1230
2  142.   2062.      4.     21.      2.   22.94   94.2616
```

```
0.0929
2  142.   2362.      4.     12.      1.   22.86   94.2052
0.0813
2  142.   2362.      4.     18.      2.   22.83   94.2824
0.0605
2  142.   2362.      4.     19.      2.   22.85   94.2396
0.0882
2  142.   2362.      4.     19.      1.   22.75   94.2087
0.0702
2  142.   2362.      4.     20.      2.   22.86   94.2937
0.0591
2  142.   2362.      4.     21.      1.   22.93   94.2330
0.0556
```

2.6.3.1.2. Measurements on wiring configurations

Check wafers were measured with the probe wired in two configurations

Measurements of resistivity (ohm.cm) were made according to an ASTM Standard Test Method (F4) to identify differences between 2 wiring configurations for probe #2362. The check standards for the study were five wafers selected at random from a batch of wafers cut from one silicon crystal doped with phosphorous to give a nominal resistivity of 100 ohm.cm.

Description of database

The data are averages of $K = 6$ days' measurements and $J = 6$ repetitions at the center of each wafer. There are $L = 2$ complete runs, separated by two months time, on each wafer.

The data recorded in the 10 columns are:

1. Wafer
2. Probe
3. Average - configuration A; run 1
4. Standard deviation - configuration A; run 1
5. Average - configuration B; run 1
6. Standard deviation - configuration B; run 1
7. Average - configuration A; run 2
8. Standard deviation - configuration A; run 2
9. Average - configuration B; run 2
10. Standard deviation - configuration B; run 2

```
Wafer Probe Config A-run1   Config B-run1   Config A-run2
Config B-run2.

138. 2362.  95.1162 0.0480  95.0993 0.0466  95.1687 0.0785
95.1589 0.0642
138. 2362.  95.0569 0.0577  95.0657 0.0450  95.1564 0.0430
95.1705 0.0730
138. 2362.  95.0598 0.0516  95.0622 0.0664  95.1354 0.0983
95.1221 0.0695
138. 2362.  95.1487 0.0386  95.1625 0.0311  95.0422 0.0773
95.0513 0.0840
138. 2362.  95.0743 0.0256  95.0599 0.0488  95.1354 0.0587
95.1531 0.0482
138. 2362.  95.1010 0.0420  95.0944 0.0393  95.1075 0.0776
95.1537 0.0230
139. 2362.  99.2939 0.0818  99.3018 0.0905  99.3408 0.1279
99.3637 0.1025
139. 2362.  99.3234 0.0723  99.3488 0.0350  99.3116 0.1131
99.3881 0.0451
```

```
139. 2362.  99.2748 0.0756  99.3571 0.1993  99.3241 0.0519
99.3737 0.0699
139. 2362.  99.3512 0.0475  99.3512 0.1286  99.2991 0.0903
99.3066 0.0709
139. 2362.  99.2350 0.0517  99.2255 0.0738  99.3049 0.0783
99.3040 0.0744
139. 2362.  99.3574 0.0485  99.3605 0.0459  99.2782 0.0718
99.3680 0.0470
140. 2362.  96.0771 0.1024  96.0915 0.1257  96.0793 0.0996
96.1041 0.0890
140. 2362.  95.9976 0.0943  96.0057 0.0806  96.1115 0.0533
96.0774 0.0983
140. 2362.  96.0148 0.0622  96.0244 0.0833  96.0803 0.0364
96.1004 0.0758
140. 2362.  96.0397 0.0702  96.0422 0.0738  96.0411 0.0768
96.0677 0.0663
140. 2362.  96.0407 0.0627  96.0738 0.0800  96.0988 0.1042
96.0585 0.0960
140. 2362.  96.0445 0.0622  96.0557 0.1129  96.0482 0.0868
96.0062 0.0895
141. 2362. 101.0471 0.0320 101.0241 0.0670 101.1353 0.0585
101.1156 0.1027
141. 2362. 101.0224 0.1020 101.0660 0.1030 101.1201 0.0868
101.1077 0.1141
141. 2362. 101.0702 0.0580 101.0509 0.0710 101.0946 0.0855
101.0455 0.1070
141. 2362. 101.0904 0.1049 101.0983 0.0894 100.9977 0.0645
101.0274 0.0666
141. 2362. 101.0626 0.0702 101.0614 0.0849 101.0963 0.0638
101.1106 0.0788
141. 2362. 101.0686 0.0661 101.0811 0.0490 101.0300 0.0549
101.1073 0.0663
142. 2362.  94.2437 0.0503  94.2088 0.0815  94.2052 0.0813
94.2487 0.0719
142. 2362.  94.2115 0.0919  94.2043 0.1176  94.2824 0.0605
94.2886 0.0499
142. 2362.  94.2348 0.0282  94.2324 0.0519  94.2396 0.0882
94.2739 0.1075
142. 2362.  94.2124 0.0513  94.2347 0.0694  94.2087 0.0702
94.2023 0.0416
142. 2362.  94.2214 0.0627  94.2416 0.0757  94.2937 0.0591
94.2600 0.0731
142. 2362.  94.1651 0.1010  94.2287 0.0919  94.2330 0.0556
94.2406 0.0651
```

2.6.3.2. Analysis and interpretation

Purpose of this page

The purpose of this page is to outline an analysis of data taken during a gauge study to quantify the type A uncertainty component for resistivity (ohm.cm) measurements on silicon wafers made with a gauge that was part of the initial study.

Summary of standard deviations at three levels

The level-1, level-2, and level-3 standard deviations for the uncertainty analysis are summarized in the table below from the gauge case study.

Standard deviations for probe #2362

Level	Symbol	Estimate	DF
Level-1	s_1	0.0729	300
Level-2	s_2	0.0362	50
Level-3	s_3	0.0197	5

Calculation of individual components for days and runs

The standard deviation that estimates the day effect is

$$s_{days} = \sqrt{s_2^2 - \frac{1}{6}s_1^2} = 0.0217 \ \ ohm.cm$$

The standard deviation that estimates the run effect is

$$s_{runs} = \sqrt{s_3^2 - \frac{1}{6}s_2^2} = 0.0130 \ \ ohm.cm$$

Calculation of the standard deviation of the certified value showing sensitivity coefficients

The certified value for each wafer is the average of $N = 6$ repeatability measurements at the center of the wafer on $M = 1$ days and over $P = 1$ runs. Notice that N, M and P are not necessarily the same as the number of measurements in the gauge study per wafer; namely, J, K and L. The standard deviation of a certified value (for time-dependent sources of error), is

$$s = \sqrt{s_{runs}^2 + s_{days}^2 - \frac{1}{6}s_1^2}$$

Standard deviations for days and runs are included in this calculation, even though there were no replications over days or runs for the certification measurements. These factors contribute

to the overall uncertainty of the measurement process even though they are not sampled for the particular measurements of interest.

The equation must be rewritten to calculate degrees of freedom

Degrees of freedom cannot be calculated from the equation above because the calculations for the individual components involve differences among variances. The table of sensitivity coefficients for a 3-level design shows that for

$$N = J,\ M = 1,\ P = 1$$

the equation above can be rewritten in the form

$$s = \sqrt{\frac{5}{6}s_2^2 + s_3^2}$$

Then the degrees of freedom can be approximated using the Welch-Satterthwaite method.

Probe bias - Graphs of probe biases

A graphical analysis shows the relative biases among the 5 probes. For each wafer, differences from the wafer average by probe are plotted versus wafer number. The graphs verify that probe #2362 (coded as 5) is biased low relative to the other probes. The bias shows up more strongly after the probes have been in use (run 2).

How to deal with bias due to the probe

Probe #2362 was chosen for the certification process because of its superior precision, but its bias relative to the other probes creates a problem. There are two possibilities for handling this problem:

1. Correct all measurements made with probe #2362 to the average of the probes.
2. Include the standard deviation for the difference among probes in the uncertainty budget.

The best strategy, as followed in the certification process, is to correct all measurements for the average bias of probe #2362 and take the standard deviation of the correction as a type A component of uncertainty.

Correction for bias or probe #2362 and uncertainty

Biases by probe and wafer are shown in the gauge case study. Biases for probe #2362 are summarized in table below for the two runs. The correction is taken to be the negative of the average bias. The standard deviation of the correction is the standard deviation of the average of the ten biases.

Estimated biases for probe #2362

Wafer	Probe	Run 1	Run 2	All
138	2362	-0.0372	-0.0507	
139	2362	-0.0094	-0.0657	
140	2362	-0.0261	-0.0398	
141	2362	-0.0252	-0.0534	

```
142   2362   -0.0383   -0.0469

Average      -0.0272   -0.0513   -0.0393
Standard deviation                0.0162
(10 values)
```

Configurations Database and plot of differences

Measurements on the check wafers were made with the probe wired in two different configurations (A, B). A plot of differences between configuration A and configuration B shows no bias between the two configurations.

Test for difference between configurations

This finding is consistent over runs 1 and 2 and is confirmed by the t-statistics in the table below where the average differences and standard deviations are computed from 6 days of measurements on 5 wafers. A t-statistic < 2 indicates no significant difference. The conclusion is that there is no bias due to wiring configuration and no contribution to uncertainty from this source.

Differences between configurations

Status	Average	Std dev	DF	t
Pre	-0.00858	0.0242	29	1.9
Post	-0.0110	0.0354	29	1.7

Error budget showing sensitivity coefficients, standard deviations and degrees of freedom

The error budget showing sensitivity coefficients for computing the standard uncertainty and degrees of freedom is outlined below.

Error budget for resistivity (ohm.cm)

Source	Type	Sensitivity	Standard Deviation	DF
Repeatability	A	$a_1 = 0$	0.0729	300
Reproducibility	A	$a_2 = \sqrt{5/6}$	0.0362	50
Run-to-run	A	$a_3 = 1$	0.0197	5
Probe #2362	A	$a_4 = \sqrt{1/10}$	0.0162	5
Wiring Configuration A	A	$a_5 = 1$	0	--

Standard uncertainty includes components for repeatability, days, runs and probe

The standard uncertainty is computed from the error budget as

$$u = \sqrt{\sum_i a_i^2 s_i^2} = \sqrt{\frac{5}{6}s_2^2 + s_3^2 + \frac{1}{10}s_{probe}^2} = 0.0388 ohm.cm$$

Approximate

The degrees of freedom associated with *u* are approximated by

degrees of freedom and expanded uncertainty

the Welch-Satterthwaite formula as:

$$\nu = \frac{u^4}{\sum_{i=1}^{5} \frac{a_i^4 s_i^4}{\nu_i}} = 42$$

where the ν_i are the degrees of freedom given in the rightmost column of the table.

The critical value at the 0.05 significance level with 42 degrees of freedom, from the t-table, is 2.018 so the expanded uncertainty is

$$U = 2.018\ u = 0.078\ ohm.cm$$

2.6.3.2.1. Difference between 2 wiring configurations

Measurements with the probe configured in two ways

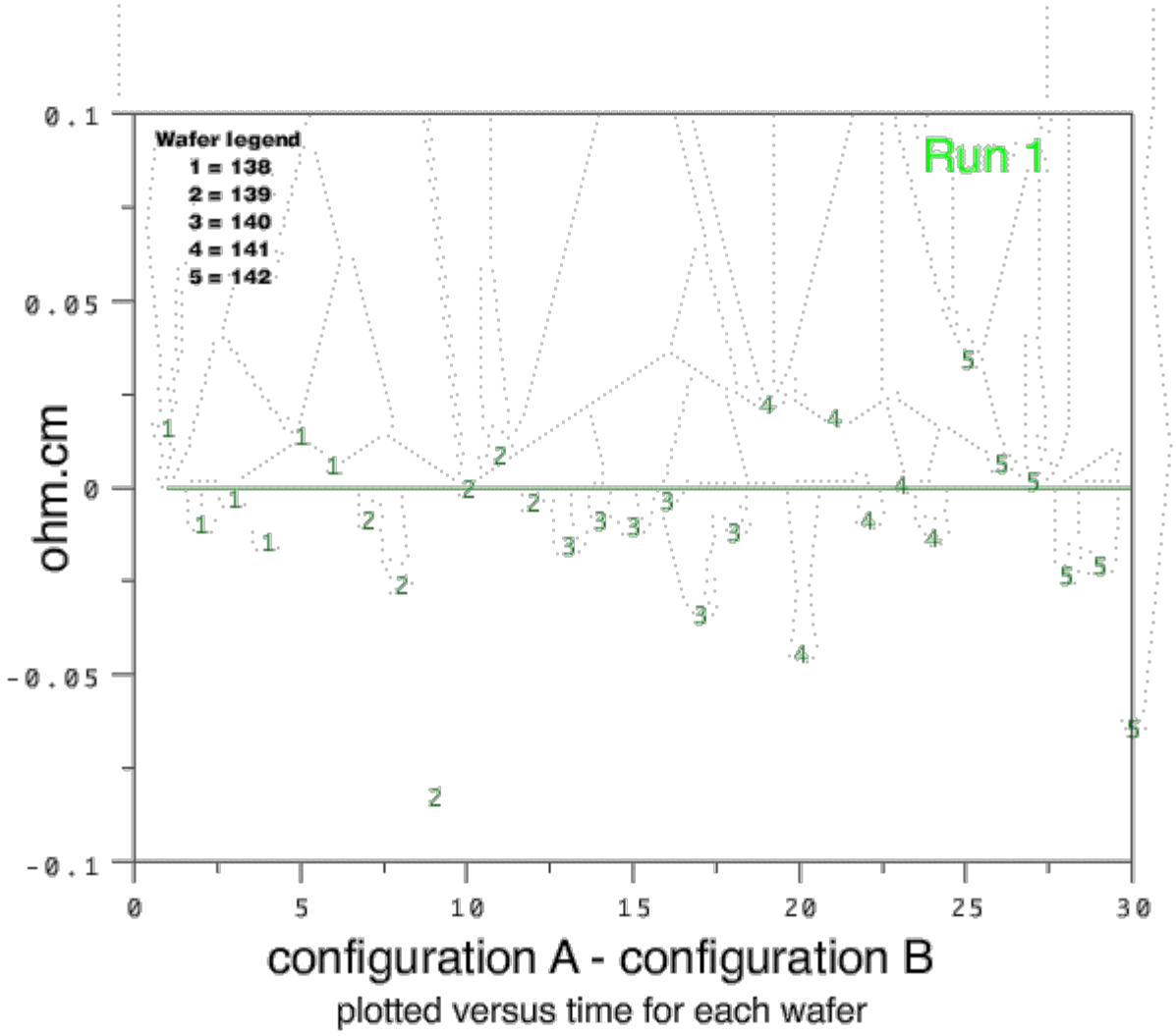
The graphs below are constructed from resistivity measurements (ohm.cm) on five wafers where the probe (#2362) was wired in two different configurations, A and B. The probe is a 4-point probe with many possible wiring configurations. For this experiment, only two configurations were tested as a means of identifying large discrepancies.

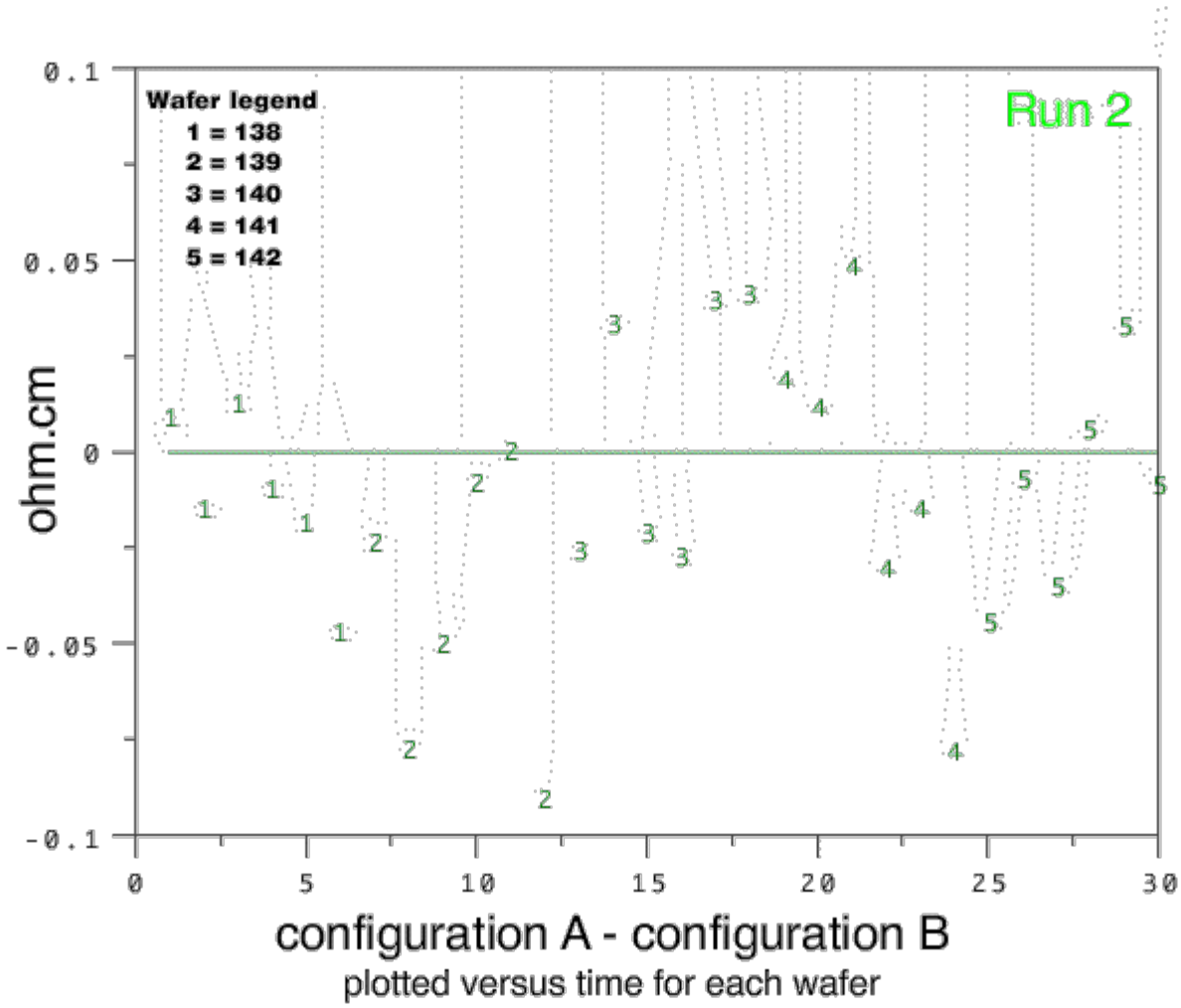
Artifacts for the study

The five wafers; namely, #138, #139, #140, #141, and #142 are coded 1, 2, 3, 4, 5, respectively, in the graphs. These wafers were chosen at random from a batch of approximately 100 wafers that were being certified for resistivity.

Interpretation

Differences between measurements in configurations A and B, made on the same day, are plotted over six days for each wafer. The two graphs represent two runs separated by approximately two months time. The dotted line in the center is the zero line. The pattern of data points scatters fairly randomly above and below the zero line -- indicating no difference between configurations for probe #2362. The conclusion applies to probe #2362 and cannot be extended to all probes of this type.


Wafer legend
1 = 138
2 = 139
3 = 140
4 = 141
5 = 142
Run 1
0.1
0.05
0
-0.05
-0.1
ohm.cm
0
5
10
15
20
25
30
configuration A - configuration B
plotted versus time for each wafer


Wafer legend
1 = 138
2 = 139
3 = 140
4 = 141
5 = 142
Run 2
ohm.cm
0.1
0.05
0
-0.05
-0.1
0
5
10
15
20
25
30
configuration A - configuration B
plotted versus time for each wafer

2.6.3.3. Run the type A uncertainty analysis using Dataplot

View of Dataplot macros for this case study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output Window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
Time-dependent components from 3-level nested design Pool repeatability standard deviations for: 1. Run 1 2. Run 2 Compute level-2 standard deviations for: 3. Run 1 4. Run 2 5. Pool level-2 standard deviations	*Database of measurements with probe #2362* 1. The repeatability standard deviation is 0.0658 ohm.cm for run 1 and 0.0758 ohm.cm for run 2. This represents the basic precision of the measuring instrument. 2. The level-2 standard deviation pooled over 5 wafers and 2 runs is 0.0362 ohm.cm. This is significant in the calculation of uncertainty. 3. The level-3 standard deviation pooled over 5 wafers is 0.0197 ohm.cm. This is small compared to the other components but is

6. Compute level-3 standard deviations	included in the uncertainty calculation for completeness.
Bias due to probe #2362 1. Plot biases for 5 NIST probes 2. Compute wafer bias and average bias for probe #2362 3. Correction for bias and standard deviation	*Database of measurements with 5 probes* 1. The plot shows that probe #2362 is biased low relative to the other probes and that this bias is consistent over 5 wafers. 2. The bias correction is the average bias = 0.0393 ohm.cm over the 5 wafers. The correction is to be subtracted from all measurements made with probe #2362. 3. The uncertainty of the bias correction = 0.0051 ohm.cm is computed from the standard deviation of the biases for the 5 wafers.
Bias due to wiring configuration A 1. Plot differences between wiring configurations 2. Averages, standard deviations and t-statistics	*Database of wiring configurations A and B* 1. The plot of measurements in wiring configurations A and B shows no difference between A and B. 2. The statistical test confirms that there is no difference between the wiring configurations.
Uncertainty 1. Standard uncertainty, df, t-value and expanded uncertainty	*Elements of error budget* 1. The uncertainty is computed from the error budget. The uncertainty for an average of 6 measurements on one day with probe #2362 is 0.078 with 42 degrees of freedom.

2.6.3.4. Dataplot macros

Reads data and plots the repeatability standard deviations for probe #2362 and pools standard deviations over days, wafers -- run 1

```
reset data
reset plot control
reset i/o
dimension 500 rows
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr    subset probe =
2362
let df = sr - sr + 5.
y1label ohm.cm
characters * all
lines blank all
x2label Repeatability standard deviations for
probe 2362 - run 1
plot sr subset run 1
let var = sr*sr
let df11 = sum df subset run 1
let s11 = sum var subset run 1
. repeatability standard deviation for run 1
let s11 = (5.*s11/df11)**(1/2)
print s11 df11
. end of calculations
```

Reads data and plots repeatability standard deviations for probe #2362 and pools standard deviations over days, wafers -- run 2

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr subset probe 2362
let df = sr - sr + 5.
y1label ohm.cm
characters * all
lines blank all
x2label Repeatability standard deviations for
probe 2362 - run 2
plot sr subset run 2
let var = sr*sr
let df11 = sum df subset run 1
let df12 = sum df subset run 2
let s11 = sum var subset run 1
let s12 = sum var subset run 2
let s11 = (5.*s11/df11)**(1/2)
let s12 = (5.*s12/df12)**(1/2)
print s11 df11
print s12 df12
let s1 = ((s11**2 + s12**2)/2.)**(1/2)
let df1=df11+df12
. repeatability standard deviation and df for
run 2
print s1 df1
. end of calculations
```

Computes level-2 standard deviations

```
reset data
reset plot control
reset i/o
dimension 500 rows
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
```

from daily averages and pools over wafers -- run 1

```
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr subset probe 2362
sd plot y wafer subset run 1
let s21 = yplot
let wafer1 = xplot
retain s21 wafer1 subset tagplot = 1
let nwaf = size s21
let df21 = 5 for i = 1 1 nwaf
. level-2 standard deviations and df for 5
wafers - run 1
print wafer1 s21 df21
. end of calculations
```

Computes level-2 standard deviations from daily averages and pools over wafers -- run 2

```
reset data
reset plot control
reset i/o
dimension 500 rows
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr subset probe 2362
sd plot y wafer subset run 2
let s22 = yplot
let wafer1 = xplot
retain s22 wafer1 subset tagplot = 1
let nwaf = size s22
let df22 = 5 for i = 1 1 nwaf
. level-2 standard deviations and df for 5
wafers - run 1
print wafer1 s22 df22
. end of calculations
```

Pools level-2 standard deviations over wafers and runs

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr subset probe 2362
sd plot y wafer subset run 1
let s21 = yplot
let wafer1 = xplot
sd plot y wafer subset run 2
let s22 = yplot
retain s21 s22 wafer1 subset tagplot = 1
let nwaf = size wafer1
let df21 = 5 for i = 1 1 nwaf
let df22 = 5 for i = 1 1 nwaf
let s2a = (s21**2)/5 + (s22**2)/5
let s2 = sum s2a
let s2 = sqrt(s2/2)
let df2a = df21 + df22
let df2 = sum df2a
. pooled level-2 standard deviation and df
across wafers and runs
print s2 df2
. end of calculations
```

Computes level-3standard deviations from run averages and pools over wafers

```
reset data
reset plot control
reset i/o
dimension 500 rows
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
retain run wafer probe y sr subset probe 2362
.
mean plot y wafer subset run 1
let m31 = yplot
let wafer1 = xplot
mean plot y wafer subset run 2
let m32 = yplot
retain m31 m32 wafer1 subset tagplot = 1
let nwaf = size m31
let s31 =(((m31-m32)**2)/2.)**(1/2)
let df31 = 1 for i = 1 1 nwaf
. level-3 standard deviations and df for 5
wafers
```

```
print wafer1 s31 df31
let s31 = (s31**2)/5
let s3 = sum s31
let s3 = sqrt(s3)
let df3=sum df31
. pooled level-3 std deviation and df over 5
wafers
print s3 df3
. end of calculations
```

Plot differences from the average wafer value for each probe showing bias for probe #2362

```
reset data
reset plot control
reset i/o
dimension 500 30
read mpc61a.dat wafer probe d1 d2
let biasrun1 = mean d1 subset probe 2362
let biasrun2 = mean d2 subset probe 2362
print biasrun1 biasrun2
title GAUGE STUDY FOR 5 PROBES
Y1LABEL OHM.CM
lines dotted dotted dotted dotted dotted solid
characters 1 2 3 4 5 blank
xlimits 137 143
let zero = pattern 0 for I = 1 1 30
x1label DIFFERENCES AMONG PROBES VS WAFER (RUN
1)
plot d1 wafer probe and
plot zero wafer
let biasrun2 = mean d2 subset probe 2362
print biasrun2
title GAUGE STUDY FOR 5 PROBES
Y1LABEL OHM.CM
lines dotted dotted dotted dotted dotted solid
characters 1 2 3 4 5 blank
xlimits 137 143
let zero = pattern 0 for I = 1 1 30
x1label DIFFERENCES AMONG PROBES VS WAFER (RUN
2)
plot d2 wafer probe and
plot zero wafer
. end of calculations
```

Compute bias for probe #2362 by wafer

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
set read format
.
cross tabulate mean y run wafer
retain run wafer probe y sr subset probe 2362
skip 1
read dpst1f.dat runid wafid ybar
print runid wafid ybar
let ngroups = size ybar
skip 0
.
let m3 = y - y
feedback off
loop for k = 1 1 ngroups
    let runa = runid(k)
    let wafera = wafid(k)
    let ytemp = ybar(k)
    let m3 = ytemp subset run = runa subset
wafer = wafera
end of loop
feedback on
.
let d = y - m3
let bias1 = average d subset run 1
let bias2 = average d subset run 2
.
mean plot d wafer subset run 1
let b1 = yplot
let wafer1 = xplot
mean plot d wafer subset run 2
let b2 = yplot
```

```
retain b1 b2 wafer1 subset tagplot = 1
let nwaf = size b1
. biases for run 1 and run 2 by wafers
print wafer1 b1 b2
. average biases over wafers for run 1 and run 2
print bias1 bias2
. end of calculations
```

Compute correction for bias for measurements with probe #2362 and the standard deviation of the correction

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
set read format f1.0,f6.0,f8.0,32x,f10.4,f10.4
read mpc633a.dat run wafer probe y sr
set read format
.
cross tabulate mean y run wafer
retain run wafer probe y sr subset probe 2362
skip 1
read dpst1f.dat runid wafid ybar
let ngroups = size ybar
skip 0
.
let m3 = y - y
feedback off
loop for k = 1 1 ngroups
    let runa = runid(k)
    let wafera = wafid(k)
    let ytemp = ybar(k)
    let m3 = ytemp subset run = runa subset
wafer = wafera
end of loop
feedback on
.
let d = y - m3
let bias1 = average d subset run 1
let bias2 = average d subset run 2
.
mean plot d wafer subset run 1
let b1 = yplot
let wafer1 = xplot
mean plot d wafer subset run 2
let b2 = yplot
retain b1 b2 wafer1 subset tagplot = 1
.
extend b1 b2
let sd = standard deviation b1
let sdcorr = sd/(10**(1/2))
let correct = -(bias1+bias2)/2.
. correction for probe #2362, standard dev, and
standard dev of corr
print correct sd sdcorr
. end of calculations
```

Plot differences between wiring configurations A and B

```
reset data
reset plot control
reset i/o
dimension 500 30
label size 3
read mpc633k.dat wafer probe a1 s1 b1 s2 a2 s3
b2 s4
let diff1 = a1 - b1
let diff2 = a2 - b2
let t = sequence 1 1 30
lines blank all
characters 1 2 3 4 5
y1label ohm.cm
x1label Config A - Config B -- Run 1
x2label over 6 days and 5 wafers
x3label legend for wafers 138, 139, 140, 141,
142: 1, 2, 3, 4, 5
plot diff1 t wafer
x1label Config A - Config B -- Run 2
plot diff2 t wafer
. end of calculations
```

```
reset data
```

Compute average differences between configuration A and B; standard deviations and t-statistics for testing significance

```
reset plot control
reset i/o
separator character @
dimension 500 rows
label size 3
read mpc633k.dat wafer probe a1 s1 b1 s2 a2 s3
b2 s4
let diff1 = a1 - b1
let diff2 = a2 - b2
let d1 = average diff1
let d2 = average diff2
let s1 = standard deviation diff1
let s2 = standard deviation diff2
let t1 = (30.)**(1/2)*(d1/s1)
let t2 = (30.)**(1/2)*(d2/s2)
. Average config A-config B; std dev difference;
t-statistic for run 1
print d1 s1 t1
. Average config A-config B; std dev difference;
t-statistic for run 2
print d2 s2 t2
separator character ;
. end of calculations
```

Compute standard uncertainty, effective degrees of freedom, t value and expanded uncertainty

```
reset data
reset plot control
reset i/o
dimension 500 rows
label size 3
read mpc633m.dat sz a df
let c = a*sz*sz
let d = c*c
let e = d/(df)
let sume = sum e
let u = sum c
let u = u**(1/2)
let effdf=(u**4)/sume
let tvalue=tppf(.975,effdf)
let expu=tvalue*u
.
. uncertainty, effective degrees of freedom,
tvalue and
. expanded uncertainty
print u effdf tvalue expu
. end of calculations
```

2.6.4. Evaluation of type B uncertainty and propagation of error

Focus of this case study

The purpose of this case study is to demonstrate uncertainty analysis using statistical techniques coupled with type B analyses and propagation of error. It is a continuation of the case study of type A uncertainties.

Background - description of measurements and constraints

The measurements in question are volume resistivities (ohm.cm) of silicon wafers which have the following definition:

$$\rho = Xo \cdot K_a \cdot F_t \cdot t \cdot F_{t/s}$$

with explanations of the quantities and their nominal values shown below:

ρ = *resistivity* = 0.00128 *ohm.cm*
X = *voltage/current* (*ohm*)
t = $thickness_{wafer}$(*cm*) = 0.628 *cm*
K_a = $factor_{electrical}$ = 4.50 *ohm.cm*
F_F = $correction_{temp} \approx 1^{\circ} C$
$F_{t/s}$ = $factor_{thickness/separation} \approx 1.0$

Type A evaluations

The resistivity measurements, discussed in the case study of type A evaluations, were replicated to cover the following sources of uncertainty in the measurement process, and the associated uncertainties are reported in units of resistivity (ohm.cm).

- Repeatability of measurements at the center of the wafer
- Day-to-day effects
- Run-to-run effects
- Bias due to probe #2362
- Bias due to wiring configuration

Need for propagation of error

Not all factors could be replicated during the gauge experiment. Wafer thickness and measurements required for the scale corrections were measured off-line. Thus, the type B evaluation of uncertainty is computed using propagation of error. The propagation of error formula in units of resistivity is as follows:

$$s_\rho = \rho\sqrt{\frac{s_Y^2}{X^2} + \frac{s_l^2}{t^2} + \frac{s_{Ka}^2}{K_a^2} + \frac{s_{FT}^2}{F_T} + \frac{s_{F(t/S)}^2}{F_{t/S}}}$$

Standard deviations for type B evaluations

Standard deviations for the type B components are summarized here. For a complete explanation, see the publication (Ehrstein and Croarkin).

Electrical measurements

There are two basic sources of uncertainty for the electrical measurements. The first is the least-count of the digital volt meter in the measurement of X with a maximum bound of

$$a = 0.0000534 \text{ ohm}$$

which is assumed to be the half-width of a uniform distribution. The second is the uncertainty of the electrical scale factor. This has two sources of uncertainty:

1. error in the solution of the transcendental equation for determining the factor
2. errors in measured voltages

The maximum bounds to these errors are assumed to be half-widths of

$$a = 0.0001 \text{ ohm.cm} \text{ and } a = 0.00038 \text{ ohm.cm}$$

respectively, from uniform distributions. The corresponding standard deviations are shown below.

$$s_X = 0.0000534/\sqrt{3} = 0.0000308 \text{ ohm}$$

$$s_{Ks} = \sqrt{\frac{0.0001^2}{3} + \frac{0.00038^2}{3}} = 0.000227 \text{ ohm.cm}$$

Thickness

The standard deviation for thickness, t, accounts for two sources of uncertainty:

1. calibration of the thickness measuring tool with precision gauge blocks
2. variation in thicknesses of the silicon wafers

The maximum bounds to these errors are assumed to be half-widths of

$$a = 0.000015 \text{ cm} \text{ and } a = 0.000001 \text{ cm}$$

respectively, from uniform distributions. Thus, the standard deviation for thickness is

$$s_t = \sqrt{\frac{.000015^2}{3} + \frac{.000001^2}{3}} = .00000868 \ cm$$

Temperature correction

The standard deviation for the temperature correction is calculated from its defining equation as shown below. Thus, the standard deviation for the correction is the standard deviation associated with the measurement of temperature multiplied by the temperature coefficient, $C(t) = 0.0083$. The maximum bound to the error of the temperature measurement is assumed to be the half-width

$$a = 0.13 \ ^\circ C$$

of a triangular distribution. Thus the standard deviation of the correction for

$$F_T = 1 - C_T(T - 23^\circ C)$$

is

$$s_{F(T)} = C_T s_T = 0.0083\sqrt{\frac{0.13^2}{6}} = 0.000441^\circ C$$

Thickness scale factor

The standard deviation for the thickness scale factor is negligible.

Associated sensitivity coefficients

Sensitivity coefficients for translating the standard deviations for the type B components into units of resistivity (ohm.cm) from the propagation of error equation are listed below and in the error budget. The sensitivity coefficient for a source is the multiplicative factor associated with the standard deviation in the formula above; i.e., the partial derivative with respect to that variable from the propagation of error equation.

$a_6 = (\rho/X) = 100/0.111 = 900.901$
$a_7 = (\rho/K_a) = 100/4.50 = 22.222$
$a_8 = (\rho/t) = 100/0.628 = 159.24$
$a_9 = (\rho/F_T) = 100$
$a_{10} = (\rho/F_{t/S}) = 100$

Sensitivity coefficients and degrees of freedom

Sensitivity coefficients for the type A components are shown in the case study of type A uncertainty analysis and repeated below. Degrees of freedom for type B uncertainties based on assumed distributions, according to the convention, are assumed to be infinite.

Error budget showing

The error budget showing sensitivity coefficients for computing the relative standard uncertainty of volume resistivity (ohm.cm)

sensitivity coefficients, standard deviations and degrees of freedom

with degrees of freedom is outlined below.

Error budget for volume resistivity (ohm.cm)

Source	Type	Sensitivity	Standard Deviation	DF
Repeatability	A	$a_1 = 0$	0.0729	300
Reproducibility	A	$a_2 = \sqrt{5/6}$	0.0362	50
Run-to-run	A	$a_3 = 1$	0.0197	5
Probe #2362	A	$a_4 = \sqrt{1/10}$	0.0162	5
Wiring Configuration A	A	$a_5 = 1$	0	--
Resistance ratio	B	$a_6 = 900.901$	0.0000308	∞
Electrical scale	B	$a_7 = 22.222$	0.000227	∞
Thickness	B	$a_8 = 159.20$	0.00000868	∞
Temperature correction	B	$a_9 = 100$	0.000441	∞
Thickness scale	B	$a_{10} = 100$	0	--

Standard uncertainty

The standard uncertainty is computed as:

$$u = \sqrt{\sum_{i=1}^{10} a_i^2 s_i^2} = 0.065 \ \ ohm.cm$$

Approximate degrees of freedom and expanded uncertainty

The degrees of freedom associated with *u* are approximated by the Welch-Satterthwaite formula as:

$$\nu = \frac{u^4}{\sum_{i=1}^{5} \frac{a_i^4 s_i^4}{\nu_i}} = 42$$

This calculation is not affected by components with infinite degrees of freedom, and therefore, the degrees of freedom for the standard uncertainty is the same as the degrees of freedom for the type A uncertainty. The critical value at the 0.05 significance level with 42 degrees of freedom, from the t-table, is 2.018 so the expanded uncertainty is

$$U = 2.018 \ u = 0.13 \ ohm.cm$$

2.7. References

Degrees of freedom	K. A. Brownlee (1960). ***Statistical Theory and Methodology in Science and Engineering,*** John Wiley & Sons, Inc., New York, p. 236.
Calibration designs	J. M. Cameron, M. C. Croarkin and R. C. Raybold (1977). ***Designs for the Calibration of Standards of Mass,*** NBS Technical Note 952, U.S. Dept. Commerce, 58 pages.
Calibration designs for eliminating drift	J. M. Cameron and G. E. Hailes (1974). ***Designs for the Calibration of Small Groups of Standards in the Presence of Drift***, Technical Note 844, U.S. Dept. Commerce, 31 pages.
Measurement assurance for measurements on ICs	Carroll Croarkin and Ruth Varner (1982). ***Measurement Assurance for Dimensional Measurements on Integrated-circuit Photomasks,*** NBS Technical Note 1164, U.S. Dept. Commerce, 44 pages.
Calibration designs for gauge blocks	Ted Doiron (1993). ***Drift Eliminating Designs for Non-Simultaneous Comparison Calibrations***, *J Research National Institute of Standards and Technology, 98,* pp. 217-224.
Type A & B uncertainty analyses for resistivities	J. R. Ehrstein and M. C. Croarkin (1998). ***Standard Reference Materials: The Certification of 100 mm Diameter Silicon Resistivity SRMs 2541 through 2547 Using Dual-Configuration Four-Point Probe Measurements,*** NIST Special Publication 260-131, Revised, 84 pages.
Calibration designs for electrical standards	W. G. Eicke and J. M. Cameron (1967). ***Designs for Surveillance of the Volt Maintained By a Group of Saturated Standard Cells,*** NBS Technical Note 430, U.S. Dept. Commerce 19 pages.
Theory of uncertainty analysis	Churchill Eisenhart (1962). ***Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems****J Research National Bureau of Standards-C. Engineering and Instrumentation, Vol. 67C, No.2*, p. 161-187.
Confidence, prediction, and	Gerald J. Hahn and William Q. Meeker (1991). ***Statistical Intervals: A Guide for Practitioners***, John Wiley & Sons,

tolerance intervals	Inc., New York.
Original calibration designs for weighings	J. A. Hayford (1893). ***On the Least Square Adjustment of Weighings,*** U.S. Coast and Geodetic Survey Appendix 10, Report for 1892.
Uncertainties for values from a calibration curve	Thomas E. Hockersmith and Harry H. Ku (1993). ***Uncertainties associated with proving ring calibrations***, NBS Special Publication 300: Precision Measurement and Calibration, Statistical Concepts and Procedures, Vol. 1, pp. 257-263, H. H. Ku, editor.
EWMA control charts	J. Stuart Hunter (1986). ***The Exponentially Weighted Moving Average***, *J Quality Technology, Vol. 18, No. 4*, pp. 203-207.
Fundamentals of mass metrology	K. B. Jaeger and R. S. Davis (1984). ***A Primer for Mass Metrology***, NBS Special Publication 700-1, 85 pages.
Fundamentals of propagation of error	Harry Ku (1966). ***Notes on the Use of Propagation of Error Formulas,*** *J Research of National Bureau of Standards-C. Engineering and Instrumentation, Vol. 70C, No.4, pp. 263-273.*
Handbook of statistical methods	Mary Gibbons Natrella (1963). ***Experimental Statistics,*** NBS Handbook 91, US Deptartment of Commerce.
Omnitab	Sally T. Peavy, Shirley G. Bremer, Ruth N. Varner, David Hogben (1986). ***OMNITAB 80: An Interpretive System for Statistical and Numerical Data Analysis,*** NBS Special Publication 701, US Deptartment of Commerce.
Uncertainties for uncorrected bias	Steve D. Phillips and Keith R. Eberhardt (1997). ***Guidelines for Expressing the Uncertainty of Measurement Results Containing Uncorrected Bias***, *NIST Journal of Research, Vol. 102, No. 5*.
Calibration of roundness artifacts	Charles P. Reeve (1979). ***Calibration designs for roundness standards***, NBSIR 79-1758, 21 pages.
Calibration designs for angle blocks	Charles P. Reeve (1967). ***The Calibration of Angle Blocks by Comparison***, NBSIR 80-19767, 24 pages.
SI units	Barry N. Taylor (1991). ***Interpretation of the SI for the United States and Metric Conversion Policy for Federal Agencies,*** NIST Special Publication 841, U.S.

Deptartment of Commerce.

Uncertainties for calibrated values — Raymond Turgel and Dominic Vecchia (1987). ***Precision Calibration of Phase Meters,*** *IEEE Transactions on Instrumentation and Measurement, Vol. IM-36, No. 4.*, pp. 918-922.

Example of propagation of error for flow measurements — James R. Whetstone et al. (1989). ***Measurements of Coefficients of Discharge for Concentric Flange-Tapped Square-Edged Orifice Meters in Water Over the Reynolds Number Range 600 to 2,700,000***, NIST Technical Note 1264. pp. 97.

Mathematica software — Stephen Wolfram (1993). ***Mathematica, A System of Doing Mathematics by Computer, 2nd edition,*** Addison-Wesley Publishing Co., New York.

Restrained least squares — Marvin Zelen (1962). ***"Linear Estimation and Related Topics"*** in *Survey of Numerical Analysis* edited by John Todd, McGraw-Hill Book Co. Inc., New York, pp. 558-577.

ASTM F84 for resistivity — ***ASTM Method F84-93, Standard Test Method for Measuring Resistivity of Silicon Wafers With an In-line Four-Point Probe.*** Annual Book of ASTM Standards, 10.05, West Conshohocken, PA 19428.

ASTM E691 for interlaboratory testing — ***ASTM Method E691-92, Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method.*** Annual Book of ASTM Standards, 10.05, West Conshohocken, PA 19428.

Guide to uncertainty analysis — ***Guide to the Expression of Uncertainty of Measurement*** (1993). ISBN 91-67-10188-9, 1st ed. ISO, Case postale 56, CH-1211, Genève 20, Switzerland, 101 pages.

ISO 5725 for interlaboratory testing — ***ISO 5725: 1997. Accuracy (trueness and precision) of measurement results, Part 2: Basic method for repeatability and reproducibility of a standard measurement method,*** ISO, Case postale 56, CH-1211, Genève 20, Switzerland.

ISO 11095 on linear calibration — ***ISO 11095: 1997. Linear Calibration using Reference Materials,*** ISO, Case postale 56, CH-1211, Genève 20, Switzerland.

MSA gauge studies manual — ***Measurement Systems Analysis Reference Manual, 2nd ed.,*** (1995). Chrysler Corp., Ford Motor Corp., General Motors Corp., 120 pages.

NCSL RP on uncertainty — ***Determining and Reporting Measurement Uncertainties,*** National Conference of Standards Laboratories RP-12,

analysis	(1994), Suite 305B, 1800 30th St., Boulder, CO 80301.
ISO Vocabulary for metrology	***International Vocabulary of Basic and General Terms in Metrology, 2nd ed.,*** (1993). ISO, Case postale 56, CH-1211, Genève 20, Switzerland, 59 pages.
Exact variance for length and width	Leo Goodman (1960). ***"On the Exact Variance of Products"*** in *Journal of the American Statistical Association*, December, 1960, pp. 708-713.

3. Production Process Characterization

The goal of this chapter is to learn how to plan and conduct a *Production Process Characterization Study* (PPC) on manufacturing processes. We will learn how to model manufacturing processes and use these models to design a data collection scheme and to guide data analysis activities. We will look in detail at how to analyze the data collected in characterization studies and how to interpret and report the results. The accompanying Case Studies provide detailed examples of several process characterization studies.

1. Introduction
   1. Definition
   2. Uses
   3. Terminology/Concepts
   4. PPC Steps

2. Assumptions
   1. General Assumptions
   2. Specific PPC Models

3. Data Collection
   1. Set Goals
   2. Model the Process
   3. Define Sampling Plan

4. Analysis
   1. First Steps
   2. Exploring Relationships
   3. Model Building
   4. Variance Components
   5. Process Stability
   6. Process Capability
   7. Checking Assumptions

5. Case Studies
   1. Furnace Case Study
   2. Machine Case Study

Detailed Chapter Table of Contents

References

3. Production Process Characterization - Detailed Table of Contents [3.]

1. Introduction to Production Process Characterization [3.1.]
    1. What is PPC? [3.1.1.]
    2. What are PPC Studies Used For? [3.1.2.]
    3. Terminology/Concepts [3.1.3.]
        1. Distribution (Location, Spread and Shape) [3.1.3.1.]
        2. Process Variability [3.1.3.2.]
            1. Controlled/Uncontrolled Variation [3.1.3.2.1.]
        3. Propagating Error [3.1.3.3.]
        4. Populations and Sampling [3.1.3.4.]
        5. Process Models [3.1.3.5.]
        6. Experiments and Experimental Design [3.1.3.6.]
    4. PPC Steps [3.1.4.]

2. Assumptions / Prerequisites [3.2.]
    1. General Assumptions [3.2.1.]
    2. Continuous Linear Model [3.2.2.]
    3. Analysis of Variance Models (ANOVA) [3.2.3.]
        1. One-Way ANOVA [3.2.3.1.]
            1. One-Way Value-Splitting [3.2.3.1.1.]
        2. Two-Way Crossed ANOVA [3.2.3.2.]
            1. Two-way Crossed Value-Splitting Example [3.2.3.2.1.]
        3. Two-Way Nested ANOVA [3.2.3.3.]
            1. Two-Way Nested Value-Splitting Example [3.2.3.3.1.]
    4. Discrete Models [3.2.4.]

3. Data Collection for PPC [3.3.]
    1. Define Goals [3.3.1.]
    2. Process Modeling [3.3.2.]
    3. Define Sampling Plan [3.3.3.]
        1. Identifying Parameters, Ranges and Resolution [3.3.3.1.]
        2. Choosing a Sampling Scheme [3.3.3.2.]
        3. Selecting Sample Sizes [3.3.3.3.]
        4. Data Storage and Retrieval [3.3.3.4.]
        5. Assign Roles and Responsibilities [3.3.3.5.]

4. Data Analysis for PPC [3.4.]
    1. First Steps [3.4.1.]
    2. Exploring Relationships [3.4.2.]
        1. Response Correlations [3.4.2.1.]
        2. Exploring Main Effects [3.4.2.2.]

3. Exploring First Order Interactions [3.4.2.3.]

3. Building Models [3.4.3.]
   1. Fitting Polynomial Models [3.4.3.1.]
   2. Fitting Physical Models [3.4.3.2.]
4. Analyzing Variance Structure [3.4.4.]
5. Assessing Process Stability [3.4.5.]
6. Assessing Process Capability [3.4.6.]
7. Checking Assumptions [3.4.7.]

5. Case Studies [3.5.]
   1. Furnace Case Study [3.5.1.]
      1. Background and Data [3.5.1.1.]
      2. Initial Analysis of Response Variable [3.5.1.2.]
      3. Identify Sources of Variation [3.5.1.3.]
      4. Analysis of Variance [3.5.1.4.]
      5. Final Conclusions [3.5.1.5.]
      6. Work This Example Yourself [3.5.1.6.]
   2. Machine Screw Case Study [3.5.2.]
      1. Background and Data [3.5.2.1.]
      2. Box Plots by Factors [3.5.2.2.]
      3. Analysis of Variance [3.5.2.3.]
      4. Throughput [3.5.2.4.]
      5. Final Conclusions [3.5.2.5.]
      6. Work This Example Yourself [3.5.2.6.]

6. References [3.6.]

3. Production Process Characterization

3.1. Introduction to Production Process Characterization

Overview Section

The goal of this section is to provide an introduction to PPC. We will define PPC and the terminology used and discuss some of the possible uses of a PPC study. Finally, we will look at the steps involved in designing and executing a PPC study.

Contents: Section 1

1. What is PPC?
2. What are PPC studies used for?
3. What terminology is used in PPC?
    1. Location, Spread and Shape
    2. Process Variability
    3. Propagating Error
    4. Populations and Sampling
    5. Process Models
    6. Experiments and Experimental Design
4. What are the steps of a PPC?
    1. Plan PPC
    2. Collect Data
    3. Analyze and Interpret Data
    4. Report Conclusions

3.1.1. What is PPC?

In PPC, we build data-based models

Process characterization is an activity in which we:

- identify the key inputs and outputs of a process
- collect data on their behavior over the entire operating range
- estimate the steady-state behavior at optimal operating conditions
- and build models describing the parameter relationships across the operating range

The result of this activity is a set of mathematical process models that we can use to monitor and improve the process.

This is a three-step process

This activity is typically a three-step process.

The Screening Step

In this phase we identify all possible significant process inputs and outputs and conduct a series of screening experiments in order to reduce that list to the key inputs and outputs. These experiments will also allow us to develop initial models of the relationships between those inputs and outputs.

The Mapping Step

In this step we map the behavior of the key outputs over their expected operating ranges. We do this through a series of more detailed experiments called Response Surface experiments.

The Passive Step

In this step we allow the process to run at nominal conditions and estimate the process stability and capability.

Not all of the steps need to be performed

The first two steps are only needed for new processes or when the process has undergone some significant engineering change. There are, however, many times throughout the life of a process when the third step is needed. Examples might be: initial process qualification, control chart development, after minor process adjustments, after schedule equipment maintenance, etc.

3.1.2. What are PPC Studies Used For?

PPC is the core of any CI program

Process characterization is an integral part of any continuous improvement program. There are many steps in that program for which process characterization is required. These might include:

When process characterization is required

- when we are bringing a new process or tool into use.
- when we are bringing a tool or process back up after scheduled/unscheduled maintenance.
- when we want to compare tools or processes.
- when we want to check the health of our process during the monitoring phase.
- when we are troubleshooting a bad process.

The techniques described in this chapter are equally applicable to the other chapters covered in this Handbook. These include:

Process characterization techniques are applicable in other areas

- calibration
- process monitoring
- process improvement
- process/product comparison
- reliability

3.1.3. Terminology/Concepts

There are just a few fundamental concepts needed for PPC. This section will review these ideas briefly and provide links to other sections in the Handbook where they are covered in more detail.

Distribution(location, spread, shape)

For basic data analysis, we will need to understand how to estimate location, spread and shape from the data. These three measures comprise what is known as the ***distribution*** of the data. We will look at both graphical and numerical techniques.

Process variability

We need to thoroughly understand the concept of process variability. This includes how variation explains the possible range of expected data values, the various classifications of variability, and the role that variability plays in process stability and capability.

Error propagation

We also need to understand how variation propagates through our manufacturing processes and how to decompose the total observed variation into components attributable to the contributing sources.

Populations and sampling

It is important to have an understanding of the various issues related to sampling. We will define a *population* and discuss how to acquire representative random samples from the population of interest. We will also discuss a useful formula for estimating the number of observations required to answer specific questions.

Modeling

For modeling, we will need to know how to identify important factors and responses. We will also need to know how to graphically and quantitatively build models of the relationships between the factors and responses.

Experiments

Finally, we will need to know about the basics of designed experiments including screening designs and response surface designs so that we can quantify these relationships. This topic will receive

only a cursory treatment in this chapter. It is covered in detail in the process improvement chapter. However, examples of its use are in the case studies.

3. Production Process Characterization
3.1. Introduction to Production Process Characterization
3.1.3. Terminology/Concepts

3.1.3.1. Distribution (Location, Spread and Shape)

Distributions are characterized by location, spread and shape

A fundamental concept in representing any of the outputs from a production process is that of a *distribution*. Distributions arise because any manufacturing process output will not yield the same value every time it is measured. There will be a natural scattering of the measured values about some central tendency value. This scattering about a central value is known as a distribution. A distribution is characterized by three values:

Location
: The location is the expected value of the output being measured. For a stable process, this is the value around which the process has stabilized.

Spread
: The spread is the expected amount of variation associated with the output. This tells us the range of possible values that we would expect to see.

Shape
: The shape shows how the variation is distributed about the location. This tells us if our variation is symmetric about the mean or if it is skewed or possibly multimodal.

A primary goal of PPC is to estimate the distributions of the process outputs

One of the primary goals of a PPC study is to characterize our process outputs in terms of these three measurements. If we can demonstrate that our process is stabilized about a constant location, with a constant variance and a known stable shape, then we have a process that is both predictable and controllable. This is required before we can set up control charts or conduct experiments.

Click on each item to read more detail

The table below shows the most common numerical and graphical measures of location, spread and shape.

Parameter	Numerical	Graphical
Location	mean median	scatter plot boxplot

		histogram
Spread	variance range inter-quartile range	boxplot histogram
Shape	skewness kurtosis	boxplot histogram probability plot

3.1.3.2. Process Variability

Variability is present everywhere

All manufacturing and measurement processes exhibit variation. For example, when we take sample data on the output of a process, such as critical dimensions, oxide thickness, or resistivity, we observe that all the values are *NOT* the same. This results in a collection of observed values distributed about some location value. This is what we call spread or variability. We represent variability numerically with the variance calculation and graphically with a histogram.

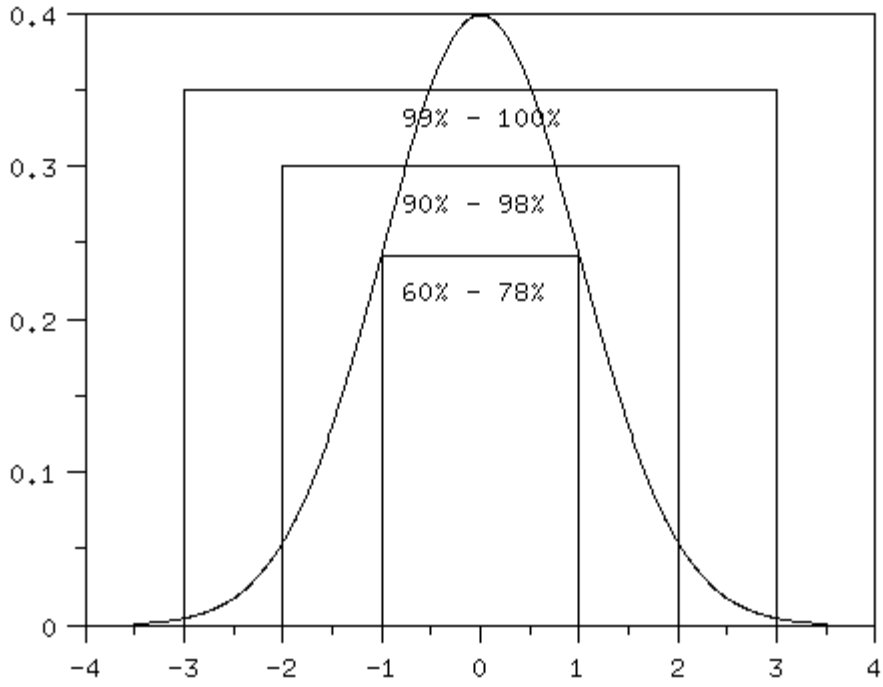
How does the standard deviation describe the spread of the data?

The standard deviation (square root of the variance) gives insight into the spread of the data through the use of what is known as the *Empirical Rule*. This rule (shown in the graph below) is:

Approximately 60-78% of the data are within a distance of one standard deviation from the average $(\bar{x} - s, \bar{x} + s)$.

Approximately 90-98% of the data are within a distance of two standard deviations from the average $(\bar{x} - 2s, \bar{x} + 2s)$.

More than 99% of the data are within a distance of three standard deviations from the average $(\bar{x} - 3s, \bar{x} + 3s)$



Variability accumulates from many sources

This observed variability is an accumulation of many different sources of variation that have occurred throughout the manufacturing process. One of the more important activities of process characterization is to identify and quantify these various sources of variation so that they may be minimized.

There are also different types

There are not only different sources of variation, but there are also different *types* of variation. Two important classifications of variation for the purposes of PPC are *controlled variation* and *uncontrolled variation*.

Click here to see examples

CONTROLLED VARIATION

Variation that is characterized by a *stable* and consistent pattern of variation over time. This type of variation will be *random* in nature and will be exhibited by a uniform fluctuation about a constant level.

UNCONTROLLED VARIATION

Variation that is characterized by a pattern of variation that *changes* over time and hence is unpredictable. This type of variation will typically contain some structure.

Stable processes only exhibit controlled variation

This concept of controlled/uncontrolled variation is important in determining if a process is *stable*. A process is deemed stable if it runs in a consistent and predictable manner. This means that the average process value is constant and the variability is controlled. If the variation is uncontrolled, then either the process average is changing or the process variation is changing or both. The first process in the example above is stable; the second is not.

In the course of process characterization we should endeavor to eliminate all sources

of uncontrolled variation.

3. Production Process Characterization
3.1. Introduction to Production Process Characterization
3.1.3. Terminology/Concepts
3.1.3.2. Process Variability

3.1.3.2.1. Controlled/Uncontrolled Variation

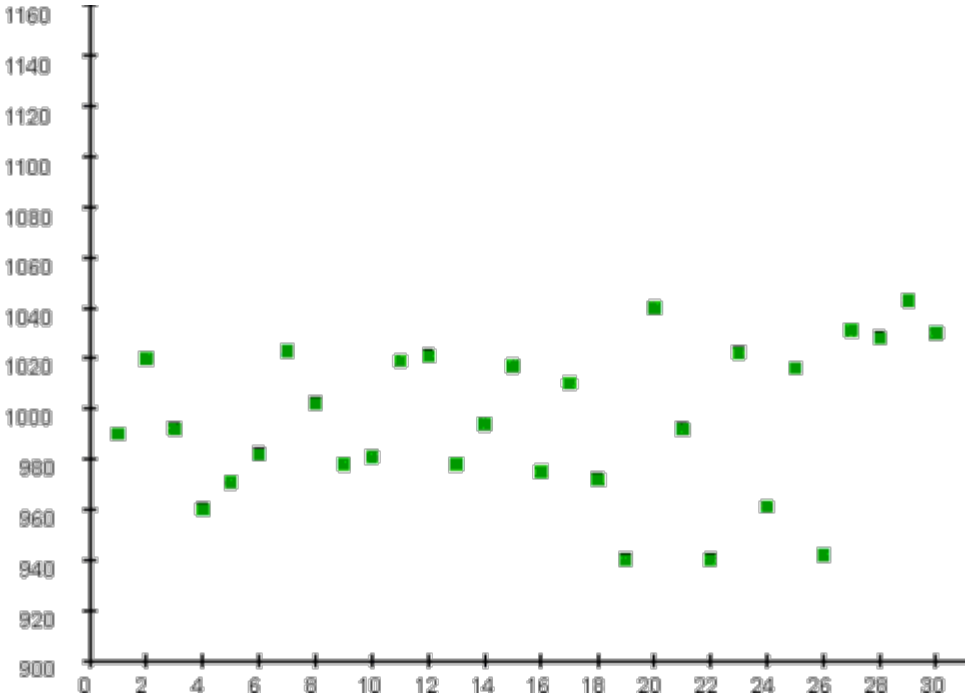
Two trend plots

The two figures below are two trend plots from two different oxide growth processes. Thirty wafers were sampled from each process: one per day over 30 days. Thickness at the center was measured on each wafer. The *x-axis* of each graph is the wafer number and the *y-axis* is the film thickness in angstroms.

Examples of"in control" and "out of control" processes

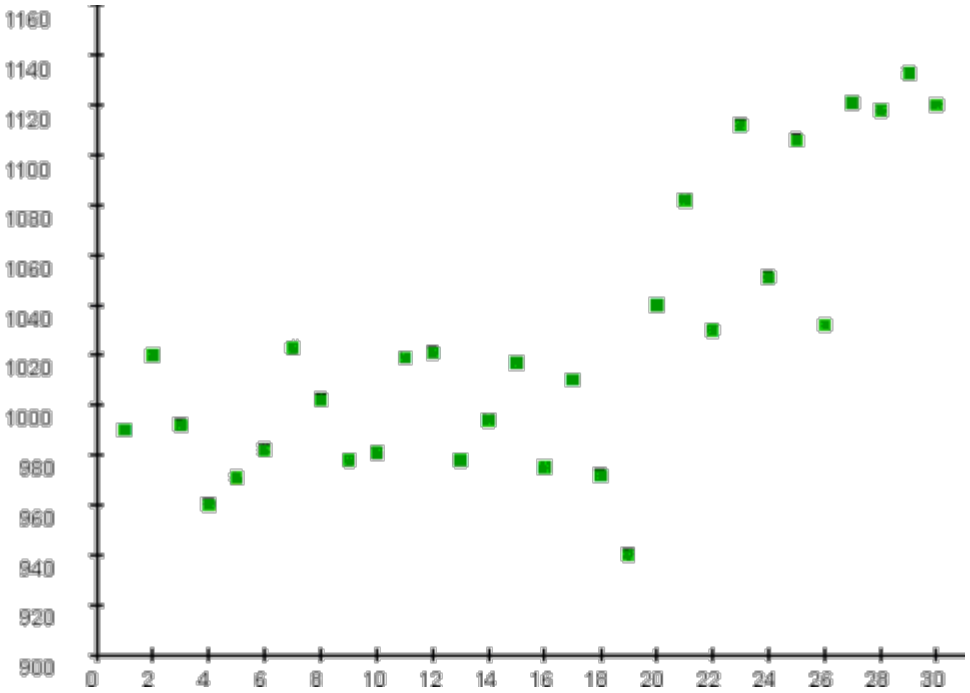
The first process is an example of a process that is "in control" with random fluctuation about a process location of approximately 990. The second process is an example of a process that is "out of control" with a process location trending upward after observation 20.

This process exhibits controlled variation. Note the random fluctuation about a constant mean.



This process exhibits uncontrolled variation. Note the structure in the

variation in the form of a linear trend.



NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

3. Production Process Characterization
3.1. Introduction to Production Process Characterization
3.1.3. Terminology/Concepts

3.1.3.3. Propagating Error

The variation we see can come from many sources

When we estimate the variance at a particular process step, this variance is typically not just a result of the current step, but rather is an accumulation of variation from previous steps and from measurement error. Therefore, an important question that we need to answer in PPC is how the variation from the different sources accumulates. This will allow us to partition the total variation and assign the parts to the various sources. Then we can attack the sources that contribute the most.

How do I partition the error?

Usually we can model the contribution of the various sources of error to the total error through a simple linear relationship. If we have a simple linear relationship between two variables, say,

$$y = \mu + \alpha y_1 + \beta y_2 + \epsilon$$

then the variance associated with, *y*, is given by,

$$\mathrm{Var}(y) = \alpha^2 \mathrm{Var}(y_1) + \beta^2 \mathrm{Var}(y_2) + 2\alpha\beta \mathrm{Cov}(y_1, y_2)$$

.

If the variables are not correlated, then there is no covariance and the last term in the above equation drops off. A good example of this is the case in which we have both process error and measurement error. Since these are usually independent of each other, the total observed variance is just the sum of the variances for process and measurement. Remember to never add standard deviations, we must add variances.

How do I calculate the individual components?

Of course, we rarely have the individual components of variation and wish to know the total variation. Usually, we have an estimate of the overall variance and wish to break that variance down into its individual components. This is known as *components of variance* estimation and is dealt with in detail in the analysis of variance page later in this chapter.

3. Production Process Characterization
3.1. Introduction to Production Process Characterization
3.1.3. Terminology/Concepts

3.1.3.4. Populations and Sampling

We take samples from a target population and make inferences

In survey sampling, if you want to know what everyone thinks about a particular topic, you can just ask everyone and record their answers. Depending on how you define the term, **everyone** (all the adults in a town, all the males in the USA, etc.), it may be impossible or impractical to survey everyone. The other option is to survey a small group (Sample) of the people whose opinions you are interested in (Target Population) , record their opinions and use that information to make inferences about what everyone thinks. Opinion pollsters have developed a whole body of tools for doing just that and many of those tools apply to manufacturing as well. We can use these sampling techniques to take a few measurements from a process and make statements about the behavior of that process.

Facts about a sample are not necessarily facts about a population

If it weren't for process variation we could just take one sample and everything would be known about the target population. Unfortunately this is never the case. We cannot take facts about the sample to be facts about the population. Our job is to reach appropriate conclusions about the population despite this variation. The more observations we take from a population, the more our sample data resembles the population. When we have reached the point at which facts about the sample are reasonable approximations of facts about the population, then we say the sample is adequate.

Four attributes of samples

Adequacy of a sample depends on the following four attributes:

- Representativeness of the sample (is it random?)
- Size of the sample
- Variability in the population
- Desired precision of the estimates

We will learn about choosing representative samples of adequate size in the section on defining sampling plans.

3.1.3.5. Process Models

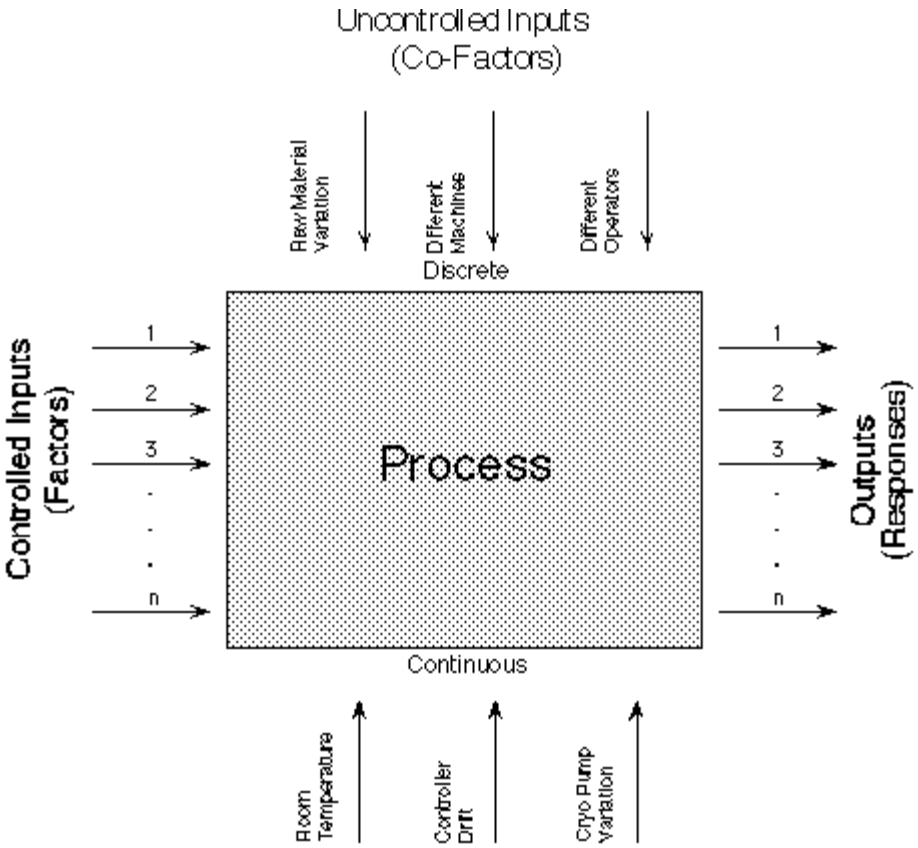
Black box model and fishbone diagram

As we will see in Section 3 of this chapter, one of the first steps in PPC is to model the process that is under investigation. Two very useful tools for doing this are the ***black-box*** model and the ***fishbone diagram***.

We use the black-box model to describe our processes

We can use the simple *black-box* model, shown below, to describe most of the tools and processes we will encounter in PPC. The process will be stimulated by inputs. These inputs can either be controlled (such as recipe or machine settings) or uncontrolled (such as humidity, operators, power fluctuations, etc.). These inputs interact with our process and produce outputs. These outputs are usually some characteristic of our process that we can measure. The measurable inputs and outputs can be sampled in order to observe and understand how they behave and relate to each other.

Diagram of the black box model



These inputs and outputs are also known as Factors and Responses, respectively.

Factors
: Observed inputs used to explain response behavior (also called explanatory variables). Factors may be fixed-level controlled inputs or sampled uncontrolled inputs.

Responses
: Sampled process outputs. Responses may also be functions of sampled outputs such as average thickness or uniformity.

Factors and Responses are further classified by variable type

We further categorize factors and responses according to their *Variable Type*, which indicates the amount of information they contain. As the name implies, this classification is useful for data modeling activities and is critical for selecting the proper analysis technique. The table below summarizes this categorization. The types are listed in order of the amount of information they contain with *Measurement* containing the most information and *Nominal* containing the least.

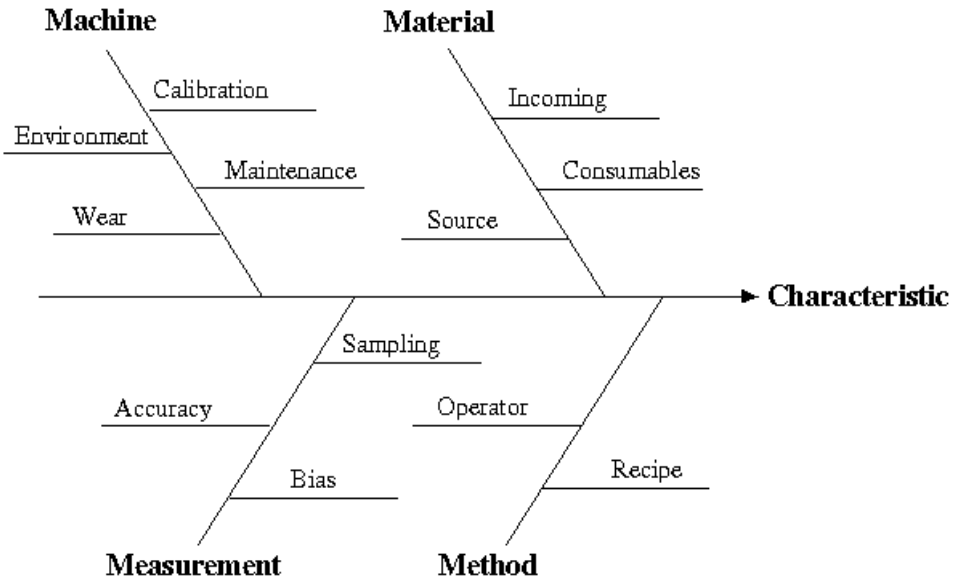
Table describing the different variable types

Type	Description	Example
Measurement	discrete/continuous, order is important, infinite range	particle count, oxide thickness, pressure, temperature
Ordinal	discrete, order is important, finite range	run #, wafer #, site, bin
Nominal	discrete, no order, very few possible values	good/bad, bin, high/medium/low, shift, operator

Fishbone diagrams help to decompose complexity

We can use the fishbone diagram to further refine the modeling process. Fishbone diagrams are very useful for decomposing the complexity of our manufacturing processes. Typically, we choose a process characteristic (either Factors or Responses) and list out the general categories that may influence the characteristic (such as material, machine method, environment, etc.), and then provide more specific detail within each category. Examples of how to do this are given in the section on Case Studies.

Sample fishbone diagram


Machine
Material
Calibration
Environment
Maintenance
Wear
Incoming
Consumables
Source
Characteristic
Sampling
Accuracy
Bias
Operator
Recipe
Measurement
Method

3.1.3.6. Experiments and Experimental Design

Factors and responses

Besides just observing our processes for evidence of stability and capability, we quite often want to know about the relationships between the various **Factors** and **Responses**.

We look for correlations and causal relationships

There are generally two types of relationships that we are interested in for purposes of PPC. They are:

Correlation
: Two variables are said to be correlated if an observed change in the level of one variable is accompanied by a change in the level of another variable. The change may be in the same direction (positive correlation) or in the opposite direction (negative correlation).

Causality
: There is a causal relationship between two variables if a change in the level of one variable causes a change in the other variable.

Note that correlation does not imply causality. It is possible for two variables to be associated with each other without one of them causing the observed behavior in the other. When this is the case it is usually because there is a third (possibly unknown) causal factor.

Our goal is to find causal relationships

Generally, our ultimate goal in PPC is to find and quantify causal relationships. Once this is done, we can then take advantage of these relationships to improve and control our processes.

Find correlations and then try to establish causal relationships

Generally, we first need to find and explore correlations and then try to establish causal relationships. It is much easier to find correlations as these are just properties of the data. It is much more difficult to prove causality as this additionally requires sound engineering judgment. There is a systematic procedure we can use to accomplish this in an efficient manner. We do this through the use of designed experiments.

First we screen, then we build

When we have many potential factors and we want to see which ones are correlated and have the potential to be involved in causal relationships with the responses, we use

models	screening designs to reduce the number of candidates. Once we have a reduced set of influential factors, we can use response surface designs to model the causal relationships with the responses across the operating range of the process factors.
Techniques discussed in process improvement chapter	The techniques are covered in detail in the process improvement section and will not be discussed much in this chapter. Examples of how the techniques are used in PPC are given in the Case Studies.

3.1.4. PPC Steps

Follow these 4 steps to ensure efficient use of resources

The primary activity of a PPC is to collect and analyze data so that we may draw conclusions about and ultimately improve our production processes. In many industrial applications, access to production facilities for the purposes of conducting experiments is very limited. Thus we must be very careful in how we go about these activities so that we can be sure of doing them in a cost-effective manner.

Step 1: Plan

The most important step by far is the planning step. By faithfully executing this step, we will ensure that we only collect data in the most efficient manner possible and still support the goals of the PPC. Planning should generate the following:

- a statement of the goals
- a descriptive process model (a list of process inputs and outputs)
- a description of the sampling plan (including a description of the procedure and settings to be used to run the process during the study with clear assignments for each person involved)
- a description of the method of data collection, tasks and responsibilities, formatting, and storage
- an outline of the data analysis

All decisions that affect how the characterization will be conducted should be made during the planning phase. The process characterization should be conducted according to this plan, with all exceptions noted.

Step 2: Collect

Data collection is essentially just the execution of the sampling plan part of the previous step. If a good job were done in the planning step, then this step should be pretty straightforward. It is important to execute to the plan as closely as possible and to note any exceptions.

Step 3: Analyze and interpret

This is the combination of quantitative (regression, ANOVA, correlation, etc.) and graphical (histograms, scatter plots, box plots, etc.) analysis techniques that are applied to the collected data in order to accomplish the goals of the PPC.

Step 4:

Reporting is an important step that should not be overlooked.

Report	By creating an informative report and archiving it in an accessible place, we can ensure that others have access to the information generated by the PPC. Often, the work involved in a PPC can be minimized by using the results of other, similar studies. Examples of PPC reports can be found in the Case Studies section.
Further information	The planning and data collection steps are described in detail in the data collection section. The analysis and interpretation steps are covered in detail in the analysis section. Examples of the reporting step can be seen in the Case Studies.

3. Production Process Characterization

3.2. Assumptions / Prerequisites

Primary goal is to identify and quantify sources of variation

The primary goal of PPC is to identify and quantify sources of variation. Only by doing this will we be able to define an effective plan for variation reduction and process improvement. Sometimes, in order to achieve this goal, we must first build mathematical/statistical models of our processes. In these models we will identify influential factors and the responses on which they have an effect. We will use these models to understand how the sources of variation are influenced by the important factors. This subsection will review many of the modeling tools we have at our disposal to accomplish these tasks. In particular, the models covered in this section are linear models, Analysis of Variance (ANOVA) models and discrete models.

Contents: Section 2

1. General Assumptions
2. Continuous Linear
3. Analysis of Variance
    1. One-Way
    2. Crossed
    3. Nested
4. Discrete

3. Production Process Characterization
3.2. Assumptions / Prerequisites

3.2.1. General Assumptions

Assumption: process is sum of a systematic component and a random component

In order to employ the modeling techniques described in this section, there are a few assumptions about the process under study that must be made. First, we must assume that the process can adequately be modeled as the sum of a systematic component and a random component. The systematic component is the mathematical model part and the random component is the error or noise present in the system. We also assume that the systematic component is fixed over the range of operating conditions and that the random component has a constant location, spread and distributional form.

Assumption: data used to fit these models are representative of the process being modeled

Finally, we assume that the data used to fit these models are representative of the process being modeled. As a result, we must additionally assume that the measurement system used to collect the data has been studied and proven to be capable of making measurements to the desired precision and accuracy. If this is not the case, refer to the Measurement Capability Section of this Handbook.

3.2.2. Continuous Linear Model

Description The continuous linear model (CLM) is probably the most commonly used model in PPC. It is applicable in many instances ranging from simple control charts to response surface models.

The CLM is a mathematical function that relates explanatory variables (either discrete or continuous) to a single continuous response variable. It is called linear because the coefficients of the terms are expressed as a linear sum. The terms themselves do not have to be linear.

Model The general form of the CLM is:

$$y = a_0 + \sum_{i=1}^{p} a_i f(x_i) + e$$

This equation just says that if we have *p* explanatory variables then the response is modeled by a constant term plus a sum of functions of those explanatory variables, plus some random error term. This will become clear as we look at some examples below.

Estimation The coefficients for the parameters in the CLM are estimated by the method of least squares. This is a method that gives estimates which minimize the sum of the squared distances from the observations to the fitted line or plane. See the chapter on Process Modeling for a more complete discussion on estimating the coefficients for these models.

Testing The tests for the CLM involve testing that the model as a whole is a good representation of the process and whether any of the coefficients in the model are zero or have no effect on the overall fit. Again, the details for testing are given in the chapter on Process Modeling.

Assumptions For estimation purposes, there are no additional assumptions necessary for the CLM beyond those stated in the assumptions section. For testing purposes, however, it is necessary to assume that the error term is adequately modeled by a Gaussian distribution.

Uses The CLM has many uses such as building predictive process models over a range of process settings that exhibit linear behavior, control charts, process capability, building models from

the data produced by designed experiments, and building response surface models for automated process control applications.

Examples

Shewhart Control Chart - The simplest example of a very common usage of the CLM is the underlying model used for Shewhart control charts. This model assumes that the process parameter being measured is a constant with additive Gaussian noise and is given by:

$$y = a_0 + e$$

Diffusion Furnace - Suppose we want to model the average wafer sheet resistance as a function of the location or zone in a furnace tube, the temperature, and the anneal time. In this case, let there be 3 distinct zones (front, center, back) and temperature and time are continuous explanatory variables. This model is given by the CLM:

$$y = a_0 + \begin{cases} a_1 & \text{if front} \\ a_2 + a_4 temp + a_5 time + e & \text{if center} \\ a_3 & \text{if back} \end{cases}$$

Diffusion Furnace (cont.) - Usually, the fitted line for the average wafer sheet resistance is not straight but has some curvature to it. This can be accommodated by adding a quadratic term for the time parameter as follows:

$$y = a_0 + \begin{cases} a_1 & \text{if front} \\ a_2 + a_4 temp + a_5 time + a_6 time^2 e & \text{if center} \\ a_3 & \text{if back} \end{cases}$$

3.2.3. Analysis of Variance Models (ANOVA)

ANOVA allows us to compare the effects of multiple levels of multiple factors

One of the most common analysis activities in PPC is comparison. We often compare the performance of similar tools or processes. We also compare the effect of different treatments such as recipe settings. When we compare two things, such as two tools running the same operation, we use comparison techniques. When we want to compare multiple things, like multiple tools running the same operation or multiple tools with multiple operators running the same operation, we turn to ANOVA techniques to perform the analysis.

ANOVA splits the data into components

The easiest way to understand ANOVA is through a concept known as value splitting. ANOVA splits the observed data values into components that are attributable to the different levels of the factors. Value splitting is best explained by example.

Example: Turned Pins

The simplest example of value splitting is when we just have one level of one factor. Suppose we have a turning operation in a machine shop where we are turning pins to a diameter of .125 +/- .005 inches. Throughout the course of a day we take five samples of pins and obtain the following measurements: .125, .127, .124, .126, .128.

We can split these data values into a common value (mean) and residuals (what's left over) as follows:

.125	.127	.124	.126	.128

=

.126	.126	.126	.126	.126

+

-.001	.001	-.002	.000	.002

From these tables, also called overlays, we can easily calculate the location and spread of the data as follows:

mean = .126

std. deviation = .0016.

Other layouts

While the above example is a trivial structural layout, it illustrates how we can split data values into its components. In the next sections, we will look at more complicated structural layouts for the data. In particular we will look at multiple levels of one factor (One-Way ANOVA) and multiple levels of two factors (Two-Way ANOVA) where the factors are crossed and nested.

3.2.3.1. One-Way ANOVA

Description

A one-way layout consists of a single factor with several levels and multiple observations at each level. With this kind of layout we can calculate the mean of the observations within each level of our factor. The residuals will tell us about the variation within each level. We can also average the means of each level to obtain a grand mean. We can then look at the deviation of the mean of each level from the grand mean to understand something about the level effects. Finally, we can compare the variation within levels to the variation across levels. Hence the name analysis of variance.

Model

It is easy to model all of this with an equation of the form:

$$y_{ij} = m + a_i + e_{ij}$$

The equation indicates that the *jth* data value, from level i, is the sum of three components: the common value (grand mean), the level effect (the deviation of each level mean from the grand mean), and the residual (what's left over).

Estimation

click here to see details of one-way value splitting

Estimation for the one-way layout can be performed one of two ways. First, we can calculate the total variation, within-level variation and across-level variation. These can be summarized in a table as shown below and tests can be made to determine if the factor levels are significant. The value splitting example illustrates the calculations involved.

ANOVA table for one-way case

In general, the ANOVA table for the one-way case is given by:

Source	Sum of Squares	DoF	Mean Square	F_0
Factor	$SS_F = J\sum(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$I-1$	$MSF = SS_F/(I-1)$	MSF/MSE
Residual	$SS_E = \sum\sum(y_{ij} - \bar{y}_{i\cdot})^2$	$I(J-1)$	$MSE = SS_E/(I(J-1))$	
Corr. Total	$SST = \sum\sum(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$IJ-1$		

where

$$\bar{y}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$$

and

$$\bar{y}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}$$

.

The row labeled, "Corr. Total", in the ANOVA table contains the corrected total sum of squares and the associated degrees of freedom (DoF).

Level effects must sum to zero

The second way to estimate effects is through the use of CLM techniques. If you look at the model above you will notice that it is in the form of a CLM. The only problem is that the model is saturated and no unique solution exists. We overcome this problem by applying a constraint to the model. Since the level effects are just deviations from the grand mean, they must sum to zero. By applying the constraint that the level effects must sum to zero, we can now obtain a unique solution to the CLM equations. Most analysis programs will handle this for you automatically. See the chapter on Process Modeling for a more complete discussion on estimating the coefficients for these models.

Testing

We are testing to see if the observed data support the hypothesis that the levels of the factor are significantly different from each other. The way we do this is by comparing the within-level variancs to the between-level variance.

If we assume that the observations within each level have the same variance, we can calculate the variance within each level and pool these together to obtain an estimate of the overall population variance. This works out to be the mean square of the residuals.

Similarly, if there really were no level effect, the mean square across levels would be an estimate of the overall variance. Therefore, if there really were no level effect, these two estimates would be just two different ways to estimate the same parameter and should be close numerically. However, if there is a level effect, the level mean square will be higher than the residual mean square.

It can be shown that given the assumptions about the data stated below, the ratio of the level mean square and the residual mean square follows an *F* distribution with degrees of freedom as shown in the ANOVA table. If the F_0 value is significant at a given significance level (greater than the cut-off value in a *F* table), then there is a level effect present in the data.

Assumptions

For estimation purposes, we assume the data can adequately be modeled as the sum of a deterministic component and a random component. We further assume that the fixed (deterministic) component can be modeled as the sum of an overall mean and some contribution from the factor level. Finally, it is assumed that the random component can be modeled with a Gaussian distribution with fixed location and spread.

Uses

The one-way ANOVA is useful when we want to compare the effect of multiple levels of one factor and we have multiple observations at each level. The factor can be either discrete (different machine, different plants, different shifts, etc.) or continuous (different gas flows, temperatures, etc.).

Example

Let's extend the machining example by assuming that we have five different machines making the same part and we take five random samples from each machine to obtain the following diameter data:

Machine				

1	2	3	4	5
0.125	0.118	0.123	0.126	0.118
0.127	0.122	0.125	0.128	0.129
0.125	0.120	0.125	0.126	0.127
0.126	0.124	0.124	0.127	0.120
0.128	0.119	0.126	0.129	0.121

Analyze Using ANOVA software or the techniques of the value-splitting example, we summarize the data in an ANOVA table as follows:

Source	Sum of Squares	Deg. of Freedom	Mean Square	F_0
Factor	0.000137	4	0.000034	4.86
Residual	0.000132	20	0.000007	
Corrected Total	0.000269	24		

Test By dividing the factor-level mean square by the residual mean square, we obtain an F_0 value of 4.86 which is greater than the cut-off value of 2.87 from the F distribution with 4 and 20 degrees of freedom and a significance level of 0.05. Therefore, there is sufficient evidence to reject the hypothesis that the levels are all the same.

Conclusion From the analysis of these data we can conclude that the factor "machine" has an effect. There is a statistically significant difference in the pin diameters across the machines on which they were manufactured.

3.2.3.1.1. One-Way Value-Splitting

Example

Let's use the data from the machining example to illustrate how to use the techniques of value-splitting to break each data value into its component parts. Once we have the component parts, it is then a trivial matter to calculate the sums of squares and form the F-value for the test.

Machine				
1	2	3	4	5
.125	.118	.123	.126	.118
.127	.122	.125	.128	.129
.125	.120	.125	.126	.127
.126	.124	.124	.127	.120
.128	.119	.126	.129	.121

Calculate level-means

Remember from our model, $y_{ij} = m + a_i + e_{ij}$, we say each observation is the sum of a common value, a level effect and a residual value. Value-splitting just breaks each observation into its component parts. The first step in value-splitting is to calculate the mean values (rounding to the nearest thousandth) within each machine to get the level means.

Machine				
1	2	3	4	5
.1262	.1206	.1246	.1272	.123

Sweep level means

We can then *sweep* (subtract the level mean from each associated data value) the means through the original data table to get the residuals:

Machine				
1	2	3	4	5
-.0012	-.0026	-.0016	-.0012	-.005
.0008	.0014	.0004	.0008	.006
-	-		-	

.0012	.0006	.0004	.0012	.004
-.0002	.0034	-.0006	-.0002	-.003
.0018	-.0016	.0014	.0018	-.002

Calculate the grand mean

The next step is to calculate the grand mean from the individual machine means as:

Grand Mean
.12432

Sweep the grand mean through the level means

Finally, we can sweep the grand mean through the individual level means to obtain the level effects:

Machine				
1	2	3	4	5
.00188	-.00372	.00028	.00288	-.00132

It is easy to verify that the original data table can be constructed by adding the overall mean, the machine effect and the appropriate residual.

Calculate ANOVA values

Now that we have the data values split and the overlays created, the next step is to calculate the various values in the One-Way ANOVA table. We have three values to calculate for each overlay. They are the sums of squares, the degrees of freedom, and the mean squares.

Total sum of squares

The total sum of squares is calculated by summing the squares of all the data values and subtracting from this number the square of the grand mean times the total number of data values. We usually don't calculate the mean square for the total sum of squares because we don't use this value in any statistical test.

Residual sum of squares, degrees of freedom and mean

The residual sum of squares is calculated by summing the squares of the residual values. This is equal to .000132. The degrees of freedom is the number of unconstrained values. Since the residuals for each level of the factor must sum to zero, once we know four of them, the last one is determined. This means we have four unconstrained values for each level,

square	or 20 degrees of freedom. This gives a mean square of .000007.
Level sum of squares, degrees of freedom and mean square	Finally, to obtain the sum of squares for the levels, we sum the squares of each value in the level effect overlay and multiply the sum by the number of observations for each level (in this case 5) to obtain a value of .000137. Since the deviations from the level means must sum to zero, we have only four unconstrained values so the degrees of freedom for level effects is 4. This produces a mean square of .000034.
Calculate F-value	The last step is to calculate the F-value and perform the test of equal level means. The F- value is just the level mean square divided by the residual mean square. In this case the F-value=4.86. If we look in an F-table for 4 and 20 degrees of freedom at 95% confidence, we see that the critical value is 2.87, which means that we have a significant result and that there is thus evidence of a strong machine effect. By looking at the level-effect overlay we see that this is driven by machines 2 and 4.

3.2.3.2. Two-Way Crossed ANOVA

Description When we have two factors with at least two levels and one or more observations at each level, we say we have a two-way layout. We say that the two-way layout is crossed when every level of Factor A occurs with every level of Factor B. With this kind of layout we can estimate the effect of each factor (Main Effects) as well as any interaction between the factors.

Model If we assume that we have K observations at each combination of I levels of Factor A and J levels of Factor B, then we can model the two-way layout with an equation of the form:

$$y_{ijk} = m + a_i + b_j + (ab)_{ij} + e_{ijk}$$

This equation just says that the *kth* data value for the *jth* level of Factor B and the *ith* level of Factor A is the sum of five components: the common value (grand mean), the level effect for Factor A, the level effect for Factor B, the interaction effect, and the residual. Note that (ab) does not mean multiplication; rather that there is interaction between the two factors.

Estimation Like the one-way case, the estimation for the two-way layout can be done either by calculating the variance components or by using CLM techniques.

Click here for the value splitting example For the two-way ANOVA, we display the data in a two-dimensional table with the levels of Factor A in columns and the levels of Factor B in rows. The replicate observations fill each cell. We can sweep out the common value, the row effects, the column effects, the interaction effects and the residuals using value-splitting techniques. Sums of squares can be calculated and summarized in an ANOVA table as shown below.

Source	Sum of Squares	DoF	Mean Square	F_0
Rows	$SSR = JK\sum(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$I-1$	$MSR = SSR/(I-1)$	MSR/MSE
Columns	$SSC = IK\sum(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$J-1$	$MSC = SSC/(J-1)$	MSC/MSE
Interaction	$SSI = K\sum\sum(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	$(I-1)(J-1)$	$MSI = SSI/((I-1)(J-1))$	MSI/MSE
Residuals	$SSE = \sum\sum\sum(y_{ijk} - \bar{y}_{ij\cdot})^2$	$IJ(K-1)$	$MSE = SSE/(IJ(K-1))$	
Corr. Total	$SST = \sum\sum\sum(y_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2$	$IJK-1$		

$$\bar{y}_{i\cdot\cdot} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}$$

$$\bar{y}_{\cdot j\cdot} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ijk}$$

$$\bar{y}_{ij\cdot} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk}$$

$$\bar{y}_{\cdot\cdot\cdot} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}$$

The row labeled, "Corr. Total", in the ANOVA table contains the corrected total sum of squares and the associated degrees of freedom (DoF).

We can use CLM techniques to do the estimation. We still have the problem that the model is saturated and no unique solution exists. We overcome this problem by applying the constraints to the model that the two main

effects and interaction effects each sum to zero.

Testing Like testing in the one-way case, we are testing that two main effects and the interaction are zero. Again we just form a ratio of each main effect mean square and the interaction mean square to the residual mean square. If the assumptions stated below are true then those ratios follow an F distribution and the test is performed by comparing the F_0 ratios to values in an F table with the appropriate degrees of freedom and confidence level.

Assumptions For estimation purposes, we assume the data can be adequately modeled as described in the model above. It is assumed that the random component can be modeled with a Gaussian distribution with fixed location and spread.

Uses The two-way crossed ANOVA is useful when we want to compare the effect of multiple levels of two factors and we can combine every level of one factor with every level of the other factor. If we have multiple observations at each level, then we can also estimate the effects of interaction between the two factors.

Example Let's extend the one-way machining example by assuming that we want to test if there are any differences in pin diameters due to different types of coolant. We still have five different machines making the same part and we take five samples from each machine for each coolant type to obtain the following data:

	Machine				
	1	**2**	**3**	**4**	**5**
Coolant A	0.125	0.118	0.123	0.126	0.118
	0.127	0.122	0.125	0.128	0.129
	0.125	0.120	0.125	0.126	0.127
	0.126	0.124	0.124	0.127	0.120
	0.128	0.119	0.126	0.129	0.121
Coolant B	0.124	0.116	0.122	0.126	0.125
	0.128	0.125	0.121	0.129	0.123
	0.127	0.119	0.124	0.125	0.114
	0.126	0.125	0.126	0.130	0.124
	0.129	0.120	0.125	0.124	0.117

Analyze For analysis details see the crossed two-way value splitting example. We can summarize the analysis results in an ANOVA table as follows:

Source	**Sum of Squares**	**Deg. of Freedom**	**Mean Square**	F_0
machine	0.000303	4	0.000076	8.8
coolant	0.00000392	1	0.00000392	0.45
interaction	0.00001468	4	0.00000367	0.42
residuals	0.000346	40	0.0000087	
corrected total	0.000668	49		

Test By dividing the mean square for machine by the mean square for residuals we obtain an F_0 value of 8.8 which is greater than the critical value of 2.61 based on 4 and 40 degrees of freedom and a 0.05 significance level. Likewise the F_0 values for Coolant and Interaction, obtained by dividing their mean squares by the residual mean square, are less than their respective critical values of 4.08 and 2.61 (0.05 significance level).

Conclusion From the ANOVA table we can conclude that machine is the most important factor and is statistically significant. Coolant is not significant and neither is the interaction. These results would lead us to believe that some tool-matching efforts would be useful for improving this process.

3.2.3.2.1. Two-way Crossed Value-Splitting Example

Example: Coolant is completely crossed with machine

The data table below is five samples each collected from five different lathes each running two different types of coolant. The measurement is the diameter of a turned pin.

	Machine				
	1	**2**	**3**	**4**	**5**
Coolant A	.125	.118	.123	.126	.118
	.127	.122	.125	.128	.129
	.125	.120	.125	.126	.127
	.126	.124	.124	.127	.120
	.128	.119	.126	.129	.121
Coolant B	.124	.116	.122	.126	.125
	.128	.125	.121	.129	.123
	.127	.119	.124	.125	.114
	.126	.125	.126	.130	.124
	.129	.120	.125	.124	.117

For the crossed two-way case, the first thing we need to do is to sweep the cell means from the data table to obtain the residual values. This is shown in the tables below.

The first step is to sweep out the cell means to obtain the residuals and means

	Machine				
	1	**2**	**3**	**4**	**5**
A	.1262	.1206	.1246	.1272	.123
B	.1268	.121	.1236	.1268	.1206
Coolant A	-.0012	-.0026	-.0016	-.0012	-.005
	.0008	.0014	.0004	.0008	.006
	-.0012	-.0006	.0004	-.0012	.004
	-.0002	.0034	-.0006	-.0002	-.003

	.0018	-.0016	.0014	.0018	-.002
Coolant B	-.0028	-.005	-.0016	-.0008	.0044
	.0012	.004	-.0026	.0022	.0024
	.0002	-.002	.0004	-.0018	-.0066
	-.0008	.004	.0024	.0032	.0034
	.0022	-.001	.0014	-.0028	-.0036

Sweep the row means

The next step is to sweep out the row means. This gives the table below.

		Machine				
		1	2	3	4	5
A	.1243	.0019	-.0037	.0003	.0029	-.0013
B	.1238	.003	-.0028	-.0002	.003	-.0032

Sweep the column means

Finally, we sweep the column means to obtain the grand mean, row (coolant) effects, column (machine) effects and the interaction effects.

		Machine				
		1	2	3	4	5
	.1241	.0025	-.0033	.00005	.003	-.0023
A	.0003	-.0006	-.0005	.00025	.0000	.001
B	-.0003	.0006	.0005	-.00025	.0000	-.001

What do these tables tell us?

By looking at the table of residuals, we see that the residuals for coolant B tend to be a little higher than for coolant A. This implies that there may be more variability in diameter when we use coolant B. From the effects table above, we see that machines 2 and 5 produce smaller pin diameters than the other machines. There is also a very slight coolant effect but the machine effect is larger. Finally, there also appears to be slight interaction effects. For instance, machines 1 and 2 had smaller diameters with coolant A but the opposite was true for machines 3,4 and 5.

Calculate

We can calculate the values for the ANOVA table according

sums of squares and mean squares

to the formulae in the table on the crossed two-way page. This gives the table below. From the F-values we see that the machine effect is significant but the coolant and the interaction are not.

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value
Machine	.000303	4	.000076	8.8 > 2.61
Coolant	.00000392	1	.00000392	.45 < 4.08
Interaction	.00001468	4	.00000367	.42 < 2.61
Residual	.000346	40	.0000087	
Corrected Total	.000668	49		

3.2.3.3. Two-Way Nested ANOVA

Description

Sometimes, constraints prevent us from crossing every level of one factor with every level of the other factor. In these cases we are forced into what is known as a *nested* layout. We say we have a nested layout when fewer than all levels of one factor occur within each level of the other factor. An example of this might be if we want to study the effects of different machines and different operators on some output characteristic, but we can't have the operators change the machines they run. In this case, each operator is not crossed with each machine but rather only runs one machine.

Model

If Factor B is nested within Factor A, then a level of Factor B can only occur within one level of Factor A and there can be no interaction. This gives the following model:

$$y_{ijk} = m + a_i + b_{j(i)} + e_{ijk}$$

This equation indicates that each data value is the sum of a common value (grand mean), the level effect for Factor A, the level effect of Factor B nested within Factor A, and the residual.

Estimation

For a nested design we typically use variance components methods to perform the analysis. We can sweep out the common value, the Factor A effects, the Factor B within A effects and the residuals using value-splitting techniques. Sums of squares can be calculated and summarized in an ANOVA table as shown below.

Click here for nested value-splitting example

It is important to note that with this type of layout, since each level of one factor is only present with one level of the other factor, we can't estimate interaction between the two.

ANOVA table for nested case

Source	Sum of Squares	DoF	Mean Square	F_0
A	$SSA = JK\sum(\bar{y}_{i..} - \bar{y}_{...})^2$	$I-1$	$MSA = SSA/(I-1)$	MSA/MSB
B within A	$SSB = K\sum\sum(\bar{y}_{ij.} - \bar{y}_{i..})^2$	$I(J-1)$	$MSB = SSB/(I(J-1))$	MSB/MSE
Residuals	$SSE = \sum\sum\sum(y_{ijk} - \bar{y}_{ij.})^2$	$IJ(K-1)$	$MSE = SSE/(IJ(K-1))$	
Corr. Total	$SST = \sum\sum\sum(y_{ijk} - \bar{y}_{...})^2$	$IJK-1$		

$$\bar{y}_{i..} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}$$

$$\bar{y}_{ij.} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk}$$

$$\bar{y}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}$$

The row labeled, "Corr. Total", in the ANOVA table contains the corrected total sum of squares and the associated degrees of freedom (DoF).

As with the crossed layout, we can also use CLM techniques. We still have the problem that the model is saturated and no unique solution exists. We overcome this problem by applying to the model the constraints that the two main effects sum to zero.

Testing

We are testing that two main effects are zero. Again we just form a ratio (F_0) of each main effect mean square to the appropriate mean-squared error term. (Note that the error term for Factor A is not *MSE*, but is *MSB*.) If the assumptions stated below are true then those ratios follow an *F* distribution and the test is performed by comparing the F_0 ratios to values in an *F* table with the appropriate degrees of freedom and confidence level.

Assumptions

For estimation purposes, we assume the data can be adequately modeled by the model above and that there is more than one variance component. It is assumed that the random component can be modeled with a Gaussian distribution with fixed location and spread.

Uses

The two-way nested ANOVA is useful when we are constrained from combining all the levels of one factor with all of the levels of the other factor. These designs are most useful when we have what is called a random effects situation. When the levels of a factor are chosen at random rather than selected intentionally, we say we have a random effects model. An example of this is when we select lots from a production run, then select units from the lot. Here the units are nested within lots and the effect of each factor is random.

Example

Let's change the two-way machining example slightly by assuming that we have five different machines making the same part and each machine has two operators, one for the day shift and one for the night shift. We take five samples from each machine for each operator to obtain the following data:

	Machine				
	1	**2**	**3**	**4**	**5**
Operator Day	0.125	0.118	0.123	0.126	0.118
	0.127	0.122	0.125	0.128	0.129
	0.125	0.120	0.125	0.126	0.127
	0.126	0.124	0.124	0.127	0.120
	0.128	0.119	0.126	0.129	0.121
Operator Night	0.124	0.116	0.122	0.126	0.125
	0.128	0.125	0.121	0.129	0.123
	0.127	0.119	0.124	0.125	0.114
	0.126	0.125	0.126	0.130	0.124
	0.129	0.120	0.125	0.124	0.117

Analyze

For analysis details see the nested two-way value splitting example. We can summarize the analysis results in an ANOVA table as follows:

Source	Sum of Squares	Deg. of Freedom	Mean Square	F_0
Machine	3.03e-4	4	7.58e-5	20.38
Operator(Machine)	1.86e-5	5	3.72e-6	0.428
Residuals	3.46e-4	40	8.70e-6	
Corrected Total	6.68e-4	49		

Test By dividing the mean square for Machine by the mean square for Operator within Machine, or Operator(Machine), we obtain an F_0 value of 20.38 which is greater than the critical value of 5.19 for 4 and 5 degrees of freedom at the 0.05 significance level. The F_0 value for Operator(Machine), obtained by dividing its mean square by the residual mean square, is less than the critical value of 2.45 for 5 and 40 degrees of freedom at the 0.05 significance level.

Conclusion From the ANOVA table we can conclude that the Machine is the most important factor and is statistically significant. The effect of Operator nested within Machine is not statistically significant. Again, any improvement activities should be focused on the tools.

3.2.3.3.1. Two-Way Nested Value-Splitting Example

Example: Operator is nested within machine.

The data table below contains data collected from five different lathes, each run by two different operators. Note we are concerned here with the effect of operators, so the layout is nested. If we were concerned with shift instead of operator, the layout would be crossed. The measurement is the diameter of a turned pin.

Machine	Operator	Sample				
		1	2	3	4	5
1	Day	.125	.127	.125	.126	.128
	Night	.124	.128	.127	.126	.129
2	Day	.118	.122	.120	.124	.119
	Night	.116	.125	.119	.125	.120
3	Day	.123	.125	.125	.124	.126
	Night	.122	.121	.124	.126	.125
4	Day	.126	.128	.126	.127	.129
	Night	.126	.129	.125	.130	.124
5	Day	.118	.129	.127	.120	.121
	Night	.125	.123	.114	.124	.117

For the nested two-way case, just as in the crossed case, the first thing we need to do is to sweep the cell means from the data table to obtain the residual values. We then sweep the nested factor (Operator) and the top level factor (Machine) to obtain the table below.

Machine	Operator	Common	Machine	Operator	Sample				
					1	2	3	4	5
1	Day	.12404	.00246	-.0003	-.0012	.0008	-.0012	-.0002	.0018
	Night			.0003	-.0028	.0012	.002	-.0008	.0022
2	Day		-.00324	-.0002	-.0026	.0014	-.0006	.0034	-.0016
	Night			.0002	-.005	.004	-.002	.004	-.001
3	Day		.00006	.0005	-.0016	.0004	.0004	-.0006	.0014

	Night			-.0005		-.0016	-.0026	.0004	.0024	.0014
4	Day		.00296	.0002		-.0012	.0008	-.0012	-.002	.0018
	Night			-.0002		-.0008	.0022	-.0018	.0032	-.0028
5	Day		-.00224	.0012		-.005	.006	.004	-.003	-.002
	Night			-.0012		.0044	.0024	-.0066	.0034	-.0036

What does this table tell us?

By looking at the residuals we see that machines 2 and 5 have the greatest variability. There does not appear to be much of an operator effect but there is clearly a strong machine effect.

Calculate sums of squares and mean squares

We can calculate the values for the ANOVA table according to the formulae in the table on the nested two-way page. This produces the table below. From the F-values we see that the machine effect is significant but the operator effect is not. (Here it is assumed that both factors are fixed).

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value
Machine	.000303	4	.0000758	8.77 > 2.61
Operator(Machine)	.0000186	5	.00000372	.428 < 2.45
Residual	.000346	40	.0000087	
Corrected Total	.000668	49		

3.2.4. Discrete Models

Description

There are many instances when we are faced with the analysis of discrete data rather than continuous data. Examples of this are yield (good/bad), speed bins (slow/fast/faster/fastest), survey results (favor/oppose), etc. We then try to explain the discrete outcomes with some combination of discrete and/or continuous explanatory variables. In this situation the modeling techniques we have learned so far (CLM and ANOVA) are no longer appropriate.

Contingency table analysis and log-linear model

There are two primary methods available for the analysis of discrete response data. The first one applies to situations in which we have discrete explanatory variables and discrete responses and is known as Contingency Table Analysis. The model for this is covered in detail in this section. The second model applies when we have both discrete and continuous explanatory variables and is referred to as a Log-Linear Model. That model is beyond the scope of this Handbook, but interested readers should refer to the reference section of this chapter for a list of useful books on the topic.

Model

Suppose we have *n* individuals that we classify according to two criteria, A and B. Suppose there are *r* levels of criterion A and *s* levels of criterion B. These responses can be displayed in an *r* **x** *s* table. For example, suppose we have a box of manufactured parts that we classify as good or bad and whether they came from supplier 1, 2 or 3.

Now, each cell of this table will have a count of the individuals who fall into its particular combination of classification levels. Let's call this count N_{ij}. The sum of all of these counts will be equal to the total number of individuals, N. Also, each row of the table will sum to $N_{i.}$ and each column will sum to $N_{.j}$.

Under the assumption that there is no interaction between the two classifying variables (like the number of good or bad parts does not depend on which supplier they came from), we can calculate the counts we would expect to see in each cell. Let's call the expected count for any cell E_{ij} . Then the expected value for a cell is $E_{ij} = N_{i.} * N_{.j} / N$. All we need to do then is to compare the expected counts to the observed

counts. If there is a consderable difference between the observed counts and the expected values, then the two variables interact in some way.

Estimation The estimation is very simple. All we do is make a table of the observed counts and then calculate the expected counts as described above.

Testing The test is performed using a Chi-Square goodness-of-fit test according to the following formula:

$$\chi^2 = \sum\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

where the summation is across all of the cells in the table.

Given the assumptions stated below, this statistic has approximately a chi-square distribution and is therefore compared against a chi-square table with $(r-1)(s-1)$ degrees of freedom, with r and s as previously defined. If the value of the test statistic is less than the chi-square value for a given level of confidence, then the classifying variables are declared independent, otherwise they are judged to be dependent.

Assumptions The estimation and testing results above hold regardless of whether the sample model is Poisson, multinomial, or product-multinomial. The chi-square results start to break down if the counts in any cell are small, say < 5.

Uses The contingency table method is really just a test of interaction between discrete explanatory variables for discrete responses. The example given below is for two factors. The methods are equally applicable to more factors, but as with any interaction, as you add more factors the interpretation of the results becomes more difficult.

Example Suppose we are comparing the yield from two manufacturing processes. We want want to know if one process has a higher yield.

Make table of counts

	Good	**Bad**	**Totals**
Process A	86	14	100
Process B	80	20	100
Totals	166	34	200

Table 1. Yields for two production processes

We obtain the expected values by the formula given above. This gives the table below.

Calculate expected counts

	Good	Bad	Totals
Process A	83	17	100
Process B	83	17	100
Totals	166	34	200

Table 2. Expected values for two production processes

Calculate chi-square statistic and compare to table value

The chi-square statistic is 1.276. This is below the chi-square value for 1 degree of freedom and 90% confidence of 2.71 . Therefore, we conclude that there is not a (significant) difference in process yield.

Conclusion

Therefore, we conclude that there is no statistically significant difference between the two processes.

3.3. Data Collection for PPC

Start with careful planning

The data collection process for PPC starts with careful planning. The planning consists of the definition of clear and concise goals, developing process models and devising a sampling plan.

Many things can go wrong in the data collection

This activity of course ends without the actual collection of the data which is usually not as straightforward as it might appear. Many things can go wrong in the execution of the sampling plan. The problems can be mitigated with the use of check lists and by carefully documenting all exceptions to the original sampling plan.

Table of Contents

1. Set Goals
2. Modeling Processes
    1. Black-Box Models
    2. Fishbone Diagrams
    3. Relationships and Sensitivities
3. Define the Sampling Plan
    1. Identify the parameters, ranges and resolution
    2. Design sampling scheme
    3. Select sample sizes
    4. Design data storage formats
    5. Assign roles and responsibilities

3. Production Process Characterization
3.3. Data Collection for PPC

3.3.1. Define Goals

State concise goals	The goal statement is one of the most important parts of the characterization plan. With clearly and concisely stated goals, the rest of the planning process falls naturally into place.
Goals usually defined in terms of key specifications	The goals are usually defined in terms of key specifications or manufacturing indices. We typically want to characterize a process and compare the results against these specifications. However, this is not always the case. We may, for instance, just want to quantify key process parameters and use our estimates of those parameters in some other activity like controller design or process improvement.
Example goal statements	Click on each of the links below to see Goal Statements for each of the case studies. 1. Furnace Case Study (Goal) 2. Machine Case Study (Goal)

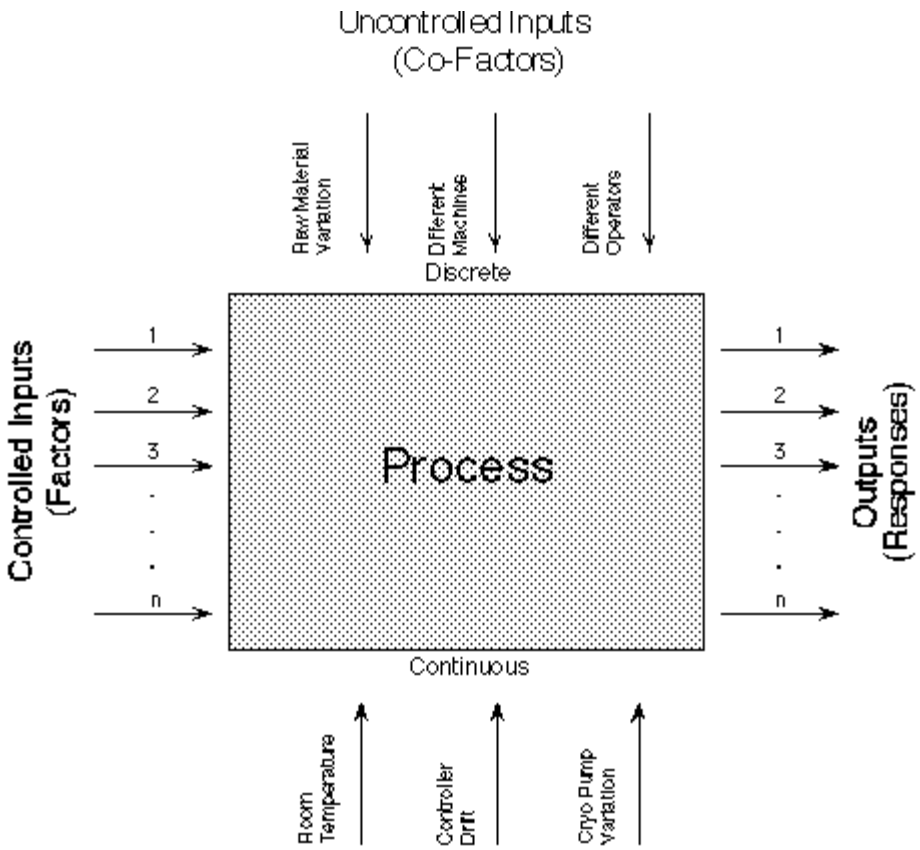
3.3.2. Process Modeling

Identify influential parameters

Process modeling begins by identifying all of the important factors and responses. This is usually best done as a team effort and is limited to the scope set by the goal statement.

Document with black-box models

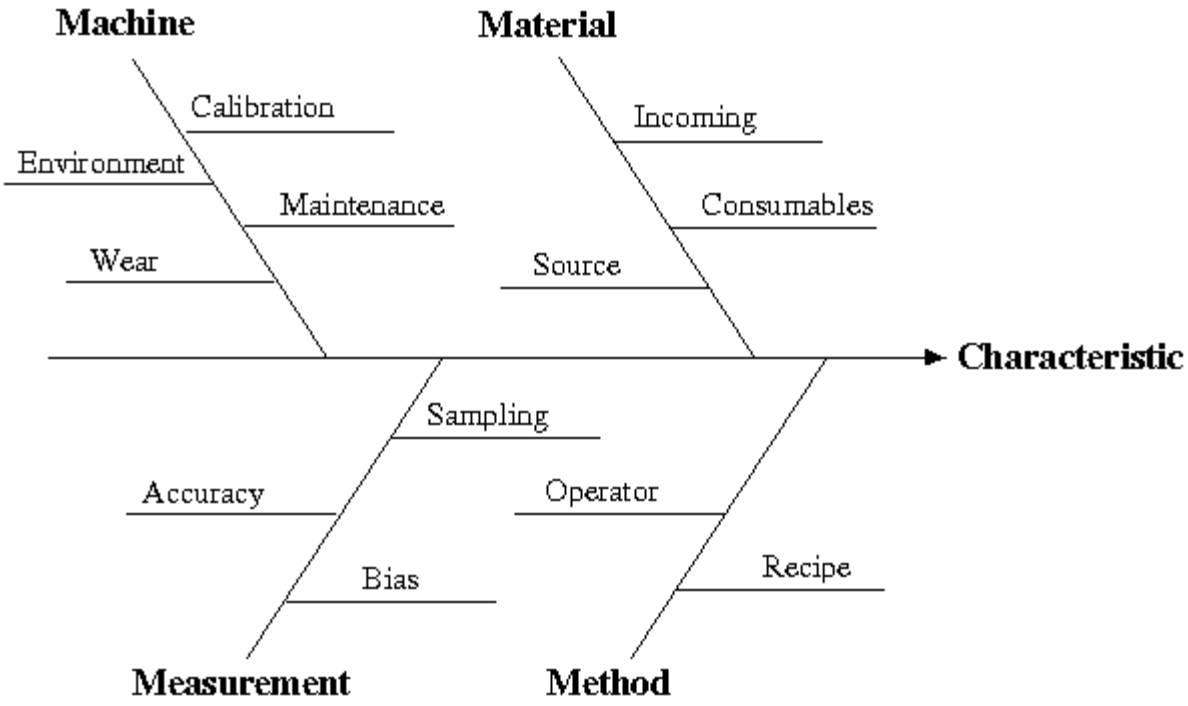
This activity is best documented in the form of a black-box model as seen in the figure below. In this figure all of the outputs are shown on the right and all of the controllable inputs are shown on the left. Any inputs or factors that may be observable but not controllable are shown on the top or bottom.



Model relationships using fishbone diagrams

The next step is to model relationships of the previously identified factors and responses. In this step we choose a parameter and identify all of the other parameters that may have an influence on it. This process is easily documented with fishbone diagrams as illustrated in the figure below.

The influenced parameter is put on the center line and the influential factors are listed off of the centerline and can be grouped into major categories like Tool, Material, Work Methods and Environment.



Document relationships and sensitivities

The final step is to document all known information about the relationships and sensitivities between the inputs and outputs. Some of the inputs may be correlated with each other as well as the outputs. There may be detailed mathematical models available from other studies or the information available may be vague such as for a machining process we know that as the feed rate increases, the quality of the finish decreases.

It is best to document this kind of information in a table with all of the inputs and outputs listed both on the left column and on the top row. Then, correlation information can be filled in for each of the appropriate cells. See the case studies for an example.

Examples

Click on each of the links below to see the process models for each of the case studies.

1. Case Study 1 (Process Model)
2. Case Study 2 (Process Model)

3.3.3. Define Sampling Plan

Sampling plan is detailed outline of measurements to be taken

A sampling plan is a detailed outline of which measurements will be taken at what times, on which material, in what manner, and by whom. Sampling plans should be designed in such a way that the resulting data will contain a representative sample of the parameters of interest and allow for all questions, as stated in the goals, to be answered.

Steps in the sampling plan

The steps involved in developing a sampling plan are:

1. identify the parameters to be measured, the range of possible values, and the required resolution
2. design a sampling scheme that details how and when samples will be taken
3. select sample sizes
4. design data storage formats
5. assign roles and responsibilities

Verify and execute

Once the sampling plan has been developed, it can be verified and then passed on to the responsible parties for execution.

3.3.3.1. Identifying Parameters, Ranges and Resolution

Our goals and the models we built in the previous steps should provide all of the information needed for selecting parameters and determining the expected ranges and the required measurement resolution.

Goals will tell us what to measure and how

The first step is to carefully examine the goals. This will tell you which response variables need to be sampled and how. For instance, if our goal states that we want to determine if an oxide film can be grown on a wafer to within 10 Angstroms of the target value with a uniformity of <2%, then we know we have to measure the film thickness on the wafers to an accuracy of at least +/- 3 Angstroms and we must measure at multiple sites on the wafer in order to calculate uniformity.

The goals and the models we build will also indicate which explanatory variables need to be sampled and how. Since the fishbone diagrams define the known important relationships, these will be our best guide as to which explanatory variables are candidates for measurement.

Ranges help screen outliers

Defining the expected ranges of values is useful for screening outliers. In the machining example , we would not expect to see many values that vary more than +/- .005" from nominal. Therefore we know that any values that are much beyond this interval are highly suspect and should be remeasured.

Resolution helps choose measurement equipment

Finally, the required resolution for the measurements should be specified. This specification will help guide the choice of metrology equipment and help define the measurement procedures. As a rule of thumb, we would like our measurement resolution to be at least 1/10 of our tolerance. For the oxide growth example, this means that we want to measure with an accuracy of 2 Angstroms. Similarly, for the turning operation we would need to measure the diameter within .001". This means that vernier calipers would be adequate as the measurement device for this application.

Examples

Click on each of the links below to see the parameter

descriptions for each of the case studies.

1. Case Study 1 (Sampling Plan)
2. Case Study 2 (Sampling Plan)

3.3.3.2. Choosing a Sampling Scheme

A sampling scheme defines what data will be obtained and how

A sampling scheme is a detailed description of what data will be obtained and how this will be done. In PPC we are faced with two different situations for developing sampling schemes. The first is when we are conducting a controlled experiment. There are very efficient and exact methods for developing sampling schemes for designed experiments and the reader is referred to the Process Improvement chapter for details.

Passive data collection

The second situation is when we are conducting a passive data collection (PDC) study to learn about the inherent properties of a process. These types of studies are usually for comparison purposes when we wish to compare properties of processes against each other or against some hypothesis. This is the situation that we will focus on here.

There are two principles that guide our choice of sampling scheme

Once we have selected our response parameters, it would seem to be a rather straightforward exercise to take some measurements, calculate some statistics and draw conclusions. There are, however, many things which can go wrong along the way that can be avoided with careful planning and knowing what to watch for. There are two overriding principles that will guide the design of our sampling scheme.

The first is precision

The first principle is that of *precision*. If the sampling scheme is properly laid out, the difference between our estimate of some parameter of interest and its *true* value will be due only to random variation. The size of this random variation is measured by a quantity called *standard error*. The magnitude of the standard error is known as precision. The smaller the standard error, the more precise are our estimates.

Precision of an estimate depends on several factors

The precision of any estimate will depend on:

- the inherent variability of the process estimator
- the measurement error
- the number of independent replications (sample size)
- the efficiency of the sampling scheme.

The second is systematic sampling error (or confounded effects)

The second principle is the avoidance of systematic errors. Systematic sampling error occurs when the levels of one explanatory variable are the same as some other unaccounted for explanatory variable. This is also referred to as confounded effects. Systematic sampling error is best seen by example.

> *Example 1:* We want to compare the effect of two different coolants on the resulting surface finish from a turning operation. It is decided to run one lot, change the coolant and then run another lot. With this sampling scheme, there is no way to distinguish the coolant effect from the lot effect or from tool wear considerations. There is systematic sampling error in this sampling scheme.
>
> *Example 2:* We wish to examine the effect of two pre-clean procedures on the uniformity of an oxide growth process. We clean one cassette of wafers with one method and another cassette with the other method. We load one cassette in the front of the furnace tube and the other cassette in the middle. To complete the run, we fill the rest of the tube with other lots. With this sampling scheme, there is no way to distinguish between the effect of the different pre-clean methods and the cassette effect or the tube location effect. Again, we have systematic sampling errors.

Stratification helps to overcome systematic error

The way to combat systematic sampling errors (and at the same time increase precision) is through stratification and randomization. Stratification is the process of segmenting our population across levels of some factor so as to minimize variability within those segments or *strata*. For instance, if we want to try several different process recipes to see which one is best, we may want to be sure to apply each of the recipes to each of the three work shifts. This will ensure that we eliminate any systematic errors caused by a shift effect. This is where the ANOVA designs are particularly useful.

Randomization helps too

Randomization is the process of randomly applying the various treatment combinations. In the above example, we would not want to apply recipe 1, 2 and 3 in the same order for each of the three shifts but would instead randomize the order of the three recipes in each shift. This will avoid any systematic errors caused by the order of the recipes.

Examples

The issues here are many and complicated. Click on each of the links below to see the sampling schemes for each of the case studies.

1. Case Study 1 (Sampling Plan)
2. Case Study 2 (Sampling Plan)

3.3.3.3. Selecting Sample Sizes

Consider these things when selecting a sample size

When choosing a sample size, we must consider the following issues:

- What population parameters we want to estimate
- Cost of sampling (importance of information)
- How much is already known
- Spread (variability) of the population
- Practicality: how hard is it to collect data
- How precise we want the final estimates to be

Cost of taking samples

The cost of sampling issue helps us determine how precise our estimates should be. As we will see below, when choosing sample sizes we need to select risk values. If the decisions we will make from the sampling activity are very valuable, then we will want low risk values and hence larger sample sizes.

Prior information

If our process has been studied before, we can use that prior information to reduce sample sizes. This can be done by using prior mean and variance estimates and by stratifying the population to reduce variation within groups.

Inherent variability

We take samples to form estimates of some characteristic of the population of interest. The variance of that estimate is proportional to the inherent variability of the population divided by the sample size:

$$\mathrm{Var}(\hat{p}) \approx \sigma^2/n .$$

with $\hat{p}$ denoting the parameter we are trying to estimate. This means that if the variability of the population is large, then we must take many samples. Conversely, a small population variance means we don't have to take as many samples.

Practicality

Of course the sample size you select must make sense. This is where the trade-offs usually occur. We want to take enough observations to obtain reasonably precise estimates

of the parameters of interest but we also want to do this within a practical resource budget. The important thing is to quantify the risks associated with the chosen sample size.

Sample size determination

In summary, the steps involved in estimating a sample size are:

1. There must be a statement about what is expected of the sample. We must determine what is it we are trying to estimate, how precise we want the estimate to be, and what are we going to do with the estimate once we have it. This should easily be derived from the goals.
2. We must find some equation that connects the desired precision of the estimate with the sample size. This is a probability statement. A couple are given below; see your statistician if these are not appropriate for your situation.
3. This equation may contain unknown properties of the population such as the mean or variance. This is where prior information can help.
4. If you are stratifying the population in order to reduce variation, sample size determination must be performed for each stratum.
5. The final sample size should be scrutinized for practicality. If it is unacceptable, the only way to reduce it is to accept less precision in the sample estimate.

Sampling proportions

When we are sampling proportions we start with a probability statement about the desired precision. This is given by:

$$Pr(|\hat{p} - P| \geq \delta) = \alpha$$

where

- $\hat{p}$ is the estimated proportion
- **P** is the unknown population parameter
- δ is the specified precision of the estimate
- α is the probability value (usually low)

This equation simply shows that we want the probability that the precision of our estimate being less than we want is α. Of course we like to set α low, usually .1 or less. Using some assumptions about the proportion being approximately normally distributed we can obtain an estimate of the required sample size as:

$$n = z_{\alpha}^2 (\frac{pq}{\delta^2})$$

where z is the ordinate on the Normal curve corresponding to α.

Example

Let's say we have a new process we want to try. We plan to run the new process and sample the output for yield (good/bad). Our current process has been yielding 65% (p=.65, q=.35). We decide that we want the estimate of the new process yield to be accurate to within δ = .10 at 95% confidence (α = .05, z_α = -2). Using the formula above we get a sample size estimate of *n*=91. Thus, if we draw 91 random parts from the output of the new process and estimate the yield, then we are 95% sure the yield estimate is within .10 of the true process yield.

Estimating location: relative error

If we are sampling continuous normally distributed variables, quite often we are concerned about the relative error of our estimates rather than the absolute error. The probability statement connecting the desired precision to the sample size is given by:

$$Pr\left(\left\|\frac{\hat{y}-\mu}{\mu}\right\| \geq \delta)\right) = \alpha$$

where μ is the (unknown) population mean and $\bar{y}$ is the sample mean.

Again, using the normality assumptions we obtain the estimated sample size to be:

$$n \approx \frac{z_\alpha^2 \sigma^2}{\delta^2 \mu^2}$$

with σ^2 denoting the population variance.

Estimating location: absolute error

If instead of relative error, we wish to use absolute error, the equation for sample size looks alot like the one for the case of proportions:

$$n \approx z_\alpha^2 \left(\frac{\sigma^2}{\delta^2}\right)$$

where σ is the population standard deviation (but in practice is usually replaced by an *engineering guesstimate*).

Example

Suppose we want to sample a stable process that deposits a 500 Angstrom film on a semiconductor wafer in order to determine the process mean so that we can set up a control chart on the process. We want to estimate the mean within 10 Angstroms (δ = 10) of the true mean with 95% confidence (α = .05, z_α = -2). Our initial guess regarding the variation in the process is that one standard deviation is about 20 Angstroms. This gives a sample size estimate of *n* = 16. Thus, if we take at least 16 samples from this process

and estimate the mean film thickness, we can be 95% sure that the estimate is within 10 angstroms of the true mean value.

3.3.3.4. Data Storage and Retrieval

Data control depends on facility size

If you are in a small manufacturing facility or a lab, you can simply design a sampling plan, run the material, take the measurements, fill in the run sheet and go back to your computer to analyze the results. There really is not much to be concerned with regarding data storage and retrieval.

In most larger facilities, however, the people handling the material usually have nothing to do with the design. Quite often the measurements are taken automatically and may not even be made in the same country where the material was produced. Your data go through a long chain of automatic acquisition, storage, reformatting, and retrieval before you are ever able to see it. All of these steps are fraught with peril and should be examined closely to ensure that valuable data are not lost or accidentally altered.

Know the process involved

In the planning phase of the PPC, be sure to understand the entire data collection process. Things to watch out for include:

- automatic measurement machines rejecting outliers
- only summary statistics (mean and standard deviation) being saved
- values for explanatory variables (location, operator, etc.) are not being saved
- how missing values are handled

Consult with support staff early on

It is important to consult with someone from the organization responsible for maintaining the data system early in the planning phase of the PPC. It can also be worthwhile to perform some "dry runs" of the data collection to ensure you will be able to actually acquire the data in the format as defined in the plan.

3.3.3.5. Assign Roles and Responsibilities

PPC is a team effort, get everyone involved early

In today's manufacturing environment, it is unusual when an investigative study is conducted by a single individual. Most PPC studies will be a team effort. It is important that all individuals who will be involved in the study become a part of the team from the beginning. Many of the various collateral activities will need approvals and sign-offs. Be sure to account for that cycle time in your plan.

Table showing roles and potential responsibilities

A partial list of these individuals along with their roles and potential responsibilities is given in the table below. There may be multiple occurrences of each of these individuals across shifts or process steps, so be sure to include everyone.

Tool Owner	Controls Tool Operations	• **S**chedules tool time • Ensures tool state • Advises on experimental design
Process Owner	Controls Process Recipe	• Advises on experimental design • Controls recipe settings
Tool Operator	Executes Experimental Plan	• Executes experimental runs • May take measurements
Metrology	Own Measurement Tools	• Maintains metrology equipment • Conducts gauge studies • May take measurements
CIM	Owns Enterprise Information System	• Maintains data collection system • Maintains equipment interfaces and data formatters • Maintains databases and information access

Statistician	Consultant	• Consults on experimental design • Consults on data analysis
Quality Control	Controls Material	• Ensures quality of incoming material • Must approve shipment of outgoing material (especially for recipe changes)

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

3. Production Process Characterization

3.4. Data Analysis for PPC

In this section we will learn how to analyze and interpret the data we collected in accordance with our data collection plan.

Click on desired topic to read more

This section discusses the following topics:

1. Initial Data Analysis
    1. Gather Data
    2. Quality Checking the Data
    3. Summary Analysis (Location, Spread and Shape)
2. Exploring Relationships
    1. Response Correlations
    2. Exploring Main Effects
    3. Exploring First-Order Interactions
3. Building Models
    1. Fitting Polynomial Models
    2. Fitting Physical Models
4. Analyzing Variance Structure
5. Assessing Process Stablility
6. Assessing Process Capability
7. Checking Assumptions

3. Production Process Characterization
3.4. Data Analysis for PPC

3.4.1. First Steps

Gather all of the data into one place	After executing the data collection plan for the characterization study, the data must be gathered up for analysis. Depending on the scope of the study, the data may reside in one place or in many different places. It may be in common factory databases, flat files on individual computers, or handwritten on run sheets. Whatever the case, the first step will be to collect all of the data from the various sources and enter it into a single data file. The most convenient format for most data analyses is the variables-in-columns format. This format has the variable names in column headings and the values for the variables in the rows.
Perform a quality check on the data using graphical and numerical techniques	The next step is to perform a quality check on the data. Here we are typically looking for data entry problems, unusual data values, missing data, etc. The two most useful tools for this step are the scatter plot and the histogram. By constructing scatter plots of all of the response variables, any data entry problems will be easily identified. Histograms of response variables are also quite useful for identifying data entry problems. Histograms of explanatory variables help identify problems with the execution of the sampling plan. If the counts for each level of the explanatory variables are not the same as called for in the sampling plan, you know you may have an execution problem. Running numerical summary statistics on all of the variables (both response and explanatory) also helps to identify data problems.
Summarize data by estimating location, spread and shape	Once the data quality problems are identified and fixed, we should estimate the location, spread and shape for all of the response variables. This is easily done with a combination of histograms and numerical summary statistics.

3.4.2. Exploring Relationships

The first analysis of our data is exploration

Once we have a data file created in the desired format, checked the data integrity, and have estimated the summary statistics on the response variables, the next step is to start exploring the data and to try to understand the underlying structure. The most useful tools will be various forms of the basic scatter plot and box plot.

These techniques will allow pairwise explorations for examining relationships between any pair of response variables, any pair of explanatory and response variables, or a response variable as a function of any two explanatory variables. Beyond three dimensions we are pretty much limited by our human frailties at visualization.

Graph everything that makes sense

In this exploratory phase, the key is to graph everything that makes sense to graph. These pictures will not only reveal any additional quality problems with the data but will also reveal influential data points and will guide the subsequent modeling activities.

Graph responses, then explanatory versus response, then conditional plots

The order that generally proves most effective for data analysis is to first graph all of the responses against each other in a pairwise fashion. Then we graph responses against the explanatory variables. This will give an indication of the main factors that have an effect on response variables. Finally, we graph response variables, conditioned on the levels of explanatory factors. This is what reveals interactions between explanatory variables. We will use nested boxplots and block plots to visualize interactions.

3.4.2.1. Response Correlations

Make scatter plots of all of the response variables

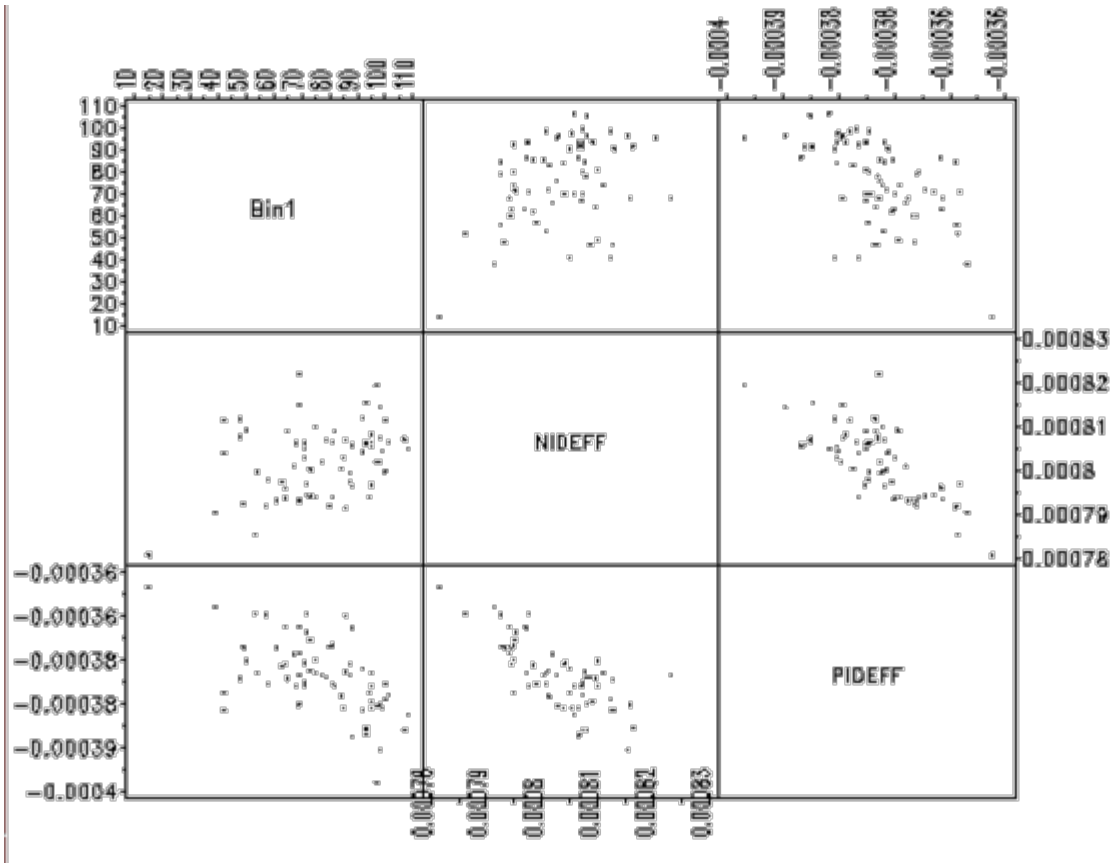
In this first phase of exploring our data, we plot all of the response variables in a pairwise fashion. The individual scatter plots are displayed in a matrix form with the y-axis scaling the same for all plots in a row of the matrix.

Check the slope of the data on the scatter plots

The scatterplot matrix shows how the response variables are related to each other. If there is a linear trend with a positive slope, this indicates that the responses are positively correlated. If there is a linear trend with a negative slope, then the variables are negatively correlated. If the data appear random with no slope, the variables are probably not correlated. This will be important information for subsequent model building steps.

This scatterplot matrix shows examples of both negatively and positively correlated variables

An example of a scatterplot matrix is given below. In this semiconductor manufacturing example, three responses, yield (Bin1), N-channel Id effective (NIDEFF), and P-channel Id effective (PIDEFF) are plotted against each other in a scatterplot matrix. We can see that Bin1 is positively correlated with NIDEFF and negatively correlated with PIDEFF. Also, as expected, NIDEFF is negatively correlated with PIDEFF. This kind of information will prove to be useful when we build models for yield improvement.


Bin1
NIDEFF
PIDEFF

3.4.2.2. Exploring Main Effects

The next step is to look for main effects

The next step in the exploratory analysis of our data is to see which factors have an effect on which response variables and to quantify that effect. Scatter plots and box plots will be the tools of choice here.

Watch out for varying sample sizes across levels

This step is relatively self explanatory. However there are two points of caution. First, be cognizant of not only the trends in these graphs but also the amount of data represented in those trends. This is especially true for categorical explanatory variables. There may be many more observations in some levels of the categorical variable than in others. In any event, take unequal sample sizes into account when making inferences.

Graph implicit as well as explicit explanatory variables

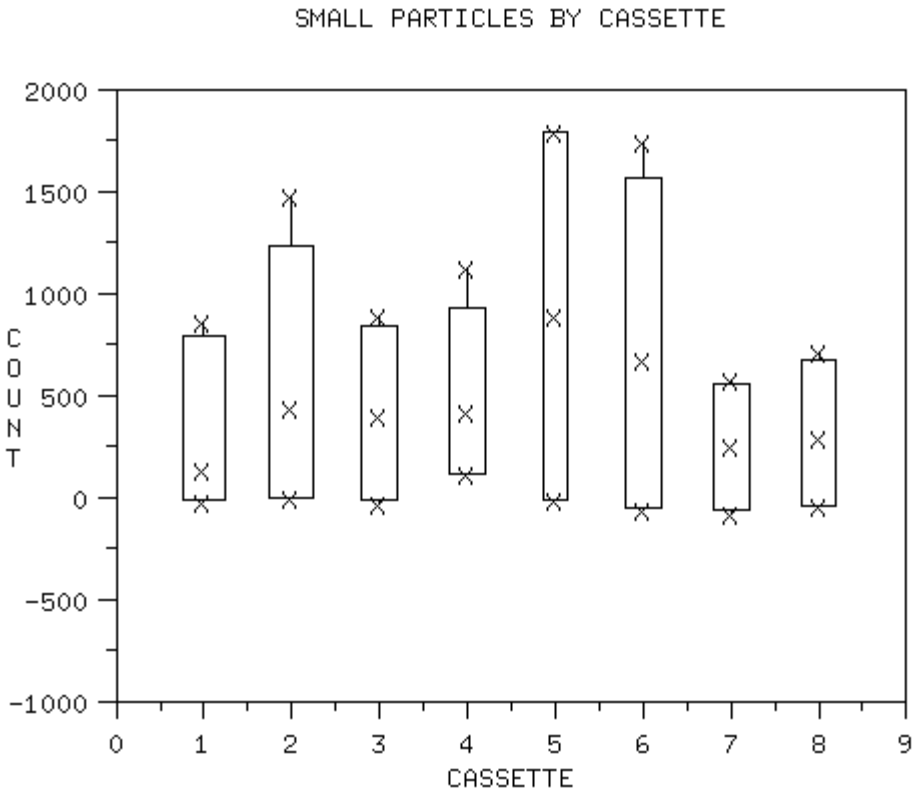
The second point is to be sure to graph the responses against implicit explanatory variables (such as observation order) as well as the explicit explanatory variables. There may be interesting insights in these *hidden* explanatory variables.

Example: wafer processing

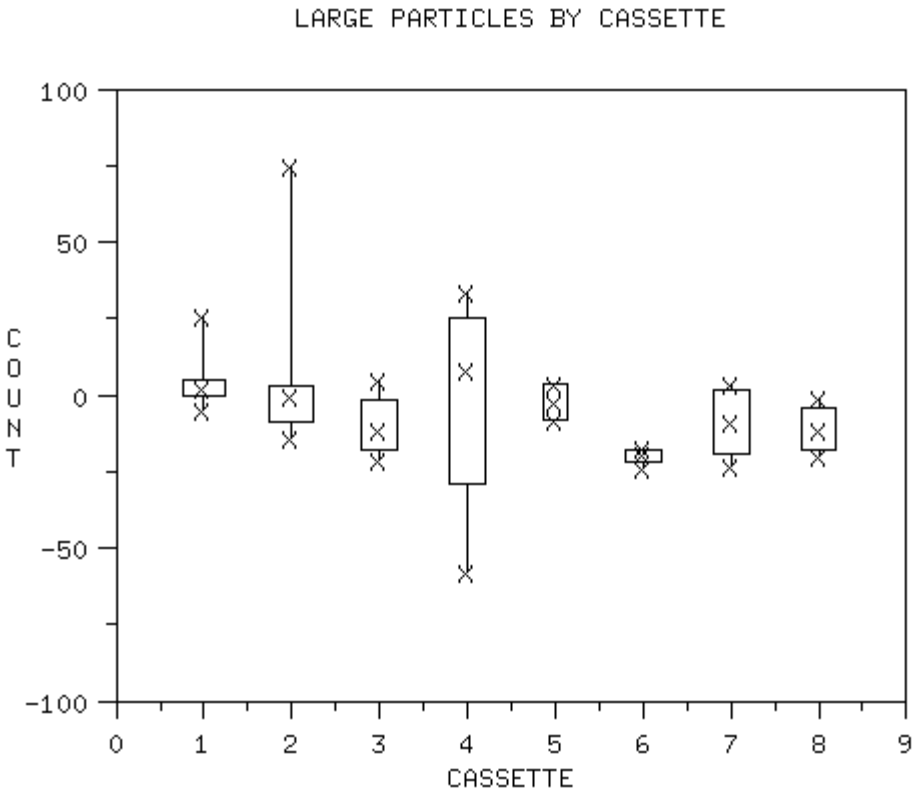
In the example below, we have collected data on the particles added to a wafer during a particular processing step. We ran a number of cassettes through the process and sampled wafers from certain slots in the cassette. We also kept track of which load lock the wafers passed through. This was done for two different process temperatures. We measured both small particles (< 2 microns) and large particles (> 2 microns). We plot the responses (particle counts) against each of the explanatory variables.

Cassette does not appear to be an important factor for small or large particles

This first graph is a box plot of the number of small particles added for each cassette type. The "X"'s in the plot represent the maximum, median, and minimum number of particles.



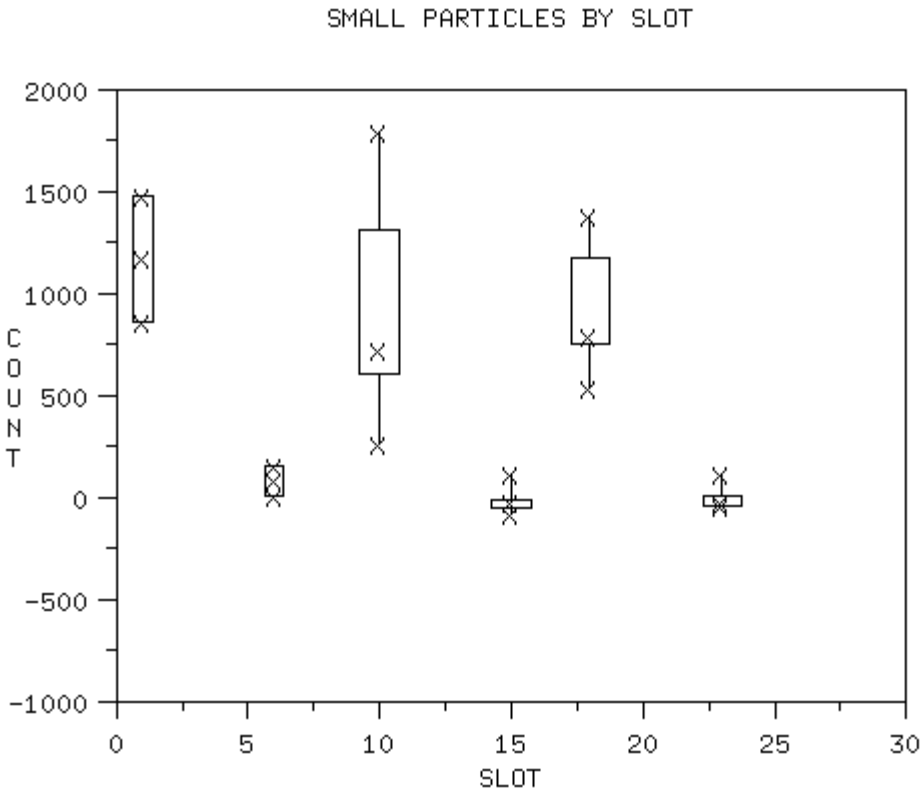
The second graph is a box plot of the number of large particles added for each cassette type.



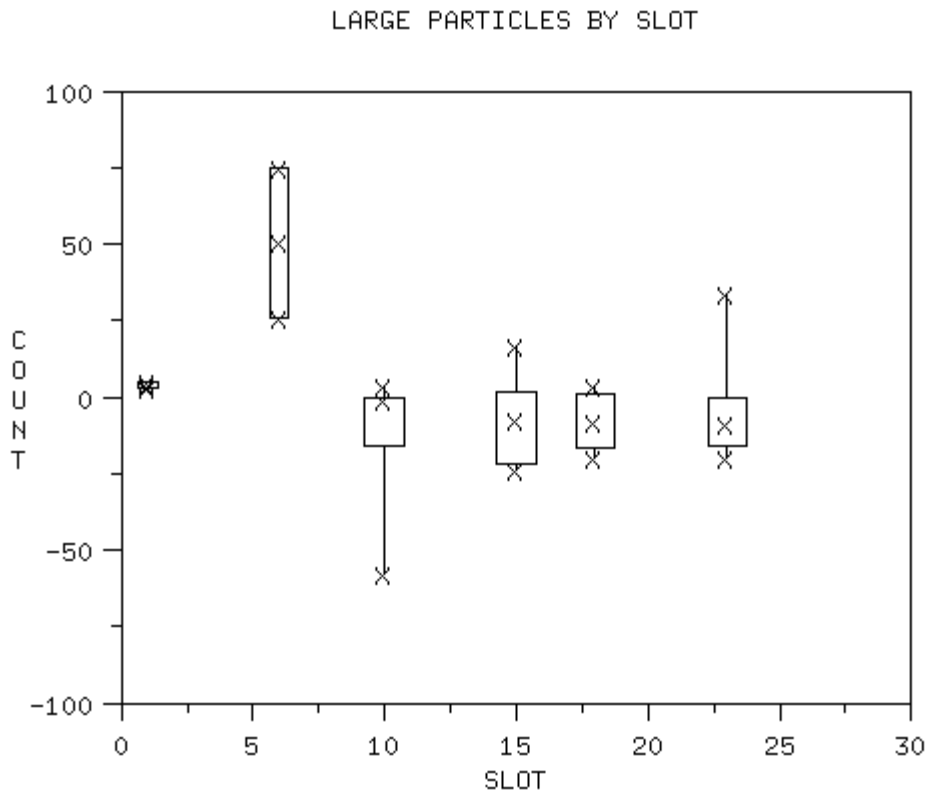
We conclude from these two box plots that cassette does not appear to be an important factor for small or large particles.

There is a difference between slots for small particles, one slot is different for large particles

We next generate box plots of small and large particles for the slot variable. First, the box plot for small particles.



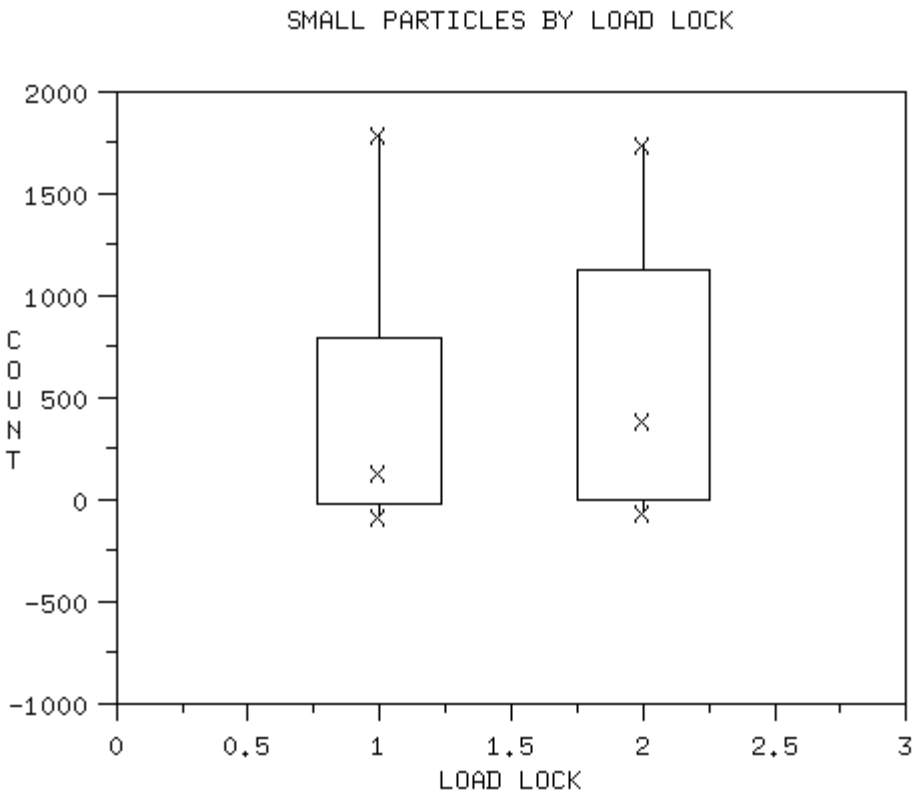
Next, the box plot for large particles.



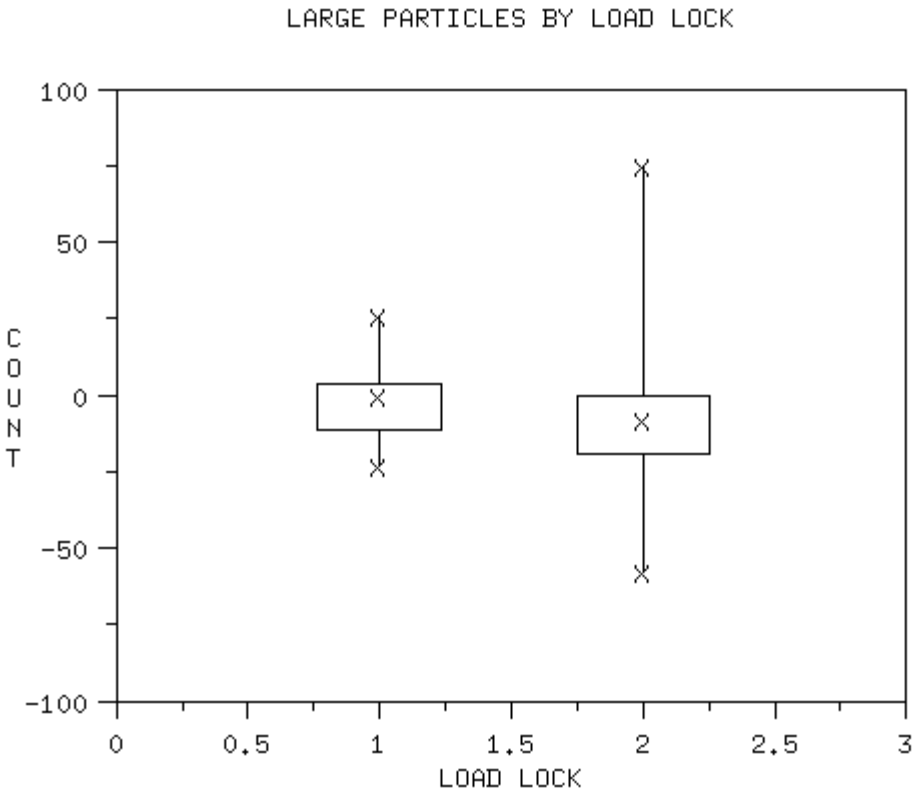
We conclude that there is a difference between slots for small particles. We also conclude that one slot appears to be different for large particles.

Load lock may have a slight effect for small and large particles

We next generate box plots of small and large particles for the load lock variable. First, the box plot for small particles.



Next, the box plot for large particles.

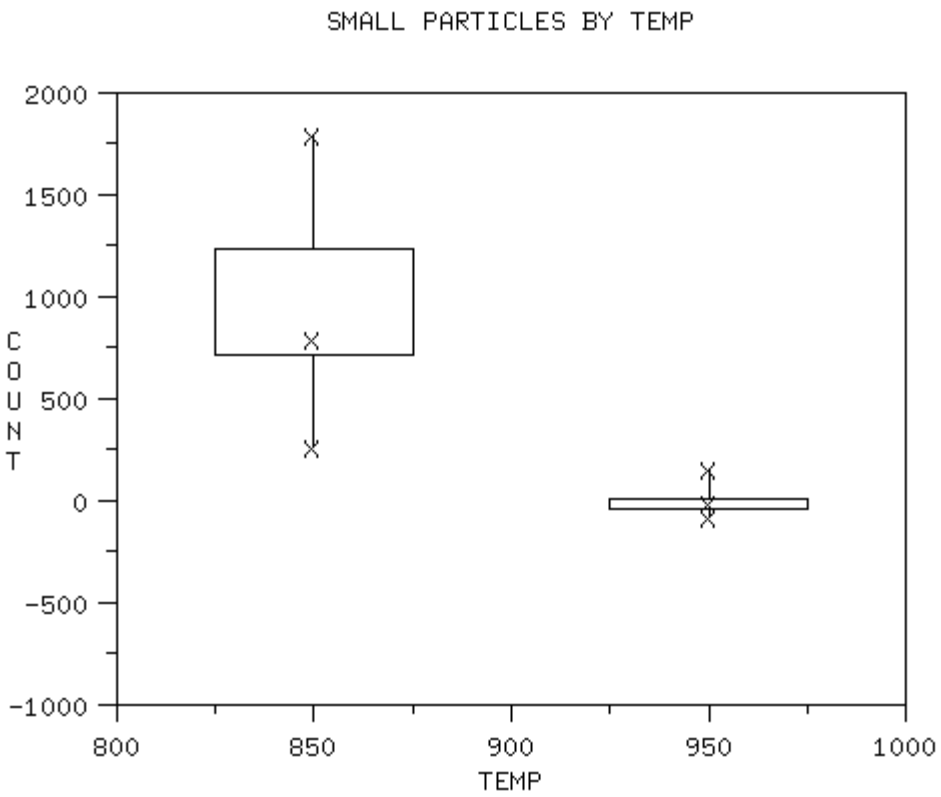


We conclude that there may be a slight effect for load lock for small and large

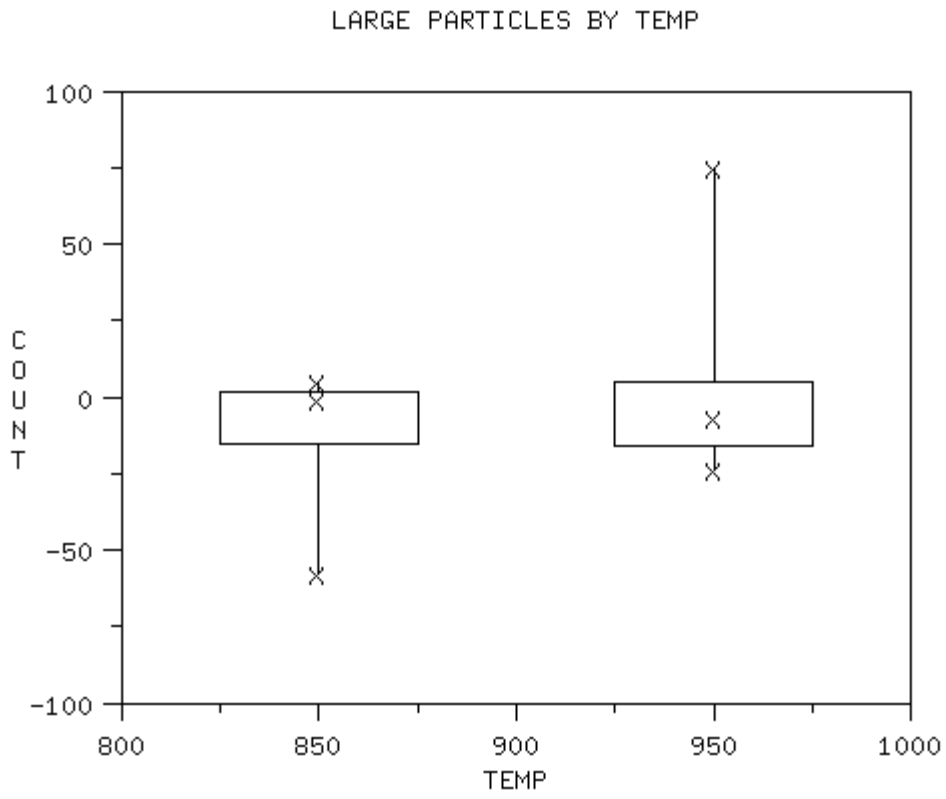
particles.

For small particles, temperature has a strong effect on both location and spread. For large particles, there may be a slight temperature effect but this may just be due to the outliers

We next generate box plots of small and large particles for the temperature variable. First, the box plot for small particles.



Next, the box plot for large particles.



'

We conclude that temperature has a strong effect on both location and spread for small particles. We conclude that there might be a small temperature effect for large particles, but this may just be due to outliers.

3.4.2.3. Exploring First Order Interactions

It is important to identify interactions

The final step (and perhaps the most important one) in the exploration phase is to find any first order interactions. When the difference in the response between the levels of one factor is not the same for all of the levels of another factor we say we have an interaction between those two factors. When we are trying to optimize responses based on factor settings, interactions provide for compromise.

The eyes can be deceiving - be careful

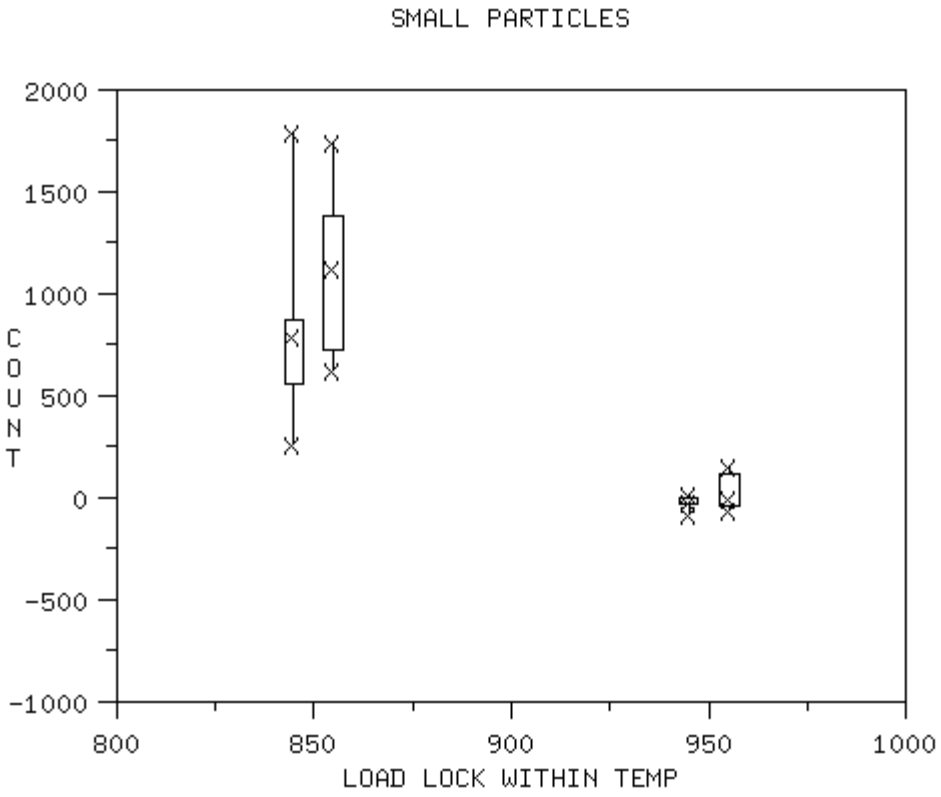
Interactions can be seen visually by using nested box plots. However, caution should be exercised when identifying interactions through graphical means alone. Any graphically identified interactions should be verified by numerical methods as well.

Previous example continued

To continue the previous example, given below are nested box plots of the small and large particles. The load lock is nested within the two temperature values. There is some evidence of possible interaction between these two factors. The effect of load lock is stronger at the lower temperature than at the higher one. This effect is stronger for the smaller particles than for the larger ones. As this example illustrates, when you have significant interactions the main effects must be interpreted conditionally. That is, the main effects do not tell the whole story by themselves.

For small particles, the load lock effect is not as strong for high temperature as it is for low temperature

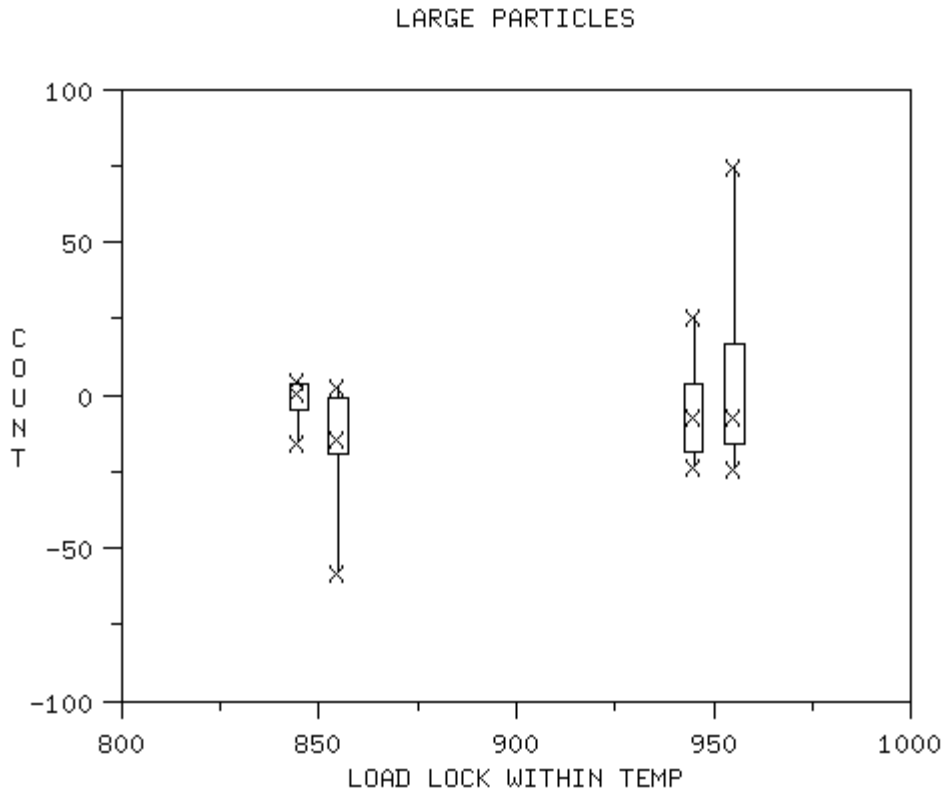
The following is the box plot of small particles for load lock nested within temperature.



We conclude from this plot that for small particles, the load lock effect is not as strong for high temperature as it is for low temperature.

The same may be true for large particles but not as strongly

The following is the box plot of large particles for load lock nested within temperature.



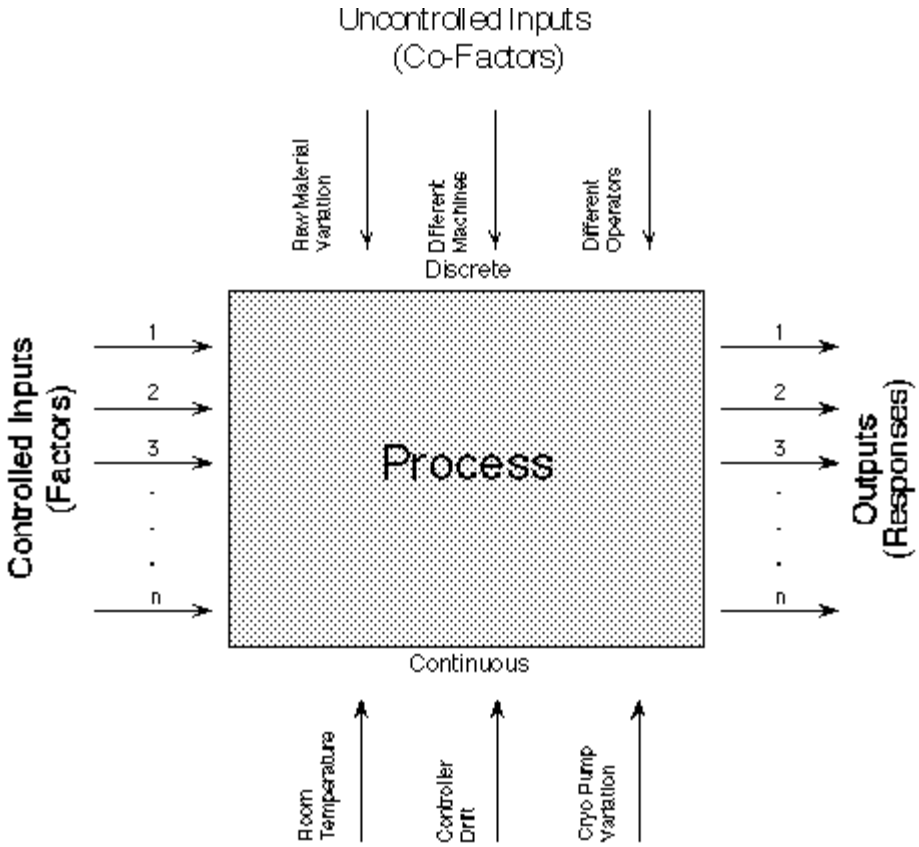
We conclude from this plot that for large particles, the load lock effect may not be as strong for high temperature as it is for low temperature. However, this effect is not as strong as it is for small particles.

3.4.3. Building Models

Black box models

When we develop a data collection plan we build *black box* models of the process we are studying like the one below:

In our data collection plan we drew process model pictures



Numerical models are explicit representations of our process model pictures

In the Exploring Relationships section, we looked at how to identify the input/output relationships through graphical methods. However, if we want to quantify the relationships and test them for statistical significance, we must resort to building mathematical models.

Polynomial models are generic descriptors of our output surface

There are two cases that we will cover for building mathematical models. If our goal is to develop an empirical prediction equation or to identify statistically significant explanatory variables and quantify their influence on output responses, we typically build polynomial models. As the name implies, these are polynomial functions (typically linear or quadratic functions) that describe the relationships between the explanatory variables and the response variable.

Physical models describe the underlying physics of our processes

On the other hand, if our goal is to fit an existing theoretical equation, then we want to build physical models. Again, as the name implies, this pertains to the case when we already have equations representing the physics involved in the process and we want to estimate specific parameter values.

3. Production Process Characterization
3.4. Data Analysis for PPC
3.4.3. Building Models

3.4.3.1. Fitting Polynomial Models

Polynomial models are a great tool for determining which input factors drive responses and in what direction

We use polynomial models to estimate and predict the shape of response values over a range of input parameter values. Polynomial models are a great tool for determining which input factors drive responses and in what direction. These are also the most common models used for analysis of designed experiments. A quadratic (second-order) polynomial model for two explanatory variables has the form of the equation below. The single x-terms are called the main effects. The squared terms are called the quadratic effects and are used to model curvature in the response surface. The cross-product terms are used to model interactions between the explanatory variables.

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{11} x_1^2 + \alpha_{22} x_2^2 + \alpha_{12} x_1 x_2 + \varepsilon$$

We generally don't need more than second-order equations

In most engineering and manufacturing applications we are concerned with at most second-order polynomial models. Polynomial equations obviously could become much more complicated as we increase the number of explanatory variables and hence the number of cross-product terms. Fortunately, we rarely see significant interaction terms above the two-factor level. This helps to keep the equations at a manageable level.

Use multiple regression to fit polynomial models

When the number of factors is small (less than 5), the complete polynomial equation can be fitted using the technique known as multiple regression. When the number of factors is large, we should use a technique known as *stepwise regression*. Most statistical analysis programs have a stepwise regression capability. We just enter all of the terms of the polynomial models and let the software choose which terms best describe the data. For a more thorough discussion of this topic and some examples, refer to the process improvement chapter.

3.4.3.2. Fitting Physical Models

Sometimes we want to use a physical model

Sometimes, rather than approximating response behavior with polynomial models, we know and can model the physics behind the underlying process. In these cases we would want to fit *physical models* to our data. This kind of modeling allows for better prediction and is less subject to variation than polynomial models (as long as the underlying process doesn't change).

We will use a CMP process to illustrate

We will illustrate this concept with an example. We have collected data on a chemical/mechanical planarization process (CMP) at a particular semiconductor processing step. In this process, wafers are polished using a combination of chemicals in a polishing slurry using polishing pads. We polished a number of wafers for differing periods of time in order to calculate material removal rates.

CMP removal rate can be modeled with a non-linear equation

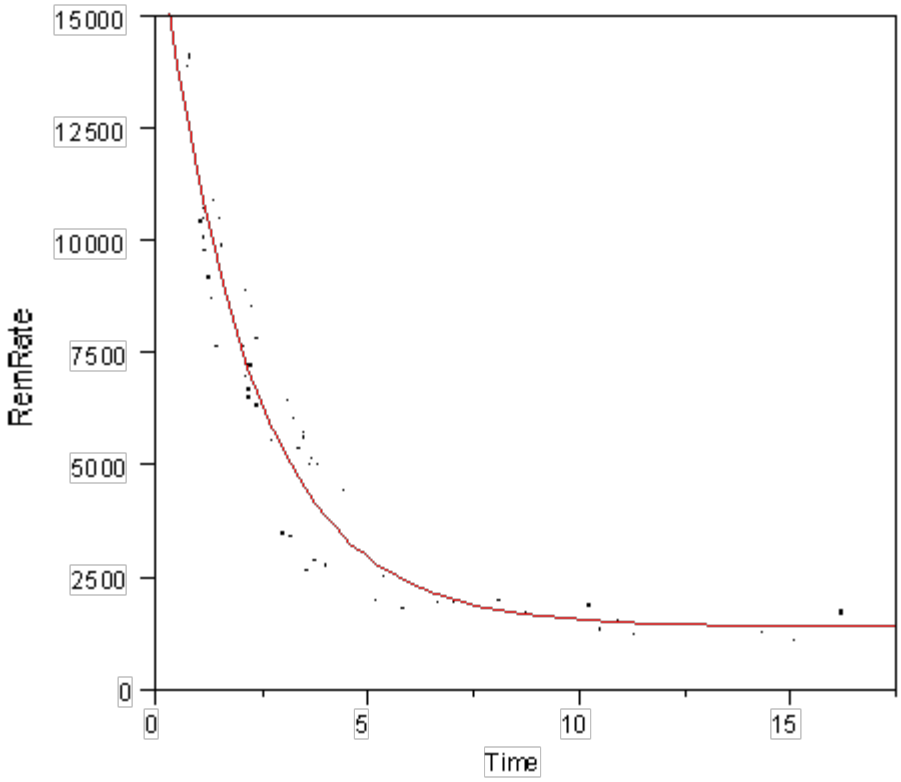
From first principles we know that removal rate changes with time. Early on, removal rate is high and as the wafer becomes more planar the removal rate declines. This is easily modeled with an exponential function of the form:

$$removal\ rate = p1 + p2 \times exp^{\,p3 \times time}$$

where $p1$, $p2$, and $p3$ are the parameters we want to estimate.

A non-linear regression routine was used to fit the data to the equation

The equation was fit to the data using a non-linear regression routine. A plot of the original data and the fitted line are given in the image below. The fit is quite good. This fitted equation was subsequently used in process optimization work.


15000
12500
10000
7500
5000
2500
0
RemRate
0
5
10
15
Time

3.4.4. Analyzing Variance Structure

Studying variation is important in PPC

One of the most common activities in process characterization work is to study the variation associated with the process and to try to determine the important sources of that variation. This is called *analysis of variance*. Refer to the section of this chapter on ANOVA models for a discussion of the theory behind this kind of analysis.

The key is to know the structure

The key to performing an analysis of variance is identifying the *structure* represented by the data. In the ANOVA models section we discussed one-way layouts and two-way layouts where the factors are either crossed or nested. Review these sections if you want to learn more about ANOVA structural layouts.

To perform the analysis, we just identify the structure, enter the data for each of the factors and levels into a statistical analysis program and then interpret the ANOVA table and other output. This is all illustrated in the example below.

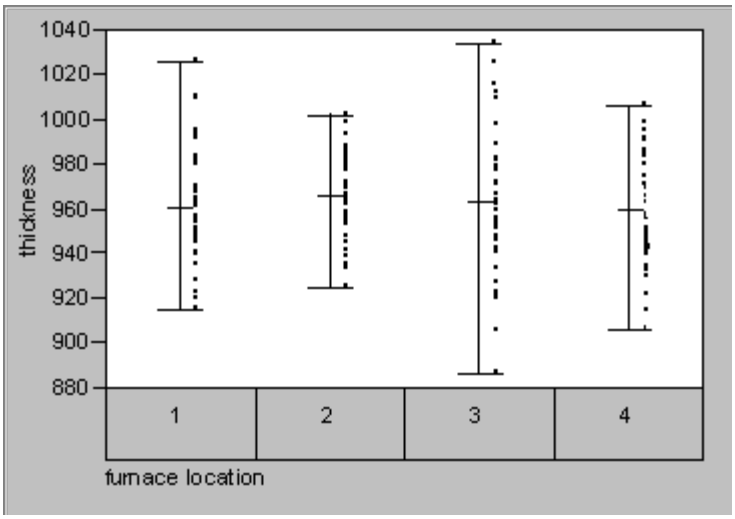
Example: furnace oxide thickness with a 1-way layout

The example is a furnace operation in semiconductor manufacture where we are growing an oxide layer on a wafer. Each lot of wafers is placed on quartz containers (boats) and then placed in a long tube-furnace. They are then raised to a certain temperature and held for a period of time in a gas flow. We want to understand the important factors in this operation. The furnace is broken down into four sections (zones) and two wafers from each lot in each zone are measured for the thickness of the oxide layer.

Look at effect of zone location on oxide thickness

The first thing to look at is the effect of zone location on the oxide thickness. This is a classic one-way layout. The factor is furnace zone and we have four levels. A plot of the data and an ANOVA table are given below.

The zone effect is masked by the lot-to-lot variation



ANOVA table

Analysis of Variance

Source	DF	SS	Mean Square	F Ratio	Prob > F
Zone	3	912.6905	304.23	0.467612	0.70527
Within	164	106699.1	650.604		

Let's account for lot with a nested layout

From the graph there does not appear to be much of a zone effect; in fact, the ANOVA table indicates that it is not significant. The problem is that variation due to lots is so large that it is masking the zone effect. We can fix this by adding a factor for lot. By treating this as a nested two-way layout, we obtain the ANOVA table below.

Now both lot and zone are revealed as important

Analysis of Variance

Source	DF	SS	Mean Square	F Ratio	Prob > F
Lot	20	61442.29	3072.11	5.37404	1.39e-7
Zone[lot]	63	36014.5	571.659	4.72864	3.9e-11
Within	84	10155	120.893		

Conclusions

Since the "Prob > F" is less than 0.05, for both lot and zone, we know that these factors are statistically significant at the 0.05 significance level.

3.4.5. Assessing Process Stability

A process is stable if it has a constant mean and a constant variance over time

A manufacturing process cannot be released to production until it has been proven to be stable. Also, we cannot begin to talk about process capability until we have demonstrated stability in our process. A process is said to be stable when all of the response parameters that we use to measure the process have both constant means and constant variances over time, and also have a constant distribution. This is equivalent to our earlier definition of controlled variation.

The graphical tool we use to assess stability is the scatter plot or the control chart

The graphical tool we use to assess process stability is the scatter plot. We collect a sufficient number of independent samples (greater than 100) from our process over a sufficiently long period of time (this can be specified in days, hours of processing time or number of parts processed) and plot them on a scatter plot with sample order on the x-axis and the sample value on the y-axis. The plot should look like constant random variation about a constant mean. Sometimes it is helpful to calculate control limits and plot them on the scatter plot along with the data. The two plots in the controlled variation example are good illustrations of stable and unstable processes.

Numerically, we assess its stationarity using the autocorrelation function

Numerically, we evaluate process stability through a times series analysis concept know as stationarity. This is just another way of saying that the process has a constant mean and a constant variance. The numerical technique used to assess stationarity is the autocovariance function.

Graphical methods usually good enough

Typically, graphical methods are good enough for evaluating process stability. The numerical methods are generally only used for modeling purposes.

3.4.6. Assessing Process Capability

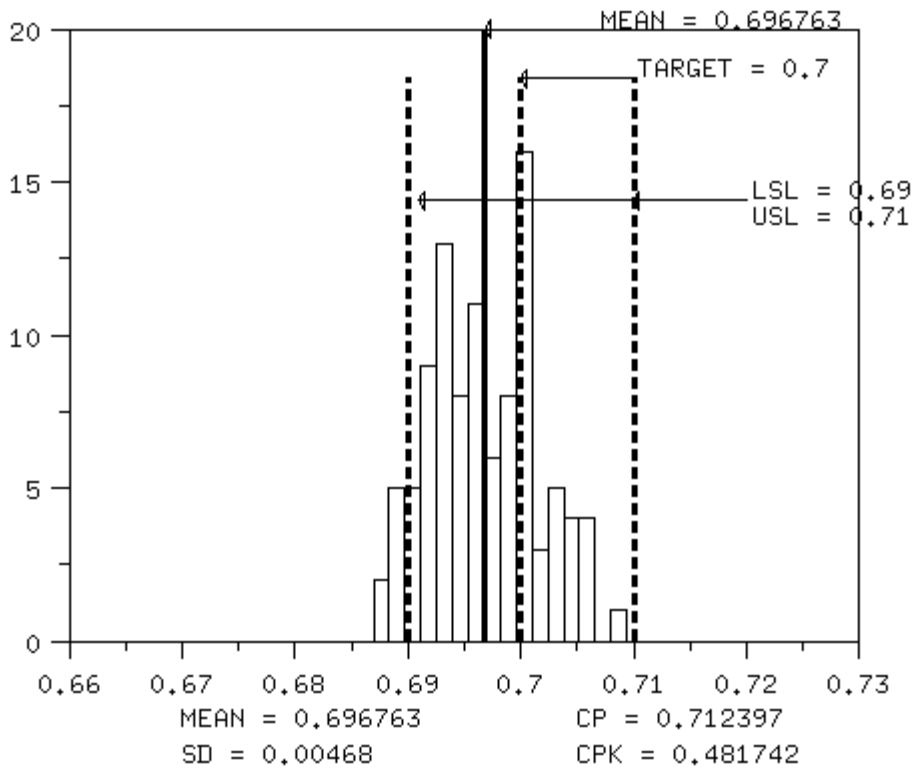
Capability compares a process against its specification

Process capability analysis entails comparing the performance of a process against its specifications. We say that a process is capable if virtually all of the possible variable values fall within the specification limits.

Use a capability chart

Graphically, we assess process capability by plotting the process specification limits on a histogram of the observations. If the histogram falls within the specification limits, then the process is capable. This is illustrated in the graph below. Note how the process is shifted below target and the process variation is too large. This is an example of an incapable process.

Notice how the process is off target and has too much variation



Numerically, we measure capability with a capability index. The general equation for the capability index, C_p, is:

Numerically, we use the C

p index

$$Cp = \frac{USL - LSL}{6s}$$

Interpretation of the C_p index

This equation just says that the measure of our process capability is how much of our observed process variation is covered by the process specifications. In this case the process variation is measured by 6 standard deviations (+/- 3 on each side of the mean). Clearly, if C_p > 1.0, then the process specification covers almost all of our process observations.

C_p does not account for process that is off center

The only problem with with the C_p index is that it does not account for a process that is off-center. We can modify this equation slightly to account for off-center processes to obtain the C_{pk} index as follows:

Or the C_{pk} index

$$Cpk = \min\left[\frac{USL - \bar{x}}{3s}, \frac{\bar{x} - LSL}{3s}\right]$$

C_{pk} accounts for a process being off center

This equation just says to take the minimum distance between our specification limits and the process mean and divide it by 3 standard deviations to arrive at the measure of process capability. This is all covered in more detail in the process capability section of the process monitoring chapter. For the example above, note how the C_{pk} value is less than the C_p value. This is because the process distribution is not centered between the specification limits.

3.4.7. Checking Assumptions

Check the normality of the data

Many of the techniques discussed in this chapter, such as hypothesis tests, control charts and capability indices, assume that the underlying structure of the data can be adequately modeled by a normal distribution. Many times we encounter data where this is not the case.

Some causes of non-normality

There are several things that could cause the data to appear non-normal, such as:

- The data come from two or more different sources. This type of data will often have a multi-modal distribution. This can be solved by identifying the reason for the multiple sets of data and analyzing the data separately.
- The data come from an unstable process. This type of data is nearly impossible to analyze because the results of the analysis will have no credibility due to the changing nature of the process.
- The data were generated by a stable, yet fundamentally non-normal mechanism. For example, particle counts are non-normal by the very nature of the particle generation process. Data of this type can be handled using transformations.

We can sometimes transform the data to make it look normal

For the last case, we could try transforming the data using what is known as a *power transformation*. The power transformation is given by the equation:

$$Y^{(\lambda)} = \begin{cases} y^{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

where Y represents the data and lambda is the transformation value. Lambda is typically any value between -2 and 2. Some of the more common values for lambda are 0, 1/2, and -1, which give the following transformations:

$$\ln(y), \quad \sqrt{y}, \quad \frac{1}{y}$$

General algorithm for trying to make non-normal data approximately normal

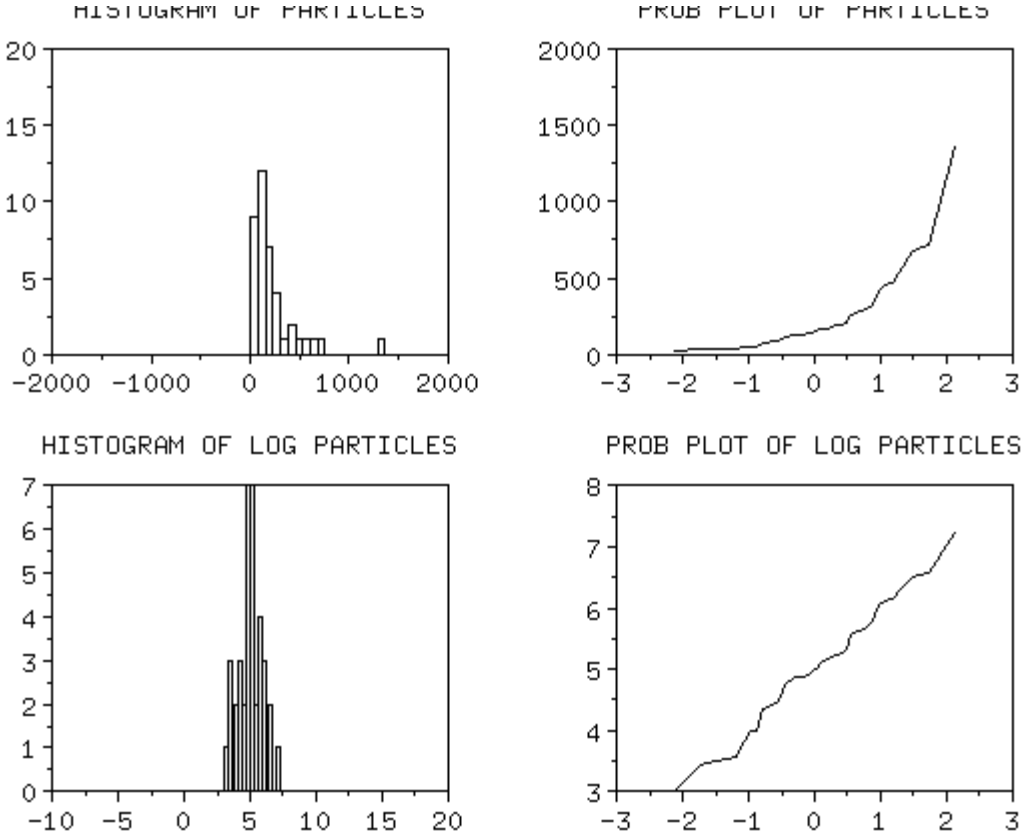
The general algorithm for trying to make non-normal data appear to be approximately normal is to:

1. Determine if the data are non-normal. (Use normal probability plot and histogram).
2. Find a transformation that makes the data look approximately normal, if possible. Some data sets may include zeros (i.e., particle data). If the data set does include zeros, you must first add a constant value to the data and then transform the results.

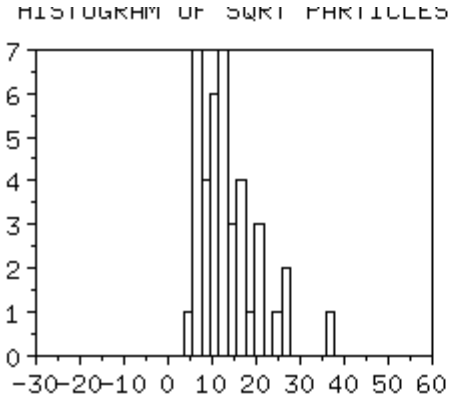
Example: particle count data

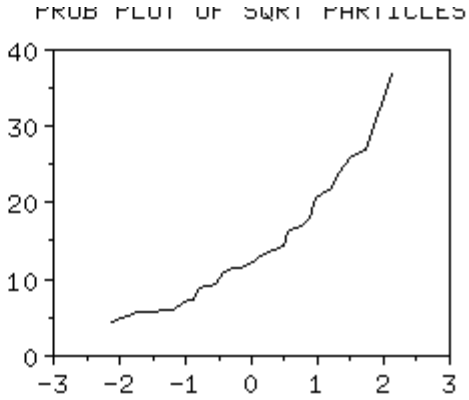
As an example, let's look at some particle count data from a semiconductor processing step. Count data are inherently non-normal. Below are histograms and normal probability plots for the original data and the ln, sqrt and inverse of the data. You can see that the log transform does the best job of making the data appear as if it is normal. All analyses can be performed on the log-transformed data and the assumptions will be approximately satisfied.

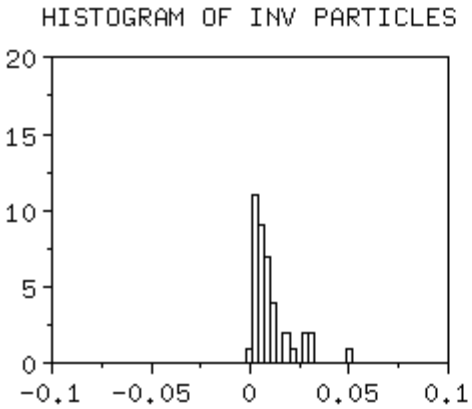
The original data is non-normal, the log transform looks fairly normal

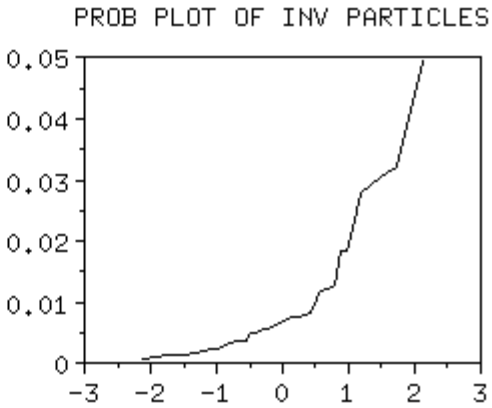


Neither the square root nor the inverse transformation looks normal


HISTOGRAM OF SQRT PARTICLES
0
1
2
3
4
5
6
7
-30 -20 -10 0 10 20 30 40 50 60


PROB PLOT OF SQRT PARTICLES
0
10
20
30
40
-3 -2 -1 0 1 2 3


HISTOGRAM OF INV PARTICLES
0
5
10
15
20
-0.1 -0.05 0 0.05 0.1


PROB PLOT OF INV PARTICLES
0
0.01
0.02
0.03
0.04
0.05
-3 -2 -1 0 1 2 3

3. Production Process Characterization

3.5. Case Studies

Summary This section presents several case studies that demonstrate the application of production process characterizations to specific problems.

Table of Contents The following case studies are available.

1. Furnace Case Study
2. Machine Case Study

3.5.1. Furnace Case Study

Introduction This case study analyzes a furnace oxide growth process.

Table of Contents The case study is broken down into the following steps.

1. Background and Data
2. Initial Analysis of Response Variable
3. Identify Sources of Variation
4. Analysis of Variance
5. Final Conclusions
6. Work This Example Yourself

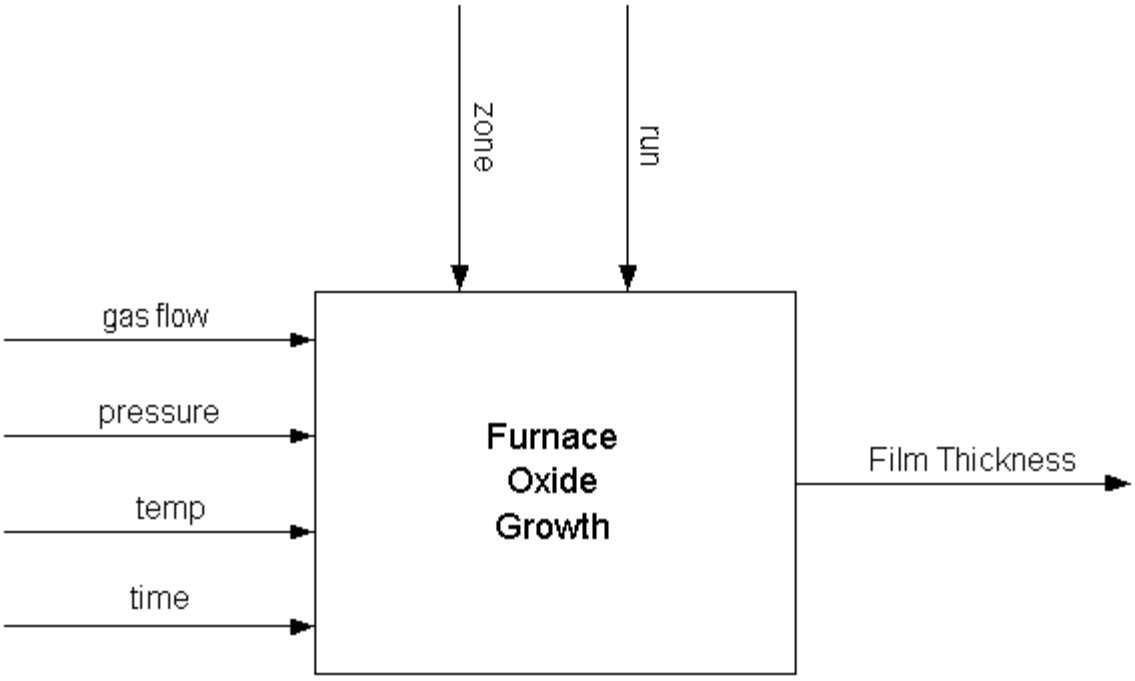
3.5.1.1. Background and Data

Introduction In a semiconductor manufacturing process flow, we have a step whereby we grow an oxide film on the silicon wafer using a furnace. In this step, a cassette of wafers is placed in a quartz "boat" and the boats are placed in the furnace. The furnace can hold four boats. A gas flow is created in the furnace and it is brought up to temperature and held there for a specified period of time (which corresponds to the desired oxide thickness). This study was conducted to determine if the process was stable and to characterize sources of variation so that a process control strategy could be developed.

Goal The goal of this study is to determine if this process is capable of consistently growing oxide films with a thickness of 560 Angstroms +/- 100 Angstroms. An additional goal is to determine important sources of variation for use in the development of a process control strategy.

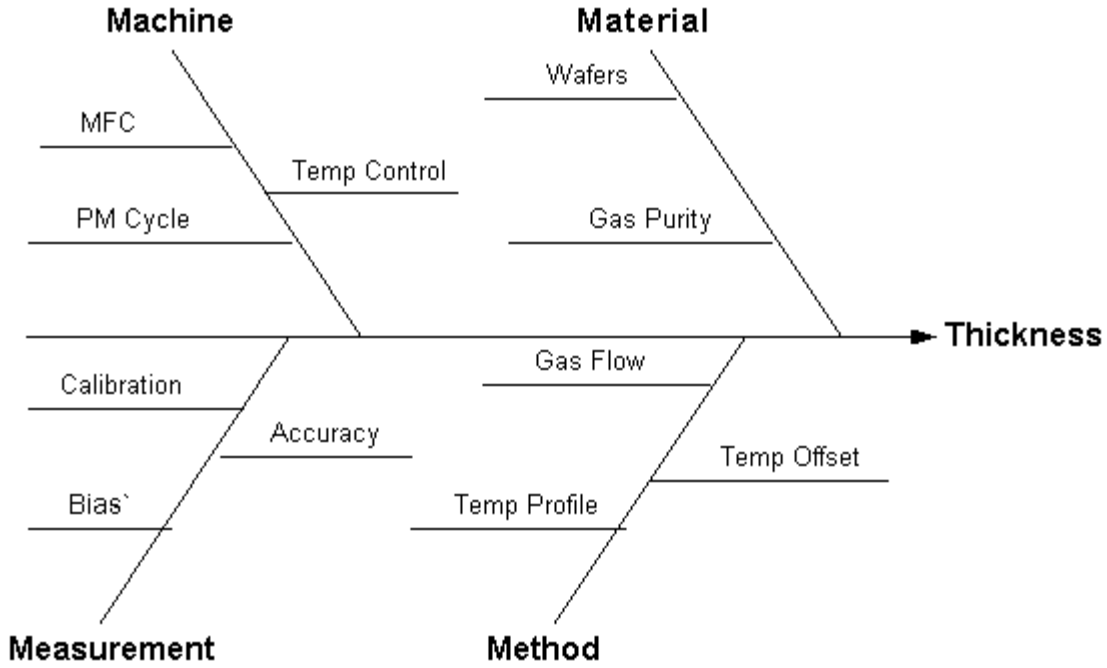
Software The analyses used in this case study can be generated using both Dataplot code and R code.

Process Model In the picture below we are modeling this process with one output (film thickness) that is influenced by four controlled factors (gas flow, pressure, temperature and time) and two uncontrolled factors (run and zone). The four controlled factors are part of our recipe and will remain constant throughout this study. We know that there is run-to-run variation that is due to many different factors (input material variation, variation in consumables, etc.). We also know that the different zones in the furnace have an effect. A zone is a region of the furnace tube that holds one boat. There are four zones in these tubes. The zones in the middle of the tube grow oxide a little bit differently from the ones on the ends. In fact, there are temperature offsets in the recipe to help minimize this problem.



Sensitivity Model

The sensitivity model for this process is fairly straightforward and is given in the figure below. The effects of the machin are mostly related to the preventative maintenance (PM) cycle. We want to make sure the quartz tube has been cleaned recently, the mass flow controllers are in good shape and the temperature controller has been calibrated recently. The same is true of the measurement equipment where the thickness readings will be taken. We want to make sure a gauge study has been performed. For material, the incoming wafers will certainly have an effect on the outgoing thickness as well as the quality of the gases used. Finally, the recipe will have an effect including gas flow, temperature offset for the different zones, and temperature profile (how quickly we raise the temperature, how long we hold it and how quickly we cool it off).



Sampling Plan

Given our goal statement and process modeling, we can now define a sampling plan. The primary goal is to determine if the process is capable. This just means that we need to monitor the process over some period of time and compare the estimates of process location and spread to the specifications. An additional goal is to identify sources of variation to aid in setting up a process control strategy. Some obvious sources of variation are incoming wafers, run-to-run variability, variation due to operators or shift, and variation due to zones within a furnace tube. One additional constraint that we must work under is that this study should not have a significant impact on normal production operations.

Given these constraints, the following sampling plan was selected. It was decided to monitor the process for one day (three shifts). Because this process is operator independent, we will not keep shift or operator information but just record run number. For each run, we will randomly assign cassettes of wafers to a zone. We will select two wafers from each zone after processing and measure two sites on each wafer. This plan should give reasonable estimates of run-to-run variation and within zone variability as well as good overall estimates of process location and spread.

We are expecting readings around 560 Angstroms. We would not expect many readings above 700 or below 400. The measurement equipment is accurate to within 0.5 Angstroms which is well within the accuracy needed for this study.

Data

The following are the data that were collected for this study.

RUN	ZONE	WAFER	THICKNESS
1	1	1	546

1	1	2	540
1	2	1	566
1	2	2	564
1	3	1	577
1	3	2	546
1	4	1	543
1	4	2	529
2	1	1	561
2	1	2	556
2	2	1	577
2	2	2	553
2	3	1	563
2	3	2	577
2	4	1	556
2	4	2	540
3	1	1	515
3	1	2	520
3	2	1	548
3	2	2	542
3	3	1	505
3	3	2	487
3	4	1	506
3	4	2	514
4	1	1	568
4	1	2	584
4	2	1	570
4	2	2	545
4	3	1	589
4	3	2	562
4	4	1	569
4	4	2	571
5	1	1	550
5	1	2	550
5	2	1	562
5	2	2	580
5	3	1	560
5	3	2	554
5	4	1	545
5	4	2	546
6	1	1	584
6	1	2	581
6	2	1	567
6	2	2	558
6	3	1	556
6	3	2	560
6	4	1	591
6	4	2	599
7	1	1	593
7	1	2	626
7	2	1	584
7	2	2	559
7	3	1	634
7	3	2	598
7	4	1	569
7	4	2	592
8	1	1	522
8	1	2	535
8	2	1	535
8	2	2	581
8	3	1	527
8	3	2	520
8	4	1	532
8	4	2	539
9	1	1	562
9	1	2	568
9	2	1	548
9	2	2	548
9	3	1	533
9	3	2	553
9	4	1	533
9	4	2	521
10	1	1	555
10	1	2	545
10	2	1	584
10	2	2	572
10	3	1	546
10	3	2	552
10	4	1	586
10	4	2	584
11	1	1	565
11	1	2	557

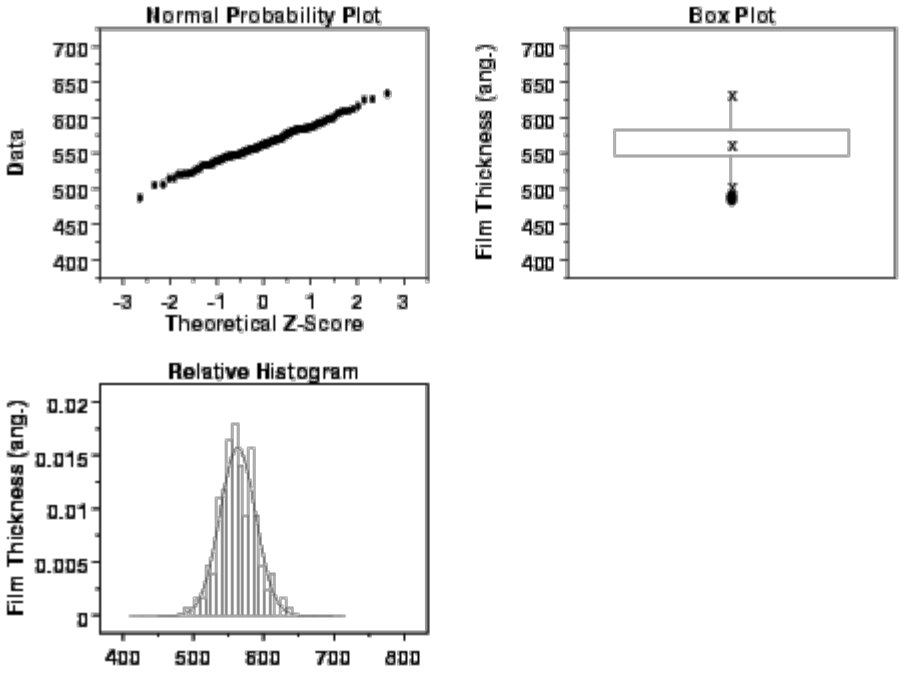
11	2	1	583
11	2	2	585
11	3	1	582
11	3	2	567
11	4	1	549
11	4	2	533
12	1	1	548
12	1	2	528
12	2	1	563
12	2	2	588
12	3	1	543
12	3	2	540
12	4	1	585
12	4	2	586
13	1	1	580
13	1	2	570
13	2	1	556
13	2	2	569
13	3	1	609
13	3	2	625
13	4	1	570
13	4	2	595
14	1	1	564
14	1	2	555
14	2	1	585
14	2	2	588
14	3	1	564
14	3	2	583
14	4	1	563
14	4	2	558
15	1	1	550
15	1	2	557
15	2	1	538
15	2	2	525
15	3	1	556
15	3	2	547
15	4	1	534
15	4	2	542
16	1	1	552
16	1	2	547
16	2	1	563
16	2	2	578
16	3	1	571
16	3	2	572
16	4	1	575
16	4	2	584
17	1	1	549
17	1	2	546
17	2	1	584
17	2	2	593
17	3	1	567
17	3	2	548
17	4	1	606
17	4	2	607
18	1	1	539
18	1	2	554
18	2	1	533
18	2	2	535
18	3	1	522
18	3	2	521
18	4	1	547
18	4	2	550
19	1	1	610
19	1	2	592
19	2	1	587
19	2	2	587
19	3	1	572
19	3	2	612
19	4	1	566
19	4	2	563
20	1	1	569
20	1	2	609
20	2	1	558
20	2	2	555
20	3	1	577
20	3	2	579
20	4	1	552
20	4	2	558
21	1	1	595
21	1	2	583
21	2	1	599

21	2	2	602
21	3	1	598
21	3	2	616
21	4	1	580
21	4	2	575

3.5.1.2. Initial Analysis of Response Variable

Initial Plots of Response Variable

The initial step is to assess data quality and to look for anomalies. This is done by generating a normal probability plot, a histogram, and a boxplot. For convenience, these are generated on a single page.



Conclusions From the Plots

We can make the following conclusions based on these initial plots.

- The box plot indicates one outlier. However, this outlier is only slightly smaller than the other numbers.
- The normal probability plot and the histogram (with an overlaid normal density) indicate that this data set is reasonably approximated by a normal distribution.

Parameter Estimates

Parameter estimates for the film thickness are summarized in the following table.

Parameter Estimates

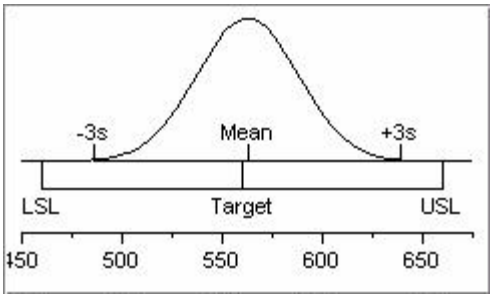
Type	Parameter	Estimate	Lower (95%) Confidence Bound	Upper (95%) Confidence Bound
Location	Mean	563.0357	559.1692	566.9023
Dispersion	Standard Deviation	25.3847	22.9297	28.4331

Quantiles Quantiles for the film thickness are summarized in the following table.

Quantiles for Film Thickness

100.0%	Maximum	634.00
99.5%		634.00
97.5%		615.10
90.0%		595.00
75.0%	Upper Quartile	582.75
50.0%	Median	562.50
25.0%	Lower Quartile	546.25
10.0%		532.90
2.5%		514.23
0.5%		487.00
0.0%	Minimum	487.00

Capability Analysis From the above preliminary analysis, it looks reasonable to proceed with the capability analysis.



The lower specification limit is 460, the upper specification limit is 660, and the target specification is 560.

Percent Defective We summarize the percent defective (i.e., the number of items outside the specification limits) in the following table.

Percentage Outside Specification Limits

Specification	Value	Percent	Actual	Theoretical (% Based On

				Normal)
Lower Specification Limit	460	Percent Below LSL = 100*Φ ((LSL - $\bar{y}$)/*s)*	0.0000	0.0025%
Upper Specification Limit	660	Percent Above USL = 100*(1 - Φ ((USL - $\bar{y}$)/*s))*	0.0000	0.0067%
Specification Target	560	Combined Percent Below LSL and Above USL	0.0000	0.0091%
Standard Deviation	25.38468			

The value Φ denotes the normal cumulative distribution function, $\bar{y}$ the sample mean, and *s* the sample standard deviation.

Capability Index Statistics

We summarize various capability index statistics in the following table.

Capability Index Statistics

Capability Statistic	Index	Lower CI	Upper CI
CP	1.313	1.172	1.454
CPK	1.273	1.128	1.419
CPM	1.304	1.165	1.442
CPL	1.353	1.218	1.488
CPU	1.273	1.142	1.404

Conclusions

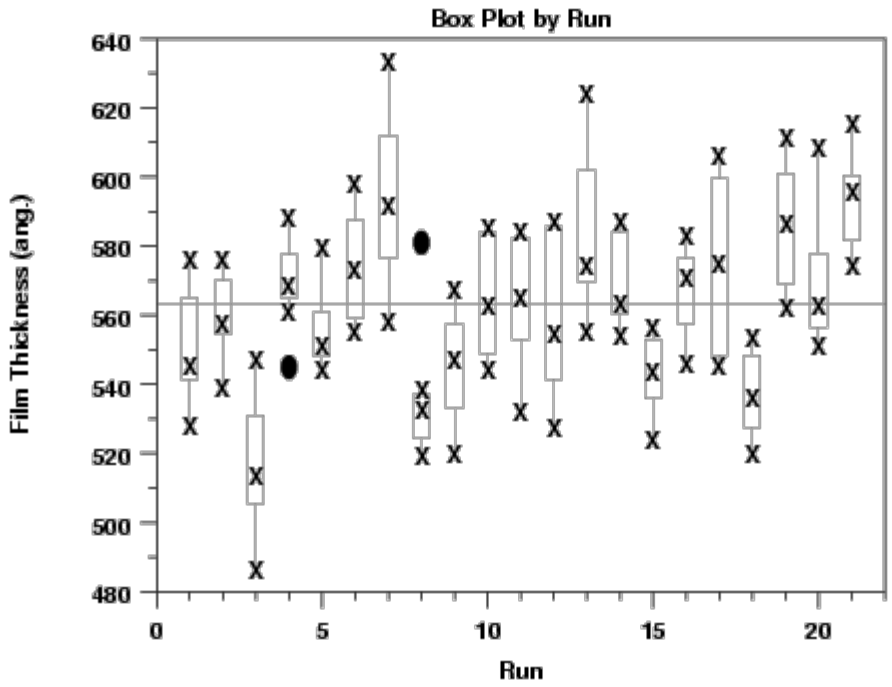
The above capability analysis indicates that the process is capable and we can proceed with the analysis.

3.5.1.3. Identify Sources of Variation

The next part of the analysis is to break down the sources of variation.

Box Plot by Run

The following is a box plot of the thickness by run number.



Conclusions From Box Plot

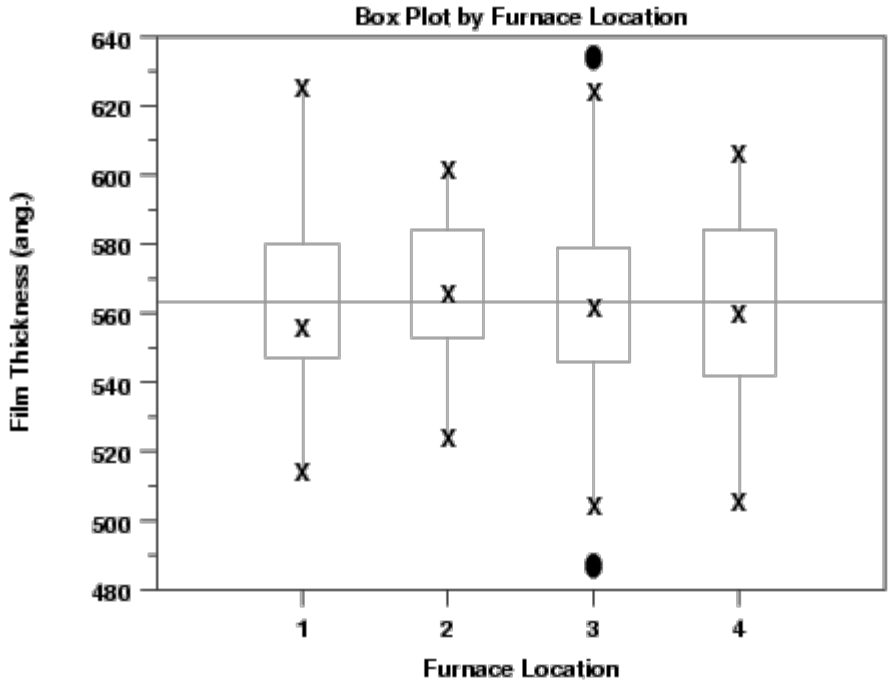
We can make the following conclusions from this box plot.

1. There is significant run-to-run variation.
2. Although the means of the runs are different, there is no discernable trend due to run.
3. In addition to the run-to-run variation, there is significant within-run variation as well. This suggests that a box plot by furnace location may be useful as well.

Box Plot by

The following is a box plot of the thickness by furnace location.

Furnace Location



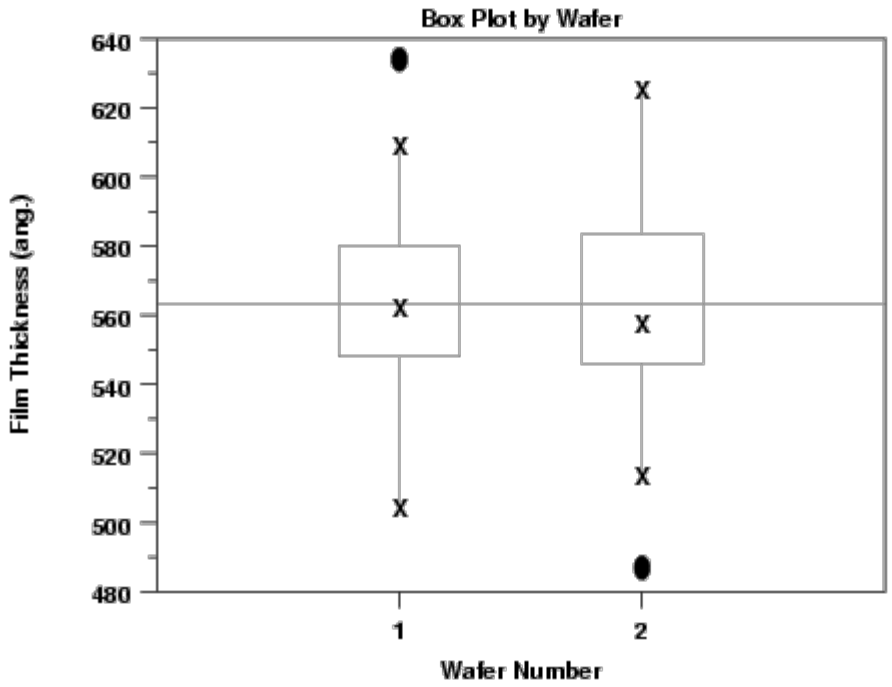
Conclusions From Box Plot

We can make the following conclusions from this box plot.

1. There is considerable variation within a given furnace location.
2. The variation between furnace locations is small. That is, the locations and scales of each of the four furnace locations are fairly comparable (although furnace location 3 seems to have a few mild outliers).

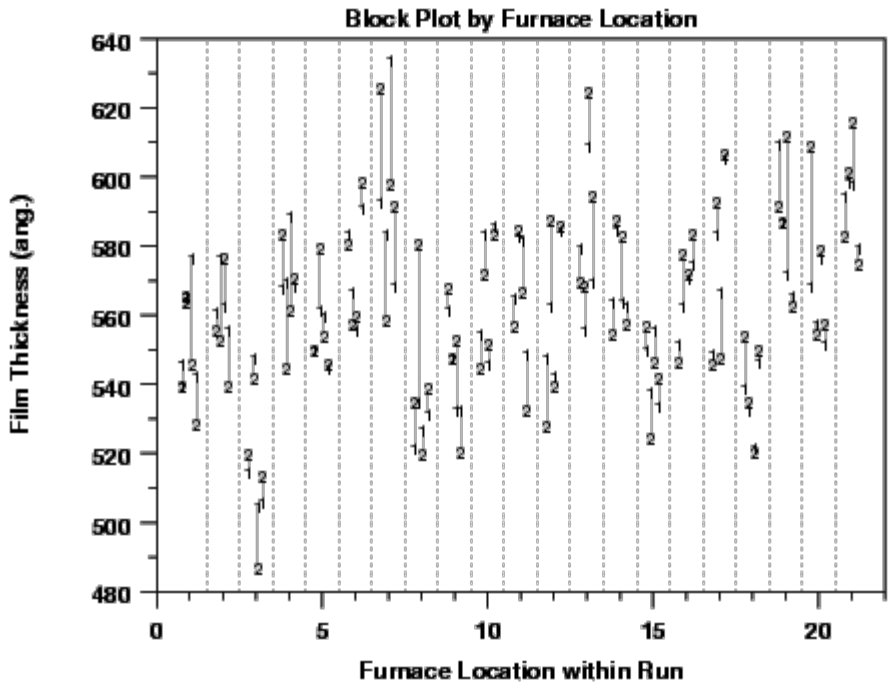
Box Plot by Wafer

The following is a box plot of the thickness by wafer.



Conclusion From Box Plot From this box plot, we conclude that wafer does not seem to be a significant factor.

Block Plot In order to show the combined effects of run, furnace location, and wafer, we draw a block plot of the thickness. Note that for aesthetic reasons, we have used connecting lines rather than enclosing boxes.



Conclusions From Block Plot

We can draw the following conclusions from this block plot.

1. There is significant variation both between runs and between furnace locations. The between-run variation appears to be greater.

2. Run 3 seems to be an outlier.

3.5.1.4. Analysis of Variance

Analysis of Variance

The next step is to confirm our interpretation of the plots in the previous section by running a nested analysis of variance.

Analysis of Variance

Source	Degrees of Freedom	Sum of Squares	Mean Square Error	F Ratio	Prob > F
Run	20	61,442.29	3,072.11	5.37404	0.0000001
Furnace Location [Run]	63	36,014.5	571.659	4.72864	3.85e-11
Within	84	10,155	120.893		
Total	167	107,611.8	644.382		

Components of Variance

From the above analysis of variance table, we can compute the components of variance. Recall that for this data set we have 2 wafers measured at 4 furnace locations for 21 runs. This leads to the following set of equations.

3072.11 = (4*2)*Var(Run) + 2*Var(Furnace Location) + Var(Within)
571.659 = 2*Var(Furnace Location) + Var(Within)
120.893 = Var(Within)

Solving these equations yields the following components of variance.

Components of Variance

Component	Variance Component	Percent of Total	Sqrt(Variance Component)
Run	312.55694	47.44	17.679
Furnace Location[Run]	225.38294	34.21	15.013
Within	120.89286	18.35	10.995

3.5.1.5. Final Conclusions

Final Conclusions

This simple study of a furnace oxide growth process indicated that the process is capable and showed that both run-to-run and zone-within-run are significant sources of variation. We should take this into account when designing the control strategy for this process. The results also pointed to where we should look when we perform process improvement activities.

3.5.1.6. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot, if you have downloaded and installed it. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window and the Data Sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Get set up and started. 1. Read in the data.	1. You have read 4 columns of numbers into Dataplot, variables run, zone, wafer, and filmthic.
2. Analyze the response variable. 1. Normal probability plot, box plot, and histogram of film thickness. 2. Compute summary statistics and quantiles of film thickness.	1. Initial plots indicate that the film thickness is reasonably approximated by a normal distribution with no significant outliers. 2. Mean is 563.04 and standard

3. Perform a capability analysis.	deviation is 25.38. Data range from 487 to 634. 3. Capability analysis indicates that the process is capable.
3. Identify Sources of Variation. 1. Generate a box plot by run. 2. Generate a box plot by furnace location. 3. Generate a box plot by wafer. 4. Generate a block plot.	1. The box plot shows significant variation both between runs and within runs. 2. The box plot shows significant variation within furnace location but not between furnace location. 3. The box plot shows no significant effect for wafer. 4. The block plot shows both run and furnace location are significant.
4. Perform an Analysis of Variance 1. Perform the analysis of variance and compute the components of variance.	1. The results of the ANOVA are summarized in an ANOVA table and a components of variance table.

3.5.2. Machine Screw Case Study

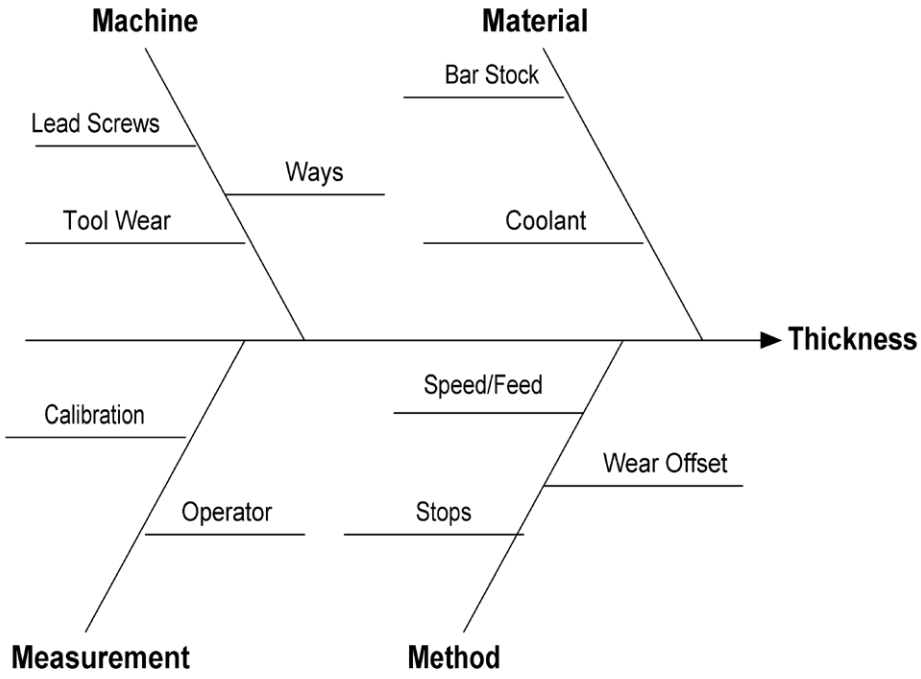
Introduction This case study analyzes three automatic screw machines with the intent of replacing one of them.

Table of Contents The case study is broken down into the following steps.

1. Background and Data
2. Box Plots by Factor
3. Analysis of Variance
4. Throughput
5. Final Conclusions
6. Work This Example Yourself

3.5.2.1. Background and Data

Introduction	A machine shop has three automatic screw machines that produce various parts. The shop has enough capital to replace one of the machines. The quality control department has been asked to conduct a study and make a recommendation as to which machine should be replaced. It was decided to monitor one of the most commonly produced parts (an $1/8^{th}$ inch diameter pin) on each of the machines and see which machine is the least stable.
Goal	The goal of this study is to determine which machine is least stable in manufacturing a steel pin with a diameter of .125 +/- .003 inches. Stability will be measured in terms of a constant variance about a constant mean. If all machines are stable, the decision will be based on process variability and throughput. Namely, the machine with the highest variability and lowest throughput will be selected for replacement.
Software	The analyses used in this case study can be generated using both Dataplot code and R code.
Process Model	The process model for this operation is trivial and need not be addressed.
Sensitivity Model	The sensitivity model, however, is important and is given in the figure below. The material is not very important. All machines will receive barstock from the same source and the coolant will be the same. The method is important. Each machine is slightly different and the operator must make adjustments to the speed (how fast the part rotates), feed (how quickly the cut is made) and stops (where cuts are finished) for each machine. The same operator will be running all three machines simultaneously. Measurement is not too important. An experienced QC engineer will be collecting the samples and making the measurements. Finally, the machine condition is really what this study is all about. The wear on the ways and the lead screws will largely determine the stability of the machining process. Also, tool wear is important. The same type of tool inserts will be used on all three machines. The tool insert wear will be monitored by the operator and they will be changed as needed.



Sampling Plan

Given our goal statement and process modeling, we can now define a sampling plan. The primary goal is to determine if the process is stable and to compare the variances of the three machines. We also need to monitor throughput so that we can compare the productivity of the three machines.

There is an upcoming three-day run of the particular part of interest, so this study will be conducted on that run. There is a suspected time-of-day effect that we must account for. It is sometimes the case that the machines do not perform as well in the morning, when they are first started up, as they do later in the day. To account for this we will sample parts in the morning and in the afternoon. So as not to impact other QC operations too severely, it was decided to sample 10 parts, twice a day, for three days from each of the three machines. Daily throughput will be recorded as well.

We are expecting readings around .125 +/- .003 inches. The parts will be measured using a standard micrometer with readings recorded to 0.0001 of an inch. Throughput will be measured by reading the part counters on the machines at the end of each day.

Data

The following are the data that were collected for this study.

```
MACHINE      DAY         TIME       SAMPLE
DIAMETER
 (1-3)      (1-3)      1 = AM      (1-10)
(inches)
                       2 = PM
-------------------------------------------------
---
   1          1           1          1
0.1247
   1          1           1          2
```

```
0.1264
   1          1          1          3
0.1252
   1          1          1          4
0.1253
   1          1          1          5
0.1263
   1          1          1          6
0.1251
   1          1          1          7
0.1254
   1          1          1          8
0.1239
   1          1          1          9
0.1235
   1          1          1         10
0.1257
   1          1          2          1
0.1271
   1          1          2          2
0.1253
   1          1          2          3
0.1265
   1          1          2          4
0.1254
   1          1          2          5
0.1243
   1          1          2          6
0.124
   1          1          2          7
0.1246
   1          1          2          8
0.1244
   1          1          2          9
0.1271
   1          1          2         10
0.1241
   1          2          1          1
0.1251
   1          2          1          2
0.1238
   1          2          1          3
0.1255
   1          2          1          4
0.1234
   1          2          1          5
0.1235
   1          2          1          6
0.1266
   1          2          1          7
0.125
   1          2          1          8
0.1246
   1          2          1          9
0.1243
   1          2          1         10
0.1248
   1          2          2          1
0.1248
   1          2          2          2
0.1235
   1          2          2          3
0.1243
   1          2          2          4
0.1265
   1          2          2          5
0.127
   1          2          2          6
0.1229
   1          2          2          7
0.125
   1          2          2          8
0.1248
   1          2          2          9
0.1252
   1          2          2         10
0.1243
   1          3          1          1
0.1255
   1          3          1          2
0.1237
```

```
   1          3          1          3
0.1235
   1          3          1          4
0.1264
   1          3          1          5
0.1239
   1          3          1          6
0.1266
   1          3          1          7
0.1242
   1          3          1          8
0.1231
   1          3          1          9
0.1232
   1          3          1         10
0.1244
   1          3          2          1
0.1233
   1          3          2          2
0.1237
   1          3          2          3
0.1244
   1          3          2          4
0.1254
   1          3          2          5
0.1247
   1          3          2          6
0.1254
   1          3          2          7
0.1258
   1          3          2          8
0.126
   1          3          2          9
0.1235
   1          3          2         10
0.1273
   2          1          1          1
0.1239
   2          1          1          2
0.1239
   2          1          1          3
0.1239
   2          1          1          4
0.1231
   2          1          1          5
0.1221
   2          1          1          6
0.1216
   2          1          1          7
0.1233
   2          1          1          8
0.1228
   2          1          1          9
0.1227
   2          1          1         10
0.1229
   2          1          2          1
0.122
   2          1          2          2
0.1239
   2          1          2          3
0.1237
   2          1          2          4
0.1216
   2          1          2          5
0.1235
   2          1          2          6
0.124
   2          1          2          7
0.1224
   2          1          2          8
0.1236
   2          1          2          9
0.1236
   2          1          2         10
0.1217
   2          2          1          1
0.1247
   2          2          1          2
0.122
   2          2          1          3
```

```
0.1218
   2          2          1          4
0.1237
   2          2          1          5
0.1234
   2          2          1          6
0.1229
   2          2          1          7
0.1235
   2          2          1          8
0.1237
   2          2          1          9
0.1224
   2          2          1         10
0.1224
   2          2          2          1
0.1239
   2          2          2          2
0.1226
   2          2          2          3
0.1224
   2          2          2          4
0.1239
   2          2          2          5
0.1237
   2          2          2          6
0.1227
   2          2          2          7
0.1218
   2          2          2          8
0.122
   2          2          2          9
0.1231
   2          2          2         10
0.1244
   2          3          1          1
0.1219
   2          3          1          2
0.1243
   2          3          1          3
0.1231
   2          3          1          4
0.1223
   2          3          1          5
0.1218
   2          3          1          6
0.1218
   2          3          1          7
0.1225
   2          3          1          8
0.1238
   2          3          1          9
0.1244
   2          3          1         10
0.1236
   2          3          2          1
0.1231
   2          3          2          2
0.1223
   2          3          2          3
0.1241
   2          3          2          4
0.1215
   2          3          2          5
0.1221
   2          3          2          6
0.1236
   2          3          2          7
0.1229
   2          3          2          8
0.1205
   2          3          2          9
0.1241
   2          3          2         10
0.1232
   3          1          1          1
0.1255
   3          1          1          2
0.1215
   3          1          1          3
0.1219
```

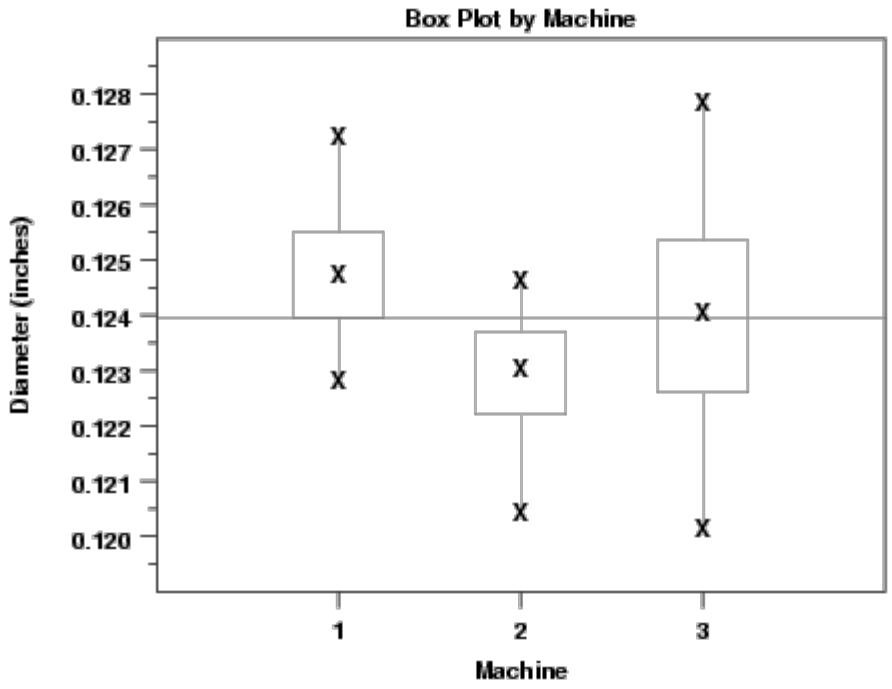
```
   3          1          1          4
0.1253
   3          1          1          5
0.1232
   3          1          1          6
0.1266
   3          1          1          7
0.1271
   3          1          1          8
0.1209
   3          1          1          9
0.1212
   3          1          1         10
0.1249
   3          1          2          1
0.1228
   3          1          2          2
0.126
   3          1          2          3
0.1242
   3          1          2          4
0.1236
   3          1          2          5
0.1248
   3          1          2          6
0.1243
   3          1          2          7
0.126
   3          1          2          8
0.1231
   3          1          2          9
0.1234
   3          1          2         10
0.1246
   3          2          1          1
0.1207
   3          2          1          2
0.1279
   3          2          1          3
0.1268
   3          2          1          4
0.1222
   3          2          1          5
0.1244
   3          2          1          6
0.1225
   3          2          1          7
0.1234
   3          2          1          8
0.1244
   3          2          1          9
0.1207
   3          2          1         10
0.1264
   3          2          2          1
0.1224
   3          2          2          2
0.1254
   3          2          2          3
0.1237
   3          2          2          4
0.1254
   3          2          2          5
0.1269
   3          2          2          6
0.1236
   3          2          2          7
0.1248
   3          2          2          8
0.1253
   3          2          2          9
0.1252
   3          2          2         10
0.1237
   3          3          1          1
0.1217
   3          3          1          2
0.122
   3          3          1          3
0.1227
   3          3          1          4
```

```
0.1202
   3          3          1          5
0.127
   3          3          1          6
0.1224
   3          3          1          7
0.1219
   3          3          1          8
0.1266
   3          3          1          9
0.1254
   3          3          1         10
0.1258
   3          3          2          1
0.1236
   3          3          2          2
0.1247
   3          3          2          3
0.124
   3          3          2          4
0.1235
   3          3          2          5
0.124
   3          3          2          6
0.1217
   3          3          2          7
0.1235
   3          3          2          8
0.1242
   3          3          2          9
0.1247
   3          3          2         10
0.125
```

3.5.2.2. Box Plots by Factors

Initial Steps The initial step is to plot box plots of the measured diameter for each of the explanatory variables.

Box Plot by Machine The following is a box plot of the diameter by machine.

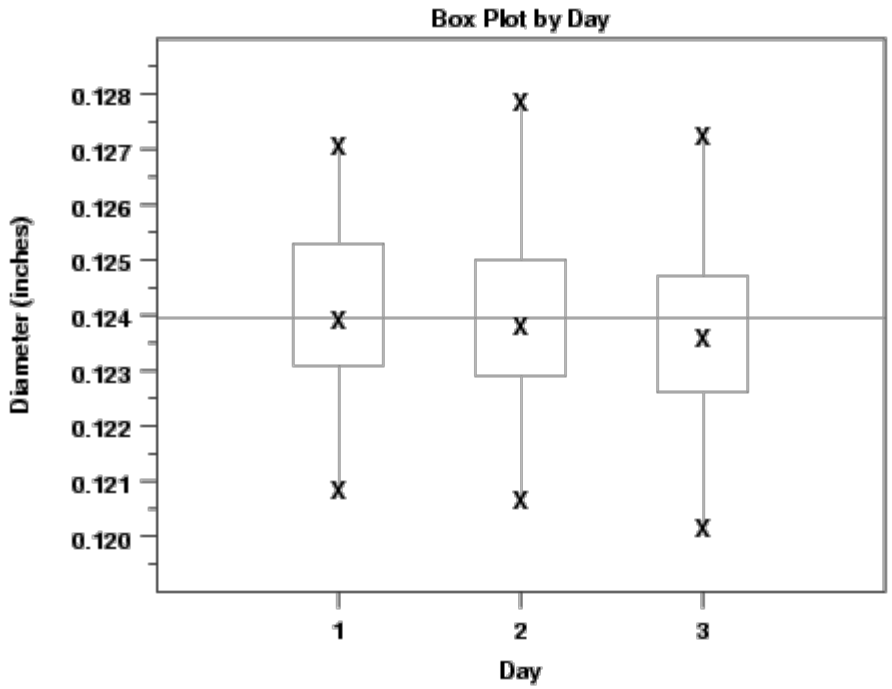


Conclusions From Box Plot We can make the following conclusions from this box plot.

1. The location appears to be significantly different for the three machines, with machine 2 having the smallest median diameter and machine 1 having the largest median diameter.

2. Machines 1 and 2 have comparable variability while machine 3 has somewhat larger variability.

Box Plot by Day The following is a box plot of the diameter by day.

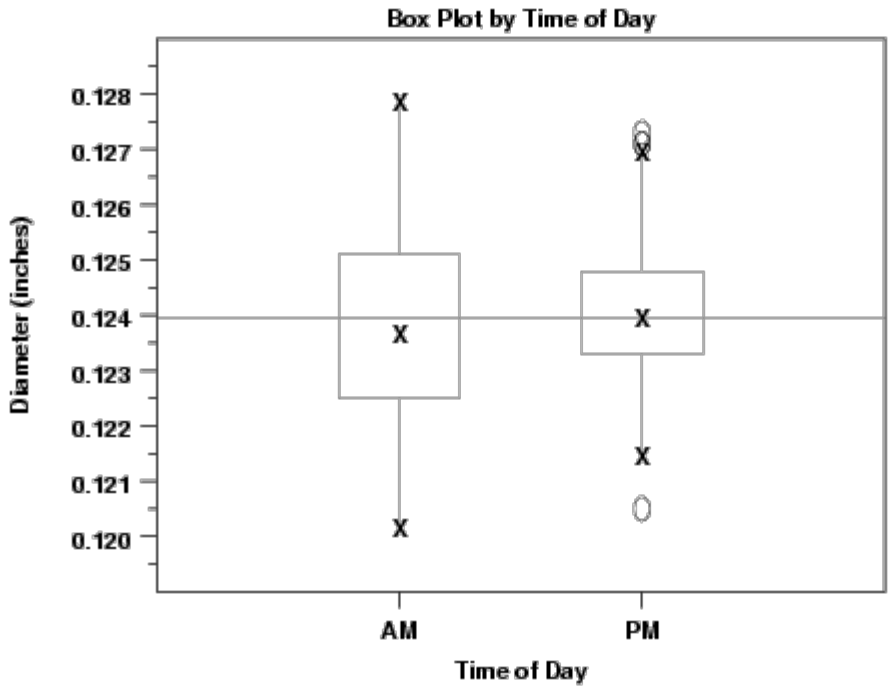


Conclusions From Box Plot

We can draw the following conclusion from this box plot. Neither the location nor the spread seem to differ significantly by day.

Box Plot by Time of Day

The following is a box plot of the time of day.

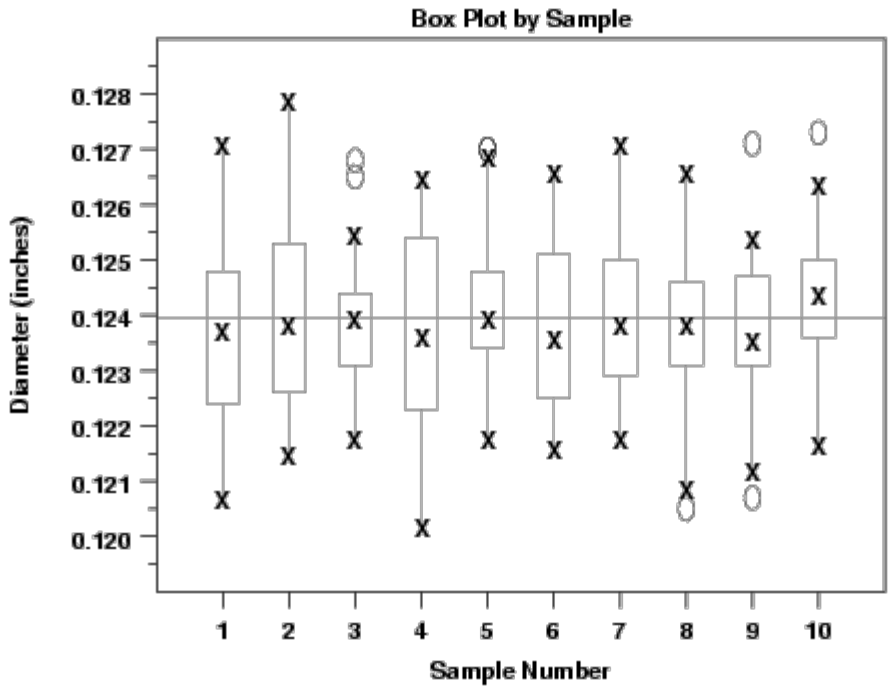


Conclusion From Box Plot

We can draw the following conclusion from this box plot. Neither the location nor the spread seem to differ significantly by time of day.

Box Plot by Sample Number

The following is a box plot of the sample number.



Conclusion From Box Plot

We can draw the following conclusion from this box plot. Although there are some minor differences in location and spread between the samples, these differences do not show a noticeable pattern and do not seem significant.

3.5.2.3. Analysis of Variance

Analysis of Variance Using All Factors

We can confirm our interpretation of the box plots by running an analysis of variance when all four factors are included.

Source	DF	Sum of Squares	Mean Square	F Statistic	Prob > F
Machine	2	0.000111	0.000055	29.3159	1.3e-11
Day	2	0.000004	0.000002	0.9884	0.37
Time	1	0.000002	0.000002	1.2478	0.27
Sample	9	0.000009	0.000001	0.5205	0.86
Residual	165	0.000312	0.000002		
Corrected Total	179	0.000437	0.000002		

Interpretation of ANOVA Output

We fit the model

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \tau_k + \phi_l + \epsilon_{ijklm}$$

which has an overall mean, as opposed to the model

$$Y_{ijklm} = A_i + B_j + C_k + D_l + \epsilon_{ijklm}$$

These models are mathematically equivalent. The effect estimates in the first model are relative to the overall mean. The effect estimates for the second model can be obtained by simply adding the overall mean to effect estimates from the first model.

Only the machine factor is statistically significant. This confirms what the box plots in the previous section had indicated graphically.

Analysis of Variance Using Only Machine

The previous analysis of variance indicated that only the machine factor was statistically significant. The following table displays the ANOVA results using only the machine factor.

Source	DF	Sum of Squares	Mean Square	F Statistic	Prob > F
Machine	2	0.000111	0.000055	30.0094	6.0E-12
Residual	177	0.000327	0.000002		
Corrected Total	179	0.000437	0.000002		

Interpretation of ANOVA Output

At this stage, we are interested in the level means for the machine variable. These can be summarized in the following table.

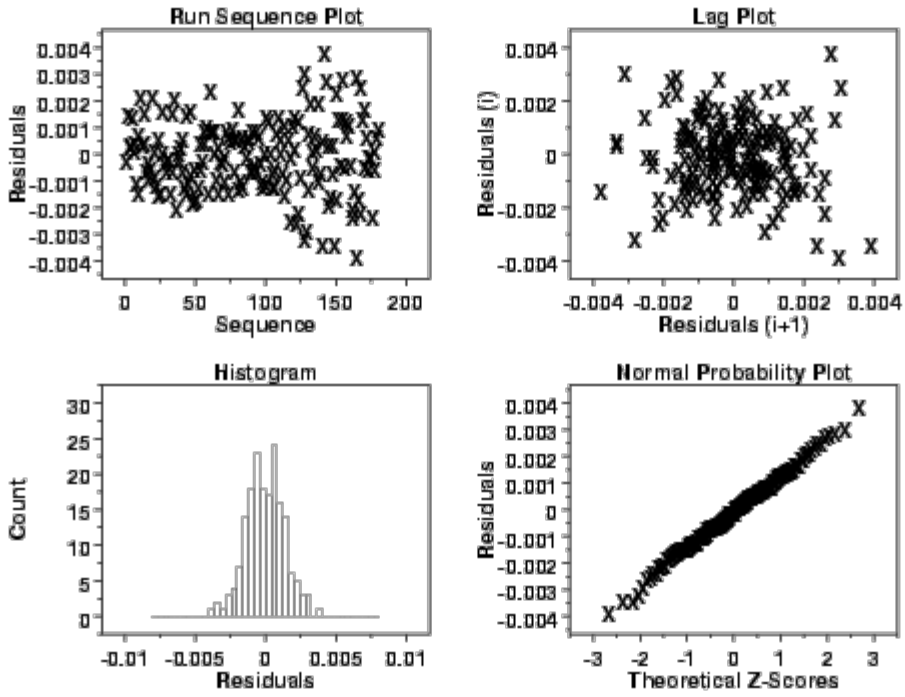
Machine Means for One-way ANOVA

Level	Number	Mean	Standard Error	Lower 95% CI	Upper 95% CI
1	60	0.124887	0.00018	0.12454	0.12523

2	60	0.122968	0.00018	0.12262	0.12331
3	60	0.124022	0.00018	0.12368	0.12437

Model Validation

As a final step, we validate the model by generating a 4-plot of the residuals.



The 4-plot does not indicate any significant problems with the ANOVA model.

3.5.2.4. Throughput

Summary of Throughput

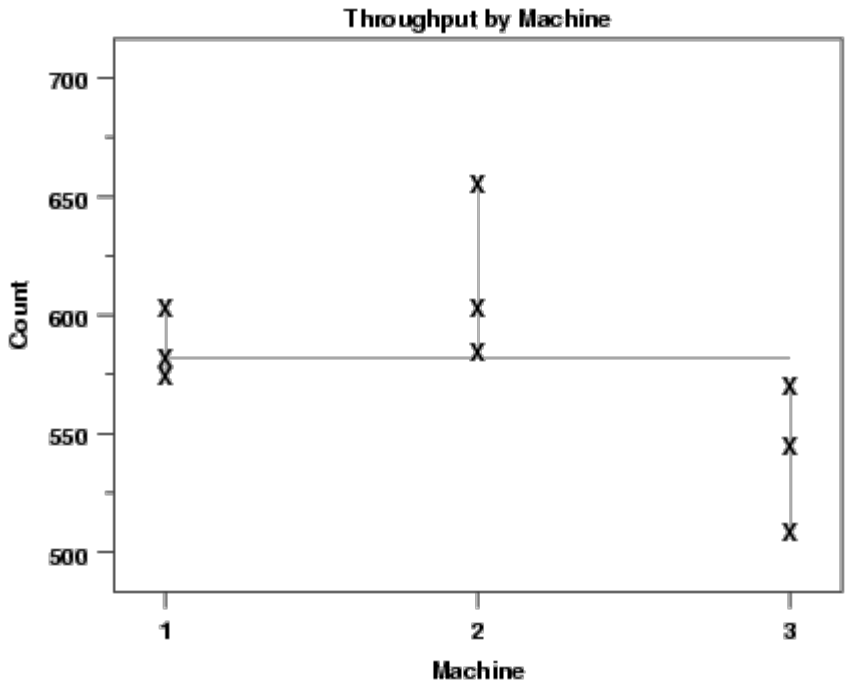
The throughput is summarized in the following table (this was part of the original data collection, not the result of analysis).

Machine	Day 1	Day 2	Day 3
1	576	604	583
2	657	604	586
3	510	546	571

This table shows that machine 3 had significantly lower throughput.

Graphical Representation of Throughput

We can show the throughput graphically.



The graph clearly shows the lower throughput for machine 3.

Analysis of Variance for Throughput

We can confirm the statistical significance of the lower throughput of machine 3 by running an analysis of variance.

```
Source            DF     Sum of       Mean    F Statistic    Prob > F
```

```
                          Squares      Square
-------------------------------------------------------------------
Machine             2    8216.89    4108.45        4.9007     0.0547
Residual            6    5030.00     838.33
-------------------------------------------------------------------
Corrected Total     8   13246.89    1655.86
```

Interpretation of ANOVA Output

We summarize the level means for machine 3 in the following table.

Machine 3 Level Means for One-way ANOVA

Level	Number	Mean	Standard Error	Lower 95% CI	Upper 95% CI
1	3	587.667	16.717	546.76	628.57
2	3	615.667	16.717	574.76	656.57
3	3	542.33	16.717	501.43	583.24

3.5.2.5. Final Conclusions

Final Conclusions

The analysis shows that machines 1 and 2 had about the same variablity but significantly different locations. The throughput for machine 2 was also higher with greater variability than for machine 1. An interview with the operator revealed that he realized the second machine was not set correctly. However, he did not want to change the settings because he knew a study was being conducted and was afraid he might impact the results by making changes. Machine 3 had significantly more variation and lower throughput. The operator indicated that the machine had to be taken down several times for minor repairs. Given the preceeding analysis results, the team recommended replacing machine 3.

3.5.2.6. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot, if you have downloaded and installed it. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window and the Data Sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Get set up and started. 1. Read in the data.	1. You have read 5 columns of numbers into Dataplot, variables machine, day, time, sample, and diameter.
2. Box Plots by Factor Variables 1. Generate a box plot by machine. 2. Generate a box plot by day. 3. Generate a box plot by time of day.	1. The box plot shows significant variation for both location and spread. 2. The box plot shows no significant location or spread effects for day.

4. Generate a box plot by sample.	3. The box plot shows no significant location or spread effects for time of day. 4. The box plot shows no significant location or spread effects for sample.
3. Analysis of Variance 1. Perform an analysis of variance with all factors. 2. Perform an analysis of variance with only the machine factor. 3. Perform model validation by generating a 4-plot of the residuals.	1. The analysis of variance shows that only the machine factor is statistically significant. 2. The analysis of variance shows the overall mean and the effect estimates for the levels of the machine variable. 3. The 4-plot of the residuals does not indicate any significant problems with the model.
4. Graph of Throughput 1. Generate a graph of the throughput. 2. Perform an analysis of variance of the throughput.	1. The graph shows the throughput for machine 3 is lower than the other machines. 2. The effect estimates from the ANIVA are given.

3. Production Process Characterization

3.6. References

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978), *Statistics for Experimenters*, John Wiley and Sons, New York.

Cleveland, W.S. (1993), *Visualizing Data*, Hobart Press, New Jersey.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1985), *Exploring Data Tables, Trends, and Shapes*, John Wiley and Sons, New York.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1991), *Fundamentals of Exploratory Analysis of Variance*, John Wiley and Sons, New York.

4. Process Modeling

The goal for this chapter is to present the background and specific analysis techniques needed to construct a statistical model that describes a particular scientific or engineering process. The types of models discussed in this chapter are limited to those based on an explicit mathematical function. These types of models can be used for prediction of process outputs, for calibration, or for process optimization.

1. **Introduction**

1. Definition
2. Terminology
3. Uses
4. Methods

2. **Assumptions**

1. Assumptions

3. **Design**

1. Definition
2. Importance
3. Design Principles
4. Optimal Designs
5. Assessment

4. **Analysis**

1. Modeling Steps
2. Model Selection
3. Model Fitting
4. Model Validation
5. Model Improvement

5. **Interpretation & Use**

1. Prediction
2. Calibration
3. Optimization

6. **Case Studies**

1. Load Cell Output
2. Alaska Pipeline
3. Ultrasonic Reference Block
4. Thermal Expansion of Copper

Detailed Table of Contents: Process Modeling

References: Process Modeling

Appendix: Some Useful Functions for Process Modeling

4. Process Modeling - Detailed Table of Contents [4.]

The goal for this chapter is to present the background and specific analysis techniques needed to construct a statistical model that describes a particular scientific or engineering process. The types of models discussed in this chapter are limited to those based on an explicit mathematical function. These types of models can be used for prediction of process outputs, for calibration, or for process optimization.

1. Introduction to Process Modeling [4.1.]
    1. What is process modeling? [4.1.1.]
    2. What terminology do statisticians use to describe process models? [4.1.2.]
    3. What are process models used for? [4.1.3.]
        1. Estimation [4.1.3.1.]
        2. Prediction [4.1.3.2.]
        3. Calibration [4.1.3.3.]
        4. Optimization [4.1.3.4.]
    4. What are some of the different statistical methods for model building? [4.1.4.]
        1. Linear Least Squares Regression [4.1.4.1.]
        2. Nonlinear Least Squares Regression [4.1.4.2.]
        3. Weighted Least Squares Regression [4.1.4.3.]
        4. LOESS (aka LOWESS) [4.1.4.4.]

2. Underlying Assumptions for Process Modeling [4.2.]
    1. What are the typical underlying assumptions in process modeling? [4.2.1.]
        1. The process is a *statistical* process. [4.2.1.1.]
        2. The means of the random errors are zero. [4.2.1.2.]
        3. The random errors have a constant standard deviation. [4.2.1.3.]
        4. The random errors follow a normal distribution. [4.2.1.4.]
        5. The data are randomly sampled from the process. [4.2.1.5.]
        6. The explanatory variables are observed without error. [4.2.1.6.]

3. Data Collection for Process Modeling [4.3.]
    1. What is design of experiments (DOE)? [4.3.1.]
    2. Why is experimental design important for process modeling? [4.3.2.]
    3. What are some general design principles for process modeling? [4.3.3.]
    4. I've heard some people refer to "optimal" designs, shouldn't I use those? [4.3.4.]
    5. How can I tell if a particular experimental design is good for my application? [4.3.5.]

4. Data Analysis for Process Modeling [4.4.]
    1. What are the basic steps for developing an effective process model? [4.4.1.]
    2. How do I select a function to describe my process? [4.4.2.]
        1. Incorporating Scientific Knowledge into Function Selection [4.4.2.1.]
        2. Using the Data to Select an Appropriate Function [4.4.2.2.]
        3. Using Methods that Do Not Require Function Specification [4.4.2.3.]
    3. How are estimates of the unknown parameters obtained? [4.4.3.]
        1. Least Squares [4.4.3.1.]
        2. Weighted Least Squares [4.4.3.2.]

4. How can I tell if a model fits my data? [4.4.4.]
    1. How can I assess the sufficiency of the functional part of the model? [4.4.4.1.]
    2. How can I detect non-constant variation across the data? [4.4.4.2.]
    3. How can I tell if there was drift in the measurement process? [4.4.4.3.]
    4. How can I assess whether the random errors are independent from one to the next? [4.4.4.4.]
    5. How can I test whether or not the random errors are distributed normally? [4.4.4.5.]
    6. How can I test whether any significant terms are missing or misspecified in the functional part of the model? [4.4.4.6.]
    7. How can I test whether all of the terms in the functional part of the model are necessary? [4.4.4.7.]
5. If my current model does not fit the data well, how can I improve it? [4.4.5.]
    1. Updating the Function Based on Residual Plots [4.4.5.1.]
    2. Accounting for Non-Constant Variation Across the Data [4.4.5.2.]
    3. Accounting for Errors with a Non-Normal Distribution [4.4.5.3.]

5. Use and Interpretation of Process Models [4.5.]
    1. What types of predictions can I make using the model? [4.5.1.]
        1. How do I estimate the average response for a particular set of predictor variable values? [4.5.1.1.]
        2. How can I predict the value and and estimate the uncertainty of a single response? [4.5.1.2.]
    2. How can I use my process model for calibration? [4.5.2.]
        1. Single-Use Calibration Intervals [4.5.2.1.]
    3. How can I optimize my process using the process model? [4.5.3.]

6. Case Studies in Process Modeling [4.6.]
    1. Load Cell Calibration [4.6.1.]
        1. Background & Data [4.6.1.1.]
        2. Selection of Initial Model [4.6.1.2.]
        3. Model Fitting - Initial Model [4.6.1.3.]
        4. Graphical Residual Analysis - Initial Model [4.6.1.4.]
        5. Interpretation of Numerical Output - Initial Model [4.6.1.5.]
        6. Model Refinement [4.6.1.6.]
        7. Model Fitting - Model #2 [4.6.1.7.]
        8. Graphical Residual Analysis - Model #2 [4.6.1.8.]
        9. Interpretation of Numerical Output - Model #2 [4.6.1.9.]
        10. Use of the Model for Calibration [4.6.1.10.]
        11. Work This Example Yourself [4.6.1.11.]
    2. Alaska Pipeline [4.6.2.]
        1. Background and Data [4.6.2.1.]
        2. Check for Batch Effect [4.6.2.2.]
        3. Initial Linear Fit [4.6.2.3.]
        4. Transformations to Improve Fit and Equalize Variances [4.6.2.4.]
        5. Weighting to Improve Fit [4.6.2.5.]
        6. Compare the Fits [4.6.2.6.]
        7. Work This Example Yourself [4.6.2.7.]
    3. Ultrasonic Reference Block Study [4.6.3.]
        1. Background and Data [4.6.3.1.]
        2. Initial Non-Linear Fit [4.6.3.2.]
        3. Transformations to Improve Fit [4.6.3.3.]
        4. Weighting to Improve Fit [4.6.3.4.]
        5. Compare the Fits [4.6.3.5.]
        6. Work This Example Yourself [4.6.3.6.]
    4. Thermal Expansion of Copper Case Study [4.6.4.]
        1. Background and Data [4.6.4.1.]

2. Rational Function Models [4.6.4.2.]
3. Initial Plot of Data [4.6.4.3.]
4. Quadratic/Quadratic Rational Function Model [4.6.4.4.]
5. Cubic/Cubic Rational Function Model [4.6.4.5.]
6. Work This Example Yourself [4.6.4.6.]

7. References For Chapter 4: Process Modeling [4.7.]

8. Some Useful Functions for Process Modeling [4.8.]
    1. Univariate Functions [4.8.1.]
        1. Polynomial Functions [4.8.1.1.]
            1. Straight Line [4.8.1.1.1.]
            2. Quadratic Polynomial [4.8.1.1.2.]
            3. Cubic Polynomial [4.8.1.1.3.]
        2. Rational Functions [4.8.1.2.]
            1. Constant / Linear Rational Function [4.8.1.2.1.]
            2. Linear / Linear Rational Function [4.8.1.2.2.]
            3. Linear / Quadratic Rational Function [4.8.1.2.3.]
            4. Quadratic / Linear Rational Function [4.8.1.2.4.]
            5. Quadratic / Quadratic Rational Function [4.8.1.2.5.]
            6. Cubic / Linear Rational Function [4.8.1.2.6.]
            7. Cubic / Quadratic Rational Function [4.8.1.2.7.]
            8. Linear / Cubic Rational Function [4.8.1.2.8.]
            9. Quadratic / Cubic Rational Function [4.8.1.2.9.]
            10. Cubic / Cubic Rational Function [4.8.1.2.10.]
            11. Determining m and n for Rational Function Models [4.8.1.2.11.]

4.1. Introduction to Process Modeling

Overview of Section 4.1

The goal for this section is to give the big picture of function-based process modeling. This includes a discussion of what process modeling is, the goals of process modeling, and a comparison of the different statistical methods used for model building. Detailed information on how to collect data, construct appropriate models, interpret output, and use process models is covered in the following sections. The final section of the chapter contains case studies that illustrate the general information presented in the first five sections using data from a variety of scientific and engineering applications.

Contents of Section 4.1

1. What is process modeling?
2. What terminology do statisticians use to describe process models?
3. What are process models used for?
    1. Estimation
    2. Prediction
    3. Calibration
    4. Optimization
4. What are some of the statistical methods for model building?
    1. Linear Least Squares Regression
    2. Nonlinear Least Squares Regression
    3. Weighted Least Squares Regression
    4. LOESS (aka LOWESS)

4.1.1. What is process modeling?

Basic Definition

Process modeling is the concise description of the total variation in one quantity, y, by partitioning it into

1. a deterministic component given by a mathematical function of one or more other quantities, $x_1, x_2, \ldots$, plus
2. a random component that follows a particular probability distribution.

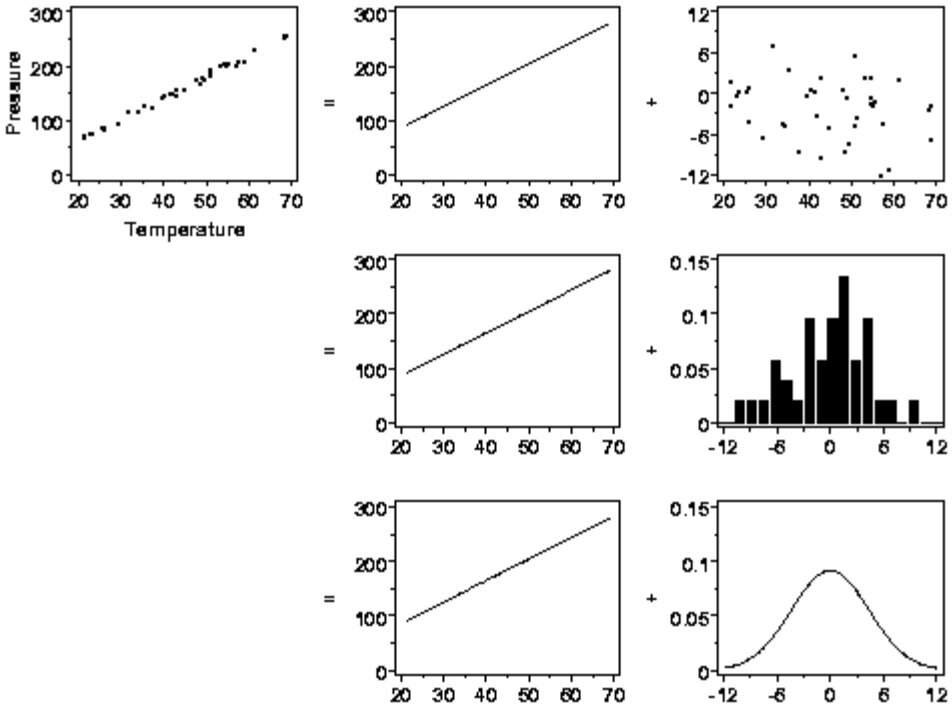
Example

For example, the total variation of the measured pressure of a fixed amount of a gas in a tank can be described by partitioning the variability into its deterministic part, which is a function of the temperature of the gas, plus some left-over random error. Charles' Law states that the pressure of a gas is proportional to its temperature under the conditions described here, and in this case most of the variation will be deterministic. However, due to measurement error in the pressure gauge, the relationship will not be purely deterministic. The random errors cannot be characterized individually, but will follow some probability distribution that will describe the relative frequencies of occurrence of different-sized errors.

Graphical Interpretation

Using the example above, the definition of process modeling can be graphically depicted like this:

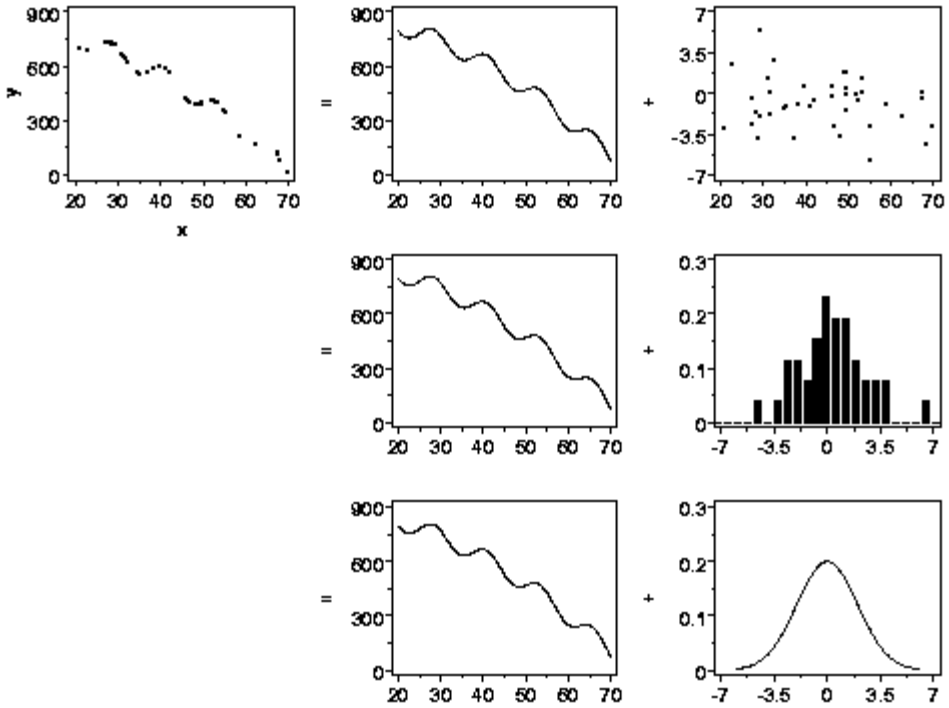
Click Figure for Full-Sized Copy



The top left plot in the figure shows pressure data that vary deterministically with temperature except for a small amount of random error. The relationship between pressure and temperature is a straight line, but not a perfect straight line. The top row plots on the right-hand side of the equals sign show a partitioning of the data into a perfect straight line and the remaining "unexplained" random variation in the data (note the different vertical scales of these plots). The plots in the middle row of the figure show the deterministic structure in the data again and a histogram of the random variation. The histogram shows the relative frequencies of observing different-sized random errors. The bottom row of the figure shows how the relative frequencies of the random errors can be summarized by a (normal) probability distribution.

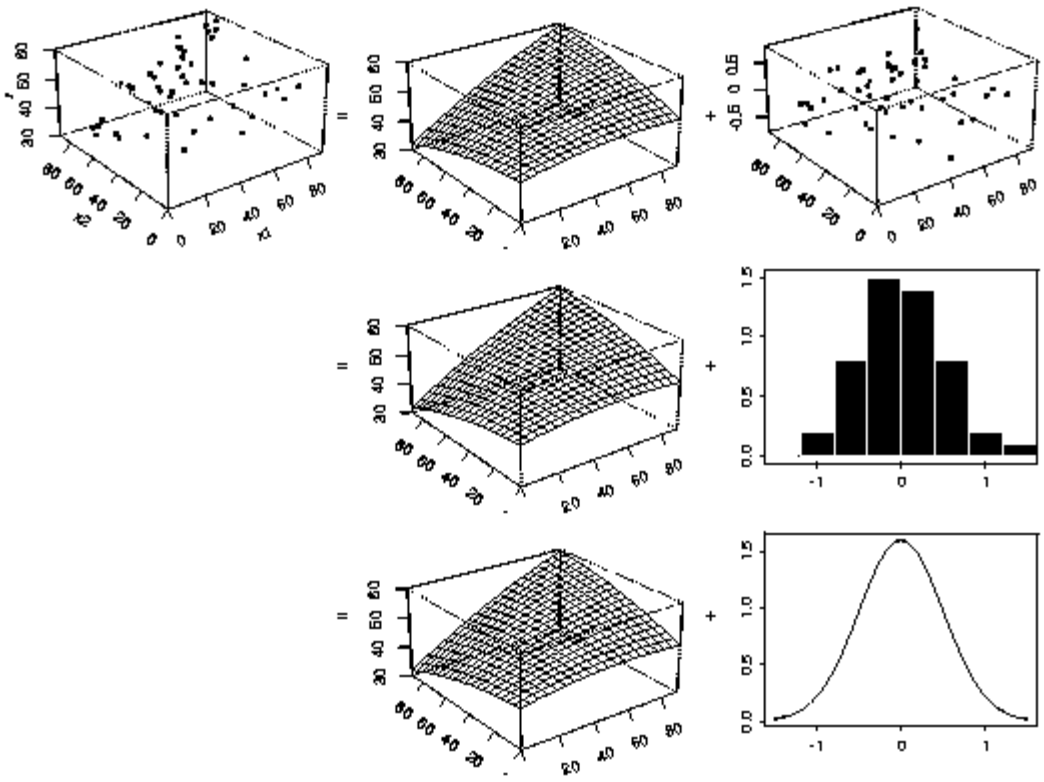
An Example from a More Complex Process

Of course, the straight-line example is one of the simplest functions used for process modeling. Another example is shown below. The concept is identical to the straight-line example, but the structure in the data is more complex. The variation in y is partitioned into a deterministic part, which is a function of another variable, x, plus some left-over random variation. (Again note the difference in the vertical axis scales of the two plots in the top right of the figure.) A probability distribution describes the leftover random variation.



An Example with Multiple Explanatory Variables

The examples of process modeling shown above have only one explanatory variable but the concept easily extends to cases with more than one explanatory variable. The three-dimensional perspective plots below show an example with two explanatory variables. Examples with three or more explanatory variables are exactly analogous, but are difficult to show graphically.



4. Process Modeling
4.1. Introduction to Process Modeling

4.1.2. What terminology do statisticians use to describe process models?

Model Components

There are three main parts to every process model. These are

1. the response variable, usually denoted by y,
2. the mathematical function, usually denoted as $f(\vec{x};\vec{\beta})$, and
3. the random errors, usually denoted by ε.

Form of Model

The general form of the model is

$$y = f(\vec{x};\vec{\beta}) + \varepsilon.$$

All process models discussed in this chapter have this general form. As alluded to earlier, the random errors that are included in the model make the relationship between the response variable and the predictor variables a "statistical" one, rather than a perfect deterministic one. This is because the functional relationship between the response and predictors holds only on average, not for each data point.

Some of the details about the different parts of the model are discussed below, along with alternate terminology for the different components of the model.

Response Variable

The response variable, y, is a quantity that varies in a way that we hope to be able to summarize and exploit via the modeling process. Generally it is known that the variation of the response variable is systematically related to the values of one or more other variables before the modeling process is begun, although testing the existence and nature of this dependence is part of the modeling process itself.

Mathematical Function

The mathematical function consists of two parts. These parts are the predictor variables, $x_1, x_2, \ldots$, and the parameters, $\beta_0, \beta_1, \ldots$. The predictor variables are observed along with the response variable. They are the quantities described on the previous page as inputs to the mathematical function, $f(\vec{x};\vec{\beta})$. The collection of all of the predictor variables is denoted by $\vec{x}$

for short.

$$\vec{x} \equiv (x_1, x_2, \ldots)$$

The parameters are the quantities that will be estimated during the modeling process. Their true values are unknown and unknowable, except in simulation experiments. As for the predictor variables, the collection of all of the parameters is denoted by $\vec{\beta}$ for short.

$$\vec{\beta} \equiv (\beta_0, \beta_1, \ldots)$$

The parameters and predictor variables are combined in different forms to give the function used to describe the deterministic variation in the response variable. For a straight line with an unknown intercept and slope, for example, there are two parameters and one predictor variable

$$f(x; \vec{\beta}) = \beta_0 + \beta_1 x.$$

For a straight line with a known slope of one, but an unknown intercept, there would only be one parameter

$$f(x; \vec{\beta}) = \beta_0 + x.$$

For a quadratic surface with two predictor variables, there are six parameters for the full model.

$$f(\vec{x}; \vec{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

.

Random Error

Like the parameters in the mathematical function, the random errors are unknown. They are simply the difference between the data and the mathematical function. They are assumed to follow a particular probability distribution, however, which is used to describe their aggregate behavior. The probability distribution that describes the errors has a mean of zero and an unknown standard deviation, denoted by σ, that is another parameter in the model, like the β's.

Alternate Terminology

Unfortunately, there are no completely standardardized names for the parts of the model discussed above. Other publications or software may use different terminology. For example, another common name for the response variable is "dependent variable". The response variable is also simply called "the response" for short. Other names for the predictor variables include "explanatory variables", "independent variables", "predictors" and "regressors". The mathematical function used to describe the deterministic variation in the response variable is sometimes called the "regression function", the "regression equation", the "smoothing function", or the "smooth".

Scope of "Model"

In its correct usage, the term "model" refers to the equation above and also includes the underlying assumptions made about the probability distribution used to describe the variation of the random errors. Often, however, people will also use the term "model" when referring specifically to the mathematical function describing the deterministic variation in the data. Since the function is part of the model, the more limited usage is not wrong, but it is important to remember that the term "model" might refer to more than just the mathematical function.

4.1.3. What are process models used for?

Three Main Purposes

Process models are used for four main purposes:

1. estimation,
2. prediction,
3. calibration, and
4. optimization.

The rest of this page lists brief explanations of the different uses of process models. More detailed explanations of the uses for process models are given in the subsections of this section listed at the bottom of this page.

Estimation

The goal of estimation is to determine the value of the regression function (i.e., the average value of the response variable), for a particular combination of the values of the predictor variables. Regression function values can be estimated for any combination of predictor variable values, including values for which no data have been measured or observed. Function values estimated for points within the observed space of predictor variable values are sometimes called interpolations. Estimation of regression function values for points outside the observed space of predictor variable values, called extrapolations, are sometimes necessary, but require caution.

Prediction

The goal of prediction is to determine either

1. the value of a new observation of the response variable, or
2. the values of a specified proportion of all future observations of the response variable

for a particular combination of the values of the predictor variables. Predictions can be made for any combination of predictor variable values, including values for which no data have been measured or observed. As in the case of estimation, predictions made outside the observed space of predictor variable values are sometimes necessary, but require caution.

Calibration

The goal of calibration is to quantitatively relate measurements made using one measurement system to those of another measurement system. This is done so that

measurements can be compared in common units or to tie results from a relative measurement method to absolute units.

Optimization Optimization is performed to determine the values of process inputs that should be used to obtain the desired process output. Typical optimization goals might be to maximize the yield of a process, to minimize the processing time required to fabricate a product, or to hit a target product specification with minimum variation in order to maintain specified tolerances.

Further Details

1. Estimation
2. Prediction
3. Calibration
4. Optimization

4.1.4. What are some of the different statistical methods for model building?

Selecting an Appropriate Stat Method: ***General Case***

For many types of data analysis problems there are no more than a couple of general approaches to be considered on the route to the problem's solution. For example, there is often a dichotomy between highly-efficient methods appropriate for data with noise from a normal distribution and more general methods for data with other types of noise. Within the different approaches for a specific problem type, there are usually at most a few competing statistical tools that can be used to obtain an appropriate solution. The bottom line for most types of data analysis problems is that selection of the best statistical method to solve the problem is largely determined by the goal of the analysis and the nature of the data.

Selecting an Appropriate Stat Method: ***Modeling***

Model building, however, is different from most other areas of statistics with regard to method selection. There are more general approaches and more competing techniques available for model building than for most other types of problems. There is often more than one statistical tool that can be effectively applied to a given modeling application. The large menu of methods applicable to modeling problems means that there is both more opportunity for effective and efficient solutions and more potential to spend time doing different analyses, comparing different solutions and mastering the use of different tools. The remainder of this section will introduce and briefly discuss some of the most popular and well-established statistical techniques that are useful for different model building situations.

Process Modeling Methods

1. Linear Least Squares Regression
2. Nonlinear Least Squares Regression
3. Weighted Least Squares Regression
4. LOESS (aka LOWESS)

4.2. Underlying Assumptions for Process Modeling

Implicit Assumptions Underlie Most Actions

Most, if not all, thoughtful actions that people take are based on ideas, or assumptions, about how those actions will affect the goals they want to achieve. The actual assumptions used to decide on a particular course of action are rarely laid out explicitly, however. Instead, they are only implied by the nature of the action itself. Implicit assumptions are inherent to process modeling actions, just as they are to most other types of action. It is important to understand what the implicit assumptions are for any process modeling method because the validity of these assumptions affect whether or not the goals of the analysis will be met.

Checking Assumptions Provides Feedback on Actions

If the implicit assumptions that underlie a particular action are not true, then that action is not likely to meet expectations either. Sometimes it is abundantly clear when a goal has been met, but unfortunately that is not always the case. In particular, it is usually not possible to obtain immediate feedback on the attainment of goals in most process modeling applications. The goals of process modeling, sucha as answering a scientific or engineering question, depend on the correctness of a process model, which can often only be directly and absolutely determined over time. In lieu of immediate, direct feedback, however, indirect information on the effectiveness of a process modeling analysis can be obtained by checking the validity of the underlying assumptions. Confirming that the underlying assumptions are valid helps ensure that the methods of analysis were appropriate and that the results will be consistent with the goals.

Overview of Section 4.2

This section discusses the specific underlying assumptions associated with most model-fitting methods. In discussing the underlying assumptions, some background is also provided on the consequences of stopping the modeling process short of completion and leaving the results of an analysis at odds with the underlying assumptions. Specific data analysis methods that can be used to check whether or not the assumptions hold in a particular case are discussed in Section 4.4.4.

Contents of Section 4.2

1. What are the typical underlying assumptions in process modeling?

1. The process is a *statistical* process.
2. The means of the random errors are zero.
3. The random errors have a constant standard deviation.
4. The random errors follow a normal distribution.
5. The data are randomly sampled from the process.
6. The explanatory variables are observed without error.

4.2.1. What are the typical underlying assumptions in process modeling?

Overview of Section 4.2.1

This section lists the typical assumptions underlying most process modeling methods. On each of the following pages, one of the six major assumptions is described individually; the reasons for it's importance are also briefly discussed; and any methods that are not subject to that particular assumption are noted. As discussed on the previous page, these are implicit assumptions based on properties inherent to the process modeling methods themselves. Successful use of these methods in any particular application hinges on the validity of the underlying assumptions, whether their existence is acknowledged or not. Section 4.4.4 discusses methods for checking the validity of these assumptions.

Typical Assumptions for Process Modeling

1. The process is a *statistical* process.
2. The means of the random errors are zero.
3. The random errors have a constant standard deviation.
4. The random errors follow a normal distribution.
5. The data are randomly sampled from the process.
6. The explanatory variables are observed without error.

4.3. Data Collection for Process Modeling

Collecting Good Data

This section lays out some general principles for collecting data for construction of process models. Using well-planned data collection procedures is often the difference between successful and unsuccessful experiments. In addition, well-designed experiments are often less expensive than those that are less well thought-out, regardless of overall success or failure.

Specifically, this section will answer the question:

> What can the analyst do even prior to collecting the data (that is, at the experimental design stage) that would allow the analyst to do an optimal job of modeling the process?

Contents: Section 3

This section deals with the following five questions:

1. What is design of experiments (DOE)?
2. Why is experimental design important for process modeling?
3. What are some general design principles for process modeling?
4. I've heard some people refer to "optimal" designs, shouldn't I use those?
5. How can I tell if a particular experimental design is good for my application?

4.3.1. What is design of experiments (DOE)?

Systematic Approach to Data Collection

Design of experiments (DOE) is a systematic, rigorous approach to engineering problem-solving that applies principles and techniques at the data collection stage so as to ensure the generation of valid, defensible, and supportable engineering conclusions. In addition, all of this is carried out under the constraint of a minimal expenditure of engineering runs, time, and money.

DOE Problem Areas

There are four general engineering problem areas in which DOE may be applied:

1. Comparative
2. Screening/Characterizing
3. Modeling
4. Optimizing

Comparative

In the first case, the engineer is interested in assessing whether a change in a single factor has in fact resulted in a change/improvement to the process as a whole.

Screening Characterization

In the second case, the engineer is interested in "understanding" the process as a whole in the sense that he/she wishes (after design and analysis) to have in hand a ranked list of important through unimportant factors (most important to least important) that affect the process.

Modeling

In the third case, the engineer is interested in functionally modeling the process with the output being a good-fitting (= high predictive power) mathematical function, and to have good (= maximal accuracy) estimates of the coefficients in that function.

Optimizing

In the fourth case, the engineer is interested in determining optimal settings of the process factors; that is, to determine for each factor the level of the factor that optimizes the process response.

In this section, we focus on case 3: modeling.

NIST
SEMATECH
HOME
TOOLS & AIDS
SEARCH
BACK
NEXT

4.3.2. Why is experimental design important for process modeling?

Output from Process Model is Fitted Mathematical Function

The output from process modeling is a fitted mathematical function with estimated coefficients. For example, in modeling resistivity, y, as a function of dopant density, x, an analyst may suggest the function

$$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

in which the coefficients to be estimated are β_0, β_1, and β_{11}. Even for a given functional form, there is an infinite number of potential coefficient values that potentially may be used. Each of these coefficient values will in turn yield predicted values.

What are Good Coefficient Values?

Poor values of the coefficients are those for which the resulting predicted values are considerably different from the observed raw data y. Good values of the coefficients are those for which the resulting predicted values are close to the observed raw data y. The best values of the coefficients are those for which the resulting predicted values are close to the observed raw data y, and the statistical uncertainty connected with each coefficient is small.

There are two considerations that are useful for the generation of "best" coefficients:

1. Least squares criterion
2. Design of experiment principles

Least Squares Criterion

For a given data set (e.g., 10 (x,y) pairs), the most common procedure for obtaining the coefficients for

$$y = f(x;\vec{\beta}) + \epsilon$$

is the least squares estimation criterion. This criterion yields coefficients with predicted values that are closest to the raw data y in the sense that the sum of the squared differences between the raw data and the predicted values is as small as possible.

The overwhelming majority of regression programs today use the least squares criterion for estimating the model coefficients. Least squares estimates are popular because

1. the estimators are statistically optimal (BLUEs: Best Linear Unbiased Estimators);
2. the estimation algorithm is mathematically tractable, in closed form, and therefore easily programmable.

How then can this be improved? For a given set of x values it cannot be; but frequently the choice of the x values is under our control. If we can select the x values, the coefficients will have less variability than if the x are not controlled.

Design of Experiment Principles

As to what values should be used for the x's, we look to established experimental design principles for guidance.

Principle 1: Minimize Coefficient Estimation Variation

The first principle of experimental design is to control the values within the x vector such that after the y data are collected, the subsequent model coefficients are as good, in the sense of having the smallest variation, as possible.

The key underlying point with respect to design of experiments and process modeling is that even though (for simple (x,y) fitting, for example) the least squares criterion may yield optimal (minimal variation) estimators for a given distribution of x values, some distributions of data in the x vector may yield better (smaller variation) coefficient estimates than other x vectors. If the analyst can specify the values in the x vector, then he or she may be able to drastically change and reduce the noisiness of the subsequent least squares coefficient estimates.

Five Designs

To see the effect of experimental design on process modeling, consider the following simplest case of fitting a line:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Suppose the analyst can afford 10 observations (that is, 10 (x,y) pairs) for the purpose of determining optimal (that is, minimal variation) estimators of β_0 and β_1. What 10 x values should be used for the purpose of collecting the corresponding 10 y values? Colloquially, where should the 10 x values be sprinkled along the horizontal axis so as to minimize the variation of the least squares estimated coefficients for β_0 and β_1? Should the 10 x values be:

1. ten equi-spaced values across the range of interest?
2. five replicated equi-spaced values across the range of

interest?
3. five values at the minimum of the x range and five values at the maximum of the x range?
4. one value at the minimum, eight values at the mid-range, and one value at the maximum?
5. four values at the minimum, two values at mid-range, and four values at the maximum?

or (in terms of "quality" of the resulting estimates for β_0 and β_1) perhaps it doesn't make any difference?

For each of the above five experimental designs, there will of course be y data collected, followed by the generation of least squares estimates for β_0 and β_1, and so each design will in turn yield a fitted line.

Are the Fitted Lines Better for Some Designs?

But are the fitted lines, i.e., the fitted process models, better for some designs than for others? Are the coefficient estimator variances smaller for some designs than for others? For given estimates, are the resulting predicted values better (that is, closer to the observed y values) than for other designs? The answer to all of the above is YES. It DOES make a difference.

The most popular answer to the above question about which design to use for linear modeling is design #1 with ten equi-spaced points. It can be shown, however, that the variance of the estimated slope parameter depends on the design according to the relationship

$$\mathrm{Var}(\hat{\beta}_1) \propto \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})}.$$

Therefore to obtain minimum variance estimators, one maximizes the denominator on the right. To maximize the denominator, it is (for an arbitrarily fixed $\bar{x}$), best to position the x's as far away from $\bar{x}$ as possible. This is done by positioning half of the x's at the lower extreme and the other half at the upper extreme. This is design #3 above, and this "dumbbell" design (half low and half high) is in fact the best possible design for fitting a line. Upon reflection, this is intuitively arrived at by the adage that "2 points define a line", and so it makes the most sense to determine those 2 points as far apart as possible (at the extremes) and as well as possible (having half the data at each extreme). Hence the design of experiment solution to model processing when the model is a line is the "dumbbell" design--half the X's at each extreme.

What is the Worst

What is the worst design in the above case? Of the five designs, the worst design is the one that has maximum

Design?

variation. In the mathematical expression above, it is the one that minimizes the denominator, and so this is design #4 above, for which almost all of the data are located at the mid-range. Clearly the estimated line in this case is going to chase the solitary point at each end and so the resulting linear fit is intuitively inferior.

Designs 1, 2, and 5

What about the other 3 designs? Designs 1, 2, and 5 are useful only for the case when we think the model may be linear, but we are not sure, and so we allow additional points that permit fitting a line if appropriate, but build into the design the "capacity" to fit beyond a line (e.g., quadratic, cubic, etc.) if necessary. In this regard, the ordering of the designs would be

- design 5 (if our worst-case model is quadratic),
- design 2 (if our worst-case model is quartic)
- design 1 (if our worst-case model is quintic and beyond)

4.3.3. What are some general design principles for process modeling?

Experimental Design Principles Applied to Process Modeling

There are six principles of experimental design as applied to process modeling:

1. Capacity for Primary Model
2. Capacity for Alternative Model
3. Minimum Variance of Coefficient Estimators
4. Sample where the Variation Is
5. Replication
6. Randomization

We discuss each in detail below.

Capacity for Primary Model

For your best-guess model, make sure that the design has the capacity for estimating the coefficients of that model. For a simple example of this, if you are fitting a quadratic model, then make sure you have at least three distinct horixontal axis points.

Capacity for Alternative Model

If your best-guess model happens to be inadequate, make sure that the design has the capacity to estimate the coefficients of your best-guess back-up alternative model (which means implicitly that you should have already identified such a model). For a simple example, if you suspect (but are not positive) that a linear model is appropriate, then it is best to employ a globally robust design (say, four points at each extreme and three points in the middle, for a ten-point design) as opposed to the locally optimal design (such as five points at each extreme). The locally optimal design will provide a best fit to the line, but have no capacity to fit a quadratic. The globally robust design will provide a good (though not optimal) fit to the line and additionally provide a good (though not optimal) fit to the quadratic.

Minimum Variance of Coefficient Estimators

For a given model, make sure the design has the property of minimizing the variation of the least squares estimated coefficients. This is a general principle that is always in effect but which in practice is hard to implement for many models beyond the simpler 1-factor

$$y = f(x; \vec{\beta}) + \epsilon$$

models. For more complicated 1-factor models, and for most multi-factor

$$y = f(\vec{x}; \vec{\beta}) + \epsilon$$

models, the expressions for the variance of the least squares estimators, although available, are complicated and assume more than the analyst typically knows. The net result is that this principle, though important, is harder to apply beyond the simple cases.

Sample Where the Variation Is (Non Constant Variance Case)

Regardless of the simplicity or complexity of the model, there are situations in which certain regions of the curve are noisier than others. A simple case is when there is a linear relationship between x and y but the recording device is proportional rather than absolute and so larger values of y are intrinsically noisier than smaller values of y. In such cases, sampling where the variation is means to have more replicated points in those regions that are noisier. The practical answer to how many such replicated points there should be is

$$n_i = \frac{1}{\sigma_i^2}$$

with σ_i denoting the theoretical standard deviation for that given region of the curve. Usually σ_i is estimated by a-priori guesses for what the local standard deviations are.

Sample Where the Variation Is (Steep Curve Case)

A common occurence for non-linear models is for some regions of the curve to be steeper than others. For example, in fitting an exponential model (small x corresponding to large y, and large y corresponding to small x) it is often the case that the y data in the steep region are intrinsically noisier than the y data in the relatively flat regions. The reason for this is that commonly the x values themselves have a bit of noise and this x-noise gets translated into larger y-noise in the steep sections than in the shallow sections. In such cases, when we know the shape of the response curve well enough to identify steep-versus-shallow regions, it is often a good idea to sample more heavily in the steep regions than in the shallow regions. A practical rule-of-thumb for where to position the x values in such situations is to

1. sketch out your best guess for what the resulting curve will be;
2. partition the vertical (that is the y) axis into n equi-spaced points (with n denoting the total number of

data points that you can afford);
3. draw horizontal lines from each vertical axis point to where it hits the sketched-in curve.
4. drop a vertical projection line from the curve intersection point to the horizontal axis.

These will be the recommended x values to use in the design.

The above rough procedure for an exponentially decreasing curve would thus yield a logarithmic preponderance of points in the steep region of the curve and relatively few points in the flatter part of the curve.

Replication

If affordable, replication should be part of every design. Replication allows us to compute a model-independent estimate of the process standard deviation. Such an estimate may then be used as a criterion in an objective lack-of-fit test to assess whether a given model is adequate. Such an objective lack-of-fit F-test can be employed only if the design has built-in replication. Some replication is essential; replication at every point is ideal.

Randomization

Just because the x's have some natural ordering does not mean that the data should be collected in the same order as the x's. Some aspect of randomization should enter into every experiment, and experiments for process modeling are no exception. Thus if your are sampling ten points on a curve, the ten y values should not be collected by sequentially stepping through the x values from the smallest to the largest. If you do so, and if some extraneous drifting or wear occurs in the machine, the operator, the environment, the measuring device, etc., then that drift will unwittingly contaminate the y values and in turn contaminate the final fit. To minimize the effect of such potential drift, it is best to randomize (use random number tables) the sequence of the x values. This will not make the drift go away, but it will spread the drift effect evenly over the entire curve, realistically inflating the variation of the fitted values, and providing some mechanism after the fact (at the residual analysis model validation stage) for uncovering or discovering such a drift. If you do not randomize the run sequence, you give up your ability to detect such a drift if it occurs.

4.3.4. I've heard some people refer to "optimal" designs, shouldn't I use those?

Classical Designs Heavily Used in Industry

The most heavily used designs in industry are the "classical designs" (full factorial designs, fractional factorial designs, Latin square designs, Box-Behnken designs, etc.). They are so heavily used because they are optimal in their own right and have served superbly well in providing efficient insight into the underlying structure of industrial processes.

Reasons Classical Designs May Not Work

Cases do arise, however, for which the tabulated classical designs do not cover a particular practical situation. That is, user constraints preclude the use of tabulated classical designs because such classical designs do not accommodate user constraints. Such constraints include:

1. Limited maximum number of runs:

   User constraints in budget and time may dictate a maximum allowable number of runs that is too small or too "irregular" (e.g., "13") to be accommodated by classical designs--even fractional factorial designs.

2. Impossible factor combinations:

   The user may have some factor combinations that are impossible to run. Such combinations may at times be specified (to maintain balance and orthogonality) as part of a recommeded classical design. If the user simply omits this impossible run from the design, the net effect may be a reduction in the quality and optimaltiy of the classical design.

3. Too many levels:

   The number of factors and/or the number of levels of some factors intended for use may not be included in tabulations of classical designs.

4. Complicated underlying model:

The user may be assuming an underlying model that is too complicated (or too non-linear), so that classical designs would be inappropriate.

What to Do If Classical Designs Do Not Exist?

If user constraints are such that classical designs do not exist to accommodate such constraints, then what is the user to do?

The previous section's list of design criteria (capability for the primary model, capability for the alternate model, minimum variation of estimated coefficients, etc.) is a good passive target to aim for in terms of desirable design properties, but provides little help in terms of an active formal construction methodology for generating a design.

Common Optimality Criteria

To satisfy this need, an "optimal design" methodology has been developed to generate a design when user constraints preclude the use of tabulated classical designs. Optimal designs may be optimal in many different ways, and what may be an optimal design according to one criterion may be suboptimal for other criteria. Competing criteria have led to a literal alphabet-soup collection of optimal design methodologies. The four most popular ingredients in that "soup" are:

D-optimal designs:	minimize the generalized variance of the parameter estimators.
A-optimal designs:	minimize the average variance of the parameter estimators.
G-optimal designs:	minimize the maximum variance of the predicted values.
V-optimal designs:	minimize the average variance of the predicted values.

Need 1: a Model

The motivation for optimal designs is the practical constraints that the user has. The advantage of optimal designs is that they do provide a reasonable design-generating methodology when no other mechanism exists. The disadvantage of optimal designs is that they require a model from the user. The user may not have this model.

All optimal designs are model-dependent, and so the quality of the final engineering conclusions that result from the ensuing design, data, and analysis is dependent on the correctness of the analyst's assumed model. For example, if the responses from a particular process are actually being drawn from a cubic model and the analyst assumes a linear model and uses the corresponding optimal design to generate data and perform the data

analysis, then the final engineering conclusions will be flawed and invalid. Hence one price for obtaining an in-hand generated design is the designation of a model. All optimal designs need a model; without a model, the optimal design-generation methodology cannot be used, and general design principles must be reverted to.

Need 2: a Candidate Set of Points

The other price for using optimal design methodology is a user-specified set of candidate points. Optimal designs will not generate the best design points from some continuous region--that is too much to ask of the mathematics. Optimal designs will generate the best subset of n points from a larger superset of candidate points. The user must specify this candidate set of points. Most commonly, the superset of candidate points is the full factorial design over a fine-enough grid of the factor space with which the analyst is comfortable. If the grid is too fine, and the resulting superset overly large, then the optimal design methodology may prove computationally challenging.

Optimal Designs are Computationally Intensive

The optimal design-generation methodology is computationally intensive. Some of the designs (e.g., D-optimal) are better than other designs (such as A-optimal and G-optimal) in regard to efficiency of the underlying search algorithm. Like most mathematical optimization techniques, there is no iron-clad guarantee that the result from the optimal design methodology is in fact the true optimum. However, the results are usually satisfactory from a practical point of view, and are far superior than any ad hoc designs.

For further details about optimal designs, the analyst is referred to Montgomery (2001).

4.3.5. How can I tell if a particular experimental design is good for my application?

Assess Relative to the Six Design Principles

If you have a design, generated by whatever method, in hand, how can you assess its after-the-fact goodness? Such checks can potentially parallel the list of the six general design principles. The design can be assessed relative to each of these six principles. For example, does it have capacity for the primary model, does it have capacity for an alternative model, etc.

Some of these checks are quantitative and complicated; other checks are simpler and graphical. The graphical checks are the most easily done and yet are among the most informative. We include two such graphical checks and one quantitative check.

Graphically Check for Univariate Balance

If you have a design that claims to be globally good in ***k*** factors, then generally that design should be locally good in each of the individual ***k*** factors. Checking high-dimensional global goodness is difficult, but checking low-dimensional local goodness is easy. Generate ***k*** counts plots, with the levels of factors x_i plotted on the horizontal axis of each plot and the number of design points for each level in factor x_i on the vertical axis. For most good designs, these counts should be about the same (= balance) for all levels of a factor. Exceptions exist, but such balance is a low-level characteristic of most good designs.

Graphically Check for Bivariate Balance

If you have a design that is purported to be globally good in ***k*** factors, then generally that design should be locally good in all pairs of the individual ***k*** factors. Graphically check for such 2-way balance by generating plots for all pairs of factors, where the horizontal axis of a given plot is x_i and the vertical axis is x_j. The response variable y does NOT come into play in these plots. We are only interested in characteristics of the design, and so only the x variables are involved. The 2-way plots of most good designs have a certain symmetric and balanced look about them--all combination points should be covered and each combination point should have about the same number of points.

Check for

For optimal designs, metrics exist (D-efficiency, A-

Minimal Variation	efficiency, etc.) that can be computed and that reflect the quality of the design. Further, relative ratios of standard deviations of the coefficient estimators and relative ratios of predicted values can be computed and compared for such designs. Such calculations are commonly performed in computer packages which specialize in the generation of optimal designs.

4.4. Data Analysis for Process Modeling

Building a Good Model

This section contains detailed discussions of the necessary steps for developing a good process model after data have been collected. A general model-building framework, applicable to multiple statistical methods, is described with method-specific points included when necessary.

Contents: Section 4

1. What are the basic steps for developing an effective process model?
2. How do I select a function to describe my process?
    1. Incorporating Scientific Knowledge into Function Selection
    2. Using the Data to Select an Appropriate Function
    3. Using Methods that Do Not Require Function Specification
3. How are estimates of the unknown parameters obtained?
    1. Least Squares
    2. Weighted Least Squares
4. How can I tell if a model fits my data?
    1. How can I assess the sufficiency of the functional part of the model?
    2. How can I detect non-constant variation across the data?
    3. How can I tell if there was drift in the measurement process?
    4. How can I assess whether the random errors are independent from one to the next?
    5. How can I test whether or not the random errors are normally distributed?
    6. How can I test whether any significant terms are missing or misspecified in the functional part of the model?
    7. How can I test whether all of the terms in the functional part of the model are necessary?
5. If my current model does not fit the data well, how can I improve it?
    1. Updating the Function Based on Residual Plots
    2. Accounting for Non-Constant Variation Across the Data
    3. Accounting for Errors with a Non-Normal Distribution

NIST
SEMATECH
HOME
TOOLS & AIDS
SEARCH
BACK
NEXT

4.4.1. What are the basic steps for developing an effective process model?

Basic Steps Provide Universal Framework

The basic steps used for model-building are the same across all modeling methods. The details vary somewhat from method to method, but an understanding of the common steps, combined with the typical underlying assumptions needed for the analysis, provides a framework in which the results from almost any method can be interpreted and understood.

Basic Steps of Model Building

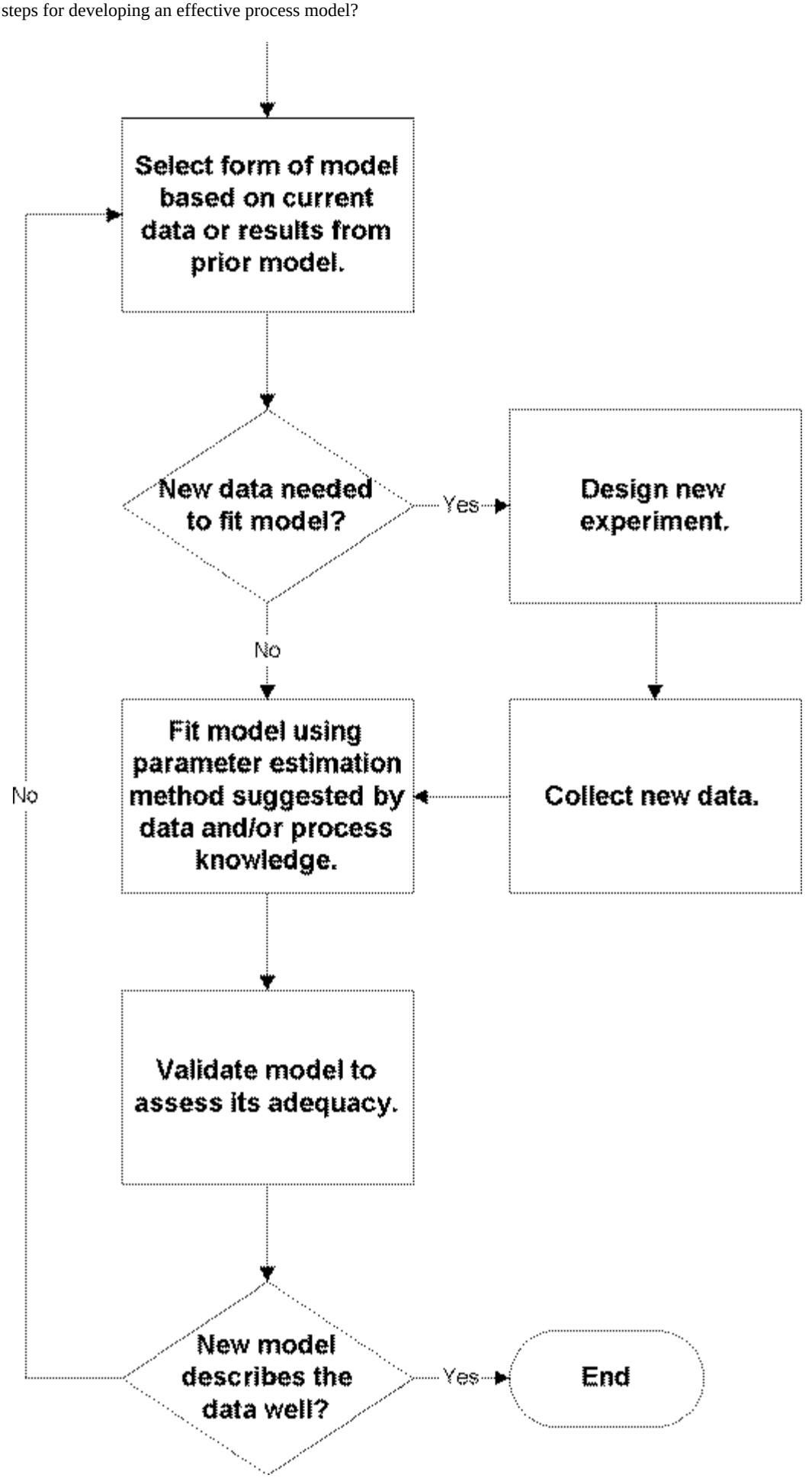
The basic steps of the model-building process are:

1. model selection
2. model fitting, and
3. model validation.

These three basic steps are used iteratively until an appropriate model for the data has been developed. In the model selection step, plots of the data, process knowledge and assumptions about the process are used to determine the form of the model to be fit to the data. Then, using the selected model and possibly information about the data, an appropriate model-fitting method is used to estimate the unknown parameters in the model. When the parameter estimates have been made, the model is then carefully assessed to see if the underlying assumptions of the analysis appear plausible. If the assumptions seem valid, the model can be used to answer the scientific or engineering questions that prompted the modeling effort. If the model validation identifies problems with the current model, however, then the modeling process is repeated using information from the model validation step to select and/or fit an improved model.

A Variation on the Basic Steps

The three basic steps of process modeling described in the paragraph above assume that the data have already been collected and that the same data set can be used to fit all of the candidate models. Although this is often the case in model-building situations, one variation on the basic model-building sequence comes up when additional data are needed to fit a newly hypothesized model based on a model fit to the initial data. In this case two additional steps, experimental design and data collection, can be added to the basic sequence between model selection and model-fitting. The flow chart below shows the basic model-fitting sequence with the integration of the related data collection steps into the model-building process.



Model Building Sequence

Examples illustrating the model-building sequence in real applications can be found in the case studies in Section 4.6. The specific tools and techniques used in the basic model-building steps are described in the remainder of this section.

Design of Initial Experiment

Of course, considering the model selection and fitting before collecting the initial data is also a good idea. Without data in hand, a hypothesis about what the data will look like is needed in order to guess what the initial model should be. Hypothesizing the outcome of an experiment is not always possible, of course, but efforts made in the earliest stages of a project often maximize the efficiency of the whole model-building process and result in the best possible models for the process. More details about experimental design can be found in Section 4.3 and in Chapter 5: Process Improvement.

4.4.2. How do I select a function to describe my process?

Synthesis of Process Information Necessary

Selecting a model of the right form to fit a set of data usually requires the use of empirical evidence in the data, knowledge of the process and some trial-and-error experimentation. As mentioned on the previous page, model building is always an iterative process. Much of the need to iterate stems from the difficulty in initially selecting a function that describes the data well. Details about the data are often not easily visible in the data as originally observed. The fine structure in the data can usually only be elicited by use of model-building tools such as residual plots and repeated refinement of the model form. As a result, it is important not to overlook any of the sources of information that indicate what the form of the model should be.

Answer Not Provided by Statistics Alone

Sometimes the different sources of information that need to be integrated to find an effective model will be contradictory. An open mind and a willingness to think about what the data are saying is important. Maintaining balance and looking for alternate sources for unusual effects found in the data are also important. For example, in the load cell calibration case study the statistical analysis pointed out that the model initially thought to be appropriate did not account for all of the structure in the data. A refined model was developed, but the appearance of an unexpected result brings up the question of whether the original understanding of the problem was inaccurate, or whether the need for an alternate model was due to experimental artifacts. In the load cell problem it was easy to accept that the refined model was closer to the truth, but in a more complicated case additional experiments might have been needed to resolve the issue.

Knowing Function Types Helps

Another helpful ingredient in model selection is a wide knowledge of the shapes that different mathematical functions can assume. Knowing something about the models that have been found to work well in the past for different application types also helps. A menu of different functions on the next page, Section 4.4.2.1. (links provided below), provides one way to learn about the function shapes and flexibility. Section 4.4.2.2. discusses how general function features and qualitative scientific information can be combined to help with model selection. Finally, Section 4.4.2.3. points to

methods that don't require specification of a particular function to be fit to the data, and how models of those types can be refined.

1. Incorporating Scientific Knowledge into Function Selection
2. Using the Data to Select an Appropriate Function
3. Using Methods that Do Not Require Function Specification

4.4.3. How are estimates of the unknown parameters obtained?

Parameter Estimation in General

After selecting the basic form of the functional part of the model, the next step in the model-building process is estimation of the unknown parameters in the function. In general, this is accomplished by solving an optimization problem in which the objective function (the function being minimized or maximized) relates the response variable and the functional part of the model containing the unknown parameters in a way that will produce parameter estimates that will be close to the true, unknown parameter values. The unknown parameters are, loosely speaking, treated as variables to be solved for in the optimization, and the data serve as known coefficients of the objective function in this stage of the modeling process.

In theory, there are as many different ways of estimating parameters as there are objective functions to be minimized or maximized. However, a few principles have dominated because they result in parameter estimators that have good statistical properties. The two major methods of parameter estimation for process models are maximum likelihood and least squares. Both of these methods provide parameter estimators that have many good properties. Both maximum likelihood and least squares are sensitive to the presence of outliers, however. There are also many newer methods of parameter estimation, called robust methods, that try to balance the efficiency and desirable properties of least squares and maximum likelihood with a lower sensitivity to outliers.

Overview of Section 4.3

Although robust techniques are valuable, they are not as well developed as the more traditional methods and often require specialized software that is not readily available. Maximum likelihood also requires specialized algorithms in general, although there are important special cases that do not have such a requirement. For example, for data with normally distributed random errors, the least squares and maximum likelihood parameter estimators are identical. As a result of these software and developmental issues, and the coincidence of maximum likelihood and least squares in many applications, this section currently focuses on parameter estimation only by least squares methods. The remainder of this section offers some intuition into how least squares works

and illustrates the effectiveness of this method.

Contents of Section 4.3

1. Least Squares
2. Weighted Least Squares

4.4.4. How can I tell if a model fits my data?

R^2 Is Not Enough!

Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. Often the validation of a model seems to consist of nothing more than quoting the R^2 statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model). Unfortunately, a high R^2 value does not guarantee that the model fits the data well. Use of a model that does not fit the data well cannot provide good answers to the underlying engineering or scientific questions under investigation.

Main Tool: Graphical Residual Analysis

There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis. Different types of plots of the residuals (see definition below) from a fitted model provide information on the adequacy of different aspects of the model. Numerical methods for model validation, such as the R^2 statistic, are also useful, but usually to a lesser degree than graphical methods. Graphical methods have an advantage over numerical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Numerical methods for model validation tend to be narrowly focused on a particular aspect of the relationship between the model and the data and often try to compress that information into a single descriptive number or test result.

Numerical Methods' Forte

Numerical methods do play an important role as confirmatory methods for graphical techniques, however. For example, the lack-of-fit test for assessing the correctness of the functional part of the model can aid in interpreting a borderline residual plot. There are also a few modeling situations in which graphical methods cannot easily be used. In these cases, numerical methods provide a fallback position for model validation. One common situation when numerical validation methods take precedence over graphical methods is when the number of parameters being estimated is relatively close to the size of the data set. In this situation residual plots are often difficult to interpret due to constraints on the residuals imposed by the estimation of the unknown parameters. One area in which this typically happens is in optimization applications using designed experiments. Logistic regression

with binary data is another area in which graphical residual analysis can be difficult.

Residuals

The residuals from a fitted model are the differences between the responses observed at each combination values of the explanatory variables and the corresponding prediction of the response computed using the regression function.

Mathematically, the definition of the residual for the i^{th} observation in the data set is written

$$e_i = y_i - f(\vec{x}_i; \hat{\vec{\beta}}),$$

with y_i denoting the i^{th} response in the data set and $\vec{x}_i$ represents the list of explanatory variables, each set at the corresponding values found in the i^{th} observation in the data set.

Example

The data listed below are from the Pressure/Temperature example introduced in Section 4.1.1. The first column shows the order in which the observations were made, the second column indicates the day on which each observation was made, and the third column gives the ambient temperature recorded when each measurement was made. The fourth column lists the temperature of the gas itself (the explanatory variable) and the fifth column contains the observed pressure of the gas (the response variable). Finally, the sixth column gives the corresponding values from the fitted straight-line regression function.

$$\hat{P} = 7.749695 + 3.930123T$$

and the last column lists the residuals, the difference between columns five and six.

Data, Fitted Values & Residuals

Run Order	Day	Ambient Temperature	Temperature	Pressure	Fitted Value	Residual
1	1	23.820	54.749	225.066	222.920	2.146
2	1	24.120	23.323	100.331	99.411	0.920
3	1	23.434	58.775	230.863	238.744	-7.881
4	1	23.993	25.854	106.160	109.359	-3.199
5	1	23.375	68.297	277.502	276.165	1.336
6	1	23.233	37.481	148.314	155.056	-6.741
7	1	24.162	49.542	197.562	202.456	-4.895
8	1	23.667	34.101	138.537	141.770	-3.232
9	1	24.056	33.901	137.969	140.983	-3.014
10	1	22.786	29.242	117.410	122.674	-5.263
11	2	23.785	39.506	164.442	163.013	1.429
12	2	22.987	43.004	181.044		

```
176.759     4.285
13      2       23.799      53.226      222.179
216.933     5.246
14      2       23.661      54.467      227.010
221.813     5.198
15      2       23.852      57.549      232.496
233.925    -1.429
16      2       23.379      61.204      253.557
248.288     5.269
17      2       24.146      31.489      139.894
131.506     8.388
18      2       24.187      68.476      273.931
276.871    -2.940
19      2       24.159      51.144      207.969
208.753    -0.784
20      2       23.803      68.774      280.205
278.040     2.165
21      3       24.381      55.350      227.060
225.282     1.779
22      3       24.027      44.692      180.605
183.396    -2.791
23      3       24.342      50.995      206.229
208.167    -1.938
24      3       23.670      21.602       91.464
92.649     -1.186
25      3       24.246      54.673      223.869
222.622     1.247
26      3       25.082      41.449      172.910
170.651     2.259
27      3       24.575      35.451      152.073
147.075     4.998
28      3       23.803      42.989      169.427
176.703    -7.276
29      3       24.660      48.599      192.561
198.748    -6.188
30      3       24.097      21.448       94.448
92.042      2.406
31      4       22.816      56.982      222.794
231.697    -8.902
32      4       24.167      47.901      199.003
196.008     2.996
33      4       22.712      40.285      168.668
166.077     2.592
34      4       23.611      25.609      109.387
108.397     0.990
35      4       23.354      22.971       98.445
98.029      0.416
36      4       23.669      25.838      110.987
109.295     1.692
37      4       23.965      49.127      202.662
200.826     1.835
38      4       22.917      54.936      224.773
223.653     1.120
39      4       23.546      50.917      216.058
207.859     8.199
40      4       24.450      41.976      171.469
172.720    -1.251
```

Why Use Residuals?

If the model fit to the data were correct, the residuals would approximate the random errors that make the relationship between the explanatory variables and the response variable a statistical relationship. Therefore, if the residuals appear to behave randomly, it suggests that the model fits the data well. On the other hand, if non-random structure is evident in the residuals, it is a clear sign that the model fits the data poorly. The subsections listed below detail the types of plots to use to test different aspects of a model and give guidance on the correct interpretations of different results that could be observed for each type of plot.

Model Validation

1. How can I assess the sufficiency of the functional part of the model?

Specifics

2. How can I detect non-constant variation across the data?
3. How can I tell if there was drift in the process?
4. How can I assess whether the random errors are independent from one to the next?
5. How can I test whether or not the random errors are distributed normally?
6. How can I test whether any significant terms are missing or misspecified in the functional part of the model?
7. How can I test whether all of the terms in the functional part of the model are necessary?

4.4.5. If my current model does not fit the data well, how can I improve it?

What Next?

Validating a model using residual plots, formal hypothesis tests and descriptive statistics would be quite frustrating if discovery of a problem meant restarting the modeling process back at square one. Fortunately, however, there are also techniques and tools to remedy many of the problems uncovered using residual analysis. In some cases the model validation methods themselves suggest appropriate changes to a model at the same time problems are uncovered. This is especially true of the graphical tools for model validation, though tests on the parameters in the regression function also offer insight into model refinement. Treatments for the various model deficiencies that were diagnosed in Section 4.4.4. are demonstrated and discussed in the subsections listed below.

Methods for Model Improvement

1. Updating the Function Based on Residual Plots
2. Accounting for Non-Constant Variation Across the Data
3. Accounting for Errors with a Non-Normal Distribution

4.5. Use and Interpretation of Process Models

Overview of Section 4.5

This section covers the interpretation and use of the models developed from the collection and analysis of data using the procedures discussed in Section 4.3 and Section 4.4. Three of the main uses of such models, estimation, prediction and calibration, are discussed in detail. Optimization, another important use of this type of model, is primarily discussed in Chapter 5: Process Improvement.

Contents of Section 4.5

1. What types of predictions can I make using the model?
    1. How do I estimate the average response for a particular set of predictor variable values?
    2. How can I predict the value and and estimate the uncertainty of a single response?
2. How can I use my process model for calibration?
    1. Single-Use Calibration Intervals
3. How can I optimize my process using the process model?

4.5.1. What types of predictions can I make using the model?

Detailed Information on Prediction

This section details some of the different types of predictions that can be made using the various process models whose development is discussed in Section 4.1 through Section 4.4. Computational formulas or algorithms are given for each different type of estimation or prediction, along with simulation examples showing its probabilisitic interpretation. An introduction to the different types of estimation and prediction can be found in Section 4.1.3.1. A brief description of estimation and prediction versus the other uses of process models is given in Section 4.1.3.

Different Types of Predictions

1. How do I estimate the average response for a particular set of predictor variable values?
2. How can I predict the value and and estimate the uncertainty of a single response?

4.5.2. How can I use my process model for calibration?

Detailed Calibration Information

This section details some of the different types of calibrations that can be made using the various process models whose development was discussed in previous sections. Computational formulas or algorithms are given for each different type of calibration, along with simulation examples showing its probabilistic interpretation. An introduction to calibration can be found in Section 4.1.3.2. A brief comparison of calibration versus the other uses of process models is given in Section 4.1.3. Additional information on calibration is available in Section 3 of Chapter 2: Measurement Process Characterization.

Calibration Procedures

1. Single-Use Calibration Intervals

4.5.3. How can I optimize my process using the process model?

Detailed Information on Process Optimization

Process optimization using models fit to data collected using response surface designs is primarily covered in Section 5.5.3 of Chapter 5: Process Improvement. In that section detailed information is given on how to determine the correct process inputs to hit a target output value or to maximize or minimize process output. Some background on the use of process models for optimization can be found in Section 4.1.3.3 of this chapter, however, and information on the basic analysis of data from optimization experiments is covered along with that of other types of models in Section 4.1 through Section 4.4 of this chapter.

Contents of Chapter 5 Section 5.5.3.

1. Optimizing a Process
    1. Single response case
        1. Path of steepest ascent
        2. Confidence region for search path
        3. Choosing the step length
        4. Optimization when there is adequate quadratic fit
        5. Effect of sampling error on optimal solution
        6. Optimization subject to experimental region constraints
    2. Multiple response case
        1. Path of steepest ascent
        2. Desirability function approach
        3. Mathematical programming approach

4.6. Case Studies in Process Modeling

Detailed, Realistic Examples

The general points of the first five sections are illustrated in this section using data from physical science and engineering applications. Each example is presented step-by-step in the text and is often cross-linked with the relevant sections of the chapter describing the analysis in general. Each analysis can also be repeated using a worksheet linked to the appropriate Dataplot macros. The worksheet is also linked to the step-by-step analysis presented in the text for easy reference.

Contents: Section 6

1. Load Cell Calibration
    1. Background & Data
    2. Selection of Initial Model
    3. Model Fitting - Initial Model
    4. Graphical Residual Analysis - Initial Model
    5. Interpretation of Numerical Output - Initial Model
    6. Model Refinement
    7. Model Fitting - Model #2
    8. Graphical Residual Analysis - Model #2
    9. Interpretation of Numerical Output - Model #2
    10. Use of the Model for Calibration
    11. Work this Example Yourself
2. Alaska Pipeline Ultrasonic Calibration
    1. Background and Data
    2. Check for Batch Effect
    3. Initial Linear Fit
    4. Transformations to Improve Fit and Equalize Variances
    5. Weighting to Improve Fit
    6. Compare the Fits
    7. Work This Example Yourself
3. Ultrasonic Reference Block Study
    1. Background and Data
    2. Initial Non-Linear Fit
    3. Transformations to Improve Fit
    4. Weighting to Improve Fit
    5. Compare the Fits
    6. Work This Example Yourself
4. Thermal Expansion of Copper Case Study
    1. Background and Data
    2. Exact Rational Models
    3. Initial Plot of Data
    4. Fit Quadratic/Quadratic Model
    5. Fit Cubic/Cubic Model

6. Work This Example Yourself

4.6.1. Load Cell Calibration

Quadratic Calibration

This example illustrates the construction of a linear regression model for load cell data that relates a known load applied to a load cell to the deflection of the cell. The model is then used to calibrate future cell readings associated with loads of unknown magnitude.

1. Background & Data
2. Selection of Initial Model
3. Model Fitting - Initial Model
4. Graphical Residual Analysis - Initial Model
5. Interpretation of Numerical Output - Initial Model
6. Model Refinement
7. Model Fitting - Model #2
8. Graphical Residual Analysis - Model #2
9. Interpretation of Numerical Output - Model #2
10. Use of the Model for Calibration
11. Work This Example Yourself

4.6.2. Alaska Pipeline

Non-Homogeneous Variances

This example illustrates the construction of a linear regression model for Alaska pipeline ultrasonic calibration data. This case study demonstrates the use of transformations and weighted fits to deal with the violation of the assumption of constant standard deviations for the random errors. This assumption is also called homogeneous variances for the errors.

1. Background and Data
2. Check for a Batch Effect
3. Fit Initial Model
4. Transformations to Improve Fit and Equalize Variances
5. Weighting to Improve Fit
6. Compare the Fits
7. Work This Example Yourself

4.6.3. Ultrasonic Reference Block Study

Non-Linear Fit with Non-Homogeneous Variances

This example illustrates the construction of a non-linear regression model for ultrasonic calibration data. This case study demonstrates fitting a non-linear model and the use of transformations and weighted fits to deal with the violation of the assumption of constant standard deviations for the errors. This assumption is also called homogeneous variances for the errors.

1. Background and Data
2. Fit Initial Model
3. Transformations to Improve Fit
4. Weighting to Improve Fit
5. Compare the Fits
6. Work This Example Yourself

4.6.4. Thermal Expansion of Copper Case Study

Rational Function Models

This case study illustrates the use of a class of nonlinear models called rational function models. The data set used is the thermal expansion of copper related to temperature.

This data set was provided by the NIST scientist Thomas Hahn.

Contents

1. Background and Data
2. Rational Function Models
3. Initial Plot of Data
4. Fit Quadratic/Quadratic Model
5. Fit Cubic/Cubic Model
6. Work This Example Yourself

4.7. References For Chapter 4: Process Modeling

Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables (1964) Abramowitz M. and Stegun I. (eds.), U.S. Government Printing Office, Washington, DC, 1046 p.

Berkson J. (1950) "Are There Two Regressions?," *Journal of the American Statistical Association*, Vol. 45, pp. 164-180.

Carroll, R.J. and Ruppert D. (1988) *Transformation and Weighting in Regression*, Chapman and Hall, New York.

Cleveland, W.S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829-836.

Cleveland, W.S. and Devlin, S.J. (1988) "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, Vol. 83, pp. 596-610.

Fuller, W.A. (1987) *Measurement Error Models*, John Wiley and Sons, New York.

Graybill, F.A. (1976) *Theory and Application of the Linear Model*, Duxbury Press, North Sciutate, Massachusetts.

Graybill, F.A. and Iyer, H.K. (1994) *Regression Analysis: Concepts and Applications*, Duxbury Press, Belmont, California.

Harter, H.L. (1983) "Least Squares," *Encyclopedia of Statistical Sciences*, Kotz, S. and Johnson, N.L., eds., John Wiley & Sons, New York, pp. 593-598.

Montgomery, D.C. (2001) *Design and Analysis of Experiments*, 5th ed., Wiley, New York.

Neter, J., Wasserman, W., and Kutner, M. (1983) *Applied Linear Regression Models*, Richard D. Irwin Inc., Homewood, IL.

Ryan, T.P. (1997) *Modern Regression Methods*, Wiley, New York

Seber, G.A.F and Wild, C.F. (1989) *Nonlinear Regression*, John Wiley and Sons, New York.

Stigler, S.M. (1978) "Mathematical Statistics in the Early States," *The Annals of Statistics*, Vol. 6, pp. 239-265.

Stigler, S.M. (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900*, The Belknap Press of Harvard University Press, Cambridge, Massachusetts.

4.8. Some Useful Functions for Process Modeling

Overview of Section 4.8

This section lists some functions commonly-used for process modeling. Constructing an exhaustive list of useful functions is impossible, of course, but the functions given here will often provide good starting points when an empirical model must be developed to describe a particular process.

Each function listed here is classified into a family of related functions, if possible. Its statistical type, linear or nonlinear in the parameters, is also given. Special features of each function, such as asymptotes, are also listed along with the function's domain (the set of allowable input values) and range (the set of possible output values). Plots of some of the different shapes that each function can assume are also included.

Contents of Section 4.8

1. Univariate Functions
    1. Polynomials
    2. Rational Functions

5. Process Improvement

1. Introduction

1. Definition of experimental design
2. Uses
3. Steps

2. Assumptions

1. Measurement system capable
2. Process stable
3. Simple model
4. Residuals well-behaved

3. Choosing an Experimental Design

1. Set objectives
2. Select process variables and levels
3. Select experimental design
    1. Completely randomized designs
    2. Randomized block designs
    3. Full factorial designs
    4. Fractional factorial designs
    5. Plackett-Burman designs
    6. Response surface designs
    7. Adding center point runs
    8. Improving fractional design resolution
    9. Three-level full factorial designs
    10. Three-level, mixed-level and fractional factorial designs

4. Analysis of DOE Data

1. DOE analysis steps
2. Plotting DOE data
3. Modeling DOE data
4. Testing and revising DOE models
5. Interpreting DOE results
6. Confirming DOE results
7. DOE examples
    1. Full factorial example
    2. Fractional factorial example
    3. Response surface example

5. Advanced Topics

1. When classical designs don't work
2. Computer-aided designs
    1. D-Optimal designs

6. Case Studies

1. Eddy current probe sensitivity study
2. Sonoluminescent light intensity study

    2. Repairing a design
3. Optimizing a process
    1. Single response case
    2. Multiple response case
4. Mixture designs
    1. Mixture screening designs
    2. Simplex-lattice designs
    3. Simplex-centroid designs
    4. Constrained mixture designs
    5. Treating mixture and process variables together
5. Nested variation
6. Taguchi designs
7. John's 3/4 fractional factorial designs
8. Small composite designs
9. An EDA approach to experiment design

7. A Glossary of DOE Terminology

8. References

Click here for a detailed table of contents

5. Process Improvement - Detailed Table of Contents [5.]

1. Introduction [5.1.]
    1. What is experimental design? [5.1.1.]
    2. What are the uses of DOE? [5.1.2.]
    3. What are the steps of DOE? [5.1.3.]

2. Assumptions [5.2.]
    1. Is the measurement system capable? [5.2.1.]
    2. Is the process stable? [5.2.2.]
    3. Is there a simple model? [5.2.3.]
    4. Are the model residuals well-behaved? [5.2.4.]

3. Choosing an experimental design [5.3.]
    1. What are the objectives? [5.3.1.]
    2. How do you select and scale the process variables? [5.3.2.]
    3. How do you select an experimental design? [5.3.3.]
        1. Completely randomized designs [5.3.3.1.]
        2. Randomized block designs [5.3.3.2.]
            1. Latin square and related designs [5.3.3.2.1.]
            2. Graeco-Latin square designs [5.3.3.2.2.]
            3. Hyper-Graeco-Latin square designs [5.3.3.2.3.]
        3. Full factorial designs [5.3.3.3.]
            1. Two-level full factorial designs [5.3.3.3.1.]
            2. Full factorial example [5.3.3.3.2.]
            3. Blocking of full factorial designs [5.3.3.3.3.]
        4. Fractional factorial designs [5.3.3.4.]
            1. A 2^{3-1} design (half of a 2^3) [5.3.3.4.1.]
            2. Constructing the 2^{3-1} half-fraction design [5.3.3.4.2.]
            3. Confounding (also called aliasing) [5.3.3.4.3.]
            4. Fractional factorial design specifications and design resolution [5.3.3.4.4.]
            5. Use of fractional factorial designs [5.3.3.4.5.]
            6. Screening designs [5.3.3.4.6.]
            7. Summary tables of useful fractional factorial designs [5.3.3.4.7.]
        5. Plackett-Burman designs [5.3.3.5.]
        6. Response surface designs [5.3.3.6.]
            1. Central Composite Designs (CCD) [5.3.3.6.1.]
            2. Box-Behnken designs [5.3.3.6.2.]
            3. Comparisons of response surface designs [5.3.3.6.3.]
            4. Blocking a response surface design [5.3.3.6.4.]
        7. Adding centerpoints [5.3.3.7.]
        8. Improving fractional factorial design resolution [5.3.3.8.]
            1. Mirror-Image foldover designs [5.3.3.8.1.]

        2. Alternative foldover designs [5.3.3.8.2.]
    9. Three-level full factorial designs [5.3.3.9.]
    10. Three-level, mixed-level and fractional factorial designs [5.3.3.10.]

4. Analysis of DOE data [5.4.]
    1. What are the steps in a DOE analysis? [5.4.1.]
    2. How to "look" at DOE data [5.4.2.]
    3. How to model DOE data [5.4.3.]
    4. How to test and revise DOE models [5.4.4.]
    5. How to interpret DOE results [5.4.5.]
    6. How to confirm DOE results (confirmatory runs) [5.4.6.]
    7. Examples of DOE's [5.4.7.]
        1. Full factorial example [5.4.7.1.]
        2. Fractional factorial example [5.4.7.2.]
        3. Response surface model example [5.4.7.3.]

5. Advanced topics [5.5.]
    1. What if classical designs don't work? [5.5.1.]
    2. What is a computer-aided design? [5.5.2.]
        1. D-Optimal designs [5.5.2.1.]
        2. Repairing a design [5.5.2.2.]
    3. How do you optimize a process? [5.5.3.]
        1. Single response case [5.5.3.1.]
            1. Single response: Path of steepest ascent [5.5.3.1.1.]
            2. Single response: Confidence region for search path [5.5.3.1.2.]
            3. Single response: Choosing the step length [5.5.3.1.3.]
            4. Single response: Optimization when there is adequate quadratic fit [5.5.3.1.4.]
            5. Single response: Effect of sampling error on optimal solution [5.5.3.1.5.]
            6. Single response: Optimization subject to experimental region constraints [5.5.3.1.6.]
        2. Multiple response case [5.5.3.2.]
            1. Multiple responses: Path of steepest ascent [5.5.3.2.1.]
            2. Multiple responses: The desirability approach [5.5.3.2.2.]
            3. Multiple responses: The mathematical programming approach [5.5.3.2.3.]
    4. What is a mixture design? [5.5.4.]
        1. Mixture screening designs [5.5.4.1.]
        2. Simplex-lattice designs [5.5.4.2.]
        3. Simplex-centroid designs [5.5.4.3.]
        4. Constrained mixture designs [5.5.4.4.]
        5. Treating mixture and process variables together [5.5.4.5.]
    5. How can I account for nested variation (restricted randomization)? [5.5.5.]
    6. What are Taguchi designs? [5.5.6.]
    7. What are John's 3/4 fractional factorial designs? [5.5.7.]
    8. What are small composite designs? [5.5.8.]
    9. An EDA approach to experimental design [5.5.9.]
        1. Ordered data plot [5.5.9.1.]
        2. DOE scatter plot [5.5.9.2.]
        3. DOE mean plot [5.5.9.3.]
        4. Interaction effects matrix plot [5.5.9.4.]
        5. Block plot [5.5.9.5.]
        6. DOE Youden plot [5.5.9.6.]
        7. |Effects| plot [5.5.9.7.]
            1. Statistical significance [5.5.9.7.1.]
            2. Engineering significance [5.5.9.7.2.]
            3. Numerical significance [5.5.9.7.3.]

4. Pattern significance [5.5.9.7.4.]
8. Half-normal probability plot [5.5.9.8.]
9. Cumulative residual standard deviation plot [5.5.9.9.]
    1. Motivation: What is a Model? [5.5.9.9.1.]
    2. Motivation: How do we Construct a Goodness-of-fit Metric for a Model? [5.5.9.9.2.]
    3. Motivation: How do we Construct a Good Model? [5.5.9.9.3.]
    4. Motivation: How do we Know When to Stop Adding Terms? [5.5.9.9.4.]
    5. Motivation: What is the Form of the Model? [5.5.9.9.5.]
    6. Motivation: Why is the 1/2 in the Model? [5.5.9.9.6.]
    7. Motivation: What are the Advantages of the LinearCombinatoric Model? [5.5.9.9.7.]
    8. Motivation: How do we use the Model to Generate Predicted Values? [5.5.9.9.8.]
    9. Motivation: How do we Use the Model Beyond the Data Domain? [5.5.9.9.9.]
    10. Motivation: What is the Best Confirmation Point for Interpolation? [5.5.9.9.10.]
    11. Motivation: How do we Use the Model for Interpolation? [5.5.9.9.11.]
    12. Motivation: How do we Use the Model for Extrapolation? [5.5.9.9.12.]
10. DOE contour plot [5.5.9.10.]
    1. How to Interpret: Axes [5.5.9.10.1.]
    2. How to Interpret: Contour Curves [5.5.9.10.2.]
    3. How to Interpret: Optimal Response Value [5.5.9.10.3.]
    4. How to Interpret: Best Corner [5.5.9.10.4.]
    5. How to Interpret: Steepest Ascent/Descent [5.5.9.10.5.]
    6. How to Interpret: Optimal Curve [5.5.9.10.6.]
    7. How to Interpret: Optimal Setting [5.5.9.10.7.]

6. Case Studies [5.6.]
    1. Eddy Current Probe Sensitivity Case Study [5.6.1.]
        1. Background and Data [5.6.1.1.]
        2. Initial Plots/Main Effects [5.6.1.2.]
        3. Interaction Effects [5.6.1.3.]
        4. Main and Interaction Effects: Block Plots [5.6.1.4.]
        5. Estimate Main and Interaction Effects [5.6.1.5.]
        6. Modeling and Prediction Equations [5.6.1.6.]
        7. Intermediate Conclusions [5.6.1.7.]
        8. Important Factors and Parsimonious Prediction [5.6.1.8.]
        9. Validate the Fitted Model [5.6.1.9.]
        10. Using the Fitted Model [5.6.1.10.]
        11. Conclusions and Next Step [5.6.1.11.]
        12. Work This Example Yourself [5.6.1.12.]
    2. Sonoluminescent Light Intensity Case Study [5.6.2.]
        1. Background and Data [5.6.2.1.]
        2. Initial Plots/Main Effects [5.6.2.2.]
        3. Interaction Effects [5.6.2.3.]
        4. Main and Interaction Effects: Block Plots [5.6.2.4.]
        5. Important Factors: Youden Plot [5.6.2.5.]
        6. Important Factors: |Effects| Plot [5.6.2.6.]
        7. Important Factors: Half-Normal Probability Plot [5.6.2.7.]
        8. Cumulative Residual Standard Deviation Plot [5.6.2.8.]
        9. Next Step: DOE Contour Plot [5.6.2.9.]
        10. Summary of Conclusions [5.6.2.10.]
        11. Work This Example Yourself [5.6.2.11.]

7. A Glossary of DOE Terminology [5.7.]

8. References [5.8.]

5.1. Introduction

This section describes the basic concepts of the Design of Experiments (DOE)

This section introduces the basic concepts, terminology, goals and procedures underlying the proper statistical design of experiments. Design of experiments is abbreviated as *DOE* throughout this chapter.

Topics covered are:

- What is experimental design or DOE?
- What are the goals or uses of DOE?
- What are the steps in DOE?

5.1.1. What is experimental design?

Experimental Design (or DOE) economically maximizes information

In an experiment, we deliberately change one or more process variables (or factors) in order to observe the effect the changes have on one or more response variables. The (statistical) design of experiments (*DOE*) is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions.

DOE begins with determining the objectives of an experiment and selecting the process factors for the study. An *Experimental Design* is the laying out of a detailed experimental plan in advance of doing the experiment. Well chosen experimental designs maximize the amount of "information" that can be obtained for a given amount of experimental effort.

The statistical theory underlying DOE generally begins with the concept of *process models*.

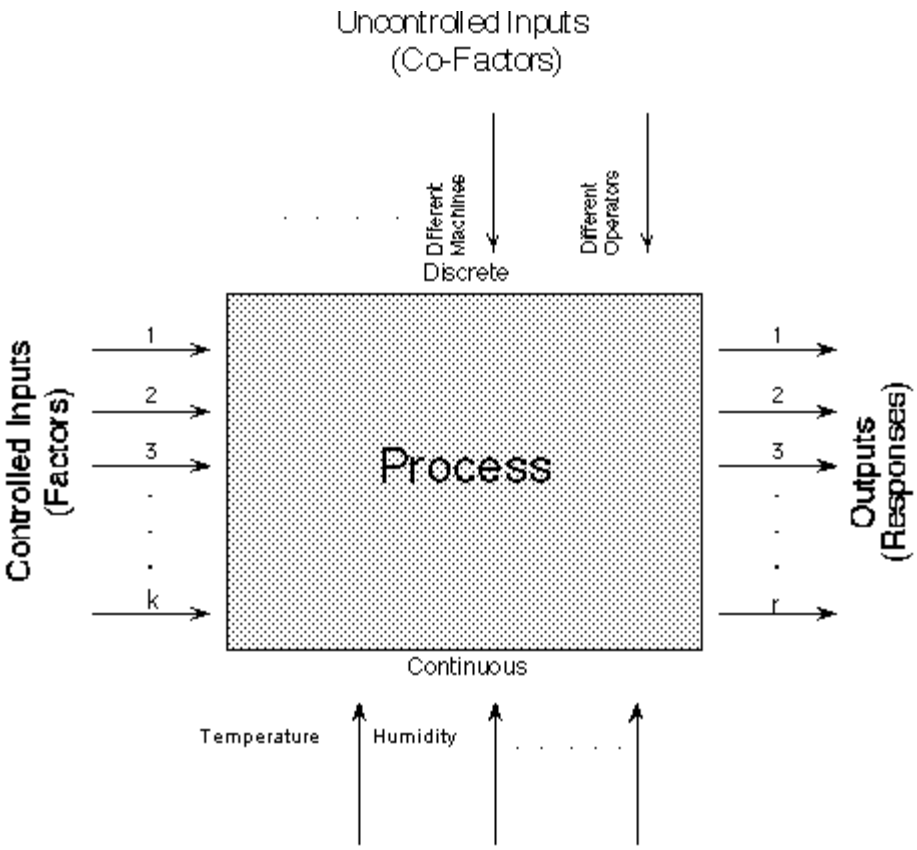
Process Models for DOE

Black box process model

It is common to begin with a process model of the `black box' type, with several discrete or continuous input factors that can be controlled--that is, varied at will by the experimenter--and one or more measured output responses. The output responses are assumed continuous. Experimental data are used to derive an empirical (approximation) model linking the outputs and inputs. These empirical models generally contain first and second-order terms.

Often the experiment has to account for a number of uncontrolled factors that may be discrete, such as different machines or operators, and/or continuous such as ambient temperature or humidity. Figure 1.1 illustrates this situation.

Schematic for a typical process with controlled inputs, outputs, discrete uncontrolled factors and continuous uncontrolled



factors

FIGURE 1.1 A `Black Box' Process Model Schematic

Models for DOE's

The most common empirical models fit to the experimental data take either a *linear* form or *quadratic* form.

Linear model

A linear model with two factors, X_1 and X_2, can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \text{experimental error}$$

Here, Y is the response for given levels of the main effects X_1 and X_2 and the X_1X_2 term is included to account for a possible interaction effect between X_1 and X_2. The constant β_0 is the response of Y when both main effects are 0.

For a more complicated example, a linear model with three factors X_1, X_2, X_3 and one response, Y, would look like (if all possible terms were included in the model)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{123} X_1 X_2 X_3 + \text{experimental error}$$

The three terms with single "X's" are the *main effects* terms. There are $k(k\text{-}1)/2 = 3*2/2 = 3$ *two-way interaction* terms and 1 *three-way* interaction term (which is often omitted, for simplicity). When the experimental data are analyzed, all the unknown "β" parameters are estimated and the coefficients of the "X" terms are tested to see which

ones are significantly different from 0.

Quadratic model

A second-order (quadratic) model (typically used in *response surface* DOE's with suspected curvature) does not include the three-way interaction term but adds three more terms to the linear model, namely

$$\beta_{11}X_1^2 + \beta_{22}X_2^2 + \beta_{33}X_3^2.$$

Note: Clearly, a full model could include many cross-product (or interaction) terms involving squared X's. However, in general these terms are not needed and most DOE software defaults to leaving them out of the model.

5.1.2. What are the uses of DOE?

DOE is a multipurpose tool that can help in many situations

Below are seven examples illustrating situations in which experimental design can be used effectively:

- Choosing Between Alternatives
- Selecting the Key Factors Affecting a Response
- Response Surface Modeling to:
  - Hit a Target
  - Reduce Variability
  - Maximize or Minimize a Response
  - Make a Process Robust (i.e., the process gets the "right" results even though there are uncontrollable "noise" factors)
  - Seek Multiple Goals
- Regression Modeling

Choosing Between Alternatives (Comparative Experiment)

A common use is planning an experiment to gather data to make a decision between two or more alternatives

Supplier A vs. supplier B? Which new additive is the most effective? Is catalyst `x' an improvement over the existing catalyst? These and countless other choices between alternatives can be presented to us in a never-ending parade. Often we have the choice made for us by outside factors over which we have no control. But in many cases we are also asked to make the choice. It helps if one has valid data to back up one's decision.

The preferred solution is to agree on a measurement by which competing choices can be compared, generate a sample of data from each alternative, and compare average results. The 'best' average outcome will be our preference. *We have performed a comparative experiment!*

Types of comparitive studies

Sometimes this comparison is performed under one common set of conditions. This is a comparative study with a narrow scope - which is suitable for some initial comparisons of possible alternatives. Other comparison studies, intended to validate that one alternative is perferred over a wide range of conditions, will purposely and systematically vary the background conditions under which the primary comparison is made in order to reach a conclusion that will be proven valid over a broad scope. We discuss experimental designs for each of these types of comparisons in Sections 5.3.3.1 and 5.3.3.2.

Selecting the Key Factors Affecting a Response (Screening Experiments)

Selecting the few that

Often there are many possible factors, some of which may be critical and others which may have little or no effect on a response. It may be desirable, as a goal by

matter from the many possible factors

itself, to reduce the number of factors to a relatively small set (2-5) so that attention can be focussed on controlling those factors with appropriate specifications, control charts, etc.

Screening experiments are an efficient way, with a minimal number of runs, of determining the important factors. They may also be used as a first step when the ultimate goal is to model a response with a response surface. We will discuss experimental designs for screening a large number of factors in Sections 5.3.3.3, 5.3.3.4 and 5.3.3.5.

Response Surface Modeling a Process

Some reasons to model a process

Once one knows the primary variables (factors) that affect the responses of interest, a number of additional objectives may be pursued. These include:

- Hitting a Target
- Maximizing or Minimizing a Response
- Reducing Variation
- Making a Process Robust
- Seeking Multiple Goals

What each of these purposes have in common is that experimentation is used to fit a model that may permit a rough, local approximation to the actual surface. Given that the particular objective can be met with such an approximate model, the experimental effort is kept to a minimum while still achieving the immediate goal.

These response surface modeling objectives will now be briefly expanded upon.

Hitting a Target

Often we want to "fine tune" a process to consistently hit a target

This is a frequently encountered goal for an experiment.

One might try out different settings until the desired target is `hit' consistently. For example, a machine tool that has been recently overhauled may require some setup `tweaking' before it runs on target. Such action is a small and common form of experimentation. However, rather than experimenting in an ad hoc manner until we happen to find a setup that hits the target, one can fit a model estimated from a small experiment and use this model to determine the necessary adjustments to hit the target.

More complex forms of experimentation, such as the determination of the correct chemical mix of a coating that will yield a desired refractive index for the dried coat (and simultaneously achieve specifications for other attributes), may involve many ingredients and be very sensitive to small changes in the percentages in the mix. Fitting suitable models, based on sequentially planned experiments, may be the only way to efficiently achieve this goal of hitting targets for multiple responses simultaneously.

Maximizing or Minimizing a Response

Optimizing a process output is a common

Many processes are being run at sub-optimal settings, some of them for years, even though each factor has been optimized individually over time. Finding settings that increase yield or decrease the amount of scrap and rework represent opportunities for substantial financial gain. Often, however, one must experiment with multiple

goal

inputs to achieve a better output. Section 5.3.3.6 on second-order designs plus material in Section 5.5.3 will be useful for these applications.

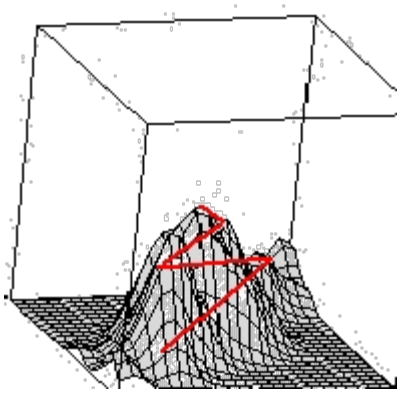


FIGURE 1.1 Pathway up the process response surface to an `optimum'

Reducing Variation

Processes that are on target, on the average, may still have too much variability

A process may be performing with unacceptable consistency, meaning its internal variation is too high.

Excessive variation can result from many causes. Sometimes it is due to the lack of having or following standard operating procedures. At other times, excessive variation is due to certain hard-to-control inputs that affect the critical output characteristics of the process. When this latter situation is the case, one may experiment with these hard-to-control factors, looking for a region where the surface is flatter and the process is easier to manage. To take advantage of such flatness in the surface, one must use designs - such as the second-order designs of Section 5.3.3.6 - that permit identification of these features. Contour or surface plots are useful for elucidating the key features of these fitted models. See also 5.5.3.1.4.

Graph of data before variation reduced

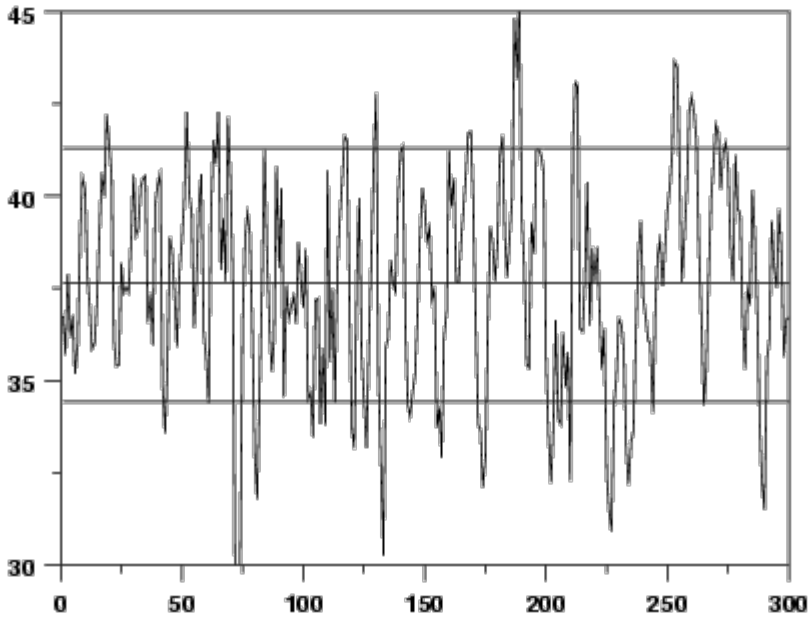


Figure 1.2 Process before variation reduced

It might be possible to reduce the variation by altering the setpoints (recipe) of the process, so that it runs in a more `stable' region.

Graph of data after process variation reduced

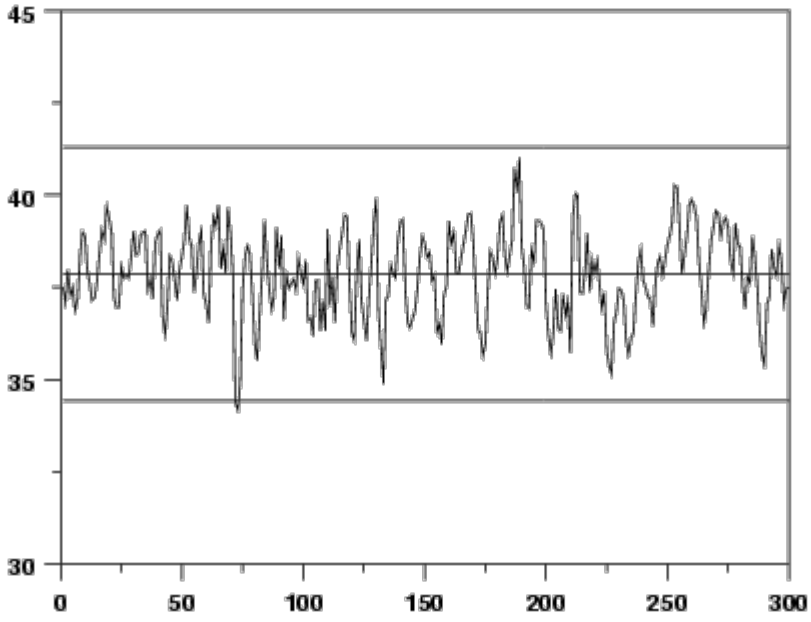


Figure 1.3 Process after variation reduced

Finding this new recipe could be the subject of an experiment, especially if there are

many input factors that could conceivably affect the output.

Making a Process Robust

The less a process or product is affected by external conditions, the better it is - this is called "Robustness"

An item designed and made under controlled conditions will be later `field tested' in the hands of the customer and may prove susceptible to failure modes not seen in the lab or thought of by design. An example would be the starter motor of an automobile that is required to operate under extremes of external temperature. A starter that performs under such a wide range is termed `robust' to temperature.

Designing an item so that it is robust calls for a special experimental effort. It is possible to stress the item in the design lab and so determine the critical components affecting its performance. A different gauge of armature wire might be a solution to the starter motor, but so might be many other alternatives. The correct combination of factors can be found only by experimentation.

Seeking Multiple Goals

Sometimes we have multiple outputs and we have to compromise to achieve desirable outcomes - DOE can help here

A product or process seldom has just one desirable output characteristic. There are usually several, and they are often interrelated so that improving one will cause a deterioration of another. For example: rate vs. consistency; strength vs. expense; etc.

Any product is a trade-off between these various desirable final characteristics. Understanding the boundaries of the trade-off allows one to make the correct choices. This is done by either constructing some weighted objective function (`*desirability function*') and optimizing it, or examining contour plots of responses generated by a computer program, as given below.

Sample contour plot of deposition rate and capability

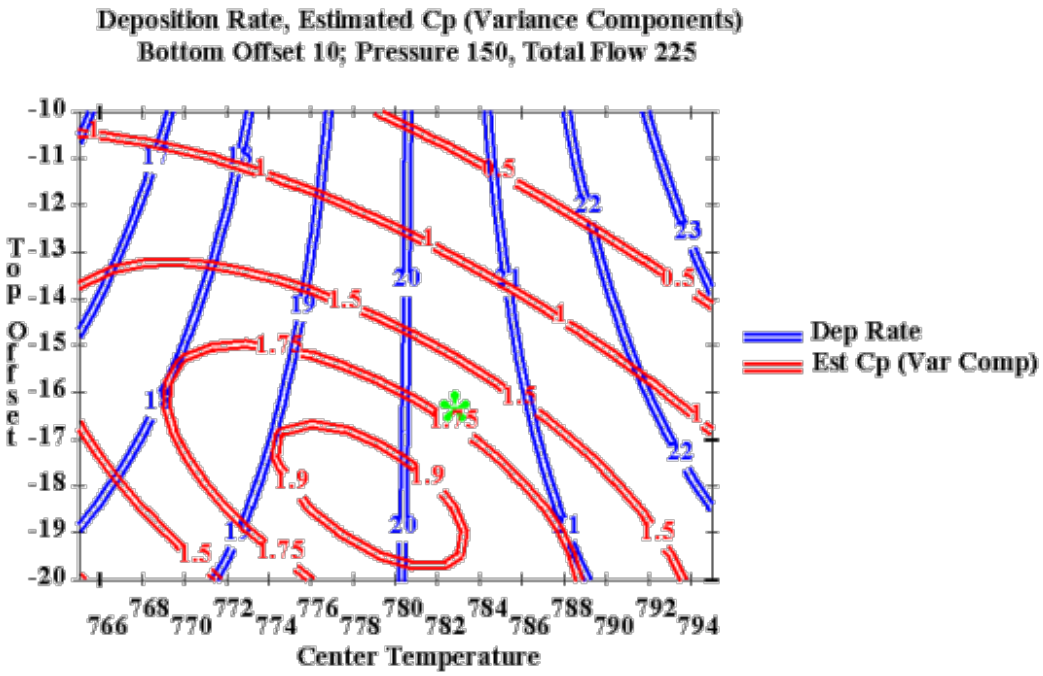


FIGURE 1.4 Overlaid contour plot of Deposition Rate and Capability (Cp)

Regression Modeling

Regression models (Chapter 4) are used to fit more precise models	Sometimes we require more than a rough approximating model over a local region. In such cases, the standard designs presented in this chapter for estimating first- or second-order polynomial models may not suffice. Chapter 4 covers the topic of experimental design and analysis for fitting general models for a single explanatory factor. If one has multiple factors, and either a nonlinear model or some other special model, the computer-aided designs of Section 5.5.2 may be useful.

5.1.3. What are the steps of DOE?

Key steps for DOE

Obtaining good results from a DOE involves these seven steps:

1. Set objectives
2. Select process variables
3. Select an experimental design
4. Execute the design
5. Check that the data are consistent with the experimental assumptions
6. Analyze and interpret the results
7. Use/present the results (may lead to further runs or DOE's).

A checklist of practical considerations

Important practical considerations in planning and running experiments are

- Check performance of gauges/measurement devices first.
- Keep the experiment as simple as possible.
- Check that all planned runs are feasible.
- Watch out for process drifts and shifts during the run.
- Avoid unplanned changes (e.g., swap operators at halfway point).
- Allow some time (and back-up material) for unexpected events.
- Obtain buy-in from all parties involved.
- Maintain effective ownership of each step in the experimental plan.
- Preserve all the raw data--do not keep only summary averages!
- Record everything that happens.
- Reset equipment to its original state after the experiment.

The *Sequential* or *Iterative* Approach to DOE

Planning to do a sequence of small experiments is

It is often a mistake to believe that `one big experiment will give the answer.'

A more useful approach to experimental design is to

often better than relying on one big experiment to give you all the answers

recognize that while one experiment might provide a useful result, it is more common to perform two or three, or maybe more, experiments before a complete answer is attained. In other words, an iterative approach is best and, in the end, most economical. Putting all one's eggs in one basket is not advisable.

Each stage provides insight for next stage

The reason an iterative approach frequently works best is because it is logical to move through stages of experimentation, each stage providing insight as to how the next experiment should be run.

5.2. Assumptions

We should check the engineering and model-building assumptions that are made in most DOE's

In all model building we make assumptions, and we also require certain conditions to be approximately met for purposes of estimation. This section looks at some of the engineering and mathematical assumptions we typically make. These are:

- Are the measurement systems capable for all of your responses?
- Is your process stable?
- Are your responses likely to be approximated well by simple polynomial models?
- Are the residuals (the difference between the model predictions and the actual observations) well behaved?

5.2.1. Is the measurement system capable?

Metrology capabilities are a key factor in most experiments	It is unhelpful to find, after you have finished all the experimental runs, that the measurement devices you have at your disposal cannot measure the changes you were hoping to see. Plan to check this out before embarking on the experiment itself. Measurement process characterization is covered in Chapter 2.
SPC check of measurement devices	In addition, it is advisable, especially if the experimental material is planned to arrive for measurement over a protracted period, that an SPC (i.e., quality control) check is kept on all measurement devices from the start to the conclusion of the whole experimental project. Strange experimental outcomes can often be traced to `hiccups' in the metrology system.

5.2.2. Is the process stable?

Plan to examine process stability as part of your experiment

Experimental runs should have control runs that are made at the `standard' process setpoints, or at least at some standard operating recipe. The experiment should start and end with such runs. A plot of the outcomes of these control runs will indicate if the underlying process itself has drifted or shifted during the experiment.

It is desirable to experiment on a stable process. However, if this cannot be achieved, then the process instability must be accounted for in the analysis of the experiment. For example, if the mean is shifting with time (or experimental trial run), then it will be necessary to include a trend term in the experimental model (i.e., include a time variable or a run number variable).

5.2.3. Is there a simple model?

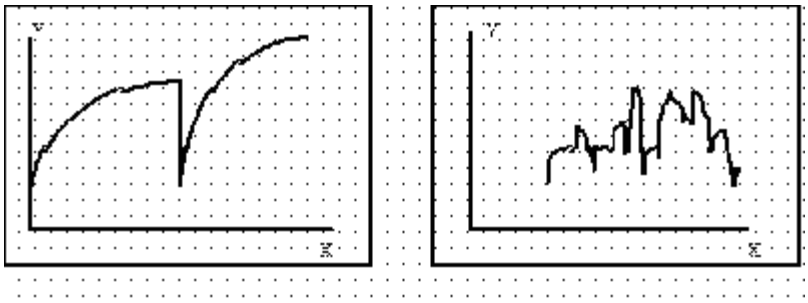
Polynomial approximation models only work for smoothly varying outputs

In this chapter we restrict ourselves to the case for which the response variable(s) are continuous outputs denoted as *Y*. Over the experimental range, the outputs must not only be continuous, but also reasonably smooth. A sharp falloff in *Y* values is likely to be missed by the approximating polynomials that we use because these polynomials assume a smoothly curving underlying response surface.

Piecewise smoothness requires separate experiments

If the surface under investigation is known to be only piecewise smooth, then the experiments will have to be broken up into separate experiments, each investigating the shape of the separate sections. A surface that is known to be very jagged (i.e., non-smooth) will not be successfully approximated by a smooth polynomial.

Examples of piecewise smooth and jagged responses



Piecewise Smooth **Jagged**

FIGURE 2.1 Examples of Piecewise Smooth and Jagged Responses

5.2.4. Are the model residuals well-behaved?

Residuals are the differences between the observed and predicted responses

Residuals are estimates of experimental error obtained by *subtracting the observed responses from the predicted responses.*

The predicted response is calculated from the chosen model, after all the unknown model parameters have been estimated from the experimental data.

Examining residuals is a key part of all statistical modeling, including DOE's. Carefully looking at residuals can tell us whether our assumptions are reasonable and our choice of model is appropriate.

Residuals are elements of variation unexplained by fitted model

Residuals can be thought of as elements of variation unexplained by the fitted model. Since this is a form of error, the same general assumptions apply to the group of residuals that we typically use for errors in general: *one expects them to be (roughly) normal and (approximately) independently distributed with a mean of 0 and some constant variance*.

Assumptions for residuals

These are the assumptions behind ANOVA and classical regression analysis. This means that an analyst should expect a regression model to err in predicting a response in a random fashion; the model should predict values higher than actual and lower than actual with equal probability. In addition, the level of the error should be independent of when the observation occurred in the study, or the size of the observation being predicted, or even the factor settings involved in making the prediction. The overall pattern of the residuals should be similar to the bell-shaped pattern observed when plotting a histogram of normally distributed data.

We emphasize the use of *graphical methods* to examine residuals.

Departures indicate inadequate model

Departures from these assumptions usually mean that the residuals contain *structure* that is not accounted for in the model. Identifying that structure and adding term(s) representing it to the original model leads to a better model.

Tests for Residual Normality

Plots for examining residuals

Any graph suitable for displaying the distribution of a set of data is suitable for judging the normality of the distribution of a group of residuals. The three most common types are:

1. histograms,
2. normal probability plots, and
3. dot plots.

Histogram

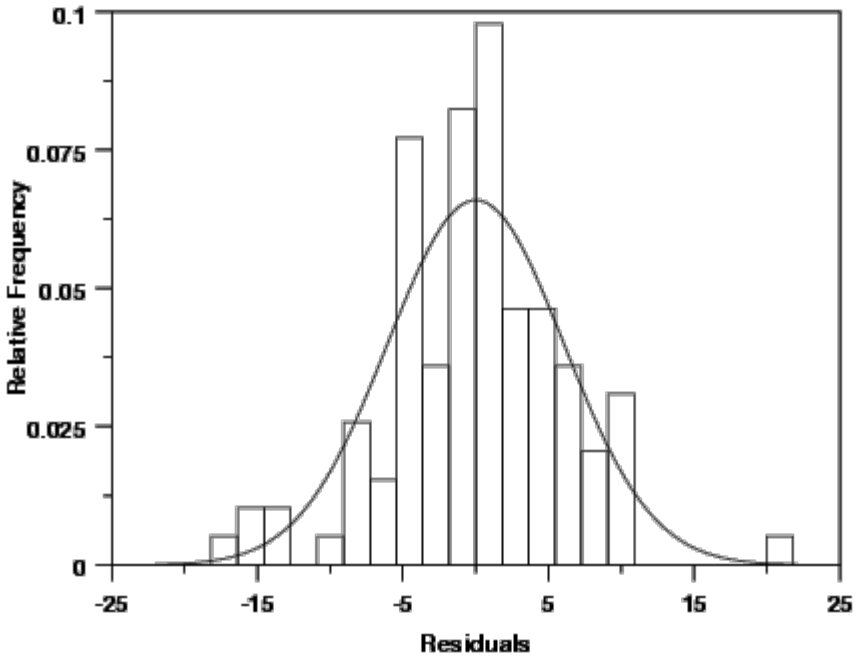


Figure 2.2

The histogram is a frequency plot obtained by placing the data in regularly spaced cells and plotting each cell frequency versus the center of the cell. Figure 2.2 illustrates an approximately normal distribution of residuals produced by a model for a calibration process. We have superimposed a normal density function on the histogram.

Small sample sizes

Sample sizes of residuals are generally small (<50) because experiments have limited treatment combinations, so a histogram is not be the best choice for judging the distribution of residuals. A more sensitive graph is the normal probability plot.

Normal probability plot

The steps in forming a normal probability plot are:

- Sort the residuals into ascending order.
- Calculate the cumulative probability of each residual using the formula:

  $$P(i\text{-th residual}) = i/(N+1)$$

  with P denoting the cumulative probability of a point, i is the order of the value in the list and N is the number of entries in the list.
- Plot the calculated p-values versus the residual value on normal probability paper.

The normal probability plot should produce an approximately straight line if the points come from a normal distribution.

Sample

Figure 2.3 below illustrates the normal probability graph created from the same

normal probability plot with overlaid dot plot

group of residuals used for Figure 2.2.

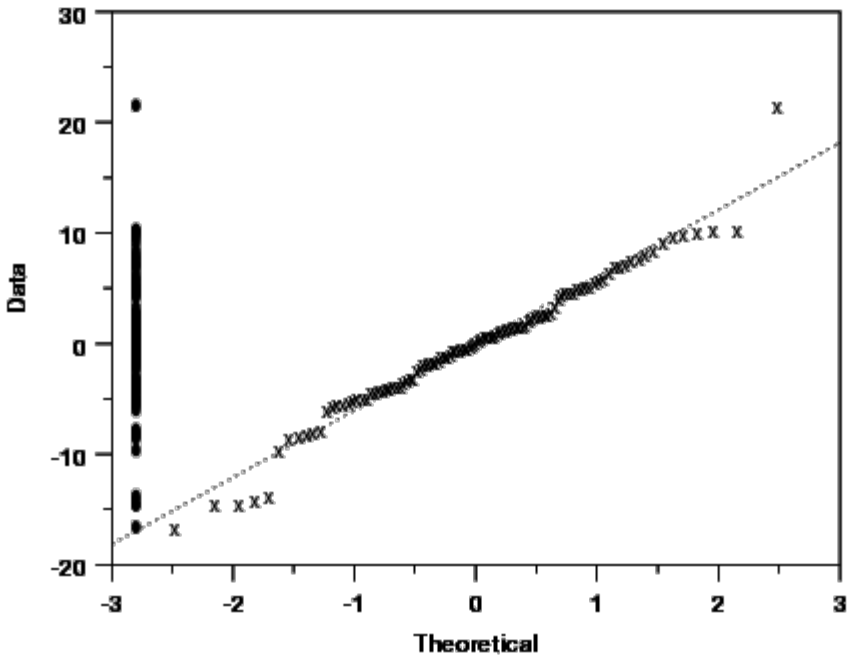


Figure 2.3

This graph includes the addition of a dot plot. The dot plot is the collection of points along the left y-axis. These are the values of the residuals. The purpose of the dot plot is to provide an indication the distribution of the residuals.

"S" shaped curves indicate bimodal distribution

Small departures from the straight line in the normal probability plot are common, but a clearly "S" shaped curve on this graph suggests a bimodal distribution of residuals. Breaks near the middle of this graph are also indications of abnormalities in the residual distribution.

NOTE: Studentized residuals are residuals converted to a scale approximately representing the standard deviation of an individual residual from the center of the residual distribution. The technique used to convert residuals to this form produces a Student's t distribution of values.

Independence of Residuals Over Time

Run sequence plot

If the order of the observations in a data table represents the order of execution of each treatment combination, then a plot of the residuals of those observations versus the case order or time order of the observations will test for any time dependency. These are referred to as run sequence plots.

Sample run sequence plot that exhibits a time trend

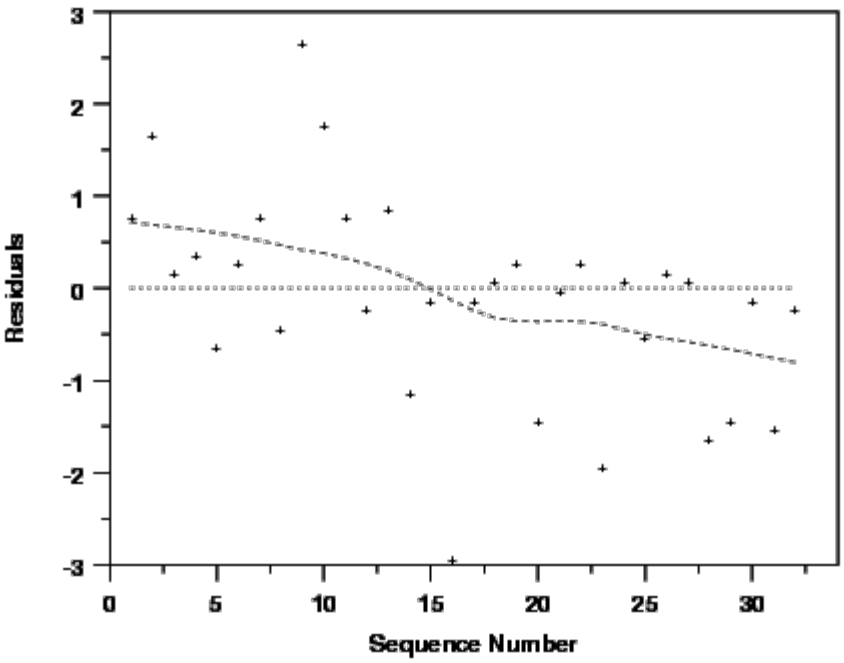


Figure 2.4

Sample run sequence plot that does not exhibit a time trend

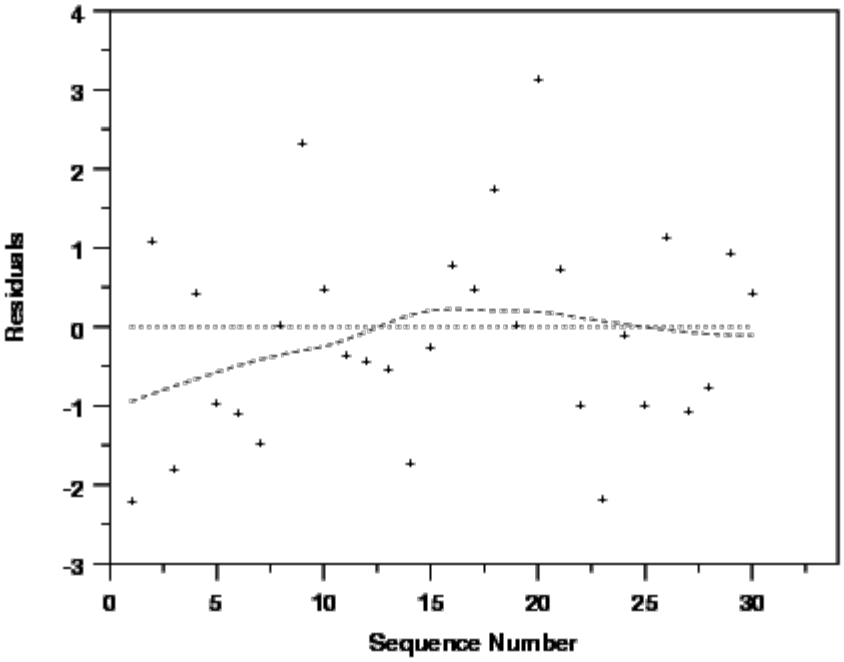


Figure 2.5

Interpretation of the sample run sequence plots

The residuals in Figure 2.4 suggest a time trend, while those in Figure 2.5 do not. Figure 2.4 suggests that the system was drifting slowly to lower values as the investigation continued. In extreme cases a drift of the equipment will produce models with very poor ability to account for the variability in the data (low R^2).

If the investigation includes centerpoints, then plotting them in time order may produce a more clear indication of a time trend if one exists. Plotting the raw responses in time sequence can also sometimes detect trend changes in a process that residual plots might not detect.

Plot of Residuals Versus Corresponding Predicted Values

Check for increasing residuals as size of fitted value increases

Plotting residuals versus the value of a fitted response should produce a distribution of points scattered randomly about 0, regardless of the size of the fitted value. Quite commonly, however, residual values may increase as the size of the fitted value increases. When this happens, the residual cloud becomes "funnel shaped" with the larger end toward larger fitted values; that is, the residuals have larger and larger scatter as the value of the response increases. Plotting the absolute values of the residuals instead of the signed values will produce a "wedge-shaped" distribution; a smoothing function is added to each graph which helps to show the trend.

Sample residuals versus fitted values plot showing increasing residuals

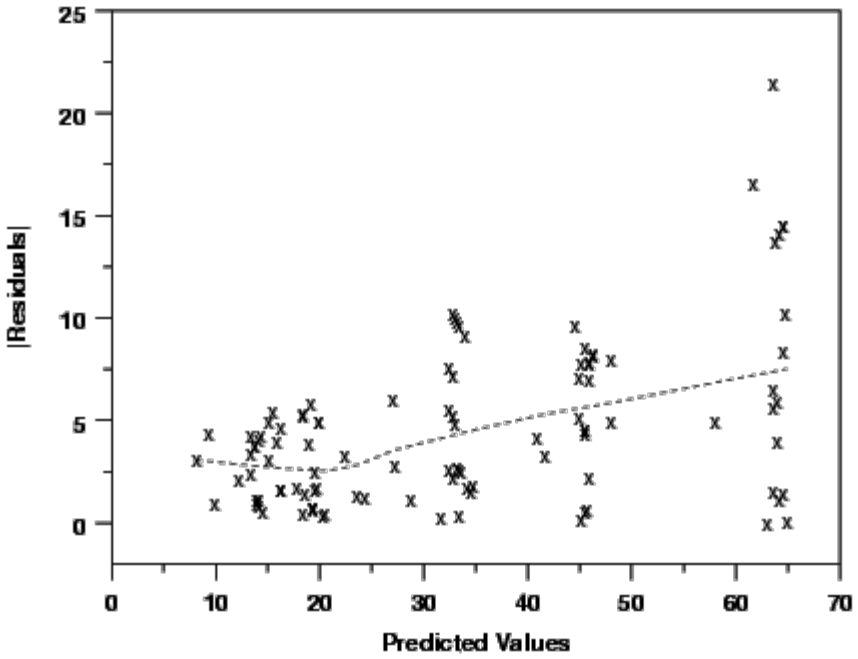


Figure 2.6

Sample residuals versus fitted values plot that does not show increasing residuals

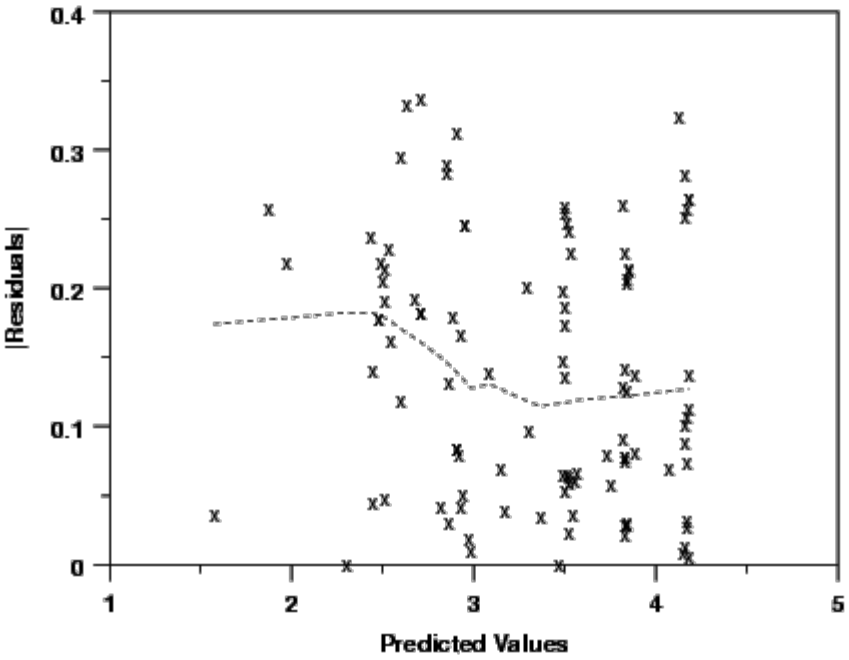


Figure 2.7

Interpretation of the residuals versus fitted values plots

A residual distribution such as that in Figure 2.6 showing a trend to higher absolute residuals as the value of the response increases suggests that one should transform the response, perhaps by modeling its logarithm or square root, etc., (contractive transformations). Transforming a response in this fashion often simplifies its relationship with a predictor variable and leads to simpler models. Later sections discuss transformation in more detail. Figure 2.7 plots the residuals after a transformation on the response variable was used to reduce the scatter. Notice the difference in scales on the vertical axes.

Independence of Residuals from Factor Settings

Sample residuals versus factor setting plot

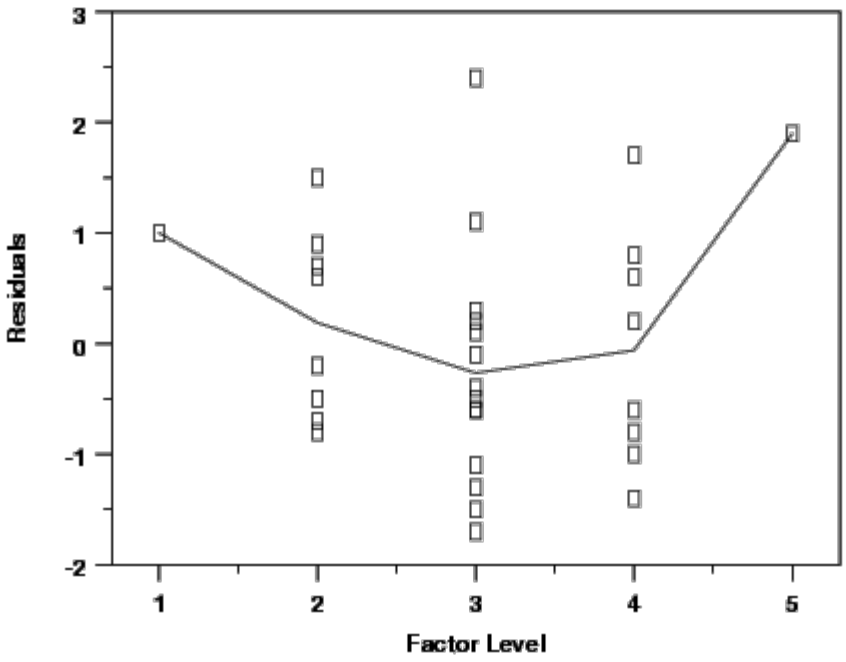


Figure 2.8

Sample residuals versus factor setting plot after adding a quadratic term

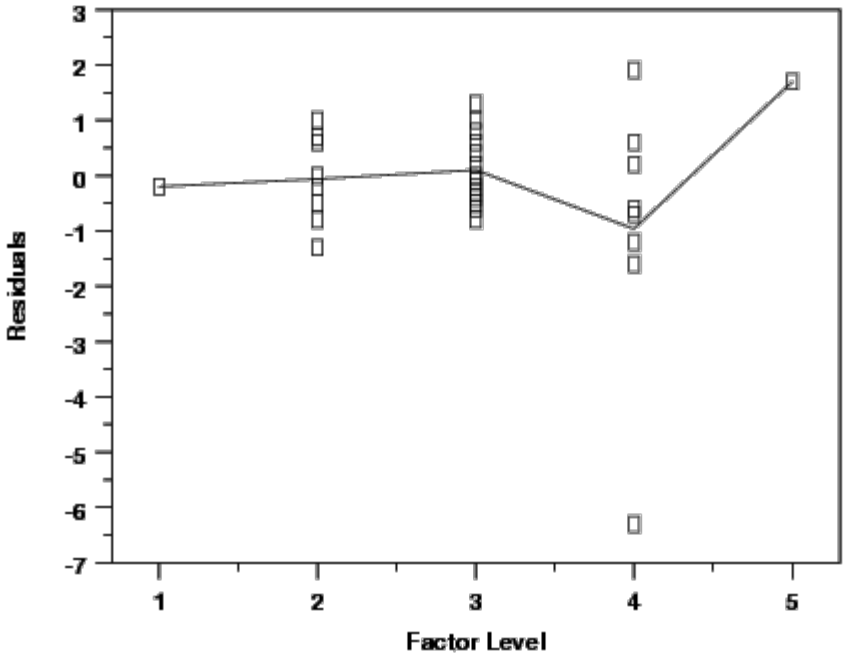


Figure 2.9

Interpreation of residuals versus factor setting plots

Figure 2.8 shows that the size of the residuals changed as a function of a predictor's settings. A graph like this suggests that the model needs a higher-order term in that predictor or that one should transform the predictor using a logarithm or square root, for example. Figure 2.9 shows the residuals for the same response after adding a

quadratic term. Notice the single point widely separated from the other residuals in Figure 2.9. This point is an "outlier." That is, its position is well within the range of values used for this predictor in the investigation, but its result was somewhat lower than the model predicted. A signal that curvature is present is a trace resembling a "frown" or a "smile" in these graphs.

Sample residuals versus factor setting plot lacking one or more higher-order terms

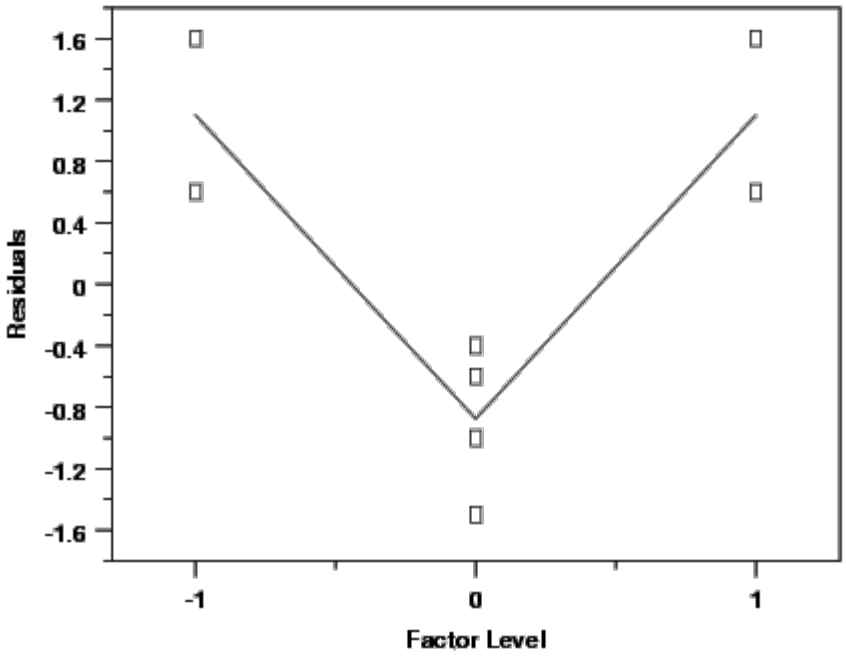


Figure 2.10

Interpretation of plot

The example given in Figures 2.8 and 2.9 obviously involves five levels of the predictor. The experiment utilized a response surface design. For the simple factorial design that includes center points, if the response model being considered lacked one or more higher-order terms, the plot of residuals versus factor settings might appear as in Figure 2.10.

Graph indicates prescence of curvature

While the graph gives a definite signal that curvature is present, identifying the source of that curvature is not possible due to the structure of the design. Graphs generated using the other predictors in that situation would have very similar appearances.

Additional discussion of residual analysis

Note: Residuals are an important subject discussed repeatedly in this Handbook. For example, graphical residual plots are discussed in Chapter 1 and the general examination of residuals as a part of model building is discussed in Chapter 4.

5.3. Choosing an experimental design

Contents of Section 3

This section describes in detail the process of choosing an experimental design to obtain the results you need. The basic designs an engineer needs to know about are described in detail.

Note that this section describes the basic designs used for most engineering and scientific applications

1. Set objectives
2. Select process variables and levels
3. Select experimental design
    1. Completely randomized designs
    2. Randomized block designs
        1. Latin squares
        2. Graeco-Latin squares
        3. Hyper-Graeco-Latin squares
    3. Full factorial designs
        1. Two-level full factorial designs
        2. Full factorial example
        3. Blocking of full factorial designs
    4. Fractional factorial designs
        1. A 2^{3-1} half-fraction design
        2. How to construct a 2^{3-1} design
        3. Confounding
        4. Design resolution
        5. Use of fractional factorial designs
        6. Screening designs
        7. Fractional factorial designs summary tables
    5. Plackett-Burman designs
    6. Response surface (second-order) designs
        1. Central composite designs
        2. Box-Behnken designs
        3. Response surface design comparisons
        4. Blocking a response surface design
    7. Adding center points
    8. Improving fractional design resolution
        1. Mirror-image foldover designs
        2. Alternative foldover designs
    9. Three-level full factorial designs
    10. Three-level, mixed level and fractional factorial designs

5.3.1. What are the objectives?

Planning an experiment begins with carefully considering what the objectives (or goals) are

The objectives for an experiment are best determined by a team discussion. All of the objectives should be written down, even the "unspoken" ones.

The group should discuss which objectives are the key ones, and which ones are "nice but not really necessary". Prioritization of the objectives helps you decide which direction to go with regard to the selection of the factors, responses and the particular design. Sometimes prioritization will force you to start over from scratch when you realize that the experiment you decided to run does not meet one or more critical objectives.

Types of designs

Examples of goals were given earlier in Section 5.1.2, in which we described four broad categories of experimental designs, with various objectives for each. These were:

- **Comparative designs** to:
  - choose between alternatives, with narrow scope, suitable for an initial comparison (see Section 5.3.3.1)
  - choose between alternatives, with broad scope, suitable for a confirmatory comparison (see Section 5.3.3.2)
- **Screening designs** to identify which factors/effects are important
  - when you have 2 - 4 factors and can perform a full factorial (Section 5.3.3.3)
  - when you have more than 3 factors and want to begin with as small a design as possible (Section 5.3.3.4 and 5.3.3.5)
  - when you have some qualitative factors, or you have some quantitative factors that are known to have a non-monotonic effect (Section 3.3.3.10)

  Note that some authors prefer to restrict the term screening design to the case where you are trying to extract the most important factors from a large (say > 5) list of initial factors (usually a fractional factorial design). We include the case with a smaller number of factors, usually a full factorial design, since the basic purpose and analysis is similar.

- **Response Surface** modeling to achieve one or more of the following objectives:
  - hit a target
  - maximize or minimize a response
  - reduce variation by locating a region where the process is easier to manage
  - make a process robust (note: this objective may often be accomplished with screening designs rather than with response surface designs - see Section 5.5.6)

- **Regression modeling**
  - to estimate a precise model, quantifying the dependence of response variable(s) on process inputs.

Based on objective, where to go next

After identifying the objective listed above that corresponds most closely to your specific goal, you can

- proceed to the next section in which we discuss selecting experimental factors

and then

- select the appropriate design named in section 5.3.3 that suits your objective (and follow the related links).

5.3.2. How do you select and scale the process variables?

Guidelines to assist the engineering judgment process of selecting process variables for a DOE

Process variables include both *inputs* and *outputs* - i.e., *factors* and *responses*. The selection of these variables is best done as a team effort. The team should

- Include all important factors (based on engineering judgment).
- Be bold, but not foolish, in choosing the low and high factor levels.
- Check the factor settings for impractical or impossible combinations - i.e., very low pressure and very high gas flows.
- Include all relevant responses.
- Avoid using only responses that combine two or more measurements of the process. For example, if interested in selectivity (the ratio of two etch rates), measure both rates, not just the ratio.

Be careful when choosing the allowable range for each factor

We have to choose the range of the settings for input factors, and it is wise to give this some thought beforehand rather than just try extreme values. In some cases, extreme values will give runs that are not feasible; in other cases, extreme ranges might move one out of a smooth area of the response surface into some jagged region, or close to an asymptote.

Two-level designs have just a "high" and a "low" setting for each factor

The most popular experimental designs are *two-level designs*. Why only two levels? There are a number of good reasons why two is the most common choice amongst engineers: one reason is that it is ideal for screening designs, simple and economical; it also gives most of the information required to go to a multilevel response surface experiment if one is needed.

Consider adding some center points to your two-level design

The term "two-level design" is something of a misnomer, however, as it is recommended to include some center points during the experiment (center points are located in the middle of the design `box').

Notation for 2-Level Designs

Matrix notation for describing an experiment

The standard layout for a 2-level design uses +1 and -1 notation to denote the "high level" and the "low level" respectively, for each factor. For example, the matrix below

	Factor 1 (X1)	Factor 2 (X2)
Trial 1	-1	-1
Trial 2	+1	-1
Trial 3	-1	+1
Trial 4	+1	+1

describes an experiment in which 4 trials (or runs) were conducted with each factor set to high or low during a run according to whether the matrix had a +1 or -1 set for the factor during that trial. If the experiment had more than 2 factors, there would be an additional column in the matrix for each additional factor.

Note: Some authors shorten the matrix notation for a two-level design by just recording the plus and minus signs, leaving out the "1's".

Coding the data

The use of +1 and -1 for the factor settings is called *coding* the data. This aids in the interpretation of the coefficients fit to any experimental model. *After factor settings are coded, center points have the value "0"*. Coding is described in more detail in the DOE glossary.

The Model or Analysis Matrix

Design matrices

If we add an "I" column and an "X1*X2" column to the matrix of 4 trials for a two-factor experiment described earlier, we obtain what is known as the *model or analysis matrix* for this simple experiment, which is shown below. The model matrix for a three-factor experiment is shown later in this section.

I	X1	X2	X1*X2
+1	-1	-1	+1
+1	+1	-1	-1
+1	-1	+1	-1
+1	+1	+1	+1

Model for the experiment

The model for this experiment is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \text{experimental error}$$

and the "I" column of the design matrix has all 1's to provide for the β_0 term. The X1*X2 column is formed by multiplying the "X1" and "X2" columns together, row element by row element. This column gives interaction term for each trial.

Model in matrix notation

In matrix notation, we can summarize this experiment by

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \text{experimental error}$$

for which ***X***is the 4 by 4 design matrix of 1's and -1's shown above, β is the vector of unknown model coefficients $(\beta_0, \beta_1, \beta_2, \beta_{12})$ and ***Y*** is a vector consisting of the four trial response observations.

Orthogonal Property of Scaling in a 2-Factor Experiment

Coding produces orthogonal columns

Coding is sometime called "*orthogonal coding*" since all the columns of a coded 2-factor design matrix (except the "I" column) are typically orthogonal. That is, the dot product for any pair of columns is zero. For example, for X1 and X2: (-1)(-1) + (+1)(-1) + (-1)(+1) + (+1)(+1) = 0.

5.3.3. How do you select an experimental design?

A design is selected based on the experimental objective and the number of factors

The choice of an experimental design depends on the objectives of the experiment and the number of factors to be investigated.

Experimental Design Objectives

Types of designs are listed here according to the experimental objective they meet

Types of designs are listed here according to the experimental objective they meet.

- **Comparative objective**: If you have one or several factors under investigation, but the primary goal of your experiment is to make a conclusion about one a-priori important factor, (in the presence of, and/or in spite of the existence of the other factors), and the question of interest is whether or not that factor is "significant", (i.e., whether or not there is a significant change in the response for different levels of that factor), then you have a *comparative problem* and you need a *comparative design* solution.

- **Screening objective**: The primary purpose of the experiment is to select or *screen out* the few important main effects from the many less important ones. These *screening designs* are also termed main effects designs.

- **Response Surface (method) objective**: The experiment is designed to allow us to estimate interaction and even quadratic effects, and therefore give us an idea of the (local) shape of the response surface we are investigating. For this reason, they are termed *response surface method (RSM) designs*. RSM designs are used to:
    - Find improved or optimal process settings
    - Troubleshoot process problems and weak points
    - Make a product or process more *robust* against

external and non-controllable influences. "Robust" means relatively insensitive to these influences.

- **Optimizing responses when factors are proportions of a mixture objective**: If you have factors that are proportions of a mixture and you want to know what the "best" proportions of the factors are so as to maximize (or minimize) a response, then you need a *mixture design*.

- **Optimal fitting of a regression model objective**: If you want to model a response as a mathematical function (either known or empirical) of a few continuous factors and you desire "good" model parameter estimates (i.e., unbiased and minimum variance), then you need a *regression design.*

Mixture and regression designs

Mixture designs are discussed briefly in section 5 (Advanced Topics) and regression designs for a single factor are discussed in chapter 4. Selection of designs for the remaining 3 objectives is summarized in the following table.

Summary table for choosing an experimental design for comparative, screening, and response surface designs

TABLE 3.1 Design Selection Guideline

Number of Factors	Comparative Objective	Screening Objective	Response Surface Objective
1	1-factor completely randomized design	–	–
2 - 4	Randomized block design	Full or fractional factorial	Central composite or Box-Behnken
5 or more	Randomized block design	Fractional factorial or Plackett-Burman	Screen first to reduce number of factors

Resources and degree of control over wrong decisions

Choice of a design from within these various types depends on the amount of resources available and the degree of control over making wrong decisions (Type I and Type II errors for testing hypotheses) that the experimenter desires.

Save some runs for center points and "redos" that might

It is a good idea to choose a design that requires somewhat fewer runs than the budget permits, so that center point runs can be added to check for curvature in a 2-level screening design and backup resources are available to redo runs that have processing mishaps.

be needed

5.3.3.1. Completely randomized designs

These designs are for studying the effects of one primary factor without the need to take other nuisance factors into account

Here we consider completely randomized designs that have one primary factor. The experiment compares the values of a response variable based on the different levels of that primary factor.

For completely randomized designs, the levels of the primary factor are randomly assigned to the experimental units. By randomization, we mean that the run sequence of the experimental units is determined randomly. For example, if there are 3 levels of the primary factor with each level to be run 2 times, then there are 6 factorial possible run sequences (or 6! ways to order the experimental trials). Because of the replication, the number of unique orderings is 90 (since 90 = 6!/(2!*2!*2!)). An example of an unrandomized design would be to always run 2 replications for the first level, then 2 for the second level, and finally 2 for the third level. To randomize the runs, one way would be to put 6 slips of paper in a box with 2 having level 1, 2 having level 2, and 2 having level 3. Before each run, one of the slips would be drawn blindly from the box and the level selected would be used for the next run of the experiment.

Randomization typically performed by computer software

In practice, the randomization is typically performed by a computer program. However, the randomization can also be generated from random number tables or by some physical mechanism (e.g., drawing the slips of paper).

Three key numbers

All completely randomized designs with one primary factor are defined by 3 numbers:

k = number of factors (= 1 for these designs)
L = number of levels
n = number of replications

and the total sample size (number of runs) is $N = k \; x \; L \; x \; n$.

Balance

Balance dictates that the number of replications be the same at each level of the factor (this will maximize the sensitivity of subsequent statistical t (or F) tests).

Typical example of a completely randomized design

A typical example of a completely randomized design is the following:

k = 1 factor ($X1$)
L = 4 levels of that single factor (called "1", "2", "3", and "4")
n = 3 replications per level
N = 4 levels * 3 replications per level = 12 runs

A sample randomized sequence of trials

The randomized sequence of trials might look like:

X1
3
1
4
2
2
1
3
4
1
2
4
3

Note that in this example there are 12!/(3!*3!*3!*3!) = 369,600 ways to run the experiment, all equally likely to be picked by a randomization procedure.

Model for a completely randomized design

The model for the response is

$$Y_{i,j} = \mu + T_i + \text{random error}$$

with

$Y_{i,j}$ being any observation for which $X1 = i$
(i and j denote the level of the factor and the replication within the level of the factor, respectively)
μ (or mu) is the general location parameter
T_i is the effect of having treatment level i

Estimates and Statistical Tests

Estimating and testing model factor

Estimate for μ : $\bar{Y}$ = the average of all the data

Estimate for T_i : $\bar{Y}_i - \bar{Y}$

levels

with $\bar{Y}_i$ = average of all Y for which $X1 = i$.

Statistical tests for levels of $X1$ are shown in the section on one-way ANOVA in Chapter 7.

5.3.3.2. Randomized block designs

Blocking to "remove" the effect of nuisance factors

For randomized block designs, there is one factor or variable that is of primary interest. However, there are also several other nuisance factors.

Nuisance factors are those that may affect the measured result, but are not of primary interest. For example, in applying a treatment, nuisance factors might be the specific operator who prepared the treatment, the time of day the experiment was run, and the room temperature. All experiments have nuisance factors. The experimenter will typically need to spend some time deciding which nuisance factors are important enough to keep track of or control, if possible, during the experiment.

Blocking used for nuisance factors that can be controlled

When we can control nuisance factors, an important technique known as blocking can be used to reduce or eliminate the contribution to experimental error contributed by nuisance factors. The basic concept is to create homogeneous blocks in which the nuisance factors are held constant and the factor of interest is allowed to vary. Within blocks, it is possible to assess the effect of different levels of the factor of interest without having to worry about variations due to changes of the block factors, which are accounted for in the analysis.

Definition of blocking factors

A nuisance factor is used as a blocking factor if every level of the primary factor occurs the same number of times with each level of the nuisance factor. The analysis of the experiment will focus on the effect of varying levels of the primary factor within each block of the experiment.

Block for a few of the most important nuisance factors

The general rule is:

"Block what you can, randomize what you cannot."

Blocking is used to remove the effects of a few of the most important nuisance variables. Randomization is then used to reduce the contaminating effects of the remaining nuisance variables.

Table of randomized

One useful way to look at a randomized block experiment is to consider it as a collection of completely randomized

block designs experiments, each run within one of the blocks of the total experiment.

Randomized Block Designs (RBD)

Name of Design	Number of Factors k	Number of Runs n
2-factor RBD	2	$L_1 * L_2$
3-factor RBD	3	$L_1 * L_2 * L_3$
4-factor RBD	4	$L_1 * L_2 * L_3 * L_4$
.	.	.
k-factor RBD	k	$L_1 * L_2 * ... * L_k$

with

L_1 = number of levels (settings) of factor 1
L_2 = number of levels (settings) of factor 2
L_3 = number of levels (settings) of factor 3
L_4 = number of levels (settings) of factor 4

.
.
.

L_k = number of levels (settings) of factor k

Example of a Randomized Block Design

Example of a randomized block design Suppose engineers at a semiconductor manufacturing facility want to test whether different wafer implant material dosages have a significant effect on resistivity measurements after a diffusion process taking place in a furnace. They have four different dosages they want to try and enough experimental wafers from the same lot to run three wafers at each of the dosages.

Furnace run is a nuisance factor The nuisance factor they are concerned with is "furnace run" since it is known that each furnace run differs from the last and impacts many process parameters.

Ideal would be to eliminate nuisance furnace factor An ideal way to run this experiment would be to run all the 4x3=12 wafers in the same furnace run. That would eliminate the nuisance furnace factor completely. However, regular production wafers have furnace priority, and only a few experimental wafers are allowed into any furnace run at the same time.

Non-Blocked method A non-blocked way to run this experiment would be to run each of the twelve experimental wafers, in random order,

one per furnace run. That would increase the experimental error of each resistivity measurement by the run-to-run furnace variability and make it more difficult to study the effects of the different dosages. The blocked way to run this experiment, assuming you can convince manufacturing to let you put four experimental wafers in a furnace run, would be to put four wafers with different dosages in each of three furnace runs. The only randomization would be choosing which of the three wafers with dosage 1 would go into furnace run 1, and similarly for the wafers with dosages 2, 3 and 4.

Description of the experiment

Let $X1$ be dosage "level" and $X2$ be the blocking factor furnace run. Then the experiment can be described as follows:

k = 2 factors (1 primary factor $X1$ and 1 blocking factor $X2$)
L_1 = 4 levels of factor $X1$
L_2 = 3 levels of factor $X2$
n = 1 replication per cell
$N = L_1 * L_2 = 4 * 3 = 12$ runs

Design trial before randomization

Before randomization, the design trials look like:

$X1$	$X2$
1	1
1	2
1	3
2	1
2	2
2	3
3	1
3	2
3	3
4	1
4	2
4	3

Matrix representation

An alternate way of summarizing the design trials would be to use a 4x3 matrix whose 4 rows are the levels of the treatment $X1$ and whose columns are the 3 levels of the blocking variable $X2$. The cells in the matrix have indices that match the $X1$, $X2$ combinations above.

By extension, note that the trials for any K-factor randomized block design are simply the cell indices of a K dimensional matrix.

Model for a Randomized Block Design

Model for a randomized block design

The model for a randomized block design with one nuisance variable is

$$Y_{i,j} = \mu + T_i + B_j + \text{random error}$$

where

$Y_{i,j}$ is any observation for which $X1 = i$ and $X2 = j$
$X1$ is the primary factor
$X2$ is the blocking factor
μ is the general location parameter (i.e., the mean)
T_i is the effect for being in treatment i (of factor $X1$)
B_j is the effect for being in block j (of factor $X2$)

Estimates for a Randomized Block Design

Estimating factor effects for a randomized block design

Estimate for μ : $\bar{Y}$ = the average of all the data

Estimate for T_i : $\bar{Y}_i - \bar{Y}$

with $\bar{Y}_i$ = average of all Y for which $X1 = i$.

Estimate for B_j : $\bar{Y}_j - \bar{Y}$

with $\bar{Y}_j$ = average of all Y for which $X2 = j$.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.2. Randomized block designs

5.3.3.2.1. Latin square and related designs

Latin square (and related) designs are efficient designs to block from 2 to 4 nuisance factors

Latin square designs, and the related Graeco-Latin square and Hyper-Graeco-Latin square designs, are a special type of comparative design.

There is a single factor of primary interest, typically called the treatment factor, and several nuisance factors. For Latin square designs there are 2 nuisance factors, for Graeco-Latin square designs there are 3 nuisance factors, and for Hyper-Graeco-Latin square designs there are 4 nuisance factors.

Nuisance factors used as blocking variables

The nuisance factors are used as blocking variables.

1. For Latin square designs, the 2 nuisance factors are divided into a tabular grid with the property that each row and each column receive each treatment exactly once.

2. As with the Latin square design, a Graeco-Latin square design is a *k*x*k* tabular grid in which *k* is the number of levels of the treatment factor. However, it uses 3 blocking variables instead of the 2 used by the standard Latin square design.

3. A Hyper-Graeco-Latin square design is also a *k*x*k* tabular grid with *k* denoting the number of levels of the treatment factor. However, it uses 4 blocking variables instead of the 2 used by the standard Latin square design.

Advantages and disadvantages of Latin square designs

The advantages of Latin square designs are:

1. They handle the case when we have several nuisance factors and we either cannot combine them into a single factor or we wish to keep them separate.

2. They allow experiments with a relatively small number of runs.

The disadvantages are:

1. The number of levels of each blocking variable must

equal the number of levels of the treatment factor.

2. The Latin square model assumes that there are no interactions between the blocking variables or between the treatment variable and the blocking variable.

Note that Latin square designs are equivalent to specific fractional factorial designs (e.g., the 4x4 Latin square design is equivalent to a 4^{3-1} fractional factorial design).

Summary of designs

Several useful designs are described in the table below.

Some Useful Latin Square, Graeco-Latin Square and Hyper-Graeco-Latin Square Designs

Name of Design	Number of Factors k	Number of Runs N
3-by-3 Latin Square	3	9
4-by-4 Latin Square	3	16
5-by-5 Latin Square	3	25
3-by-3 Graeco-Latin Square	4	9
4-by-4 Graeco-Latin Square	4	16
5-by-5 Graeco-Latin Square	4	25
4-by-4 Hyper-Graeco-Latin Square	5	16
5-by-5 Hyper-Graeco-Latin Square	5	25

Model for Latin Square and Related Designs

Latin square design model and estimates for effect levels

The model for a response for a latin square design is

$$Y_{ijk} = \mu + R_i + C_j + T_k + \text{random error}$$

with

Y_{ijk} denoting any observation for which
$X1 = i$, $X2 = j$, $X3 = k$
$X1$ and $X2$ are blocking factors
$X3$ is the primary factor

μ denoting the general location parameter
R_i denoting the effect for block i
C_j denoting the effect for block j
T_k denoting the effect for treatment k

Models for Graeco-Latin and Hyper-Graeco-Latin squares are the obvious extensions of the Latin square model, with additional blocking variables added.

Estimates for Latin Square Designs

Estimates

Estimate for μ: $\bar{Y}$ = the average of all the data

Estimate for R_i: $\bar{Y}_i - \bar{Y}$

$\bar{Y}_i$ = average of all Y for which $X1 = i$

Estimate for C_j: $\bar{Y}_j - \bar{Y}$

$\bar{Y}_j$ = average of all Y for which $X2 = j$

Estimate for T_k: $\bar{Y}_k - \bar{Y}$

$\bar{Y}_k$ = average of all Y for which $X3 = k$

Randomize as much as design allows

Designs for Latin squares with 3-, 4-, and 5-level factors are given next. These designs show what the treatment combinations should be for each run. ***When using any of these designs, be sure to randomize the treatment units and trial order, as much as the design allows***.

For example, one recommendation is that a Latin square design be randomly selected from those available, then randomize the run order.

Latin Square Designs for 3-, 4-, and 5-Level Factors

Designs for 3-level factors (and 2 nuisance or blocking factors)

3-Level Factors

$X1$ row blocking factor	$X2$ column blocking factor	$X3$ treatment factor
1	1	1
1	2	2
1	3	3
2	1	3
2	2	1
2	3	2
3	1	2
3	2	3
3	3	1

with

k = 3 factors (2 blocking factors and 1 primary factor)
L_1 = 3 levels of factor $X1$ (block)
L_2 = 3 levels of factor $X2$ (block)
L_3 = 3 levels of factor $X3$ (primary)
$N = L1 * L2 = 9$ runs

This can alternatively be represented as

A	B	C
C	A	B
B	C	A

Designs for 4-level factors (and 2 nuisance or blocking factors)

4-Level Factors

*X*1 row blocking factor	*X*2 column blocking factor	*X*3 treatment factor
1	1	1
1	2	2
1	3	4
1	4	3
2	1	4
2	2	3
2	3	1
2	4	2
3	1	2
3	2	4
3	3	3
3	4	1
4	1	3
4	2	1
4	3	2
4	4	4

with

k = 3 factors (2 blocking factors and 1 primary factor)
L_1 = 4 levels of factor $X1$ (block)
L_2 = 4 levels of factor $X2$ (block)
L_3 = 4 levels of factor $X3$ (primary)
$N = L1 * L2 = 16$ runs

This can alternatively be represented as

A	B	D	C
D	C	A	B
B	D	C	A
C	A	B	D

Designs for

5-Level Factors

5-level factors (and 2 nuisance or blocking factors)

$X1$ row blocking factor	$X2$ column blocking factor	$X3$ treatment factor
1	1	1
1	2	2
1	3	3
1	4	4
1	5	5
2	1	3
2	2	4
2	3	5
2	4	1
2	5	2
3	1	5
3	2	1
3	3	2
3	4	3
3	5	4
4	1	2
4	2	3
4	3	4
4	4	5
4	5	1
5	1	4
5	2	5
5	3	1
5	4	2
5	5	3

with

k = 3 factors (2 blocking factors and 1 primary factor)
L_1 = 5 levels of factor $X1$ (block)
L_2 = 5 levels of factor $X2$ (block)
L_3 = 5 levels of factor $X3$ (primary)
$N = L1 * L2$ = 25 runs

This can alternatively be represented as

A	B	C	D	E
C	D	E	A	B
E	A	B	C	D
B	C	D	E	A
D	E	A	B	C

Further information

More details on Latin square designs can be found in Box, Hunter, and Hunter (1978).

5.3.3.2.2. Graeco-Latin square designs

These designs handle 3 nuisance factors

Graeco-Latin squares, as described on the previous page, are efficient designs to study the effect of one treatment factor in the presence of 3 nuisance factors. They are restricted, however, to the case in which all the factors have the same number of levels.

Randomize as much as design allows

Designs for 3-, 4-, and 5-level factors are given on this page. These designs show what the treatment combinations would be for each run. ***When using any of these designs, be sure to randomize the treatment units and trial order, as much as the design allows***.

For example, one recommendation is that a Graeco-Latin square design be randomly selected from those available, then randomize the run order.

Graeco-Latin Square Designs for 3-, 4-, and 5-Level Factors

Designs for 3-level factors

3-Level Factors

$X1$ row blocking factor	$X2$ column blocking factor	$X3$ blocking factor	$X4$ treatment factor
1	1	1	1
1	2	2	2
1	3	3	3
2	1	2	3
2	2	3	1
2	3	1	2
3	1	3	2
3	2	1	3
3	3	2	1

with

k = 4 factors (3 blocking factors and 1 primary factor)
L_1 = 3 levels of factor $X1$ (block)

L_2 = 3 levels of factor $X2$ (block)
L_3 = 3 levels of factor $X3$ (primary)
L_4 = 3 levels of factor $X4$ (primary)
$N = L1 * L2 = 9$ runs

This can alternatively be represented as (A, B, and C represent the treatment factor and 1, 2, and 3 represent the blocking factor):

A1	B2	C3
C2	A3	B1
B3	C1	A2

Designs for 4-level factors

4-Level Factors

$X1$ row blocking factor	$X2$ column blocking factor	$X3$ blocking factor	$X4$ treatment factor
1	1	1	1
1	2	2	2
1	3	3	3
1	4	4	4
2	1	2	4
2	2	1	3
2	3	4	2
2	4	3	1
3	1	3	2
3	2	4	1
3	3	1	4
3	4	2	3
4	1	4	3
4	2	3	4
4	3	2	1
4	4	1	2

with

k = 4 factors (3 blocking factors and 1 primary factor)
L_1 = 3 levels of factor $X1$ (block)
L_2 = 3 levels of factor $X2$ (block)
L_3 = 3 levels of factor $X3$ (primary)
L_4 = 3 levels of factor $X4$ (primary)
$N = L1 * L2 = 16$ runs

This can alternatively be represented as (A, B, C, and D represent the treatment factor and 1, 2, 3, and 4 represent the

blocking factor):

A1	B2	C3	D4
D2	C1	B4	A3
B3	A4	D1	C2
C4	D3	A2	B1

Designs for 5-level factors

5-Level Factors

*X*1 row blocking factor	*X*2 column blocking factor	*X*3 blocking factor	*X*4 treatment factor
1	1	1	1
1	2	2	2
1	3	3	3
1	4	4	4
1	5	5	5
2	1	2	3
2	2	3	4
2	3	4	5
2	4	5	1
2	5	1	2
3	1	3	5
3	2	4	1
3	3	5	2
3	4	1	3
3	5	2	4
4	1	4	2
4	2	5	3
4	3	1	4
4	4	2	5
4	5	3	1
5	1	5	4
5	2	1	5
5	3	2	1
5	4	3	2
5	5	4	3

with

k = 4 factors (3 blocking factors and 1 primary factor)
L_1 = 3 levels of factor $X1$ (block)
L_2 = 3 levels of factor $X2$ (block)
L_3 = 3 levels of factor $X3$ (primary)

L_4 = 3 levels of factor $X4$ (primary)

$N = L1 * L2 = 25$ runs

This can alternatively be represented as (A, B, C, D, and E represent the treatment factor and 1, 2, 3, 4, and 5 represent the blocking factor):

A1	B2	C3	D4	E5
C2	D3	E4	A5	B1
E3	A4	B5	C1	D2
B4	C5	D1	E2	A3
D5	E1	A2	B3	C4

Further information

More designs are given in Box, Hunter, and Hunter (1978).

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.2. Randomized block designs

5.3.3.2.3. Hyper-Graeco-Latin square designs

These designs handle 4 nuisance factors

Hyper-Graeco-Latin squares, as described earlier, are efficient designs to study the effect of one treatment factor in the presence of 4 nuisance factors. They are restricted, however, to the case in which all the factors have the same number of levels.

Randomize as much as design allows

Designs for 4- and 5-level factors are given on this page. These designs show what the treatment combinations should be for each run. ***When using any of these designs, be sure to randomize the treatment units and trial order, as much as the design allows***.

For example, one recommendation is that a hyper-Graeco-Latin square design be randomly selected from those available, then randomize the run order.

Hyper-Graeco-Latin Square Designs for 4- and 5-Level Factors

Designs for 4-level factors (there are no 3-level factor Hyper-Graeco Latin square designs)

4-Level Factors

*X*1 row blocking factor	*X*2 column blocking factor	*X*3 blocking factor	*X*4 blocking factor	*X*5 treatment factor
1	1	1	1	1
1	2	2	2	2
1	3	3	3	3
1	4	4	4	4
2	1	4	2	3
2	2	3	1	4
2	3	2	4	1
2	4	1	3	2
3	1	2	3	4
3	2	1	4	3
3	3	4	1	2
3	4	3	2	1

4	1	3	4	2
4	2	4	3	1
4	3	1	2	4
4	4	2	1	3

with

k = 5 factors (4 blocking factors and 1 primary factor)
L_1 = 4 levels of factor $X1$ (block)
L_2 = 4 levels of factor $X2$ (block)
L_3 = 4 levels of factor $X3$ (primary)
L_4 = 4 levels of factor $X4$ (primary)
L_5 = 4 levels of factor $X5$ (primary)
$N = L1 * L2 = 16$ runs

This can alternatively be represented as (A, B, C, and D represent the treatment factor and 1, 2, 3, and 4 represent the blocking factors):

A11	B22	C33	D44
C42	D31	A24	B13
D23	C14	B41	A32
B34	A43	D12	C21

Designs for 5-level factors

5-Level Factors

$X1$ row blocking factor	$X2$ column blocking factor	$X3$ blocking factor	$X4$ blocking factor	$X5$ treatment factor
1	1	1	1	1
1	2	2	2	2
1	3	3	3	3
1	4	4	4	4
1	5	5	5	5
2	1	2	3	4
2	2	3	4	5
2	3	4	5	1
2	4	5	1	2
2	5	1	2	3
3	1	3	5	2
3	2	4	1	3
3	3	5	2	4
3	4	1	3	5
3	5	2	4	1
4	1	4	2	5

4	2	5	3	1
4	3	1	4	2
4	4	2	5	3
4	5	3	1	4
5	1	5	4	3
5	2	1	5	4
5	3	2	1	5
5	4	3	2	1
5	5	4	3	2

with

k = 5 factors (4 blocking factors and 1 primary factor)
L_1 = 5 levels of factor $X1$ (block)
L_2 = 5 levels of factor $X2$ (block)
L_3 = 5 levels of factor $X3$ (primary)
L_4 = 5 levels of factor $X4$ (primary)
L_5 = 5 levels of factor $X5$ (primary)
$N = L1 * L2 = 25$ runs

This can alternatively be represented as (A, B, C, D, and E represent the treatment factor and 1, 2, 3, 4, and 5 represent the blocking factors):

A11	B22	C33	D44	E55
D23	E34	A45	B51	C12
B35	C41	D52	E31	A24
E42	A53	B14	C25	D31
C54	D15	E21	A32	B43

Further information More designs are given in Box, Hunter, and Hunter (1978).

5.3.3.3. Full factorial designs

Full factorial designs in two levels

A design in which every setting of every factor appears with every setting of every other factor is a full factorial design

A common experimental design is one with all input factors set at two levels each. These levels are called `high' and `low' or `+1' and `-1', respectively. A design with all possible high/low combinations of all the input factors is called a full factorial design in two levels.

If there are k factors, each at 2 levels, a full factorial design has 2^k runs.

TABLE 3.2 Number of Runs for a 2^k Full Factorial

Number of Factors	Number of Runs
2	4
3	8
4	16
5	32
6	64
7	128

Full factorial designs not recommended for 5 or more factors

As shown by the above table, when the number of factors is 5 or greater, a full factorial design requires a large number of runs and is not very efficient. As recommended in the Design Guideline Table, a fractional factorial design or a Plackett-Burman design is a better choice for 5 or more factors.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.3. Full factorial designs

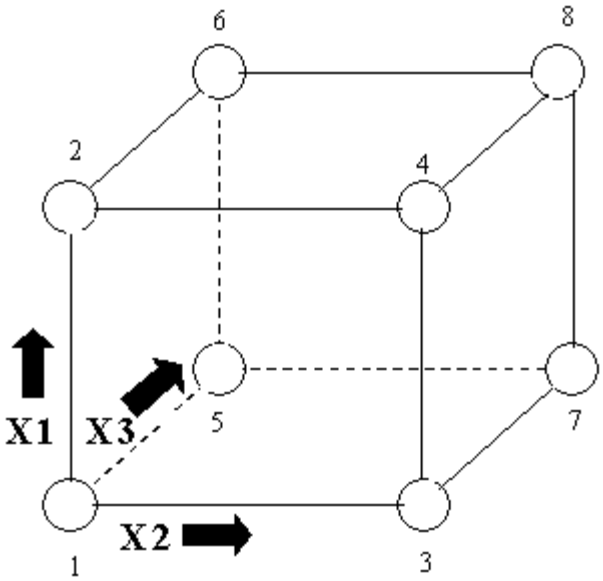
5.3.3.3.1. Two-level full factorial designs

Description

Graphical representation of a two-level design with 3 factors

Consider the two-level, full factorial design for three factors, namely the 2^3 design. This implies eight runs (not counting replications or center point runs). Graphically, we can represent the 2^3 design by the cube shown in Figure 3.1. The arrows show the direction of increase of the factors. The numbers `1' through `8' at the corners of the design box reference the `Standard Order' of runs (see Figure 3.1).

FIGURE 3.1 A 2^3 two-level, full factorial design; factors X1, X2, X3



The design matrix

In tabular form, this design is given by:

TABLE 3.3 A 2^3 two-level, full factorial design table showing runs in `Standard Order'

run	*X*1	*X*2	*X*3
1	-1	-1	-1

2	1	-1	-1
3	-1	1	-1
4	1	1	-1
5	-1	-1	1
6	1	-1	1
7	-1	1	1
8	1	1	1

The left-most column of Table 3.3, numbers 1 through 8, specifies a (non-randomized) run order called the `Standard Order.' These numbers are also shown in Figure 3.1. For example, run 1 is made at the `low' setting of all three factors.

Standard Order for a 2^k Level Factorial Design

Rule for writing a 2^k full factorial in "standard order"

We can readily generalize the 2^3 standard order matrix to a 2-level full factorial with k factors. The first ($X1$) column starts with -1 and alternates in sign for all 2^k runs. The second ($X2$) column starts with -1 repeated twice, then alternates with 2 in a row of the opposite sign until all 2^k places are filled. The third ($X3$) column starts with -1 repeated 4 times, then 4 repeats of +1's and so on. In general, the i-th column (X_i) starts with 2^{i-1} repeats of -1 folowed by 2^{i-1} repeats of +1.

Example of a 2^3 Experiment

Analysis matrix for the 3-factor complete factorial

An engineering experiment called for running three factors; namely, Pressure (*factor X*1), Table speed (*factor X*2) and Down force (*factor X*3), each at a `high' and `low' setting, on a production tool to determine which had the greatest effect on product uniformity. Two replications were run at each setting. A (full factorial) 2^3 design with 2 replications calls for 8*2=16 runs.

TABLE 3.4 Model or Analysis Matrix for a 2^3 Experiment

Model Matrix								Response Variables	
I	***X1***	***X2***	***X1*X2***	***X3***	***X1*X3***	***X2*X3***	***X1*X2*X3***	**Rep 1**	**Rep 2**
+1	-1	-1	+1	-1	+1	+1	-1	-3	-1
+1	+1	-1	-1	-1	-1	+1	+1	0	-1
+1	-1	+1	-1	-1	+1	-1	+1	-1	0
+1	+1	+1	+1	-1	-1	-1	-1	+2	+3
+1	-1	-1	+1	+1	-1	-1	+1	-1	0

+1	+1	-1	-1	+1	+1	-1	-1	+2	+1
+1	-1	+1	-1	+1	-1	+1	-1	+1	+1
+1	+1	+1	+1	+1	+1	+1	+1	+6	+5

The block with the 1's and -1's is called the *Model Matrix* or the *Analysis Matrix*. The table formed by the columns X1, X2 and X3 is called the *Design Table* or *Design Matrix*.

Orthogonality Properties of Analysis Matrices for 2-Factor Experiments

Eliminate correlation between estimates of main effects and interactions

When all factors have been coded so that the high value is "1" and the low value is "-1", the design matrix for any full (or suitably chosen fractional) factorial experiment has columns that are all pairwise orthogonal and all the columns (except the "I" column) sum to 0.

The orthogonality property is important because it eliminates correlation between the estimates of the main effects and interactions.

5.3.3.3.2. Full factorial example

A Full Factorial Design Example

An example of a full factorial design with 3 factors

The following is an example of a full factorial design with 3 factors that also illustrates replication, randomization, and added center points.

Suppose that we wish to improve the yield of a polishing operation. The three inputs (factors) that are considered important to the operation are Speed (**X1**), Feed (**X2**), and Depth **(X3**). We want to ascertain the relative importance of each of these factors on Yield (**Y**).

Speed, Feed and Depth can all be varied continuously along their respective scales, from a low to a high setting. Yield is observed to vary smoothly when progressive changes are made to the inputs. This leads us to believe that the ultimate response surface for **Y** will be smooth.

Table of factor level settings

TABLE 3.5 High (+1), Low (-1), and Standard (0) Settings for a Polishing Operation

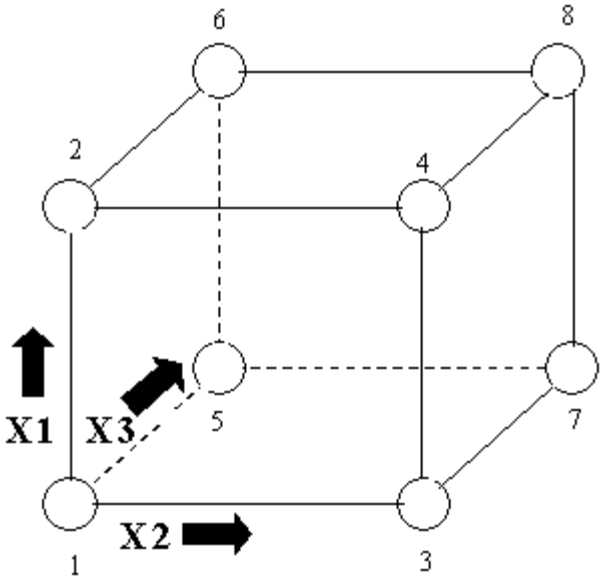
	Low (-1)	Standard (0)	High (+1)	Units
Speed	16	20	24	rpm
Feed	0.001	0.003	0.005	cm/sec
Depth	0.01	0.015	0.02	cm/sec

Factor Combinations

Graphical representation of the factor level settings

We want to try various combinations of these settings so as to establish the best way to run the polisher. There are eight different ways of combining high and low settings of Speed, Feed, and Depth. These eight are shown at the corners of the following diagram.

FIGURE 3.2 A 2^3 Two-level, Full Factorial Design; Factors X1, X2, X3. (The arrows show the direction of increase of the factors.)



2^3 implies 8 runs

Note that if we have *k* factors, each run at two levels, there will be 2^k different combinations of the levels. In the present case, $k = 3$ and $2^3 = 8$.

Full Model

Running the full complement of all possible factor combinations means that we can estimate all the main and interaction effects. There are three main effects, three two-factor interactions, and a three-factor interaction, all of which appear in the full model as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{123} X_1 X_2 X_3 + \epsilon$$

A full factorial design allows us to estimate all eight `beta' coefficients $\{\beta_0, ..., \beta_{123}\}$.

Standard order

Coded variables in standard order

The numbering of the corners of the box in the last figure refers to a standard way of writing down the settings of an experiment called `standard order'. We see standard order displayed in the following tabular representation of the eight-cornered box. Note that the factor settings have been coded, replacing the low setting by -1 and the high setting by 1.

Factor settings in tabular form

TABLE 3.6 A 2^3 Two-level, Full Factorial Design Table Showing Runs in `Standard Order'

	*X*1	*X*2	*X*3

1	-1	-1	-1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	-1
5	-1	-1	+1
6	+1	-1	+1
7	-1	+1	+1
8	+1	+1	+1

Replication

Replication provides information on variability

Running the entire design more than once makes for easier data analysis because, for each run (i.e., `corner of the design box') we obtain an average value of the response as well as some idea about the dispersion (variability, consistency) of the response at that setting.

Homogeneity of variance

One of the usual analysis assumptions is that the response dispersion is uniform across the experimental space. The technical term is `homogeneity of variance'. Replication allows us to check this assumption and possibly find the setting combinations that give inconsistent yields, allowing us to avoid that area of the factor space.

Factor settings in standard order with replication

We now have constructed a design table for a two-level full factorial in three factors, replicated twice.

TABLE 3.7 The 2^3 Full Factorial Replicated Twice and Presented in Standard Order

	Speed, *X*1	**Feed, *X*2**	**Depth, *X*3**
1	16, -1	.001, -1	.01, -1
2	24, +1	.001, -1	.01, -1
3	16, -1	.005, +1	.01, -1
4	24, +1	.005, +1	.01, -1
5	16, -1	.001, -1	.02, +1
6	24, +1	.001, -1	.02, +1
7	16, -1	.005, +1	.02, +1
8	24, +1	.005, +1	.02, +1
9	16, -1	.001, -1	.01, -1
10	24, +1	.001, -1	.01, -1
11	16, -1	.005, +1	.01, -1
12	24, +1	.005, +1	.01, -1
13	16, -1	.001, -1	.02, +1
14	24, +1	.001, -1	.02, +1

15	16, -1	.005, +1	.02, +1
16	24, +1	.005, +1	.02, +1

Randomization

No randomization and no center points

If we now ran the design as is, in the order shown, we would have two deficiencies, namely:

1. no randomization, and
2. no center points.

Randomization provides protection against extraneous factors affecting the results

The more freely one can randomize experimental runs, the more insurance one has against extraneous factors possibly affecting the results, and hence perhaps wasting our experimental time and effort. For example, consider the `Depth' column: the settings of Depth, in standard order, follow a `four low, four high, four low, four high' pattern.

Suppose now that four settings are run in the day and four at night, and that (unknown to the experimenter) ambient temperature in the polishing shop affects Yield. We would run the experiment over two days and two nights and conclude that Depth influenced Yield, when in fact ambient temperature was the significant influence. So the moral is: Randomize experimental runs as much as possible.

Table of factor settings in randomized order

Here's the design matrix again with the rows randomized. The old standard order column is also shown for comparison and for re-sorting, if desired, after the runs are in.

TABLE 3.8 The 2^3 Full Factorial Replicated Twice with Random Run Order Indicated

Random Order	Standard Order	*X*1	*X*2	*X*3
1	5	-1	-1	+1
2	15	-1	+1	+1
3	9	-1	-1	-1
4	7	-1	+1	+1
5	3	-1	+1	-1
6	12	+1	+1	-1
7	6	+1	-1	+1
8	4	+1	+1	-1
9	2	+1	-1	-1
10	13	-1	-1	+1
11	8	+1	+1	+1

12	16	+1	+1	+1
13	1	-1	-1	-1
14	14	+1	-1	+1
15	11	-1	+1	-1
16	10	+1	-1	-1

Table showing design matrix with randomization and center points

This design would be improved by adding at least 3 centerpoint runs placed at the beginning, middle and end of the experiment. The final design matrix is shown below:

TABLE 3.9 The 2^3 Full Factorial Replicated Twice with Random Run Order Indicated and Center Point Runs Added

Random Order	Standard Order	*X*1	*X*2	*X*3
1		0	0	0
2	5	-1	-1	+1
3	15	-1	+1	+1
4	9	-1	-1	-1
5	7	-1	+1	+1
6	3	-1	+1	-1
7	12	+1	+1	-1
8	6	+1	-1	+1
9		0	0	0
10	4	+1	+1	-1
11	2	+1	-1	-1
12	13	-1	-1	+1
13	8	+1	+1	+1
14	16	+1	+1	+1
15	1	-1	-1	-1
16	14	+1	-1	+1
17	11	-1	+1	-1
18	10	+1	-1	-1
19		0	0	0

5.3.3.3.3. Blocking of full factorial designs

Eliminate the influence of extraneous factors by "blocking"

We often need to eliminate the influence of extraneous factors when running an experiment. We do this by "blocking".

Previously, blocking was introduced when randomized block designs were discussed. There we were concerned with one factor in the presence of one of more nuisance factors. In this section we look at a general approach that enables us to divide 2-level factorial experiments into blocks.

For example, assume we anticipate predictable shifts will occur while an experiment is being run. This might happen when one has to change to a new batch of raw materials halfway through the experiment. The effect of the change in raw materials is well known, and we want to eliminate its influence on the subsequent data analysis.

Blocking in a 2^3 factorial design

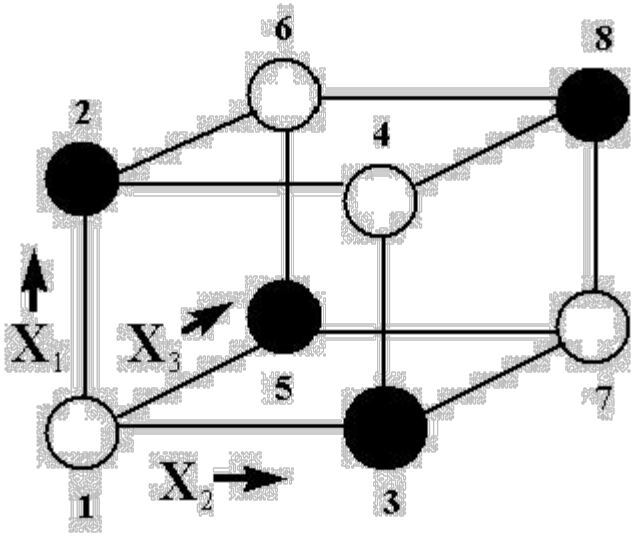
In this case, we need to divide our experiment into two halves (2 *blocks*), one with the first raw material batch and the other with the new batch. The division has to balance out the effect of the materials change in such a way as to eliminate its influence on the analysis, and we do this by *blocking.*

Example

Example: An eight-run 2^3 full factorial has to be blocked into two groups of four runs each. Consider the design `box' for the 2^3 full factorial. Blocking can be achieved by assigning the first block to the dark-shaded corners and the second block to the open circle corners.

Graphical representation of blocking scheme

FIGURE 3.3 Blocking Scheme for a 2^3 Using Alternate Corners



Three-factor interaction confounded with the block effect

This works because we are in fact assigning the `estimation' of the (unwanted) blocking effect to the three-factor interaction, and because of the special property of two-level designs called orthogonality. That is, the three-factor interaction is "confounded" with the block effect as will be seen shortly.

Orthogonality

Orthogonality guarantees that we can always estimate the effect of one factor or interaction clear of any influence due to any other factor or interaction. Orthogonality is a very desirable property in DOE and this is a major reason why two-level factorials are so popular and successful.

Table showing blocking scheme

Formally, consider the 2^3 design table with the three-factor interaction column added.

TABLE 3.10 Two Blocks for a 2^3 Design

SPEED *X1*	FEED *X2*	DEPTH *X3*	*X1*X2*X3*	BLOCK
-1	-1	-1	-1	**I**
+1	-1	-1	+1	**II**
-1	+1	-1	+1	**II**
+1	+1	-1	-1	**I**
-1	-1	+1	+1	**II**
+1	-1	+1	-1	**I**
-1	+1	+1	-1	**I**
+1	+1	+1	+1	**II**

Block by assigning the "Block effect"

Rows that have a `-1' in the three-factor interaction column are assigned to `Block I' (rows 1, 4, 6, 7), while the other rows are assigned to `Block II' (rows 2, 3, 5, 8). Note that

to a high-order interaction	the Block I rows are the open circle corners of the design `box' above; Block II are dark-shaded corners.
Most DOE software will do blocking for you	The general rule for blocking is: use one or a combination of high-order interaction columns to construct blocks. This gives us a formal way of blocking complex designs. Apart from simple cases in which you can design your own blocks, your statistical/DOE software will do the blocking if asked, but you do need to understand the principle behind it.
Block effects are confounded with higher-order interactions	The price you pay for blocking by using high-order interaction columns is that you can no longer distinguish the high-order interaction(s) from the blocking effect - they have been `confounded,' or `aliased.' In fact, the blocking effect is now the sum of the blocking effect and the high-order interaction effect. This is fine as long as our assumption about negligible high-order interactions holds true, which it usually does.
Center points within a block	Within a block, center point runs are assigned as if the block were a separate experiment - which in a sense it is. Randomization takes place within a block as it would for any non-blocked DOE.

5.3.3.4. Fractional factorial designs

Full factorial experiments can require many runs

The ASQC (1983) Glossary & Tables for Statistical Quality Control defines fractional factorial design in the following way: "*A factorial experiment in which only an adequately chosen fraction of the treatment combinations required for the complete factorial experiment is selected to be run.*"

A carefully chosen fraction of the runs may be all that is necessary

Even if the number of factors, k, in a design is small, the 2^k runs specified for a full factorial can quickly become very large. For example, $2^6 = 64$ runs is for a two-level, full factorial design with six factors. To this design we need to add a good number of centerpoint runs and we can thus quickly run up a very large resource requirement for runs with only a modest number of factors.

Later sections will show how to choose the "right" fraction for 2-level designs - these are both balanced and orthogonal

The solution to this problem is to use only a fraction of the runs specified by the full factorial design. Which runs to make and which to leave out is the subject of interest here. In general, we pick a fraction such as ½, ¼, etc. of the runs called for by the full factorial. We use various strategies that ensure an appropriate choice of runs. The following sections will show you how to choose an appropriate fraction of a full factorial design to suit your purpose at hand. *Properly chosen fractional factorial designs for 2-level experiments have the desirable properties of being both balanced and orthogonal.*

2-Level fractional factorial designs emphasized

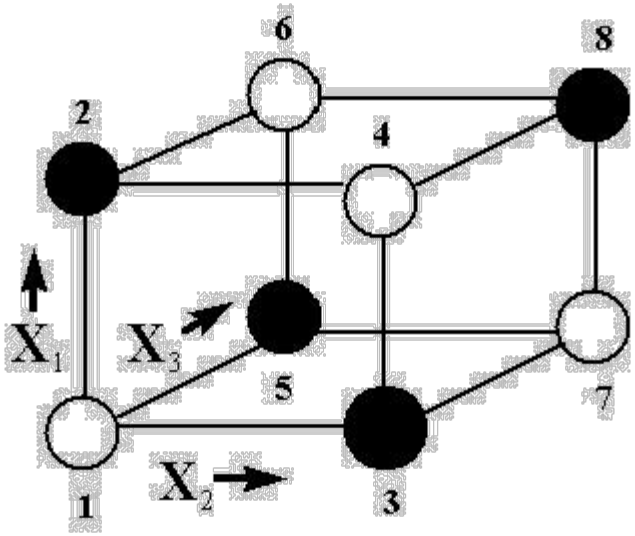
Note: We will be emphasizing fractions of two-level designs only. This is because two-level fractional designs are, in engineering at least, by far the most popular fractional designs. Fractional factorials where some factors have three levels will be covered briefly in Section 5.3.3.10.

5.3.3.4.1. A 2^{3-1} design (half of a 2^3)

We can run a fraction of a full factorial experiment and still be able to estimate main effects

Consider the two-level, full factorial design for three factors, namely the 2^3 design. This implies eight runs (not counting replications or center points). Graphically, as shown earlier, we can represent the 2^3 design by the following cube:

FIGURE 3.4 A 2^3 Full Factorial Design; Factors X_1, X_2, X_3. (The arrows show the direction of increase of the factors. Numbers `1' through `8' at the corners of the design cube reference the `Standard Order' of runs)



Tabular representation of the design

In tabular form, this design (also showing eight observations `y_j' (j = 1,...,8) is given by

TABLE 3.11 A 2^3 Two-level, Full Factorial Design Table Showing Runs in `Standard Order,' Plus Observations (y_j)

	*X*1	*X*2	*X*3	*Y*
1	-1	-1	-1	y_1 = 33

2	+1	-1	-1	$y_2 = 63$
3	-1	+1	-1	$y_3 = 41$
4	+1	+1	-1	$Y_4 = 57$
5	-1	-1	+1	$y_5 = 57$
6	+1	-1	+1	$y_6 = 51$
7	-1	+1	+1	$y_7 = 59$
8	+1	+1	+1	$y_8 = 53$

Responses in standard order

The right-most column of the table lists `$y_1$' through `y_8' to indicate the responses measured for the experimental runs when listed in standard order. For example, `$y_1$' is the response (i.e., output) observed when the three factors were all run at their `low' setting. The numbers entered in the "y" column will be used to illustrate calculations of effects.

Computing X1 main effect

From the entries in the table we are able to compute all `effects' such as main effects, first-order `interaction' effects, etc. For example, to compute the main effect estimate `$c_1$' of factor $X_1$, we compute the average response at all runs with $X_1$ at the `high' setting, namely (1/4)(y_2 + y_4 + y_6 + y_8), minus the average response of all runs with X_1 set at `low,' namely (1/4)(y_1 + y_3 + y_5 + y_7). That is,

c_1 = (1/4) (y_2 + y_4 + y_6 + y_8) - (1/4)(y_1 + y_3 + y_5 + y_7) or
c_1 = (1/4)(63+57+51+53) - (1/4)(33+41+57+59) = 8.5

Can we estimate X1 main effect with four runs?

Suppose, however, that we only have enough resources to do four runs. Is it still possible to estimate the main effect for X_1? Or any other main effect? The answer is yes, and there are even different choices of the four runs that will accomplish this.

Example of computing the main effects using only four runs

For example, suppose we select only the four light (unshaded) corners of the design cube. Using these four runs (1, 4, 6 and 7), we can still compute c_1 as follows:

c_1 = (1/2) (y_4 + y_6) - (1/2) (y_1 + y_7) or
c_1 = (1/2) (57+51) - (1/2) (33+59) = 8.

Simarly, we would compute c_2, the effect due to X_2, as

c_2 = (1/2) (y_4 + y_7) - (1/2) (y_1 + y_6) or
c_2 = (1/2) (57+59) - (1/2) (33+51) = 16.

Finally, the computation of c_3 for the effect due to X_3 would be

$$c_3 = (1/2)\ (y_6 + y_7) - (1/2)\ (y_1 + y_4) \text{ or}$$
$$c_3 = (1/2)\ (51+59) - (1/2)\ (33+57) = 10.$$

Alternative runs for computing main effects

We could also have used the four dark (shaded) corners of the design cube for our runs and obtained similiar, but slightly different, estimates for the main effects. In either case, we would have used half the number of runs that the full factorial requires. *The half fraction we used is a new design written as* 2^{3-1}. Note that $2^{3-1} = 2^3/2 = 2^2 = 4$, which is the number of runs in this half-fraction design. In the next section, a general method for choosing fractions that "work" will be discussed.

Example of how fractional factorial experiments often arise in industry

Example: An engineering experiment calls for running three factors, namely Pressure, Table speed, and Down force, each at a `high' and a `low' setting, on a production tool to determine which has the greatest effect on product uniformity. Interaction effects are considered negligible, but uniformity measurement error requires that at least two separate runs (replications) be made at each process setting. In addition, several `standard setting' runs (centerpoint runs) need to be made at regular intervals during the experiment to monitor for process drift. As experimental time and material are limited, no more than 15 runs can be planned.

A full factorial 2^3 design, replicated twice, calls for 8x2 = 16 runs, even without centerpoint runs, so this is not an option. However a 2^{3-1} design replicated twice requires only 4x2 = 8 runs, and then we would have 15-8 = 7 spare runs: 3 to 5 of these spare runs can be used for centerpoint runs and the rest saved for backup in case something goes wrong with any run. As long as we are confident that the interactions are negligbly small (compared to the main effects), and as long as complete replication is required, then the above replicated 2^{3-1} fractional factorial design (with center points) is a very reasonable choice.

On the other hand, if interactions are potentially large (and if the replication required could be set aside), then the usual 2^3 full factorial design (with center points) would serve as a good design.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.4. Fractional factorial designs

5.3.3.4.2. Constructing the 2^{3-1} half-fraction design

Construction of a 2^{3-1} half fraction design by staring with a 2^2 full factorial design

First note that, mathematically, $2^{3-1} = 2^2$. This gives us the first step, which is to start with a regular 2^2 full factorial design. That is, we start with the following design table.

TABLE 3.12 A Standard Order 2^2 Full Factorial Design Table

	*X*1	*X*2
1	-1	-1
2	+1	-1
3	-1	+1
4	+1	+1

Assign the third factor to the interaction column of a 2^2 design

This design has four runs, the right number for a half-fraction of a 2^3, but there is no column for factor *X*3. We need to add a third column to take care of this, and we do it by adding the *X*1**X*2 interaction column. This column is, as you will recall from full factorial designs, constructed by multiplying the row entry for *X*1 with that of *X*2 to obtain the row entry for *X*1**X*2.

TABLE 3.13 A 2^2 Design Table Augmented with the X1*X2 Interaction Column `X1*X2'

	*X*1	*X*2	*X*1**X*2
1	-1	-1	+1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	+1

*Design table with X3 set to X1*X2*

We may now substitute `X3' in place of `X1*X2' in this table.

TABLE 3.15 A 2^{3-1} Design Table with Column X3 set to X1*X2

	*X*1	*X*2	*X*3
1	-1	-1	+1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	+1

*Design table with X3 set to -X1*X2*

Note that the rows of Table 3.14 give the dark-shaded corners of the design in Figure 3.4. If we had set X3 = -X1*X2 as the rule for generating the third column of our 2^{3-1} design, we would have obtained:

TABLE 3.15 A 2^{3-1} Design Table with Column X3 set to - X1*X2

	*X*1	*X*2	*X*3
1	-1	-1	-1
2	+1	-1	+1
3	-1	+1	+1
4	+1	+1	-1

Main effect estimates from fractional factorial not as good as full factorial

This design gives the light-shaded corners of the box of Figure 3.4. Both 2^{3-1} designs that we have generated are equally good, and both save half the number of runs over the original 2^3 full factorial design. If c_1, c_2, and c_3 are our estimates of the main effects for the factors *X*1, *X*2, *X*3 (i.e., the difference in the response due to going from "low" to "high" for an effect), then the precision of the estimates c_1, c_2, and c_3 are not quite as good as for the full 8-run factorial because we only have four observations to construct the averages instead of eight; this is one price we have to pay for using fewer runs.

Example

Example: For the `Pressure (P), Table speed (T), and Down force (D)' design situation of the previous example, here's a replicated $2^{3-1}$ in randomized run order, with five centerpoint runs (`000') interspersed among the runs. This design table was constructed using the technique discussed above, with D = P*T.

Design table for the example

TABLE 3.16 A 2^{3-1} Design Replicated Twice, with Five Centerpoint Runs Added

	Pattern	P	T	D	Center Point

1	000	0	0	0	1
2	+--	+1	-1	-1	0
3	-+-	-1	+1	-1	0
4	000	0	0	0	1
5	+++	+1	+1	+1	0
6	--+	-1	-1	+1	0
7	000	0	0	0	1
8	+--	+1	-1	-1	0
9	--+	-1	-1	+1	0
10	000	0	0	0	1
11	+++	+1	+1	+1	0
12	-+-	-1	+1	-1	0
13	000	0	0	0	1

5.3.3.4.3. Confounding (also called aliasing)

Confounding means we have lost the ability to estimate some effects and/or interactions

One price we pay for using the design table column $X1*X2$ to obtain column $X3$ in Table 3.14 is, clearly, our inability to obtain an estimate of the interaction effect for $X1*X2$ (i.e., c_{12}) that is separate from an estimate of the main effect for $X3$. In other words, we have confounded the main effect estimate for factor $X3$ (i.e., c_3) with the estimate of the interaction effect for $X1$ and $X2$ (i.e., with c_{12}). The whole issue of confounding is fundamental to the construction of fractional factorial designs, and we will spend time discussing it below.

Sparsity of effects assumption

In using the 2^{3-1} design, we also assume that c_{12} is small compared to c_3; this is called a `*sparsity of effects*' assumption. Our computation of c_3 is in fact a computation of $c_3 + c_{12}$. If the desired effects are only confounded with non-significant interactions, then we are OK.

A Notation and Method for Generating Confounding or Aliasing

A short way of writing factor column multiplication

A short way of writing `$X3 = X1*X2$' (understanding that we are talking about multiplying columns of the design table together) is: `3 = 12' (similarly 3 = -12 refers to $X3 = -X1*X2$). Note that `12' refers to column multiplication of the kind we are using to construct the fractional design and any column multiplied by itself gives the identity column of all 1's.

Next we multiply both sides of 3=12 by 3 and obtain 33=123, or I=123 since 33=I (or a column of all 1's). Playing around with this "algebra", we see that 2I=2123, or 2=2123, or 2=1223, or 2=13 (since 2I=2, 22=I, and 1I3=13). Similarly, 1=23.

Definition of "design generator" or "generating relation" and "defining relation"

I=123 is called a *design generator* or a *generating relation* for this 2^{3-1}design (the dark-shaded corners of Figure 3.4). Since there is only one design generator for this design, it is also the *defining relation* for the design. Equally, I=-123 is the design generator (and defining relation) for the light-shaded corners of Figure 3.4. *We call I=123 the defining relation for the 2^{3-1} design because with it we can generate (by "multiplication") the*

complete confounding pattern for the design. That is, given I=123, we can generate the set of {1=23, 2=13, 3=12, I=123}, which is the complete set of *aliases*, as they are called, for this 2^{3-1} fractional factorial design. With I=123, we can easily generate all the columns of the half-fraction design 2^{3-1}.

Principal fraction

Note: We can replace any design generator by its negative counterpart and have an equivalent, but different fractional design. The fraction generated by positive design generators is sometimes called the *principal fraction*.

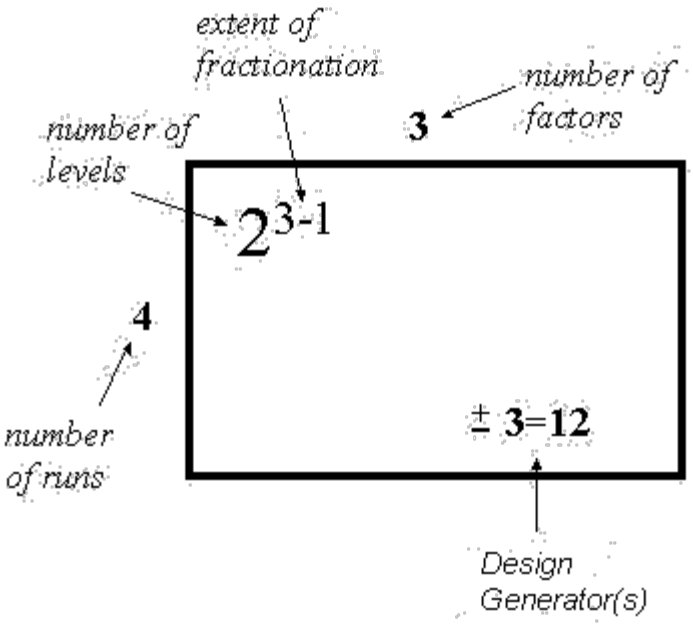
All main effects of 2^{3-1} design confounded with two-factor interactions

The confounding pattern described by 1=23, 2=13, and 3=12 tells us that all the main effects of the 2^{3-1} design are confounded with two-factor interactions. That is the price we pay for using this fractional design. Other fractional designs have different confounding patterns; for example, in the typical quarter-fraction of a 2^6 design, i.e., in a 2^{6-2} design, main effects are confounded with three-factor interactions (e.g., 5=123) and so on. In the case of 5=123, we can also readily see that 15=23 (etc.), which alerts us to the fact that certain two-factor interactions of a 2^{6-2} are confounded with other two-factor interactions.

A useful summary diagram for a fractional factorial design

Summary: A convenient summary diagram of the discussion so far about the 2^{3-1} design is as follows:

FIGURE 3.5 Essential Elements of a 2^{3-1} Design



The next section will add one more item to the above box, and then we will be able to select the right two-level fractional factorial design for a wide range of experimental tasks.

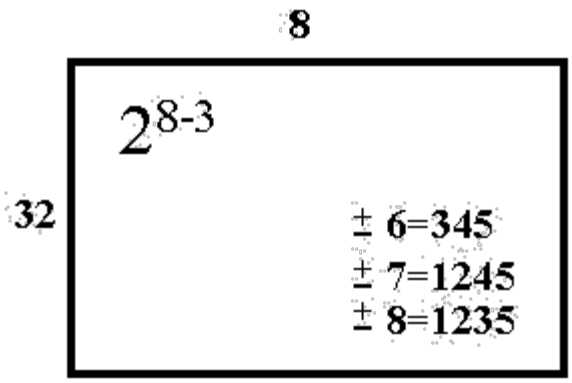
5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.4. Fractional factorial designs

5.3.3.4.4. Fractional factorial design specifications and design resolution

Generating relation and diagram for the 2^{8-3} fractional factorial design

We considered the 2^{3-1} design in the previous section and saw that its generator written in "I = ... " form is {I = +123}. Next we look at a one-eighth fraction of a 2^8 design, namely the 2^{8-3} fractional factorial design. Using a diagram similar to Figure 3.5, we have the following:

FIGURE 3.6 Specifications for a 2^{8-3} Design



2^{8-3} design has 32 runs

Figure 3.6 tells us that a 2^{8-3} design has 32 runs, not including centerpoint runs, and eight factors. There are three generators since this is a $1/8 = 2^{-3}$ fraction (in general, a 2^{k-p} fractional factorial needs p generators which define the settings for p additional factor columns to be added to the 2^{k-p} full factorial design columns - see the following detailed description for the 2^{8-3} design).

How to Construct a Fractional Factorial Design From the Specification

Rule for constructing a fractional factorial design

In order to construct the design, we do the following:

1. Write down a full factorial design in standard order for k-p factors (8-3 = 5 factors for the example

above). In the specification above we start with a 2 full factorial design. Such a design has $2^5 = 32$ rows.

2. Add a sixth column to the design table for factor 6, using 6 = 345 (or 6 = -345) to manufacture it (i.e., create the new column by multiplying the indicated old columns together).

3. Do likewise for factor 7 and for factor 8, using the appropriate design generators given in Figure 3.6.

4. The resultant design matrix gives the 32 trial runs for an 8-factor fractional factorial design. (When actually running the experiment, we would of course randomize the run order.

Design generators

We note further that the design generators, written in `I = ...' form, for the principal 2^{8-3} fractional factorial design are:

{ I = + 3456; I = + 12457; I = +12358 }.

These design generators result from multiplying the "6 = 345" generator by "6" to obtain "I = 3456" and so on for the other two generqators.

"Defining relation" for a fractional factorial design

The total collection of design generators for a factorial design, *including all new generators that can be formed as products of these generators*, is called a *defining relation*. There are seven "*words"*, or strings of numbers, in the defining relation for the 2^{8-3} design, starting with the original three generators and adding all the new "words" that can be formed by multiplying together any two or three of these original three words. These seven turn out to be I = 3456 = 12457 = 12358 = 12367 = 12468 = 3478 = 5678. In general, there will be $(2^p - 1)$ words in the defining relation for a 2^{k-p} fractional factorial.

Definition of "Resolution"

The length of the shortest word in the defining relation is called the resolution of the design. Resolution describes the degree to which estimated main effects are aliased (or confounded) with estimated 2-level interactions, 3-level interactions, etc.

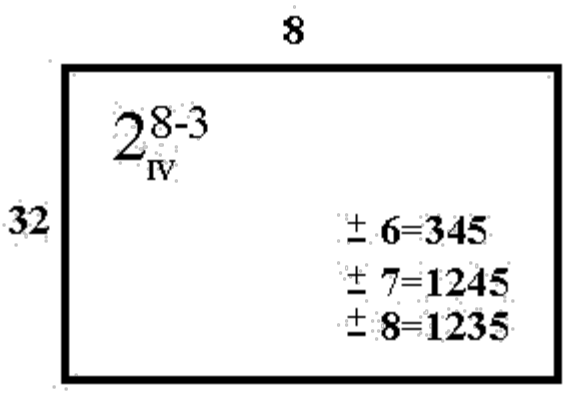
Notation for resolution (Roman numerals)

The length of the shortest word in the defining relation for the 2^{8-3} design is four. This is written in Roman numeral script, and subscripted as 2^{8-3}_{IV}. Note that the 2^{3-1} design has only one word, "I = 123" (or "I = -123"), in its defining relation since there is only one design generator, and so this fractional factorial design has resolution three; that is, we may write 2^{3-1}_{III}.

Diagram for a 2^{8-3} design showing resolution

Now Figure 3.6 may be completed by writing it as:

FIGURE 3.7 Specifications for a 2^{8-3}, Showing Resolution IV



Resolution and confounding

The design resolution tells us how badly the design is confounded. Previously, in the 2^{3-1} design, we saw that the main effects were confounded with two-factor interactions. However, main effects were not confounded with other main effects. So, at worst, we have 3=12, or 2=13, etc., but we do not have 1=2, etc. In fact, a resolution II design would be pretty useless for any purpose whatsoever!

Similarly, in a resolution IV design, main effects are confounded with at worst three-factor interactions. We can see, in Figure 3.7, that 6=345. We also see that 36=45, 34=56, etc. (i.e., some two-factor interactions are confounded with certain other two-factor interactions) etc.; but we never see anything like 2=13, or 5=34, (i.e., main effects confounded with two-factor interactions).

The complete first-order interaction confounding for the given 2^{8-3} design

The complete confounding pattern, for confounding of up to two-factor interactions, arising from the design given in Figure 3.7 is

34 = 56 = 78
35 = 46
36 = 45
37 = 48
38 = 47
57 = 68
58 = 67

All of these relations can be easily verified by multiplying the indicated two-factor interactions by the generators. For example, to verify that 38= 47, multiply both sides of 8=1235 by 3 to get 38=125. Then, multiply 7=1245 by 4 to get 47=125. From that it follows that 38=47.

One or two factors suspected of possibly having significant first-order interactions can be assigned in such a way as to avoid having them aliased

For this 2^{8-3}_V fractional factorial design, 15 two-factor interactions are aliased (confounded) in pairs or in a group of three. The remaining 28 - 15 = 13 two-factor interactions are only aliased with higher-order interactions (which are generally assumed to be negligible). This is verified by noting that factors "1" and "2" never appear in a length-4 word in the defining relation. So, all 13 interactions involving "1" and "2" are clear of aliasing with any other two factor interaction.

If one or two factors are suspected of possibly having significant first-order interactions, they can be assigned in such a way as to avoid having them aliased.

Higher resoulution designs have less severe confounding, but require more runs

A resolution IV design is "better" than a resolution III design because we have less-severe confounding pattern in the `IV' than in the `III' situation; higher-order interactions are less likely to be significant than low-order interactions.

A higher-resolution design for the same number of factors will, however, require more runs and so it is `worse' than a lower order design in that sense.

Resolution V designs for 8 factors

Similarly, with a resolution V design, main effects would be confounded with four-factor (and possibly higher-order) interactions, and two-factor interactions would be confounded with certain three-factor interactions. To obtain a resolution V design for 8 factors requires more runs than the 2^{8-3} design. One option, if estimating all main effects and two-factor interactions is a requirement, is a 2^{8-3}_V design. However, a 48-run alternative (John's 3/4 fractional factorial) is also available.

There are many choices of fractional factorial designs - some may have the same number of runs and resolution, but different aliasing patterns.

Note: There are other 2^{8-3}_V fractional designs that can be derived starting with different choices of design generators for the "6", "7" and "8" factor columns. However, they are either equivalent (in terms of the number of words of length of length of four) to the fraction with generators 6 = 345, 7 = 1245, 8 = 1235 (obtained by relabeling the factors), or they are inferior to the fraction given because their defining relation contains more words of length four (and therefore more confounded two-factor interactions). For example, the 2^{8-3}_V design with generators 6 = 12345, 7 = 135, and 8 = 245 has five length-four words in the defining relation (the defining relation is I = 123456 = 1357 = 2458 = 2467 = 1368 = 123478 = 5678). As a result, this design would confound more two factor-interactions (23 out of 28 possible two-factor interactions are confounded, leaving only "12", "14", "23", "27" and "34" as estimable two-factor interactions).

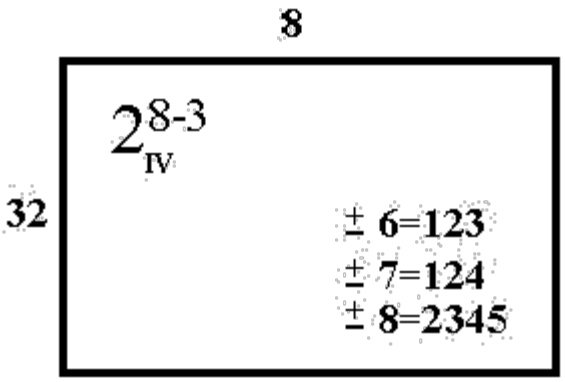
Diagram of

As an example of an equivalent "best" fractional

an alternative way for generating the 2^{8-3} design

2^{8-3}_{V} factorial design, obtained by "relabeling", consider the design specified in Figure 3.8.

FIGURE 3.8 Another Way of Generating the 2^{8-3} Design



This design is equivalent to the design specified in Figure 3.7 after relabeling the factors as follows: 1 becomes 5, 2 becomes 8, 3 becomes 1, 4 becomes 2, 5 becomes 3, 6 remains 6, 7 becomes 4 and 8 becomes 7.

Minimum aberration

A table given later in this chapter gives a collection of useful fractional factorial designs that, for a given *k* and *p*, maximize the possible resolution and minimize the number of short words in the defining relation (which minimizes two-factor aliasing). The term for this is "minimum aberration".

Design Resolution Summary

Commonly used design Resolutions

The meaning of the most prevalent resolution levels is as follows:

Resolution III Designs

Main effects are confounded (aliased) with two-factor interactions.

Resolution IV Designs

No main effects are aliased with two-factor interactions, but two-factor interactions are aliased with each other.

Resolution V Designs

No main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.4. Fractional factorial designs

5.3.3.4.5. Use of fractional factorial designs

Use low-resolution designs for screening among main effects and use higher-resolution designs when interaction effects and response surfaces need to be investigated

The basic purpose of a fractional factorial design is to economically investigate cause-and-effect relationships of significance in a given experimental setting. This does not differ in essence from the purpose of any experimental design. However, because we are able to choose fractions of a full design, and hence be more economical, we also have to be aware that different factorial designs serve different purposes.

Broadly speaking, with designs of resolution three, and sometimes four, we seek to screen out the few important main effects from the many less important others. For this reason, these designs are often termed main effects designs, or screening designs.

On the other hand, designs of resolution five, and higher, are used for focusing on more than just main effects in an experimental situation. These designs allow us to estimate interaction effects and such designs are easily augmented to complete a second-order design - a design that permits estimation of a full second-order (quadratic) model.

Different purposes for screening/RSM designs

Within the screening/RSM strategy of design, there are a number of functional purposes for which designs are used. For example, an experiment might be designed to determine how to make a product better or a process more robust against the influence of external and non-controllable influences such as the weather. Experiments might be designed to troubleshoot a process, to determine bottlenecks, or to specify which component(s) of a product are most in need of improvement. Experiments might also be designed to optimize yield, or to minimize defect levels, or to move a process away from an unstable operating zone. All these aims and purposes can be achieved using fractional factorial designs and their appropriate design enhancements.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.4. Fractional factorial designs

5.3.3.4.6. Screening designs

Screening designs are an efficient way to identify significant main effects

The term `Screening Design' refers to an experimental plan that is intended to find the few significant factors from a list of many potential ones. Alternatively, we refer to a design as a screening design if its primary purpose is to identify significant main effects, rather than interaction effects, the latter being assumed an order of magnitude less important.

Use screening designs when you have many factors to consider

Even when the experimental goal is to eventually fit a response surface model (an RSM analysis), the first experiment should be a screening design when there are many factors to consider.

Screening designs are usually resolution III or IV

Screening designs are typically of resolution III. The reason is that resolution III designs permit one to explore the effects of many factors with an efficient number of runs.

Sometimes designs of resolution IV are also used for screening designs. In these designs, main effects are confounded with, at worst, three-factor interactions. This is better from the confounding viewpoint, but the designs require more runs than a resolution III design.

Plackett-Burman designs

Another common family of screening designs is the Plackett-Burman set of designs, so named after its inventors. These designs are of resolution III and will be described later.

Economical plans for determing significant main effects

In short, screening designs are economical experimental plans that focus on determining the relative significance of many main effects.

5.3.3.4.7. Summary tables of useful fractional factorial designs

Useful fractional factorial designs for up to 10 factors are summarized here

There are very useful summaries of two-level fractional factorial designs for up to 11 factors, originally published in the book *Statistics for Experimenters* by G.E.P. Box, W.G. Hunter, and J.S. Hunter (New York, John Wiley & Sons, 1978). and also given in the book *Design and Analysis of Experiments, 5th edition* by Douglas C. Montgomery (New York, John Wiley & Sons, 2000).

Generator column notation can use either numbers or letters for the factor columns

They differ in the notation for the design generators. Box, Hunter, and Hunter use numbers (as we did in our earlier discussion) and Montgomery uses capital letters according to the following scheme:

1	2	3	4	5	6	7	8	9	10	11
A	B	C	D	E	F	G	H	J	K	L

Notice the absence of the letter I. This is usually reserved for the intercept column that is identically 1. As an example of the letter notation, note that the design generator "6 = 12345" is equivalent to "F = ABCDE".

Details of the design generators, the defining relation, the confounding structure, and the design matrix

TABLE 3.17 catalogs these useful fractional factorial designs using the notation previously described in FIGURE 3.7.

Clicking on the 2_R^{k-p} *specification for a given design provides details (courtesy of Dataplot files) of the design generators, the defining relation, the confounding structure (as far as main effects and two-level interactions are concerned), and the design matrix.* The notation used follows our previous labeling of factors with numbers, not letters.

Click on the design specification in the table below and a

TABLE 3.17 Summary of Useful Fractional Factorial Designs

Number of Factors, k	Design Specification	Number of Runs N

text file with details about the design can be viewed or saved

3	2_{III}^{3-1}	**4**
4	2_{IV}^{4-1}	**8**
5	2_{V}^{5-1}	**16**
5	2_{III}^{5-2}	**8**
6	2_{VI}^{6-1}	**32**
6	2_{IV}^{6-2}	**16**
6	2_{III}^{6-3}	**8**
7	2_{VII}^{7-1}	**64**
7	2_{IV}^{7-2}	**32**
7	2_{IV}^{7-3}	**16**
7	2_{III}^{7-4}	**8**
8	2_{VIII}^{8-1}	**128**
8	2_{V}^{8-2}	**64**
8	2_{IV}^{8-3}	**32**
8	2_{IV}^{8-4}	**16**
9	2_{VI}^{9-2}	**128**
9	2_{IV}^{9-3}	**64**
9	2_{IV}^{9-4}	**32**
9	2_{III}^{9-5}	**16**
10	2_{V}^{10-3}	**128**
10	2_{IV}^{10-4}	**64**
10	2_{IV}^{10-5}	**32**
10	2_{III}^{10-6}	**16**
11	2_{V}^{11-4}	**128**
11	2_{IV}^{11-5}	**64**
11	2_{IV}^{11-6}	**32**
11	2_{III}^{11-7}	**16**
15	2_{III}^{15-11}	**16**
31	2_{III}^{31-26}	**32**

5.3.3.5. Plackett-Burman designs

Plackett-Burman designs

In 1946, R.L. Plackett and J.P. Burman published their now famous paper "The Design of Optimal Multifactorial Experiments" in *Biometrika* (vol. 33). This paper described the construction of very economical designs with the run number a multiple of four (rather than a power of 2). Plackett-Burman designs are very efficient screening designs when only main effects are of interest.

These designs have run numbers that are a multiple of 4

Plackett-Burman (PB) designs are used for screening experiments because, in a PB design, main effects are, in general, heavily confounded with two-factor interactions. The PB design in 12 runs, for example, may be used for an experiment containing up to 11 factors.

12-Run Plackett-Burnam design

TABLE 3.18 Plackett-Burman Design in 12 Runs for up to 11 Factors

	Pattern	*X*1	*X*2	*X*3	*X*4	*X*5	*X*6	*X*7	*X*8	*X*9	*X*10	*X*11
1	++++++++++	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
2	-+-+++---+-	-1	+1	-1	+1	+1	+1	-1	-1	-1	+1	-1
3	--+-+++---+	-1	-1	+1	-1	+1	+1	+1	-1	-1	-1	+1
4	+--+-+++---	+1	-1	-1	+1	-1	+1	+1	+1	-1	-1	-1
5	-+--+-+++--	-1	+1	-1	-1	+1	-1	+1	+1	+1	-1	-1
6	--+--+-+++-	-1	-1	+1	-1	-1	+1	-1	+1	+1	+1	-1
7	---+--+-+++	-1	-1	-1	+1	-1	-1	+1	-1	+1	+1	+1
8	+---+--+-++	+1	-1	-1	-1	+1	-1	-1	+1	-1	+1	+1
9	++---+--+-+	+1	+1	-1	-1	-1	+1	-1	-1	+1	-1	+1
10	+++---+--+-	+1	+1	+1	-1	-1	-1	+1	-1	-1	+1	-1
11	-+++---+--+	-1	+1	+1	+1	-1	-1	-1	+1	-1	-1	+1
12	+-+++---+--	+1	-1	+1	+1	+1	-1	-1	-1	+1	-1	-1

Saturated Main Effect designs

PB designs also exist for 20-run, 24-run, and 28-run (and higher) designs. With a 20-run design you can run a screening experiment for up to 19 factors, up to 23 factors in a 24-run design, and up to 27 factors in a 28-run design. These Resolution III designs are known as *Saturated Main Effect* designs because all degrees of freedom are utilized to estimate main effects. The designs for 20 and 24 runs are shown below.

20-Run

TABLE 3.19 A 20-Run Plackett-Burman Design

Plackett-Burnam design

	*X*1	*X*2	*X*3	*X*4	*X*5	*X*6	*X*7	*X*8	*X*9	*X*10	*X*11	*X*12	*X*13	*X*14	*X*15	*X*16	*X*17	*X*18	*X*19
1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
2	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1
3	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1
4	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1
5	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1
6	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1
7	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1
8	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1
9	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1
10	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1
11	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1
12	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1	-1
13	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1	+1
14	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1	+1
15	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1	+1
16	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1	+1
17	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1	-1
18	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1	-1
19	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1	+1
20	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	-1	-1	+1	+1	-1	-1

24-Run Plackett-Burnam design

TABLE 3.20 A 24-Run Plackett-Burman Design

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1
3	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1
4	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1
5	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1
6	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1
7	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1
8	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1
9	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1
10	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1
11	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1
12	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1	1
13	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	-1
14	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1
15	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1
16	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1
17	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1	-1
18	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	1
19	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1	-1
20	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	1
21	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1	1
22	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1	1
23	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	1
24	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1

No defining relation	These designs do not have a defining relation since interactions are not identically equal to main effects. With the 2^{k-p}_{III} designs, a main effect column X_i is either orthogonal to X_iX_j or identical to plus or minus X_iX_j. For Plackett-Burman designs, the two-factor interaction column X_iX_j is correlated with every X_k (for k not equal to i or j).
Economical for detecting large main effects	However, these designs are very useful for economically detecting large main effects, assuming all interactions are negligible when compared with the few important main effects.

5.3.3.6. Response surface designs

Response surface models may involve just main effects and interactions or they may also have quadratic and possibly cubic terms to account for curvature

Earlier, we described the response surface method (RSM) objective. Under some circumstances, a model involving only main effects and interactions may be appropriate to describe a response surface when

1. Analysis of the results revealed no evidence of "pure quadratic" curvature in the response of interest (i.e., the response at the center approximately equals the average of the responses at the factorial runs).
2. The design matrix originally used included the limits of the factor settings available to run the process.

Equations for quadratic and cubic models

In other circumstances, a complete description of the process behavior might require a quadratic or cubic model:

Quadratic

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2$$

Cubic

$$\hat{y} = \text{quadratic model} + b_{123}x_1x_2x_3 + b_{112}x_1^2x_2 + b_{113}x_1^2x_3 + b_{122}x_1x_2^2 + b_{133}x_1x_3^2 + b_{223}x_2^2x_3 + b_{233}x_2x_3^2 + b_{111}x_1^3 + b_{222}x_2^3 + b_{333}x_3^3$$

These are the full models, with all possible terms, rarely would all of the terms be needed in an application.

Quadratic models almost always sufficient for industrial applications

If the experimenter has defined factor limits appropriately and/or taken advantage of all the tools available in multiple regression analysis (transformations of responses and factors, for example), then finding an industrial process that requires a third-order model is highly unusual. Therefore, we will only focus on designs that are useful for fitting quadratic models. As we will see, these designs often provide lack of fit detection that will help determine when a higher-order model is needed.

General quadratic surface types

Figures 3.9 to 3.12 identify the general quadratic surface types that an investigator might encounter

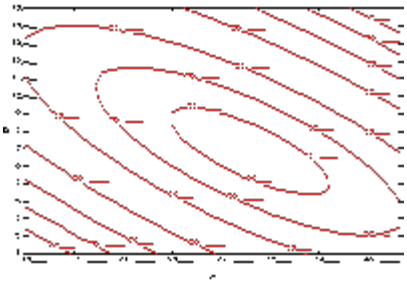


FIGURE 3.9 A Response Surface "Peak"

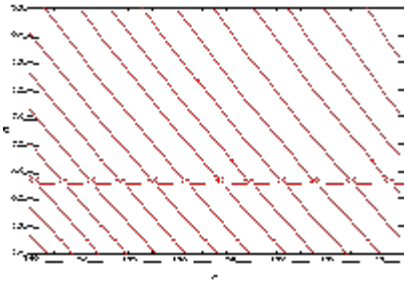


FIGURE 3.10 A Response Surface "Hillside"

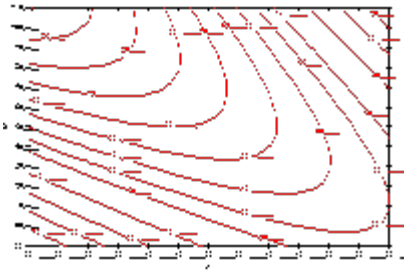


FIGURE 3.11 A Response Surface "Rising Ridge"

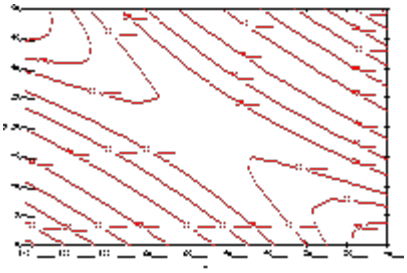


FIGURE 3.12 A Response Surface "Saddle"

Factor Levels for Higher-Order Designs

Possible behaviors of responses as functions of factor settings

Figures 3.13 through 3.15 illustrate possible behaviors of responses as functions of factor settings. In each case, assume the value of the response increases from the bottom of the figure to the top and that the factor settings increase from left to right.

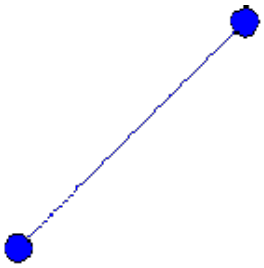


FIGURE 3.13 Linear Function

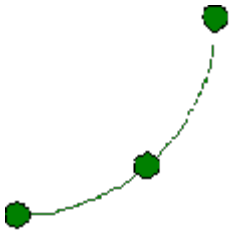


FIGURE 3.14 Quadratic Function

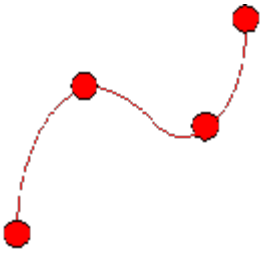


FIGURE 3.15 Cubic Function

A two-level experiment with center points can detect, but not fit, quadratic

If a response behaves as in Figure 3.13, the design matrix to quantify that behavior need only contain factors with two levels -- low and high. This model is a basic assumption of simple two-level factorial and fractional factorial designs. If a response behaves as in Figure 3.14, the minimum number of levels required for a factor to quantify that behavior is three. One might logically assume that adding center points to a two-level design would satisfy that

effects

requirement, but the arrangement of the treatments in such a matrix confounds all quadratic effects with each other. *While a two-level design with center points cannot estimate individual pure quadratic effects, it can detect them effectively*.

Three-level factorial design

A solution to creating a design matrix that permits the estimation of simple curvature as shown in Figure 3.14 would be to use a three-level factorial design. Table 3.21 explores that possibility.

Four-level factorial design

Finally, in more complex cases such as illustrated in Figure 3.15, the design matrix must contain at least four levels of each factor to characterize the behavior of the response adequately.

3-level factorial designs can fit quadratic models but they require many runs when there are more than 4 factors

TABLE 3.21 Three-level Factorial Designs

Number of Factors	Treatment Combinations 3^k Factorial	Number of Coefficients Quadratic Empirical Model
2	9	6
3	27	10
4	81	15
5	243	21
6	729	28

Fractional factorial designs created to avoid such a large number of runs

Two-level factorial designs quickly become too large for practical application as the number of factors investigated increases. This problem was the motivation for creating `fractional factorial' designs. Table 3.21 shows that the number of runs required for a 3^k factorial becomes unacceptable even more quickly than for 2^k designs. The last column in Table 3.21 shows the number of terms present in a quadratic model for each case.

Number of runs large even for modest number of factors

With only a modest number of factors, the number of runs is very large, even an order of magnitude greater than the number of parameters to be estimated when k isn't small. For example, the absolute minimum number of runs required to estimate all the terms present in a four-factor quadratic model is 15: the intercept term, 4 main effects, 6 two-factor interactions, and 4 quadratic terms.

The corresponding 3^k design for $k = 4$ requires 81 runs.

Complex alias structure and lack of rotatability for 3-level fractional factorial

Considering a fractional factorial at three levels is a logical step, given the success of fractional designs when applied to two-level designs. Unfortunately, the alias structure for the three-level fractional factorial designs is considerably more complex and harder to define than in the two-level case.

Additionally, the three-level factorial designs suffer a major flaw in their lack of `*rotatability*.'

designs

Rotatability of Designs

"Rotatability" is a desirable property not present in 3-level factorial designs

In a rotatable design, the variance of the predicted values of *y* is a function of the distance of a point from the center of the design and is not a function of the direction the point lies from the center. Before a study begins, little or no knowledge may exist about the region that contains the optimum response. Therefore, the experimental design matrix should not bias an investigation in any direction.

Contours of variance of predicted values are concentric circles

In a rotatable design, the contours associated with the variance of the predicted values are concentric circles. Figures 3.16 and 3.17 (adapted from Box and Draper, `Empirical Model Building and Response Surfaces,' page 485) illustrate a three-dimensional plot and contour plot, respectively, of the `information function' associated with a 3^2 design.

Information function

The information function is:

$$\frac{1}{V(\hat{y})}$$

with V denoting the variance (of the predicted value $\hat{y}$).

Each figure clearly shows that the information content of the design is not only a function of the distance from the center of the design space, but also a function of direction.

Graphs of the information function for a rotatable quadratic design

Figures 3.18 and 3.19 are the corresponding graphs of the information function for a rotatable quadratic design. In each of these figures, the value of the information function depends only on the distance of a point from the center of the space.

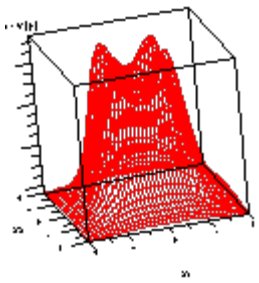


FIGURE 3.16 Three-Dimensional Illustration for the Information Function of a 3^2 Design

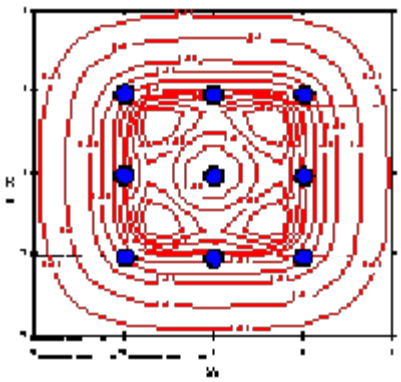


FIGURE 3.17 Contour Map of the Information Function for a 3^2 Design

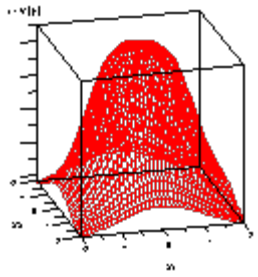


FIGURE 3.18 Three-Dimensional Illustration of the Information Function for a Rotatable Quadratic Design for Two Factors

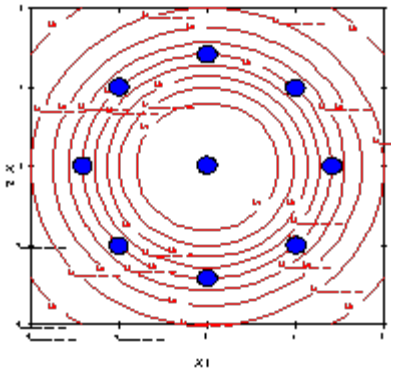


FIGURE 3.19 Contour Map of the Information Function for a Rotatable Quadratic Design for Two Factors

Classical Quadratic Designs

Central composite and Box-Behnken designs

Introduced during the 1950's, classical quadratic designs fall into two broad categories: Box-Wilson central composite designs and Box-Behnken designs. The next sections describe these design classes and their properties.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.6. Response surface designs

5.3.3.6.1. Central Composite Designs (CCD)

Box-Wilson Central Composite Designs

CCD designs start with a factorial or fractional factorial design (with center points) and add "star" points to estimate curvature

A Box-Wilson Central Composite Design, commonly called `a central composite design,' contains an imbedded factorial or fractional factorial design with center points that is augmented with a group of `star points' that allow estimation of curvature. If the distance from the center of the design space to a factorial point is ±1 unit for each factor, the distance from the center of the design space to a star point is $\pm\alpha$ with $|\alpha| > 1$. The precise value of α depends on certain properties desired for the design and on the number of factors involved.

Similarly, the number of centerpoint runs the design is to contain also depends on certain properties required for the design.

Diagram of central composite design generation for two factors

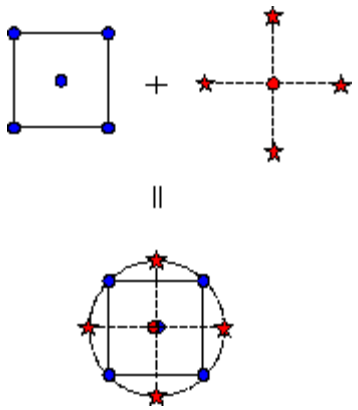


FIGURE 3.20 Generation of a Central Composite Design for Two Factors

A CCD design with k factors has 2k star points

A central composite design always contains twice as many star points as there are factors in the design. The star points represent new extreme values (low and high) for each factor in the design. Table 3.22 summarizes the properties of the three varieties of central composite designs. Figure 3.21 illustrates the relationships among these varieties.

Description of

TABLE 3.22 Central Composite Designs

3 types of CCD designs, which depend on where the star points are placed

Central Composite Design Type	Terminology	Comments
Circumscribed	CCC	CCC designs are the original form of the central composite design. The star points are at some distance α from the center based on the properties desired for the design and the number of factors in the design. The star points establish new extremes for the low and high settings for all factors. Figure 5 illustrates a CCC design. These designs have circular, spherical, or hyperspherical symmetry and require 5 levels for each factor. Augmenting an existing factorial or resolution V fractional factorial design with star points can produce this design.
Inscribed	CCI	For those situations in which the limits specified for factor settings are truly limits, the CCI design uses the factor settings as the star points and creates a factorial or fractional factorial design within those limits (in other words, a CCI design is a scaled down CCC design with each factor level of the CCC design divided by α to generate the CCI design). This design also requires 5 levels of each factor.
Face Centered	CCF	In this design the star points are at the center of each face of the factorial space, so $\alpha = \pm 1$. This variety requires 3 levels of each factor. Augmenting an existing factorial or resolution V design with appropriate star points can

		also produce this design.

Pictorial representation of where the star points are placed for the 3 types of CCD designs

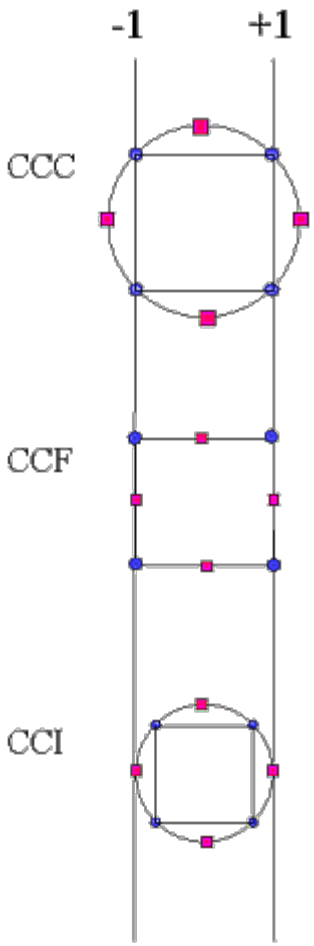


FIGURE 3.21 Comparison of the Three Types of Central Composite Designs

Comparison of the 3 central composite designs

The diagrams in Figure 3.21 illustrate the three types of central composite designs for two factors. Note that the CCC explores the largest process space and the CCI explores the smallest process space. Both the CCC and CCI are rotatable designs, but the CCF is not. In the CCC design, the design points describe a circle *circumscribed* about the factorial square. For three factors, the CCC design points describe a sphere around the factorial cube.

Determining α in Central Composite Designs

The value of α is chosen to maintain rotatability

To maintain rotatability, the value of α depends on the number of experimental runs in the factorial portion of the central composite design:

$$\alpha = [\text{number of factorial runs}]^{1/4}$$

If the factorial is a full factorial, then

$$\alpha = \left[2^k\right]^{1/4}$$

However, the factorial portion can also be a fractional factorial design of resolution V.

Table 3.23 illustrates some typical values of α as a function of the number of factors.

Values of α depending on the number of factors in the factorial part of the design

TABLE 3.23 Determining α for Rotatability

Number of Factors	Factorial Portion	Scaled Value for α Relative to ±1
2	2^2	$2^{2/4} = 1.414$
3	2^3	$2^{3/4} = 1.682$
4	2^4	$2^{4/4} = 2.000$
5	2^{5-1}	$2^{4/4} = 2.000$
5	2^5	$2^{5/4} = 2.378$
6	2^{6-1}	$2^{5/4} = 2.378$
6	2^6	$2^{6/4} = 2.828$

Orthogonal blocking

The value of α also depends on whether or not the design is orthogonally blocked. That is, the question is whether or not the design is divided into blocks such that the block effects do not affect the estimates of the coefficients in the 2nd order model.

Example of both rotatability and orthogonal blocking for two factors

Under some circumstances, the value of α allows simultaneous rotatability and orthogonality. One such example for $k = 2$ is shown below:

BLOCK	X1	X2
1	**-1**	**-1**
1	**1**	**-1**
1	**-1**	**1**
1	**1**	**1**
1	**0**	**0**
1	**0**	**0**
2	**-1.414**	**0**
2	**1.414**	**0**
2	**0**	**-1.414**
2	**0**	**1.414**

2	**0**	**0**
2	**0**	**0**

Additional central composite designs

Examples of other central composite designs will be given after Box-Behnken designs are described.

5.3.3.6.2. Box-Behnken designs

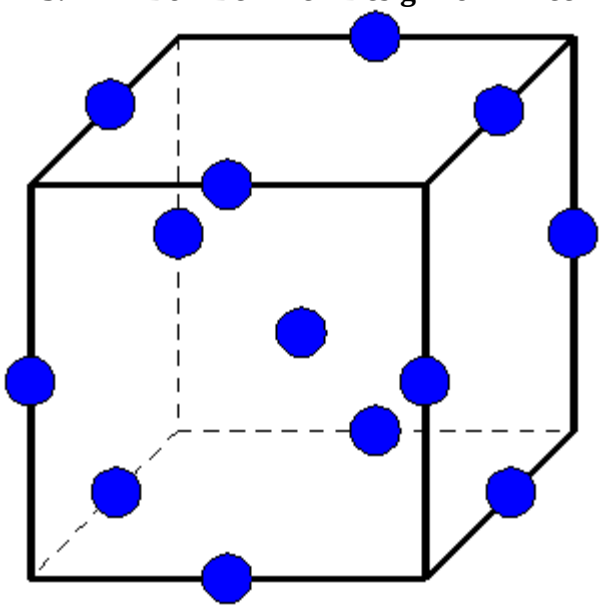
An alternate choice for fitting quadratic models that requires 3 levels of each factor and is rotatable (or "nearly" rotatable)

The Box-Behnken design is an independent quadratic design in that it does not contain an embedded factorial or fractional factorial design. In this design the treatment combinations are at the midpoints of edges of the process space and at the center. These designs are rotatable (or near rotatable) and require 3 levels of each factor. The designs have limited capability for orthogonal blocking compared to the central composite designs.

Figure 3.22 illustrates a Box-Behnken design for three factors.

Box-Behnken design for 3 factors

FIGURE 3.22 A Box-Behnken Design for Three Factors



Geometry of the design

The geometry of this design suggests a sphere within the process space such that the surface of the sphere protrudes through each face with the surface of the sphere tangential to the midpoint of each edge of the space.

Examples of Box-Behnken designs are given on the next page.

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.6. Response surface designs

5.3.3.6.3. Comparisons of response surface designs

Choosing a Response Surface Design

Various CCD designs and Box-Behnken designs are compared and their properties discussed

Table 3.24 contrasts the structures of four common quadratic designs one might use when investigating three factors. The table combines CCC and CCI designs because they are structurally identical.

For three factors, the Box-Behnken design offers some advantage in requiring a fewer number of runs. For 4 or more factors, this advantage disappears.

Structural comparisons of CCC (CCI), CCF, and Box-Behnken designs for three factors

TABLE 3.24 Structural Comparisons of CCC (CCI), CCF, and Box-Behnken Designs for Three Factors

CCC (CCI)				CCF				Box-Behnken			
Rep	***X*1**	***X*2**	***X*3**	**Rep**	***X*1**	***X*2**	***X*3**	**Rep**	***X*1**	***X*2**	***X*3**
1	-1	-1	-1	1	-1	-1	-1	1	-1	-1	0
1	+1	-1	-1	1	+1	-1	-1	1	+1	-1	0
1	-1	+1	-1	1	-1	+1	-1	1	-1	+1	0
1	+1	+1	-1	1	+1	+1	-1	1	+1	+1	0
1	-1	-1	+1	1	-1	-1	+1	1	-1	0	-1
1	+1	-1	+1	1	+1	-1	+1	1	+1	0	-1
1	-1	+1	+1	1	-1	+1	+1	1	-1	0	+1
1	+1	+1	+1	1	+1	+1	+1	1	+1	0	+1
1	-1.682	0	0	1	-1	0	0	1	0	-1	-1
1	1.682	0	0	1	+1	0	0	1	0	+1	-1
1	0	-1.682	0	1	0	-1	0	1	0	-1	+1
1	0	1.682	0	1	0	+1	0	1	0	+1	+1
1	0	0	-1.682	1	0	0	-1	3	0	0	0
1	0	0	1.682	1	0	0	+1				
6	0	0	0	6	0	0	0				
Total Runs = 20				**Total Runs = 20**				**Total Runs = 15**			

Factor settings for CCC and CCI three factor designs

Table 3.25 illustrates the factor settings required for a central composite circumscribed (CCC) design and for a central composite inscribed (CCI) design (standard order), assuming three factors, each with low and high settings of 10 and 20, respectively. Because the CCC design generates new extremes for all factors, the investigator must inspect any worksheet generated for such a design to make certain that the factor settings called for are reasonable.

In Table 3.25, treatments 1 to 8 in each case are the factorial points in the design; treatments 9 to 14 are the star points; and 15 to 20 are the system-recommended center points. Notice in the CCC design how the low and high values of each factor have been extended to create the star points. In the CCI design, the specified low and high values become the star points, and the system computes appropriate settings for the factorial part of the design inside those boundaries.

TABLE 3.25 Factor Settings for CCC and CCI Designs for Three Factors

Central Composite Circumscribed CCC					**Central Composite Inscribed CCI**			
Sequence Number	***X*1**	***X*2**	***X*3**		**Sequence Number**	***X*1**	***X*2**	***X*3**
1	10	10	10		**1**	12	12	12
2	20	10	10		**2**	18	12	12
3	10	20	10		**3**	12	18	12
4	20	20	10		**4**	18	18	12
5	10	10	20		**5**	12	12	18
6	20	10	20		**6**	18	12	18
7	10	20	20		**7**	12	12	18
8	20	20	20		**8**	18	18	18
9	6.6	15	15	*	**9**	10	15	15
10	23.4	15	15	*	**10**	20	15	15
11	15	6.6	15	*	**11**	15	10	15
12	15	23.4	15	*	**12**	15	20	15
13	15	15	6.6	*	**13**	15	15	10
14	15	15	23.4	*	**14**	15	15	20
15	15	15	15		**15**	15	15	15
16	15	15	15		**16**	15	15	15
17	15	15	15		**17**	15	15	15
18	15	15	15		**18**	15	15	15
19	15	15	15		**19**	15	15	15
20	15	15	15		**20**	15	15	15

* are star points

Factor settings for CCF and Box-Behnken three factor designs

Table 3.26 illustrates the factor settings for the corresponding central composite face-centered (CCF) and Box-Behnken designs. Note that each of these designs provides three levels for each factor and that the Box-Behnken design requires fewer runs in the three-factor case.

TABLE 3.26 Factor Settings for CCF and Box-Behnken Designs for Three Factors

Central Composite Face-Centered CCC					**Box-Behnken**			
Sequence Number	***X*1**	***X*2**	***X*3**		**Sequence Number**	***X*1**	***X*2**	***X*3**
1	10	10	10		**1**	10	10	15
2	20	10	10		**2**	20	10	15
3	10	20	10		**3**	10	20	15
4	20	20	10		**4**	20	20	15
5	10	10	20		**5**	10	15	10
6	20	10	20		**6**	20	15	10
7	10	20	20		**7**	10	15	20
8	20	20	20		**8**	20	15	20
9	10	15	15	*	**9**	15	10	10
10	20	15	15	*	**10**	15	20	10
11	15	10	15	*	**11**	15	10	20
12	15	20	15	*	**12**	15	20	20
13	15	15	10	*	**13**	15	15	15
14	15	15	20	*	**14**	15	15	15
15	15	15	15		**15**	15	15	15
16	15	15	15					
17	15	15	15					
18	15	15	15					
19	15	15	15					
20	15	15	15					

* are star points for the CCC

Properties of classical response surface designs

Table 3.27 summarizes properties of the classical quadratic designs. Use this table for broad guidelines when attempting to choose from among available designs.

TABLE 3.27 Summary of Properties of Classical Response Surface Designs

Design Type	Comment
	CCC designs provide high quality predictions over the entire design space, but require factor settings outside the range of the factors in the factorial part. **Note:** When the

CCC	possibility of running a CCC design is recognized before starting a factorial experiment, factor spacings can be reduced to ensure that $\pm \alpha$ for each coded factor corresponds to feasible (reasonable) levels. Requires 5 levels for each factor.
CCI	CCI designs use only points within the factor ranges originally specified, but do not provide the same high quality prediction over the entire space compared to the CCC. Requires 5 levels of each factor.
CCF	CCF designs provide relatively high quality predictions over the entire design space and do not require using points outside the original factor range. However, they give poor precision for estimating pure quadratic coefficients. Requires 3 levels for each factor.
Box-Behnken	These designs require fewer treatment combinations than a central composite design in cases involving 3 or 4 factors. The Box-Behnken design is rotatable (or nearly so) but it contains regions of poor prediction quality like the CCI. Its "missing corners" may be useful when the experimenter should avoid combined factor extremes. This property prevents a potential loss of data in those cases. Requires 3 levels for each factor.

Number of runs required by central composite and Box-Behnken designs

Table 3.28 compares the number of runs required for a given number of factors for various Central Composite and Box-Behnken designs.

TABLE 3.28 Number of Runs Required by Central Composite and Box-Behnken Designs

Number of Factors	Central Composite	Box-Behnken
2	13 (5 center points)	-
3	20 (6 centerpoint runs)	15
4	30 (6 centerpoint runs)	27
5	33 (fractional factorial) or 52 (full factorial)	46
6	54 (fractional factorial) or 91 (full factorial)	54

Desirable Features for Response Surface Designs

A summary of desirable properties for response

G. E. P. Box and N. R. Draper in "Empirical Model Building and Response Surfaces," John Wiley and Sons, New York, 1987, page 477, identify desirable properties for a response surface design:

- Satisfactory distribution of information across the experimental

surface designs

region.
- *rotatability*

- Fitted values are as close as possible to observed values.
  - *minimize residuals or error of prediction*
- Good lack of fit detection.
- Internal estimate of error.
- Constant variance check.
- Transformations can be estimated.
- Suitability for blocking.
- Sequential construction of higher order designs from simpler designs
- Minimum number of treatment combinations.
- Good graphical analysis through simple data patterns.
- Good behavior when errors in settings of input variables occur.

5.3.3.6.4. Blocking a response surface design

How can we block a response surface design?

When augmenting a resolution V design to a CCC design by adding star points, it may be desirable to block the design

If an investigator has run either a 2^k full factorial or a 2^{k-p} fractional factorial design of at least resolution V, augmentation of that design to a central composite design (either CCC of CCF) is easily accomplished by adding an additional set (block) of star and centerpoint runs. If the factorial experiment indicated (via the *t* test) curvature, this composite augmentation is the best follow-up option (follow-up options for other situations will be discussed later).

An orthogonal blocked response surface design has advantages

An important point to take into account when choosing a response surface design is the possibility of running the design in blocks. Blocked designs are better designs if the design allows the estimation of individual and interaction factor effects independently of the block effects. This condition is called orthogonal blocking. Blocks are assumed to have no impact on the nature and shape of the response surface.

CCF designs cannot be orthogonally blocked

The CCF design does not allow orthogonal blocking and the Box-Behnken designs offer blocking only in limited circumstances, whereas the CCC does permit orthogonal blocking.

Axial and factorial blocks

In general, when two blocks are required there should be an axial block and a factorial block. For three blocks, the factorial block is divided into two blocks and the axial block is not split. The blocking of the factorial design points should result in orthogonality between blocks and individual factors and between blocks and the two factor interactions.

The following Central Composite design in two factors is broken into two blocks.

Table of

TABLE 3.29 CCD: 2 Factors, 2 Blocks

CCD design with 2 factors and 2 blocks

Pattern	Block	*X*1	*X*2	Comment
--	1	-1	-1	Full Factorial
-+	1	-1	+1	Full Factorial
+-	1	+1	-1	Full Factorial
++	1	+1	+1	Full Factorial
00	1	0	0	Center-Full Factorial
00	1	0	0	Center-Full Factorial
00	1	0	0	Center-Full Factorial
-0	2	-1.414214	0	Axial
+0	2	+1.414214	0	Axial
0-	2	0	-1.414214	Axial
0+	2	0	+1.414214	Axial
00	2	0	0	Center-Axial
00	2	0	0	Center-Axial
00	2	0	0	Center-Axial

Note that the first block includes the full factorial points and three centerpoint replicates. The second block includes the axial points and another three centerpoint replicates. Naturally these two blocks should be run as two separate random sequences.

Table of CCD design with 3 factors and 3 blocks

The following three examples show blocking structure for various designs.

TABLE 3.30 CCD: 3 Factors 3 Blocks, Sorted by Block

Pattern	Block	*X*1	*X*2	*X*3	Comment
---	1	-1	-1	-1	Full Factorial
-++	1	-1	+1	+1	Full Factorial
+-+	1	+1	-1	+1	Full Factorial
++-	1	+1	+1	-1	Full Factorial
000	1	0	0	0	Center-Full Factorial
000	1	0	0	0	Center-Full Factorial
--+	2	-1	-1	+1	Full Factorial
-+-	2	-1	+1	-1	Full Factorial
+--	2	+1	-1	-1	Full Factorial
+++	2	+1	+1	+1	Full Factorial
000	2	0	0	0	Center-Full Factorial
000	2	0	0	0	Center-Full Factorial
-00	3	-1.63299	0	0	Axial
+00	3	+1.63299	0	0	Axial
0-0	3	0	-1.63299	0	Axial
0+0	3	0	+1.63299	0	Axial
00-	3	0	0	-1.63299	Axial
00+	3	0	0	+1.63299	Axial

000	3	0	0	0	Axial
000	3	0	0	0	Axial

Table of CCD design with 4 factors and 3 blocks

TABLE 3.31 CCD: 4 Factors, 3 Blocks

Pattern	Block	*X*1	*X*2	*X*3	*X*4	Comment
---+	1	-1	-1	-1	+1	Full Factorial
--+-	1	-1	-1	+1	-1	Full Factorial
-+--	1	-1	+1	-1	-1	Full Factorial
-+++	1	-1	+1	+1	+1	Full Factorial
+---	1	+1	-1	-1	-1	Full Factorial
+-++	1	+1	-1	+1	+1	Full Factorial
++-+	1	+1	+1	-1	+1	Full Factorial
+++-	1	+1	+1	+1	-1	Full Factorial
0000	1	0	0	0	0	Center-Full Factorial
0000	1	0	0	0	0	Center-Full Factorial
----	2	-1	-1	-1	-1	Full Factorial
--++	2	-1	-1	+1	+1	Full Factorial
-+-+	2	-1	+1	-1	+1	Full Factorial
-++-	2	-1	+1	+1	-1	Full Factorial
+--+	2	+1	-1	-1	+1	Full Factorial
+-+-	2	+1	-1	+1	-1	Full Factorial
++--	2	+1	+1	-1	-1	Full Factorial
++++	2	+1	+1	+1	+1	Full Factorial
0000	2	0	0	0	0	Center-Full Factorial
0000	2	0	0	0	0	Center-Full Factorial
-000	3	-2	0	0	0	Axial
+000	3	+2	0	0	0	Axial
+000	3	+2	0	0	0	Axial
0-00	3	0	-2	0	0	Axial
0+00	3	0	+2	0	0	Axial
00-0	3	0	0	-2	0	Axial
00+0	3	0	0	+2	0	Axial
000-	3	0	0	0	-2	Axial
000+	3	0	0	0	+2	Axial
0000	3	0	0	0	0	Center-Axial

Table of CCD design with 5 factors and 2 blocks

TABLE 3.32 CCD: 5 Factors, 2 Blocks

Pattern	Block	*X*1	*X*2	*X*3	*X*4	*X*5	Comment
----+	1	-1	-1	-1	-1	+1	Fractional Factorial
---+-	1	-1	-1	-1	+1	-1	Fractional Factorial

--+--	1	-1	-1	+1	-1	-1	Fractional Factorial
--+++	1	-1	-1	+1	+1	+1	Fractional Factorial
-+---	1	-1	+1	-1	-1	-1	Fractional Factorial
-+-++	1	-1	+1	-1	+1	+1	Fractional Factorial
-++-+	1	-1	+1	+1	-1	+1	Fractional Factorial
-+++-	1	-1	+1	+1	+1	-1	Fractional Factorial
+----	1	+1	-1	-1	-1	-1	Fractional Factorial
+--++	1	+1	-1	-1	+1	+1	Fractional Factorial
+-+-+	1	+1	-1	+1	-1	+1	Fractional Factorial
+-++-	1	+1	-1	+1	+1	-1	Fractional Factorial
++--+	1	+1	+1	-1	-1	+1	Fractional Factorial
++-+-	1	+1	+1	-1	+1	-1	Fractional Factorial
+++--	1	+1	+1	+1	-1	-1	Fractional Factorial
+++++	1	+1	+1	+1	+1	+1	Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
00000	1	0	0	0	0	0	Center-Fractional Factorial
-0000	2	-2	0	0	0	0	Axial
+0000	2	+2	0	0	0	0	Axial
0-000	2	0	-2	0	0	0	Axial

0+000	2	0	+2	0	0	0	Axial
00-00	2	0	0	-2	0	0	Axial
00+00	2	0	0	+2	0	0	Axial
000-0	2	0	0	0	-2	0	Axial
000+0	2	0	0	0	+2	0	Axial
0000-	2	0	0	0	0	-2	Axial
0000+	2	0	0	0	0	+2	Axial
00000	2	0	0	0	0	0	Center-Axial

5.3.3.7. Adding centerpoints

Center point, or `Control' Runs

Centerpoint runs provide a check for both process stability and possible curvature

As mentioned earlier in this section, we add centerpoint runs interspersed among the experimental setting runs for two purposes:

1. To provide a measure of process stability and inherent variability
2. To check for curvature.

Centerpoint runs are not randomized

Centerpoint runs should begin and end the experiment, and should be dispersed as evenly as possible throughout the design matrix. The centerpoint runs are not randomized! There would be no reason to randomize them as they are there as guardians against process instability and the best way to find instability is to sample the process on a regular basis.

Rough rule of thumb is to add 3 to 5 center point runs to your design

With this in mind, we have to decide on how many centerpoint runs to do. This is a tradeoff between the resources we have, the need for enough runs to see if there is process instability, and the desire to get the experiment over with as quickly as possible. *As a rough guide, you should generally add approximately 3 to 5 centerpoint runs to a full or fractional factorial design.*

Table of randomized, replicated 2^3 full factorial design with centerpoints

In the following Table we have added three centerpoint runs to the otherwise randomized design matrix, making a total of nineteen runs.

TABLE 3.32 Randomized, Replicated 2^3 Full Factorial Design Matrix with Centerpoint Control Runs Added

	Random Order	Standard Order	SPEED	FEED	DEPTH
1	not applicable	not applicable	0	0	0
2	1	5	-1	-1	1
3	2	15	-1	1	1
4	3	9	-1	-1	-1

5	4	7	-1	1	1
6	5	3	-1	1	-1
7	6	12	1	1	-1
8	7	6	1	-1	1
9	8	4	1	1	-1
10	not applicable	not applicable	0	0	0
11	9	2	1	-1	-1
12	10	13	-1	-1	1
13	11	8	1	1	1
14	12	16	1	1	1
15	13	1	-1	-1	-1
16	14	14	1	-1	1
17	15	11	-1	1	-1
18	16	10	1	-1	-1
19	not applicable	not applicable	0	0	0

Preparing a worksheet for operator of experiment

To prepare a worksheet for an operator to use when running the experiment, delete the columns `RandOrd' and `Standard Order.' Add an additional column for the output (Yield) on the right, and change all `-1', `0', and `1' to original factor levels as follows.

Operator worksheet

TABLE 3.33 DOE Worksheet Ready to Run

Sequence Number	Speed	Feed	Depth	Yield
1	20	0.003	0.015	
2	16	0.001	0.02	
3	16	0.005	0.02	
4	16	0.001	0.01	
5	16	0.005	0.02	
6	16	0.005	0.01	
7	24	0.005	0.01	
8	24	0.001	0.02	
9	24	0.005	0.01	
10	20	0.003	0.015	
11	24	0.001	0.01	
12	16	0.001	0.02	
13	24	0.005	0.02	
14	24	0.005	0.02	
15	16	0.001	0.01	
16	24	0.001	0.02	
17	16	0.005	0.01	

18	24	0.001	0.01	
19	20	0.003	0.015	

Note that the control (centerpoint) runs appear at rows 1, 10, and 19.

This worksheet can be given to the person who is going to do the runs/measurements and asked to proceed through it from first row to last in that order, filling in the Yield values as they are obtained.

Pseudo Center points

Center points for discrete factors

One often runs experiments in which some factors are nominal. For example, Catalyst "A" might be the (-1) setting, catalyst "B" might be coded (+1). The choice of which is "high" and which is "low" is arbitrary, but one must have some way of deciding which catalyst setting is the "standard" one.

These standard settings for the discrete input factors together with center points for the continuous input factors, will be regarded as the "center points" for purposes of design.

Center Points in Response Surface Designs

Uniform precision

In an unblocked response surface design, the number of center points controls other properties of the design matrix. The number of center points can make the design orthogonal or have "uniform precision." We will only focus on uniform precision here as classical quadratic designs were set up to have this property.

Variance of prediction

Uniform precision ensures that the variance of prediction is the same at the center of the experimental space as it is at a unit distance away from the center.

Protection against bias

In a response surface context, to contrast the virtue of uniform precision designs over replicated center-point orthogonal designs one should also consider the following guidance from Montgomery ("Design and Analysis of Experiments," Wiley, 1991, page 547), "*A uniform precision design offers more protection against bias in the regression coefficients than does an orthogonal design because of the presence of third-order and higher terms in the true surface*.

Controlling α and the number of center points

Myers, Vining, et al, ["Variance Dispersion of Response Surface Designs," Journal of Quality Technology, 24, pp. 1-11 (1992)] have explored the options regarding the number of center points and the value of α somewhat further: An investigator may control two parameters, α and the number of center points (n_c), given k factors. Either set $\alpha = 2^{(k/4)}$

(for rotatability) or $\sqrt{k}$ -- an axial point on perimeter of design region. Designs are similar in performance with $\sqrt{k}$ preferable as k increases. Findings indicate that the best overall design performance occurs with $\alpha \approx \sqrt{k}$ and $2 \leq n_c \leq 5$.

5.3.3.8. Improving fractional factorial design resolution

Foldover designs increase resolution

Earlier we saw how fractional factorial designs resulted in an alias structure that confounded main effects with certain interactions. Often it is useful to know how to run a few additional treatment combinations to remove alias structures that might be masking significant effects or interactions.

Partial foldover designs break up specific alias patterns

Two methods will be described for selecting these additional treatment combinations:

- Mirror-image foldover designs (to build a resolution IV design from a resolution III design)
- Alternative foldover designs (to break up specific alias patterns).

5. Process Improvement
5.3. Choosing an experimental design
5.3.3. How do you select an experimental design?
5.3.3.8. Improving fractional factorial design resolution

5.3.3.8.1. Mirror-Image foldover designs

A foldover design is obtained from a fractional factorial design by reversing the signs of all the columns

A mirror-image fold-over (or foldover, without the hyphen) design is used to augment fractional factorial designs to increase the resolution of 2^{5-1}_{III} and Plackett-Burman designs. It is obtained by reversing the signs of all the columns of the original design matrix. The original design runs are combined with the mirror-image fold-over design runs, and this combination can then be used to estimate all main effects clear of any two-factor interaction. This is referred to as: *breaking the alias link between main effects and two-factor interactions.*

Before we illustrate this concept with an example, we briefly review the basic concepts involved.

Review of Fractional 2^{k-p} Designs

A resolution III design, combined with its mirror-image foldover, becomes resolution IV

In general, a design type that uses a specified fraction of the runs from a full factorial and is balanced and orthogonal is called a *fractional factorial*.

A 2-level fractional factorial is constructed as follows: *Let the number of runs be 2^{k-p}. Start by constructing the full factorial for the k-p variables. Next associate the extra factors with higher-order interaction columns. The Table shown previously details how to do this to achieve a minimal amount of confounding.*

For example, consider the 2^{5-2} design (a resolution III design). The full factorial for $k = 5$ requires $2^5 = 32$ runs. The fractional factorial can be achieved in $2^{5-2} = 8$ runs, called a quarter (1/4) fractional design, by setting $X4 = X1*X2$ and $X5 = X1*X3$.

Design matrix for a 2^{5-2} fractional factorial

The design matrix for a 2^{5-2} fractional factorial looks like:

TABLE 3.34 Design Matrix for a 2^{5-2} Fractional Factorial

run	X1	X2	X3	X4 = X1X2	X5 = X1X3
1	-1	-1	-1	+1	+1

2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	+1
4	+1	+1	-1	+1	-1
5	-1	-1	+1	+1	-1
6	+1	-1	+1	-1	+1
7	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1

Design Generators, Defining Relation and the Mirror-Image Foldover

Increase to resolution IV design by augmenting design matrix

In this design the *X*1*X*2 column was used to generate the *X*4 main effect and the *X*1*X*3 column was used to generate the *X*5 main effect. The design generators are: 4 = 12 and 5 = 13 and the defining relation is I = 124 = 135 = 2345. Every main effect is confounded (aliased) with at least one first-order interaction (see the confounding structure for this design).

We can increase the resolution of this design to IV if we augment the 8 original runs, adding on the 8 runs from the mirror-image fold-over design. These runs make up another 1/4 fraction design with design generators 4 = -12 and 5 = -13 and defining relation I = -124 = -135 = 2345. The augmented runs are:

Augmented runs for the design matrix

run	***X*1**	***X*2**	***X*3**	***X*4 = -*X*1*X*2**	***X*5 = -*X*1*X*3**
9	+1	+1	+1	-1	-1
10	-1	+1	+1	+1	+1
11	+1	-1	+1	+1	-1
12	-1	-1	+1	-1	+1
13	+1	+1	-1	-1	+1
14	-1	+1	-1	+1	-1
15	+1	-1	-1	+1	+1
16	-1	-1	-1	-1	-1

Mirror-image foldover design reverses all signs in original design matrix

A *mirror-image foldover design* is the original design with *all signs reversed*. It breaks the alias chains between *every main factor and two-factor interaction* of a resolution III design. That is, we can estimate *all the main effects clear of any two-factor interaction*.

A 1/16 Design Generator Example

2^{7-3} example

Now we consider a more complex example.

We would like to study the effects of 7 variables. A full 2-level factorial, 2^7, would require 128 runs.

Assume economic reasons restrict us to 8 runs. We will build a $2^{7-4} = 2^3$ full factorial and assign certain products of columns to the $X4$, $X5$, $X6$ and $X7$ variables. This will generate a resolution III design in which all of the main effects are aliased with first-order and higher interaction terms. The design matrix (see the previous Table for a complete description of this fractional factorial design) is:

Design matrix for 2^{7-3} fractional factorial

Design Matrix for a 2^{7-3} Fractional Factorial

run	*X*1	*X*2	*X*3	*X*4 = *X*1*X*2	*X*5 = *X*1*X*3	*X*6 = *X*2*X*3	*X*7 = *X*1*X*2*X*3
1	-1	-1	-1	+1	+1	+1	-1
2	+1	-1	-1	-1	-1	+1	+1
3	-1	+1	-1	-1	+1	-1	+1
4	+1	+1	-1	+1	-1	-1	-1
5	-1	-1	+1	+1	-1	-1	+1
6	+1	-1	+1	-1	+1	-1	-1
7	-1	+1	+1	-1	-1	+1	-1
8	+1	+1	+1	+1	+1	+1	+1

Design generators and defining relation for this example

The design generators for this 1/16 fractional factorial design are:

4 = 12, 5 = 13, 6 = 23 and 7 = 123

From these we obtain, by multiplication, the defining relation:

I = 124 = 135 = 236 = 347 = 257 = 167 = 456 = 1237 =

2345 = 1346 = 1256 = 1457 = 2467 = 3567 = 1234567.

Computing alias structure for complete design

Using this defining relation, we can easily compute the alias structure for the complete design, as shown previously in the link to the fractional design Table given earlier. For example, to figure out which effects are aliased (confounded) with factor $X1$ we multiply the defining relation by 1 to obtain:

1 = 24 = 35 = 1236 = 1347 = 1257 = 67 = 1456 = 237
= 12345 = 346 = 256 = 457 = 12467 = 13567 = 234567

In order to simplify matters, let us ignore all interactions with 3 or more factors; we then have the following 2-factor alias pattern for $X1$: 1 = 24 = 35 = 67 or, using the full notation, $X1 = X2*X4 = X3*X5 = X6*X7$.

The same procedure can be used to obtain all the other aliases for each of the main effects, generating the following list:

1 = 24 = 35 = 67
2 = 14 = 36 = 57
3 = 15 = 26 = 47
4 = 12 = 37 = 56
5 = 13 = 27 = 46
6 = 17 = 23 = 45
7 = 16 = 25 = 34

Signs in every column of original design matrix reversed for mirror-image foldover design

The chosen design used a set of generators with all positive signs. The mirror-image foldover design uses generators with negative signs for terms with an even number of factors or, 4 = -12, 5 = -13, 6 = -23 and 7 = 123. This generates a design matrix that is equal to the original design matrix with every sign in every column reversed.

If we augment the initial 8 runs with the 8 mirror-image foldover design runs (with all column signs reversed), we can de-alias all the main effect estimates from the 2-way interactions. The additional runs are:

Design matrix for mirror-image foldover runs

Design Matrix for the Mirror-Image Foldover Runs of the 2^{7-3} Fractional Factorial

run	*X*1	*X*2	*X*3	*X*4 = *X*1*X*2	*X*5 = *X*1*X*3	*X*6 = *X*2*X*3	*X*7 = *X*1*X*2*X*3
1	+1	+1	+1	-1	-1	-1	+1
2	-1	+1	+1	+1	+1	-1	-1
3	+1	-1	+1	+1	-1	+1	-1
4	-1	-1	+1	-1	+1	+1	+1
5	+1	+1	-1	-1	+1	+1	-1
6	-1	+1	-1	+1	-1	+1	+1
7	+1	-1	-1	+1	+1	-1	+1
8	-1	-1	-1	-1	-1	-1	-1

Alias structure for augmented runs

Following the same steps as before and making the same assumptions about the omission of higher-order interactions in the alias structure, we arrive at:

1 = -24 = -35 = -67
2 = -14 = -36 =- 57
3 = -15 = -26 = -47
4 = -12 = -37 = -56
5 = -13 = -27 = -46
6 = -17 = -23 = -45
7 = -16 = -25 = -34

With both sets of runs, we can now estimate all the main effects free from two factor interactions.

Build a resolution IV design from a resolution III design

Note: *In general, a mirror-image foldover design is a method to build a resolution IV design from a resolution III design. It is never used to follow-up a resolution IV design*.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

5.3.3.8.2. Alternative foldover designs

Alternative foldover designs can be an economical way to break up a selected alias pattern

The mirror-image foldover (in which signs in all columns are reversed) is only one of the possible follow-up fractions that can be run to augment a fractional factorial design. It is the most common choice when the original fraction is resolution III. However, alternative foldover designs with fewer runs can often be utilized to break up selected alias patterns. We illustrate this by looking at what happens when the signs of a single factor column are reversed.

Example of de-aliasing a single factor

Previously, we described how we de-alias all the factors of a 2^{7-4} experiment. Suppose that we only want to de-alias the $X4$ factor. This can be accomplished by only changing the sign of $X4 = X1X2$ to $X4 = -X1X2$. The resulting design is:

Table showing design matrix of a reverse X4 foldover design

TABLE 3.36 A "Reverse X4" Foldover Design

run	*X*1	*X*2	*X*3	*X*4 = -*X*1*X*2	*X*5 = -*X*1*X*3	*X*6 = *X*2*X*3	*X*7 = *X*1*X*2*X*3
1	-1	-1	-1	-1	+1	+1	-1
2	+1	-1	-1	+1	-1	+1	+1
3	-1	+1	-1	+1	+1	-1	+1
4	+1	+1	-1	-1	-1	-1	-1
5	-1	-1	+1	-1	-1	-1	+1
6	+1	-1	+1	+1	+1	-1	-1
7	-1	+1	+1	+1	-1	+1	-1
8	+1	+1	+1	-1	+1	+1	+1

Alias patterns and effects that can be estimated in the example design

The two-factor alias patterns for $X4$ are: Original experiment: $X4 = X1X2 = X3X7 = X5X6$; "Reverse $X4$" foldover experiment: $X4 = -X1X2 = -X3X7 = -X5X6$.

The following effects can be estimated by combining the original 2^{7-4}_{III} with the "Reverse $X4$" foldover fraction:

$X1 + X3X5 + X6X7$
$X2 + X3X6 + X5X7$
$X3 + X1X5 + X2X6$
$X4$

$X5 + X1X3 + X2X7$
$X6 + X2X3 + X1X7$
$X7 + X2X5 + X1X6$
$X1X4$
$X2X4$
$X3X4$
$X4X5$
$X4X6$
$X4X7$
$X1X2 + X3X7 + X5X6$

Note: The 16 runs allow estimating the above 14 effects, with one degree of freedom left over for a possible block effect.

Advantage and disadvantage of this example design

The advantage of this follow-up design is that it permits estimation of the $X4$ effect and each of the six two-factor interaction terms involving $X4$.

The disadvantage is that the combined fractions still yield a resolution III design, with all main effects other than $X4$ aliased with two-factor interactions.

Case when purpose is simply to estimate all two-factor interactions of a single factor

Reversing a single factor column to obtain de-aliased two-factor interactions for that one factor works for any resolution III or IV design. When used to follow-up a resolution IV design, there are relatively few new effects to be estimated (as compared to 2^{k-p}_{III} designs). When the original resolution IV fraction provides sufficient precision, and the purpose of the follow-up runs is simply to estimate all two-factor interactions for one factor, the *semifolding* option should be considered.

Semifolding

Number of runs can be reduced for resolution IV designs

For resolution IV fractions, it is possible to economize on the number of runs that are needed to break the alias chains for all two-factor interactions of a single factor. In the above case we needed 8 additional runs, which is the same number of runs that were used in the original experiment. This can be improved upon.

Additional information on John's 3/4 designs

We can repeat only the points that were set at the high levels of the factor of choice and then run them at their low settings in the next experiment. For the given example, this means an additional 4 runs instead 8. We mention this technique only in passing, more details may be found in the references (or see John's 3/4 designs).

5.3.3.9. Three-level full factorial designs

Three-level designs are useful for investigating quadratic effects

The three-level design is written as a 3^k factorial design. It means that k factors are considered, each at 3 levels. These are (usually) referred to as low, intermediate and high levels. These levels are numerically expressed as 0, 1, and 2. One could have considered the digits -1, 0, and +1, but this may be confusing with respect to the 2-level designs since 0 is reserved for center points. Therefore, we will use the 0, 1, 2 scheme. The reason that the three-level designs were proposed is to model possible curvature in the response function and to handle the case of nominal factors at 3 levels. A third level for a continuous factor facilitates investigation of a quadratic relationship between the response and each of the factors.

Three-level design may require prohibitive number of runs

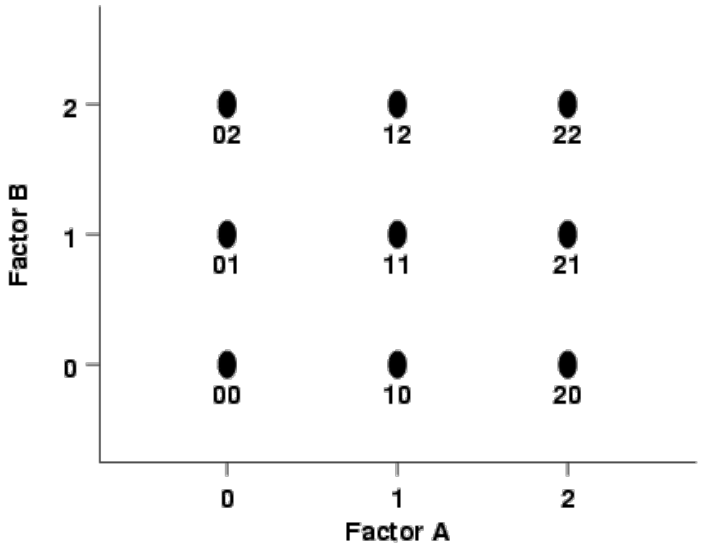
Unfortunately, the three-level design is prohibitive in terms of the number of runs, and thus in terms of cost and effort. For example a two-level design with center points is much less expensive while it still is a very good (and simple) way to establish the presence or absence of curvature.

The 3^2 design

The simplest 3-level design - with only 2 factors

This is the simplest three-level design. It has two factors, each at three levels. The 9 treatment combinations for this type of design can be shown pictorially as follows:

FIGURE 3.23 A 3^2 Design Schematic



A notation such as "20" means that factor A is at its high level (2) and factor B is at its low level (0).

The 3^3 design

The model and treatment runs for a 3 factor, 3-level design

This is a design that consists of three factors, each at three levels. It can be expressed as a 3 x 3 x 3 = 3^3 design. The model for such an experiment is

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + C_k + AC_{ik} + BC_{jk} + ABC_{ijk} + \varepsilon_{ijk}$$

where each factor is included as a nominal factor rather than as a continuous variable. In such cases, main effects have 2 degrees of freedom, two-factor interactions have $2^2 = 4$ degrees of freedom and k-factor interactions have 2^k degrees of freedom. The model contains 2 + 2 + 2 + 4 + 4 + 4 + 8 = 26 degrees of freedom. Note that if there is no replication, the fit is exact and there is no error term (the epsilon term) in the model. In this no replication case, if one assumes that there are no three-factor interactions, then one can use these 8 degrees of freedom for error estimation.

In this model we see that i = 1, 2, 3, and similarly for j and k, making 27 treatments.

Table of treatments for the 3^3 design

These treatments may be displayed as follows:

TABLE 3.37 The 3^3 Design

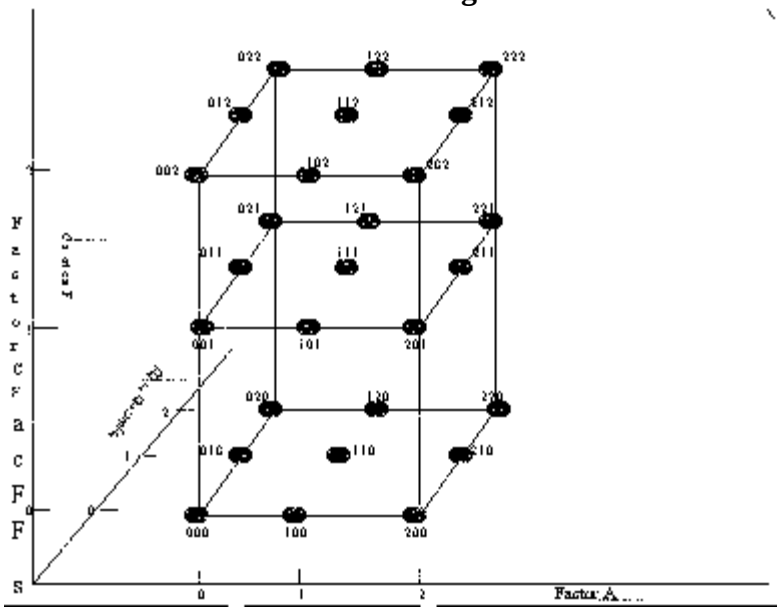
		Factor A		
Factor B	**Factor C**	**0**	**1**	**2**

0	0	000	100	200
0	1	001	101	201
0	2	002	102	202
1	0	010	110	210
1	1	011	111	211
1	2	012	112	212
2	0	020	120	220
2	1	021	121	221
2	2	022	122	222

Pictorial representation of the 3^3 design

The design can be represented pictorially by

FIGURE 3.24 A 3^3 Design Schematic



Two types of 3^k designs

Two types of fractions of 3^k designs are employed:

- Box-Behnken designs whose purpose is to estimate a second-order model for quantitative factors (discussed earlier in section 5.3.3.6.2)
- 3^{k-p} orthogonal arrays.

5.3.3.10. Three-level, mixed-level and fractional factorial designs

Mixed level designs have some factors with, say, 2 levels, and some with 3 levels or 4 levels

The 2^k and 3^k experiments are special cases of factorial designs. In a factorial design, one obtains data at every combination of the levels. The importance of factorial designs, especially 2-level factorial designs, was stated by Montgomery (1991): *It is our belief that the two-level factorial and fractional factorial designs should be the cornerstone of industrial experimentation for product and process development and improvement.* He went on to say: *There are, however, some situations in which it is necessary to include a factor (or a few factors) that have more than two levels.*

This section will look at how to add three-level factors starting with two-level designs, obtaining what is called a *mixed-level* design. We will also look at how to add a four-level factor to a two-level design. The section will conclude with a listing of some useful orthogonal three-level and mixed-level designs (a few of the so-called Taguchi "L" orthogonal array designs), and a brief discussion of their benefits and disadvantages.

Generating a Mixed Three-Level and Two-Level Design

Montgomery scheme for generating a mixed design

Montgomery (1991) suggests how to derive a variable at three levels from a 2^3 design, using a rather ingenious scheme. The objective is to generate a design for one variable, *A*, at 2 levels and another, *X*, at three levels. This will be formed by combining the -1 and 1 patterns for the *B* and *C* factors to form the levels of the three-level factor *X*:

TABLE 3.38 Generating a Mixed Design

Two-Level		Three-Level
B	**C**	**X**
-1	-1	x_1
+1	-1	x_2
-1	+1	x_2
+1	+1	x_3

Similar to the 3^k case, we observe that *X* has 2 degrees of freedom, which can be broken out into a linear and a quadratic component. To illustrate how the 2^3 design leads to the design with one factor at two levels and one factor at three levels, consider the following table, with particular attention focused on the column labels.

Table illustrating the generation of a design with one factor at 2 levels and another at 3 levels from a 2^3 design

	A	**X_L**	**X_L**	**AX_L**	**AX_L**	**X_Q**	**AX_Q**	**TRT**	**MNT**
Run	**A**	**B**	**C**	**AB**	**AC**	**BC**	**ABC**	**A**	**X**
1	-1	-1	-1	+1	+1	+1	-1	Low	Low
2	+1	-1	-1	-1	-1	+1	+1	High	Low
3	-1	+1	-1	-1	+1	-1	+1	Low	Medium
4	+1	+1	-1	+1	-1	-1	-1	High	Medium
5	-1	-1	+1	+1	-1	-1	+1	Low	Medium
6	+1	-1	+1	-1	+1	-1	-1	High	Medium
7	-1	+1	+1	-1	-1	+1	-1	Low	High
8	+1	+1	+1	+1	+1	+1	+1	High	High

If quadratic effect negligble, we may include a second two-level factor

If we believe that the quadratic effect is negligible, we may include a second two-level factor, D, with D = ABC. In fact, we can convert the design to exclusively a main effect (resolution III) situation consisting of four two-level factors and one three-level factor. This is accomplished by equating the second two-level factor to AB, the third to AC and the fourth to ABC. Column BC cannot be used in this manner because it contains the quadratic effect of the three-level factor X.

More than one three-level factor

3-Level factors from 2^4 and 2^5 designs

We have seen that in order to create one three-level factor, the starting design can be a 2^3 factorial. Without proof we state that a 2^4 can split off 1, 2 or 3 three-level factors; a 2^5 is able to generate 3 three-level factors and still maintain a full factorial structure. For more on this, see Montgomery (1991).

Generating a Two- and Four-Level Mixed Design

Constructing a design with one 4-level factor and two 2-level factors

We may use the same principles as for the three-level factor example in creating a four-level factor. We will assume that the goal is to construct a design with one four-level and two two-level factors.

Initially we wish to estimate all main effects and interactions. It has been shown (see Montgomery, 1991) that this can be accomplished via a 2^4 (16 runs) design, with columns A and B used to create the four level factor *X*.

Table showing

TABLE 3.39 A Single Four-level Factor and Two Two-level Factors in 16 runs

design with 4-level, two 2-level factors in 16 runs

Run	(A	B)	= X	C	D
1	-1	-1	x_1	-1	-1
2	+1	-1	x_2	-1	-1
3	-1	+1	x_3	-1	-1
4	+1	+1	x_4	-1	-1
5	-1	-1	x_1	+1	-1
6	+1	-1	x_2	+1	-1
7	-1	+1	x_3	+1	-1
8	+1	+1	x_4	+1	-1
9	-1	-1	x_1	-1	+1
10	+1	-1	x_2	-1	+1
11	-1	+1	x_3	-1	+1
12	+1	+1	x_4	-1	+1
13	-1	-1	x_1	+1	+1
14	+1	-1	x_2	+1	+1
15	-1	+1	x_3	+1	+1
16	+1	+1	x_4	+1	+1

Some Useful (Taguchi) Orthogonal "L" Array Designs

L_9 design

L_9 - A 3^{4-2} Fractional Factorial Design 4 Factors at Three Levels (9 runs)

Run	X1	X2	X3	X4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

L_{18} design

L_{18} - A 2 x 3^{7-5} Fractional Factorial (Mixed-Level) Design 1 Factor at Two Levels and Seven Factors at 3 Levels (18 Runs)

Run	X1	X2	X3	X4	X5	X6	X7	X8
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	1	1	2	2	3	3
5	1	2	2	2	3	3	1	1
6	1	2	3	3	1	1	2	2

7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	2	1	2	3	1	3	2
14	2	2	2	3	1	2	1	3
15	2	2	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

L_{27} design

L_{27} - A 3^{13-10} Fractional Factorial Design
Thirteen Factors at Three Levels (27 Runs)

Run	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2	2	2	2	2	2
3	1	1	1	1	3	3	3	3	3	3	3	3	3
4	1	2	2	2	1	1	1	2	2	2	3	3	3
5	1	2	2	2	2	2	2	3	3	3	1	1	1
6	1	2	2	2	3	3	3	1	1	1	2	2	2
7	1	3	3	3	1	1	1	3	3	3	2	2	2
8	1	3	3	3	2	2	2	1	1	1	3	3	3
9	1	3	3	3	3	3	3	2	2	2	1	1	1
10	2	1	2	3	1	2	3	1	2	3	1	2	3
11	2	1	2	3	2	3	1	2	3	1	2	3	1
12	2	1	2	3	3	1	2	3	1	2	3	1	2
13	2	2	3	1	1	2	3	2	3	1	3	1	2
14	2	2	3	1	2	3	1	3	1	2	1	2	3
15	2	2	3	1	3	1	2	1	2	3	2	3	1
16	2	3	1	2	1	2	3	3	1	2	2	3	1
17	2	3	1	2	2	3	1	1	2	3	3	1	2
18	2	3	1	2	3	1	2	2	3	1	1	2	3
19	3	1	3	2	1	3	2	1	3	2	1	3	2
20	3	1	3	2	2	1	3	2	1	3	2	1	3
21	3	1	3	2	3	2	1	3	2	1	3	2	1
22	3	2	1	3	1	3	2	2	1	3	3	2	1
23	3	2	1	3	2	1	3	3	2	1	1	3	2
24	3	2	1	3	3	2	1	1	3	2	2	1	3
25	3	3	2	1	1	3	2	3	2	1	2	1	3
26	3	3	2	1	2	1	3	1	3	2	3	2	1
27	3	3	2	1	3	2	1	2	1	3	1	3	2

L_{36} design

L36 - A Fractional Factorial (Mixed-Level) Design Eleven Factors at Two Levels and Twelve Factors at 3 Levels (36 Runs)

Run	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2
3	1	1	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	3	3	3	3	3
4	1	1	1	1	1	2	2	2	2	2	2	1	1	1	1	2	2	2	2	3	3	3	3
5	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	1	1	1	1
6	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	1	1	1	1	2	2	2	2
7	1	1	2	2	2	1	1	1	2	2	2	1	1	2	3	1	2	3	3	1	2	2	3
8	1	1	2	2	2	1	1	1	2	2	2	2	2	3	1	2	3	1	1	2	3	3	1
9	1	1	2	2	2	1	1	1	2	2	2	3	3	1	2	3	1	2	2	3	1	1	2
10	1	2	1	2	2	1	2	2	1	1	2	1	1	3	2	1	3	2	3	2	1	3	2
11	1	2	1	2	2	1	2	2	1	1	2	2	2	1	3	2	1	3	1	3	2	1	3
12	1	2	1	2	2	1	2	2	1	1	2	3	3	2	1	3	2	1	2	1	3	2	1
13	1	2	2	1	2	2	1	2	1	2	1	1	2	3	1	3	2	1	3	3	2	1	2
14	1	2	2	1	2	2	1	2	1	2	1	2	3	1	2	1	3	2	1	1	3	2	3
15	1	2	2	1	2	2	1	2	1	2	1	3	1	2	3	2	1	3	2	2	1	3	1
16	1	2	2	2	1	2	2	1	2	1	1	1	2	3	2	1	1	3	2	3	3	2	1
17	1	2	2	2	1	2	2	1	2	1	1	2	3	1	3	2	2	1	3	1	1	3	2
18	1	2	2	2	1	2	2	1	2	1	1	3	1	2	1	3	3	2	1	2	2	1	3
19	2	1	2	2	1	1	2	2	1	2	1	1	2	1	3	3	3	1	2	2	1	2	3
20	2	1	2	2	1	1	2	2	1	2	1	2	3	2	1	1	1	2	3	3	2	3	1
21	2	1	2	2	1	1	2	2	1	2	1	3	1	3	2	2	2	3	1	1	3	1	2
22	2	1	2	1	2	2	2	1	1	1	2	1	2	2	3	3	1	2	1	1	3	3	2
23	2	1	2	1	2	2	2	1	1	1	2	2	3	3	1	1	2	3	2	2	1	1	3
24	2	1	2	1	2	2	2	1	1	1	2	3	1	1	2	2	3	1	3	3	2	2	1
25	2	1	1	2	2	2	1	2	2	1	1	1	3	2	1	2	3	3	1	3	1	2	2
26	2	1	1	2	2	2	1	2	2	1	1	2	1	3	2	3	1	1	2	1	2	3	3
27	2	1	1	2	2	2	1	2	2	1	1	3	2	1	3	1	2	2	3	2	3	1	1
28	2	2	2	1	1	1	1	2	2	1	2	1	3	2	2	2	1	1	3	2	3	1	3
29	2	2	2	1	1	1	1	2	2	1	2	2	1	3	3	3	2	2	1	3	1	2	1
30	2	2	2	1	1	1	1	2	2	1	2	3	2	1	1	1	3	3	2	1	2	3	2
31	2	2	1	2	1	2	1	1	1	2	2	1	3	3	3	2	3	2	2	1	2	1	1
32	2	2	1	2	1	2	1	1	1	2	2	2	1	1	1	3	1	3	3	2	3	2	2
33	2	2	1	2	1	2	1	1	1	2	2	3	2	2	2	1	2	1	1	3	1	3	3
34	2	2	1	1	2	1	2	1	2	2	1	1	3	1	2	3	2	3	1	2	2	3	1
35	2	2	1	1	2	1	2	1	2	2	1	2	1	2	3	1	3	1	2	3	3	1	2
36	2	2	1	1	2	1	2	1	2	2	1	3	2	3	1	2	1	2	3	1	1	2	3

Advantages and Disadvantages of Three-Level and Mixed-Level "L" Designs

Advantages The good features of these designs are:

and disadvantages of three-level mixed-level designs

- They are orthogonal arrays. Some analysts believe this simplifies the analysis and interpretation of results while other analysts believe it does not.
- They obtain a lot of information about the main effects in a relatively few number of runs.
- You can test whether non-linear terms are needed in the model, at least as far as the three-level factors are concerned.

On the other hand, there are several undesirable features of these designs to consider:

- They provide limited information about interactions.
- They require more runs than a comparable 2^{k-p} design, and a two-level design will often suffice when the factors are continuous and monotonic (many three-level designs are used when two-level designs would have been adequate).

5.4. Analysis of DOE data

Contents of this section

Assuming you have a starting model that you want to fit to your experimental data and the experiment was designed correctly for your objective, most DOE software packages will analyze your DOE data. This section will illustrate how to analyze DOE's by first going over the generic basic steps and then showing software examples. The contents of the section are:

- DOE analysis steps
- Plotting DOE data
- Modeling DOE data
- Testing and revising DOE models
- Interpreting DOE results
- Confirming DOE results
- DOE examples
  - Full factorial example
  - Fractional factorial example
  - Response surface example

Prerequisite statistical tools and concepts needed for DOE analyses

The examples in this section assume the reader is familiar with the concepts of

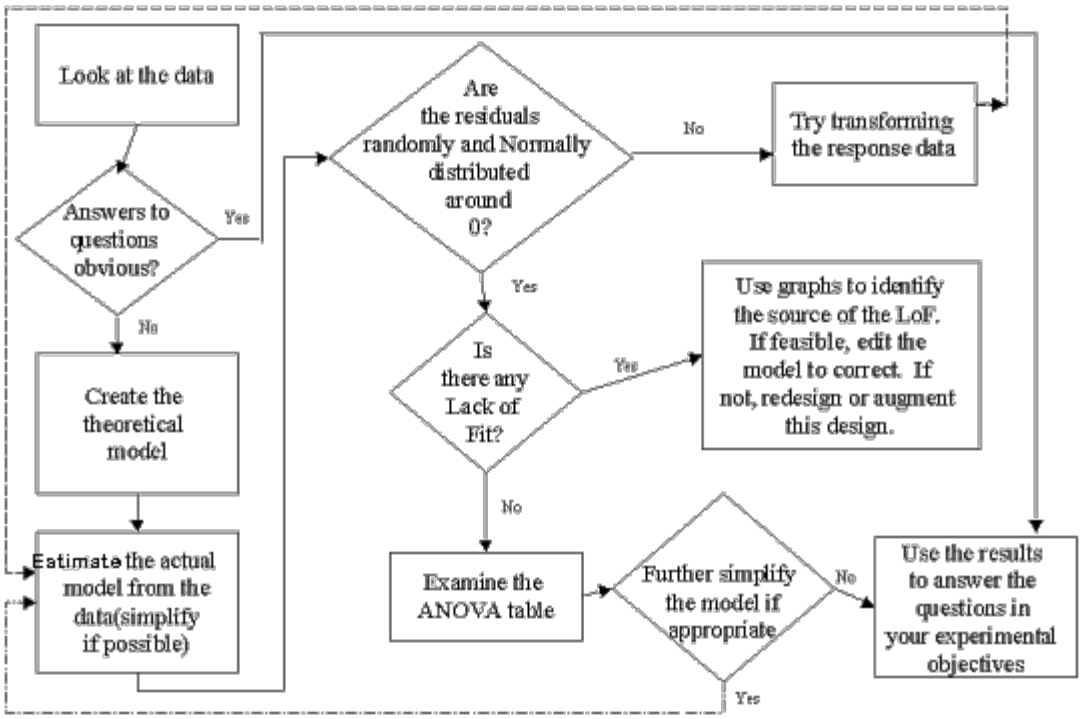
- ANOVA tables (see Chapter 3 or Chapter 7)
- p-values
- Residual analysis
- Model Lack of Fit tests
- Data transformations for normality and linearity

5. Process Improvement
5.4. Analysis of DOE data

5.4.1. What are the steps in a DOE analysis?

General flowchart for analyzing DOE data

Flowchart of DOE Analysis Steps



DOE Analysis Steps

Analysis steps: graphics, theoretical model, actual model, validate model, use model

The following are the basic steps in a DOE analysis.

1. Look at the data. Examine it for outliers, typos and obvious problems. Construct as many graphs as you can to get the big picture.
   - Response distributions (histograms, box plots, etc.)
   - Responses versus time order scatter plot (a check for possible time effects)
   - Responses versus factor levels (first look at magnitude of factor effects)
   - Typical DOE plots (which assume standard models for effects and errors)
     - Main effects mean plots
     - Block plots
     - Normal or half-normal plots of the effects
     - Interaction plots
   - Sometimes the right graphs and plots of the data lead to obvious answers for your experimental objective questions and you can skip to step 5. In most cases, however, you will want to continue by fitting and validating a model that can be used to answer your questions.

2. Create the theoretical model (the experiment should have been designed with this model in mind!).
3. Create a model from the data. Simplify the model, if possible, using stepwise regression methods and/or parameter p-value significance information.
4. Test the model assumptions using residual graphs.
    - If none of the model assumptions were violated, examine the ANOVA.
        - Simplify the model further, if appropriate. If reduction is appropriate, then return to step 3 with a new model.
    - If model assumptions were violated, try to find a cause.
        - Are necessary terms missing from the model?
        - Will a transformation of the response help? If a transformation is used, return to step 3 with a new model.
5. Use the results to answer the questions in your experimental objectives -- finding important factors, finding optimum settings, etc.

Flowchart is a guideline, not a hard-and-fast rule

Note: The above flowchart and sequence of steps should not be regarded as a "hard-and-fast rule" for analyzing all DOE's. Different analysts may prefer a different sequence of steps and not all types of experiments can be analyzed with one set procedure. There still remains some *art* in both the design and the analysis of experiments, which can only be learned from experience. In addition, the role of engineering judgment should not be underestimated.

5.4.2. How to "look" at DOE data

The importance of looking at the data with a wide array of plots or visual displays cannot be over-stressed

The right graphs, plots or visual displays of a dataset can uncover anomalies or provide insights that go beyond what most quantitative techniques are capable of discovering. Indeed, in many cases quantitative techniques and models are tools used to confirm and extend the conclusions an analyst has already formulated after carefully "looking" at the data.

Most software packages have a selection of different kinds of plots for displaying DOE data. Some of these useful ways of looking at data are mentioned below, with links to detailed explanations in Chapter 1 (Exploratory Data Analysis or EDA) or to other places where they are illustrated and explained. In addition, examples and detailed explanations of visual (EDA) DOE techniques can be found in section 5.5.9.

Plots for viewing the response data

First "Look" at the Data

- Histogram of responses
- Run-sequence plot (pay special attention to results at center points)
- Scatter plot (for pairs of response variables)
- Lag plot
- Normal probability plot
- Autocorrelation plot

Plots for viewing main effects and 2-factor interactions, explanation of normal or half-normal plots to detect possible important effects

Subsequent Plots: Main Effects, Comparisons and 2-Way Interactions

- Quantile-quantile (q-q) plot
- Block plot
- Box plot
- Bi-histogram
- DOE scatter plot
- DOE mean plot
- DOE standard deviation plot
- DOE interaction plots
- Normal or half-normal probability plots for effects. **Note**: these links show how to generate plots to test for normal (or half-normal) data with points lining up along a straight line, approximately, if the plotted points were from the assumed normal (or half-normal) distribution. For two-level full factorial and fractional

factorial experiments, the points plotted are the estimates of all the model effects, including possible interactions. Those effects that are really negligible should have estimates that resemble normally distributed noise, with mean zero and a constant variance. Significant effects can be picked out as the ones that do not line up along the straight line. Normal effect plots use the effect estimates directly, while half-normal plots use the absolute values of the effect estimates.

- Youden plots

Plots for testing and validating models

Model testing and Validation

- Response vs predictions
- Residuals vs predictions
- Residuals vs independent variables
- Residuals lag plot
- Residuals histogram
- Normal probability plot of residuals

Plots for model prediction

Model Predictions

- Contour plots

5.4.3. How to model DOE data

DOE models should be consistent with the goal of the experiment	In general, the trial model that will be fit to DOE data should be consistent with the goal of the experiment and has been predetermined by the goal of the experiment and the experimental design and data collection methodology.
Comparative designs	Models were given earlier for comparative designs (completely randomized designs, randomized block designs and Latin square designs).
Full factorial designs	For full factorial designs with k factors (2^k runs, not counting any center points or replication runs), the full model contains all the main effects and all orders of interaction terms. Usually, higher-order (three or more factors) interaction terms are included initially to construct the normal (or half-normal) plot of effects, but later dropped when a simpler, adequate model is fit. Depending on the software available or the analyst's preferences, various techniques such as normal or half-normal plots, Youden plots, p-value comparisons and stepwise regression routines are used to reduce the model to the minimum number of needed terms. An example of model selection is shown later in this section and an example of Yates algorithm is given as a case study.
Fractional factorial designs	For fractional factorial screening designs, it is necessary to know the alias structure in order to write an appropriate starting model containing only the interaction terms the experiment was designed to estimate (assuming all terms confounded with these selected terms are insignificant). This is illustrated by the fractional factorial example later in this section. The starting model is then possibly reduced by the same techniques described above for full factorial models.
Response surface designs	Response surface initial models include quadratic terms and may occasionally also include cubic terms. These models were described in section 3.
Model validation	Of course, as in all cases of model fitting, residual analysis and other tests of model fit are used to confirm or adjust

models, as needed.

5.4.4. How to test and revise DOE models

Tools for testing, revising, and selecting models

All the tools and procedures for testing, revising and selecting final DOE models are covered in various sections of the Handbook. The outline below gives many of the most common and useful techniques and has links to detailed explanations.

Outline of Model Testing and Revising: Tools and Procedures

An outline (with links) covers most of the useful tools and procedures for testing and revising DOE models

- Graphical Indicators for testing models (using residuals)
  - Response vs predictions
  - Residuals vs predictions
  - Residuals vs independent variables
  - Residuals lag plot
  - Residuals histogram
  - Normal probability plot of residuals
- Overall numerical indicators for testing models and model terms
  - R Squared and R Squared adjusted
  - Model Lack of Fit tests
  - ANOVA tables (see Chapter 3 or Chapter 7)
  - *p*-values
- Model selection tools or procedures
  - ANOVA tables (see Chapter 3 or Chapter 7)
  - *p*-values
  - Residual analysis
  - Model Lack of Fit tests
  - Data transformations for normality and linearity
  - Stepwise regression procedures
  - Normal or half-normal plots of effects (primarily for two-level full and fractional factorial experiments)
  - Youden plots
  - Other methods

5.4.5. How to interpret DOE results

Final model used to make conclusions and decisions

Assume that you have a final model that has passed all the relevant tests (visual and quantitative) and you are ready to make conclusions and decisions. These should be responses to the questions or outputs dictated by the original experimental goals.

Checklist relating DOE conclusions or outputs to experimental goals or experimental purpose:

A checklist of how to compare DOE results to the experimental goals

- Do the responses differ significantly over the factor levels? (comparative experiment goal)
- Which are the significant effects or terms in the final model? (screening experiment goal)
- What is the model for estimating responses?
    - Full factorial case (main effects plus significant interactions)
    - Fractional factorial case (main effects plus significant interactions that are not confounded with other possibly real effects)
    - RSM case (allowing for quadratic or possibly cubic models, if needed)
- What responses are predicted and how can responses be optimized? (RSM goal)
    - Contour plots
    - Settings for confirmation runs and prediction intervals for results

5.4.6. How to confirm DOE results (confirmatory runs)

Definition of confirmation runs

When the analysis of the experiment is complete, one must verify that the predictions are good. These are called confirmation runs.

The interpretation and conclusions from an experiment may include a "best" setting to use to meet the goals of the experiment. Even if this "best" setting were included in the design, you should run it again as part of the confirmation runs to make sure nothing has changed and that the response values are close to their predicted values. would get.

At least 3 confirmation runs should be planned

In an industrial setting, it is very desirable to have a stable process. Therefore, one should run more than one test at the "best" settings. A minimum of 3 runs should be conducted (allowing an estimate of variability at that setting).

If the time between actually running the experiment and conducting the confirmation runs is more than a few hours, the experimenter must be careful to ensure that nothing else has changed since the original data collection.

Carefully duplicate the original environment

The confirmation runs should be conducted in an environment as similar as possible to the original experiment. For example, if the experiment were conducted in the afternoon and the equipment has a warm-up effect, the confirmation runs should be conducted in the afternoon after the equipment has warmed up. Other extraneous factors that may change or affect the results of the confirmation runs are: person/operator on the equipment, temperature, humidity, machine parameters, raw materials, etc.

Checks for when confirmation runs give surprises

What do you do if you don't obtain the results you expected? If the confirmation runs don't produce the results you expected:

1. check to see that nothing has changed since the original data collection
2. verify that you have the correct settings for the confirmation runs
3. revisit the model to verify the "best" settings from the analysis

4. verify that you had the correct predicted value for the confirmation runs.

If you don't find the answer after checking the above 4 items, the model may not predict very well in the region you decided was "best". You still learned from the experiment and you should use the information gained from this experiment to design another follow-up experiment.

Even when the experimental goals are not met, something was learned that can be used in a follow-up experiment

Every well-designed experiment is a success in that you learn something from it. However, every experiment will not necessarily meet the goals established before experimentation. That is why it makes sense to plan to experiment sequentially in order to meet the goals.

5.4.7. Examples of Designed Experiments

Three detailed examples

Perhaps one of the best ways to illustrate how to analyze data from a designed experiment is to work through a detailed example, explaining each step in the analysis.

Detailed analyses are presented for three basic types of designed experiments:

1. A full factorial experiment
2. A fractional factorial experiment
3. A response surface experiment

Software

Most analyses of designed experiments are performed by statistical software packages. Good statistical software enables the analyst to view graphical displays, build models, and test assumptions. Occasionally, the goals of the experiment can be achieved by simply examining appropriate graphical displays of the experimental responses. In other cases, a satisfactory model has to be fit in order to determine the most significant factors or the optimal contours of the response surface. In any case, the software will perform the appropriate calculations as long as the analyst knows what to request and how to interpret the program outputs.

5.4.7.1. Full factorial example

This example uses data from a NIST high performance ceramics experiment

Data Source

This data set was taken from an experiment that was performed a few years ago at NIST by Said Jahanmir of the Ceramics Division in the Material Science and Engineering Laboratory. The original analysis was performed primarily by Lisa Gill of the Statistical Engineering Division. The example shown here is an independent analysis of a modified portion of the original data set.

The original data set was part of a high performance ceramics experiment with the goal of characterizing the effect of grinding parameters on sintered reaction-bonded silicon nitride, reaction bonded silicone nitride, and sintered silicon nitride.

Only modified data from the first of the three ceramic types (sintered reaction-bonded silicon nitride) will be discussed in this illustrative example of a full factorial data analysis.

The reader can download the data as a text file.

Description of Experiment: Response and Factors

Response and factor variables

Purpose: To determine the effect of machining factors on ceramic strength
Response variable = mean (over 15 repetitions) of the ceramic strength
Number of observations = 32 (a complete 2^5 factorial design)

Response Variable Y = Mean (over 15 reps) of Ceramic Strength
Factor 1 = Table Speed (2 levels: slow (.025 m/s) and fast (.125 m/s))
Factor 2 = Down Feed Rate (2 levels: slow (.05 mm) and fast (.125 mm))
Factor 3 = Wheel Grit (2 levels: 140/170 and 80/100)
Factor 4 = Direction (2 levels: longitudinal and transverse)
Factor 5 = Batch (2 levels: 1 and 2)

Since two factors were qualitative (direction and batch) and it was reasonable to expect monotone effects from the quantitative factors, no centerpoint runs were included.

The data

The design matrix, with measured ceramic strength responses, appears below. The actual randomized run order is given in the last column. (The interested reader may download the data as a text file.)

	speed	rate	grit	direction	batch	strength	order
1	-1	-1	-1	-1	-1	680.45	17
2	1	-1	-1	-1	-1	722.48	30
3	-1	1	-1	-1	-1	702.14	14
4	1	1	-1	-1	-1	666.93	8
5	-1	-1	1	-1	-1	703.67	32
6	1	-1	1	-1	-1	642.14	20

7	-1	1	1	-1	-1	692.98	26
8	1	1	1	-1	-1	669.26	24
9	-1	-1	-1	1	-1	491.58	10
10	1	-1	-1	1	-1	475.52	16
11	-1	1	-1	1	-1	478.76	27
12	1	1	-1	1	-1	568.23	18
13	-1	-1	1	1	-1	444.72	3
14	1	-1	1	1	-1	410.37	19
15	-1	1	1	1	-1	428.51	31
16	1	1	1	1	-1	491.47	15
17	-1	-1	-1	-1	1	607.34	12
18	1	-1	-1	-1	1	620.80	1
19	-1	1	-1	-1	1	610.55	4
20	1	1	-1	-1	1	638.04	23
21	-1	-1	1	-1	1	585.19	2
22	1	-1	1	-1	1	586.17	28
23	-1	1	1	-1	1	601.67	11
24	1	1	1	-1	1	608.31	9
25	-1	-1	-1	1	1	442.90	25
26	1	-1	-1	1	1	434.41	21
27	-1	1	-1	1	1	417.66	6
28	1	1	-1	1	1	510.84	7
29	-1	-1	1	1	1	392.11	5
30	1	-1	1	1	1	343.22	13
31	-1	1	1	1	1	385.52	22
32	1	1	1	1	1	446.73	29

Analysis of the Experiment

Five basic steps

The experimental data will be analyzed following the previously described five basic steps. The analyses shown in this page can be generated using R code.

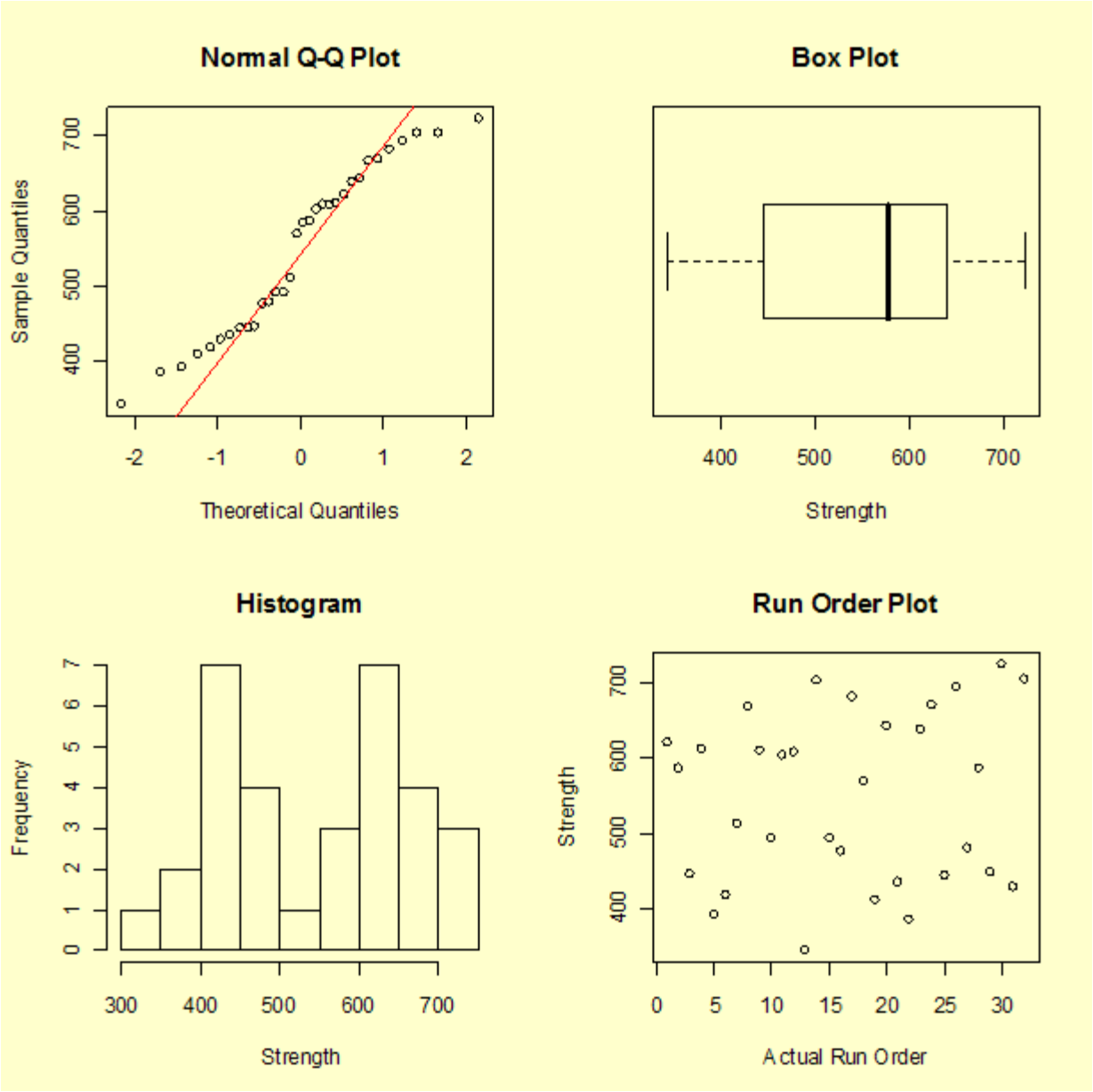
Step 1: Look at the data

Plot the response variable

We start by plotting the response data several ways to see if any trends or anomalies appear that would not be accounted for by the standard linear response models.

First, we look at the distribution of the response variable regardless of factor levels by generating the following four plots.

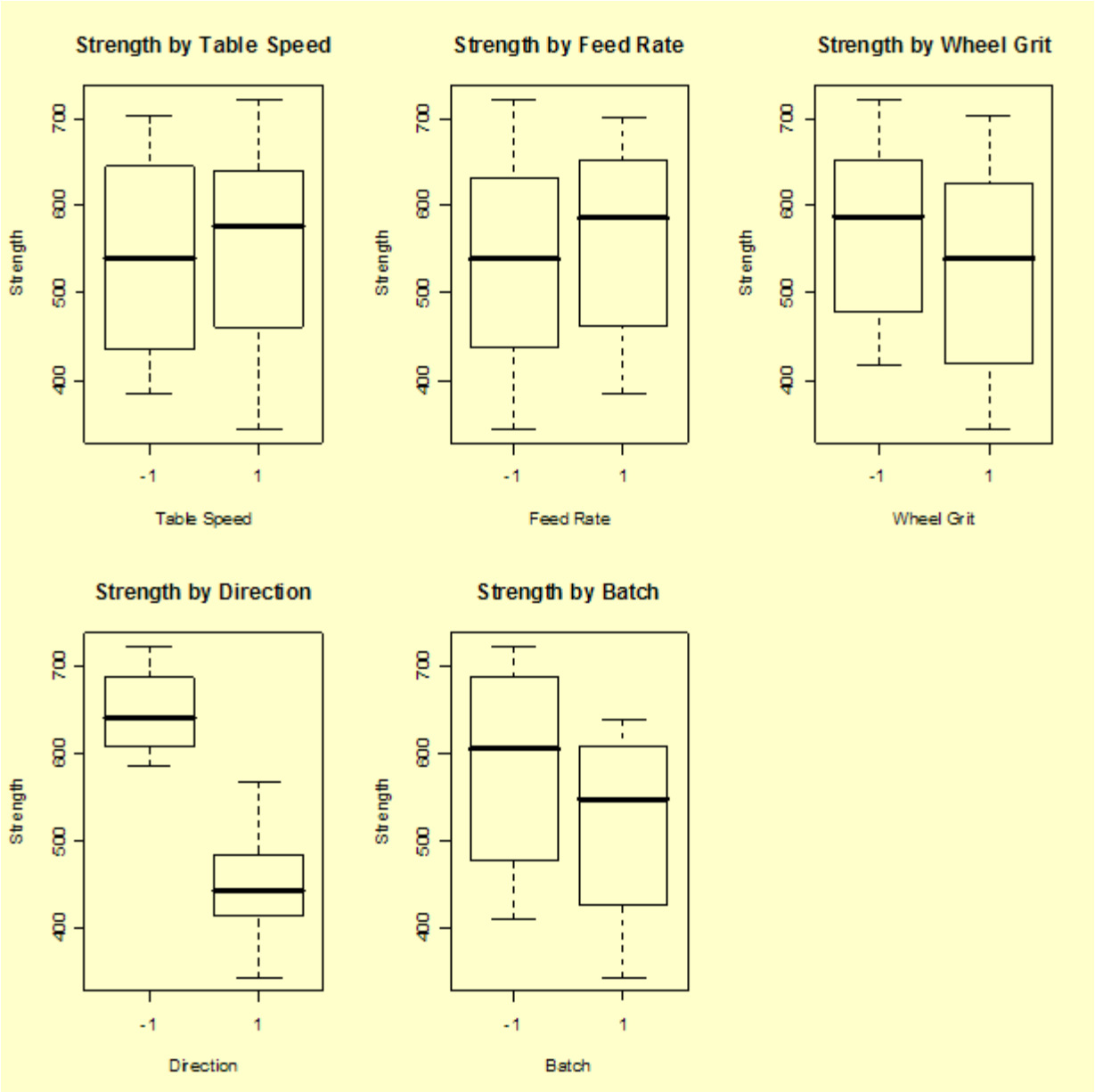
1. The first plot is a normal probability plot of the response variable. The red line is the theoretical normal distribution.
2. The second plot is a box plot of the response variable.
3. The third plot is a histogram of the response variable.
4. The fourth plot is the response versus the run order.



Clearly there is "structure" that we hope to account for when we fit a response model. For example, the response variable is separated into two roughly equal-sized clumps in the histogram. The first clump is centered approximately around the value 450 while the second clump is centered approximately around the value 650. As hoped for, the run-order plot does not indicate a significant time effect.

Box plots of response by factor variables

Next, we look at box plots of the response for each factor.



Several factors, most notably "Direction" followed by "Batch" and possibly "Wheel Grit", appear to change the average response level.

Step 2: Create the theoretical model

Theoretical model: assume all four-factor and higher interaction terms are not significant

For a 2^5 full factorial experiment we can fit a model containing a mean term, five main effect terms, ten two-factor interaction terms, ten three-factor interaction terms, five four-factor interaction terms, and a five-factor interaction term (32 parameters). However, we start by assuming all four-factor and higher interaction terms are non-existent. It's very rare for such high-order interactions to be significant, and they are very difficult to interpret from an engineering viewpoint. The assumption allows us to accumulate the sums of squares for these terms and use them to estimate an error term. We start with a theoretical model with 26 unknown constants, hoping the data will clarify which of these are the significant main effects and interactions we need for a final model.

Step 3: Fit model to the data

Results from

The ANOVA table for the 26-parameter model (intercept not shown) follows.

fitting up to and including third-order interaction terms

Summary of Fit

RSquare	0.995127
RSquare Adj	0.974821
Root Mean Square Error	17.81632
Mean of Response	546.8959
Observations	32

Source	DF	Sum of Squares	F Ratio	Prob>F
X1: Table Speed	1	894.33	2.8175	0.1442
X2: Feed Rate	1	3497.20	11.0175	0.0160
X1: Table Speed* X2: Feed Rate	1	4872.57	15.3505	0.0078
X3: Wheel Grit	1	12663.96	39.8964	0.0007
X1: Table Speed* X3: Wheel Grit	1	1838.76	5.7928	0.0528
X2: Feed Rate* X3: Wheel Grit	1	307.46	0.9686	0.3630
X1:Table Speed* X2: Feed Rate* X3: Wheel Grit	1	357.05	1.1248	0.3297
X4: Direction	1	315132.65	992.7901	<.0001
X1: Table Speed* X4: Direction	1	1637.21	5.1578	0.0636
X2: Feed Rate* X4: Direction	1	1972.71	6.2148	0.0470
X1: Table Speed X2: Feed Rate* X4: Direction	1	5895.62	18.5735	0.0050
X3: Wheel Grit* X4: Direction	1	3158.34	9.9500	0.0197
X1: Table Speed* X3: Wheel Grit* X4: Direction	1	2.12	0.0067	0.9376
X2: Feed Rate* X3: Wheel Grit* X4: Direction	1	44.49	0.1401	0.7210
X5: Batch	1	33653.91	106.0229	<.0001
X1: Table Speed* X5: Batch	1	465.05	1.4651	0.2716
X2: Feed Rate* X5: Batch	1	199.15	0.6274	0.4585
X1: Table Speed* X2: Feed Rate* X5: Batch	1	144.71	0.4559	0.5247
X3: Wheel Grit* X5: Batch	1	29.36	0.0925	0.7713
X1: Table Speed* X3: Wheel Grit* X5: Batch	1	30.36	0.0957	0.7676
X2: Feed Rate* X3: Wheel Grit* X5: Batch	1	25.58	0.0806	0.7860
X4: Direction * X5: Batch	1	1328.83	4.1863	0.0867
X1: Table Speed* X4: Directio* X5: Batch	1	544.58	1.7156	0.2382
X2: Feed Rate* X4: Direction* X5: Batch	1	167.31	0.5271	0.4952
X3: Wheel Grit* X4: Direction* X5: Batch	1	32.46	0.1023	0.7600

This fit has a large R^2 and adjusted R^2, but the high number of large (>0.10) *p*-values (in the "Prob>F" column) makes it clear that the model has many unnecessary terms.

Stepwise regression

Starting with the 26 terms, we use stepwise regression to eliminate unnecessary terms. By a combination of stepwise regression and the removal of remaining terms with a *p*-value larger than 0.05, we quickly arrive at a model with an intercept and 12 significant effect terms.

Results from fitting the 12-

Summary of Fit

RSquare	0.989114
RSquare Adj	0.982239

term model

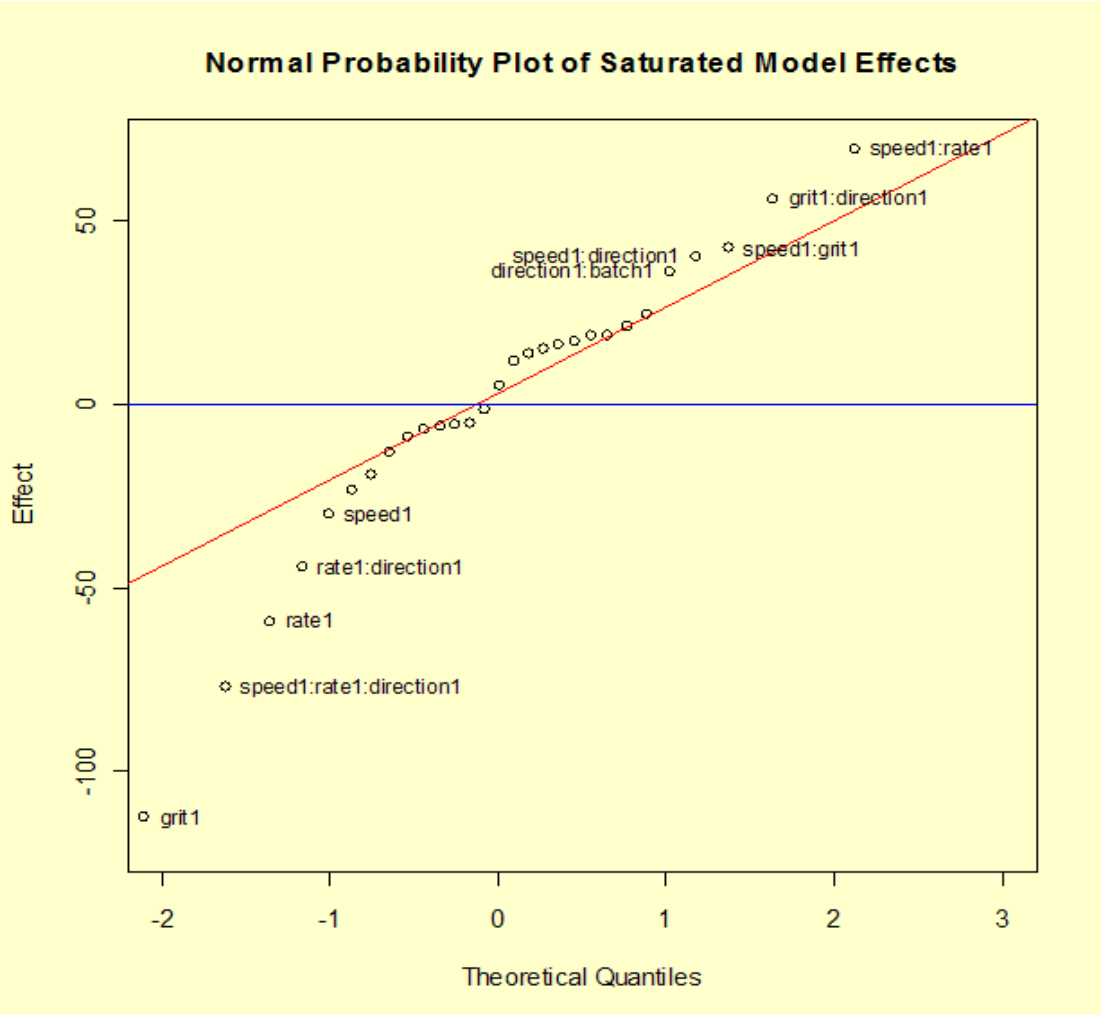
```
Root Mean Square Error 14.96346
 Mean of Response 546.8959
Observations (or Sum Wgts) 32
```

Source	DF	Sum of Squares	F Ratio	Prob>F
X1: Table Speed	1	894.33	3.9942	0.0602
X2: Feed Rate	1	3497.20	15.6191	0.0009
X1: Table Speed* X2: Feed Rate	1	4872.57	21.7618	0.0002
X3: Wheel Grit	1	12663.96	56.5595	<.0001
X1: Table Speed* X3: Wheel Grit	1	1838.76	8.2122	0.0099
X4: Direction	1	315132.65	1407.4390	<.0001
X1: Table Speed* X4: Direction	1	1637.21	7.3121	0.0141
X2: Feed Rate* X4: Direction	1	1972.71	8.8105	0.0079
X1: Table Speed* X2: Feed Rate* X4:Direction	1	5895.62	26.3309	<.0001
X3: Wheel Grit* X4: Direction	1	3158.34	14.1057	0.0013
X5: Batch	1	33653.91	150.3044	<.0001
X4: Direction* X5: Batch	1	1328.83	5.9348	0.0249

Normal plot of the effects

Non-significant effects should effectively follow an approximately normal distribution with the same location and scale. Significant effects will vary from this normal distribution. Therefore, another method of determining significant effects is to generate a normal probability plot of all 31 effects. The effects that deviate substantially from the straight line fit to the data are considered significant. Although this is a somewhat subjective criteria, it tends to work well in practice. It is helpful to use both the numerical output from the fit and graphical techniques such as the normal probability plot in deciding which terms to keep in the model.

A normal probability plot of the effects is shown below. (To reduce the scale of the y-axis, the largest two effects, X4: Direction and X5: Batch, are not shown on the plot. In addition, these two effects were not used to compute the normal reference line.) The effects we consider to be significant are labeled. In this case, we have arrived at the exact same 12 terms by looking at the normal probability plot as we did from the stepwise regression.



Most of the effects cluster close to the center (zero) line and follow the fitted normal model straight line. The effects that appear to be above or below the line by more than a small amount are the same effects identified using the stepwise routine, with the exception of $X1$. Some analysts prefer to include a main effect term when it has several significant interactions even if the main effect term itself does not appear to be significant.

Model appears to account for most of the variability

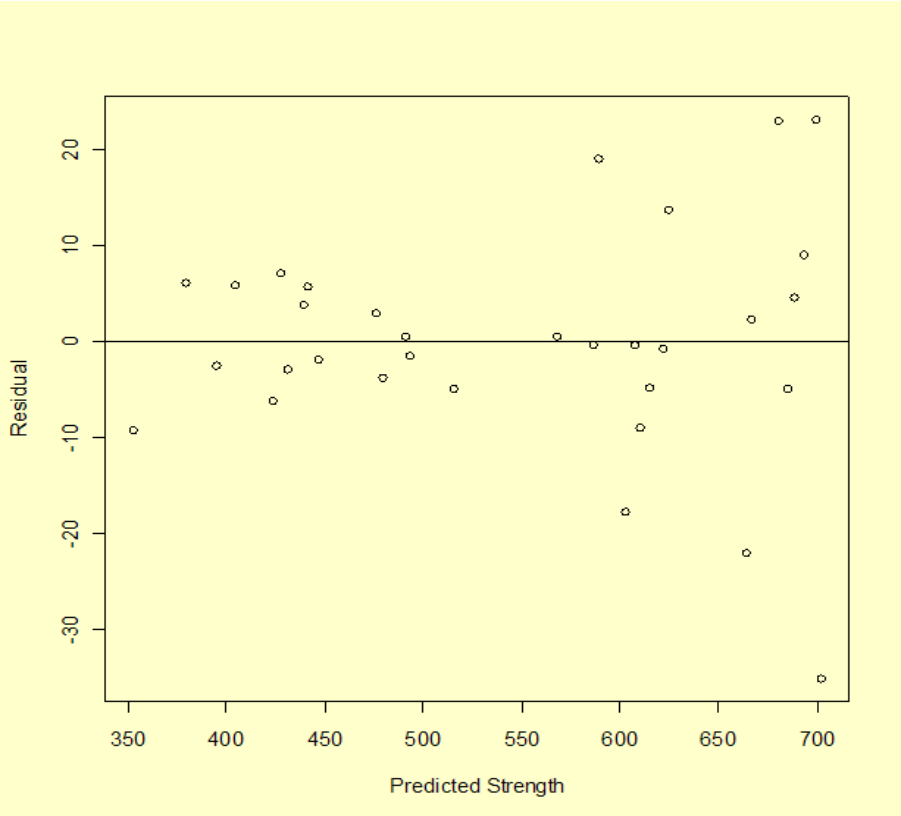
At this stage, the model appears to account for most of the variability in the response, achieving an adjusted R^2 of 0.982. All the main effects are significant, as are six 2-factor interactions and one 3-factor interaction. The only interaction that makes little physical sense is the " $X4$: Direction*$X5$: Batch" interaction - why would the response using one batch of material react differently when the batch is cut in a different direction as compared to another batch of the same formulation?

However, before accepting any model, residuals need to be examined.

Step 4: Test the model assumptions using residual graphs (adjust and simplify as needed)

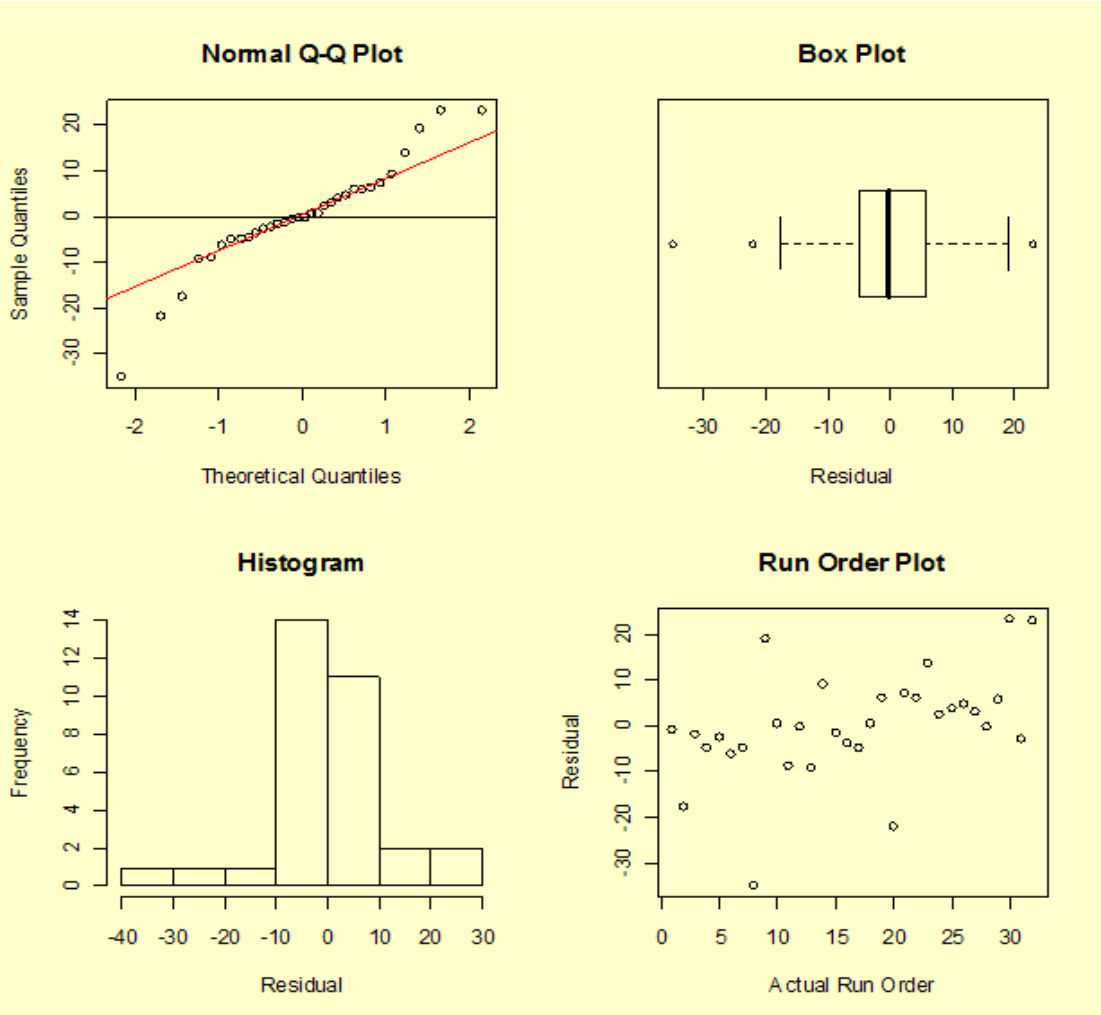
Plot of residuals versus predicted responses

First we look at the residuals plotted versus the predicted responses.



The residuals appear to spread out more with larger values of predicted strength, which should not happen when there is a common variance.

Next we examine the distribution of the residuals with a normal quantile plot, a box plot, a histogram, and a run-order plot.

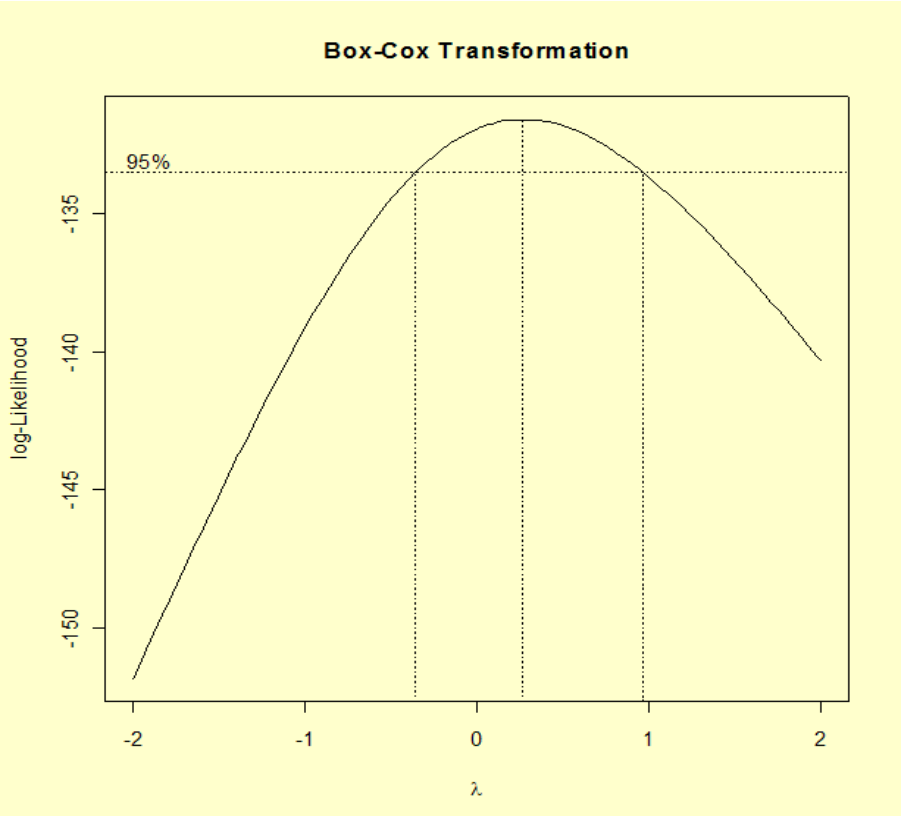


None of these plots appear to show typical normal residuals and the boxplot indicates that there may be outliers.

Step 4 continued: Transform the data and fit the model again

Box-Cox Transformation

We next look at whether we can model a transformation of the response variable and obtain residuals with the assumed properties. We calculate an optimum Box-Cox transformation by finding the value of λ that maximizes the negative log likelihood.



The optimum is found at λ = 0.2. A new Y: Strength variable is calculated using:

$$\frac{(y_i)^{\lambda} - 1}{\lambda \left[\left(\prod_{i-1}^{n} y_i \right)^{\frac{1}{n}} \right]^{\lambda - 1}}$$

Fit model to transformed data

When the 12-effect model is fit to the transformed data, the "X4: Direction * X5: Batch" interaction term is no longer significant. The 11-effect model fit is shown below, with parameter estimates and *p*-values.

The fitted model after applying Box-Cox transformation

The 11-Effect Model Fit to Tranformed Response Data

Response: Y:NewStrength

Summary of Fit

RSquare 0.99041
RSquare Adj 0.985135
Root Mean Square Error 13.81065
Mean of Response 1917.115
Observations (or Sum Wgts) 32

Effect	Parameter Estimate	p-value
Intercept	1917.115	<.0001
X1: Table Speed	5.777	0.0282
X2: Feed Rate	11.691	0.0001
X1: Table Speed* X2: Feed Rate	-14.467	<.0001
X3: Wheel Grit	-21.649	<.0001
X1: Table Speed* X3: Wheel Grit	7.339	0.007
X4: Direction	-99.272	<.0001
X1: Table Speed* X4: Direction	-7.188	0.0080
X2: Feed Rate*	-9.160	0.0013

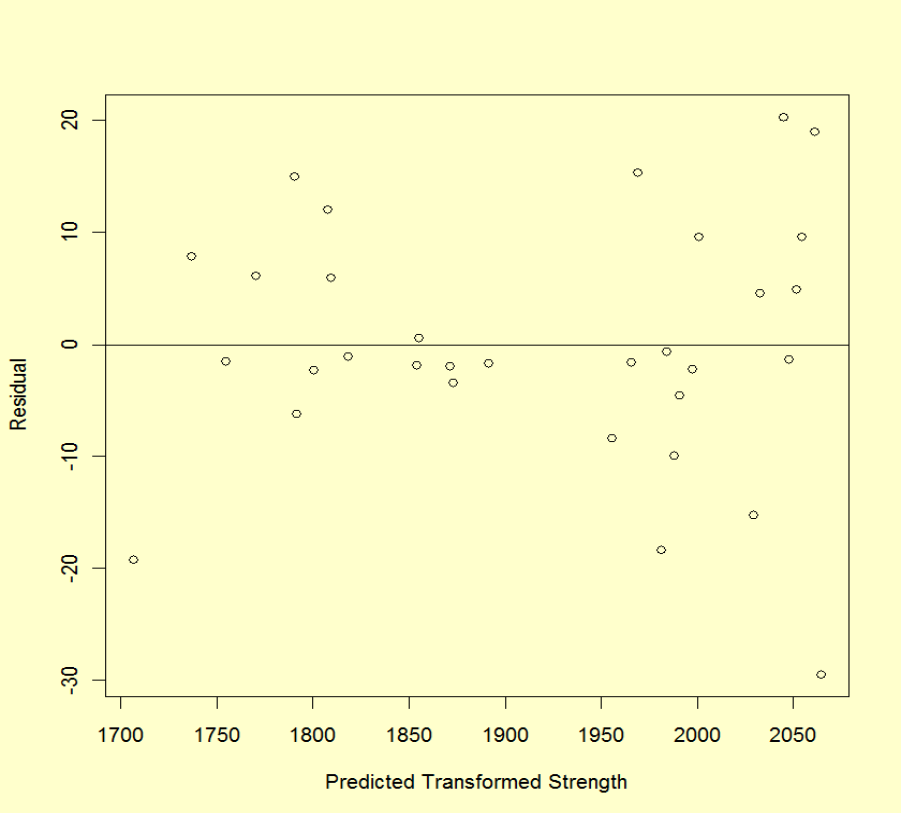
```
    X4: Direction
X1: Table Speed*          15.325          <.0001
    X2: Feed Rate*
    X4:Direction
X3: Wheel Grit*           12.965          <.0001
    X4: Direction
X5: Batch                -31.871          <.0001
```

Model has high R^2

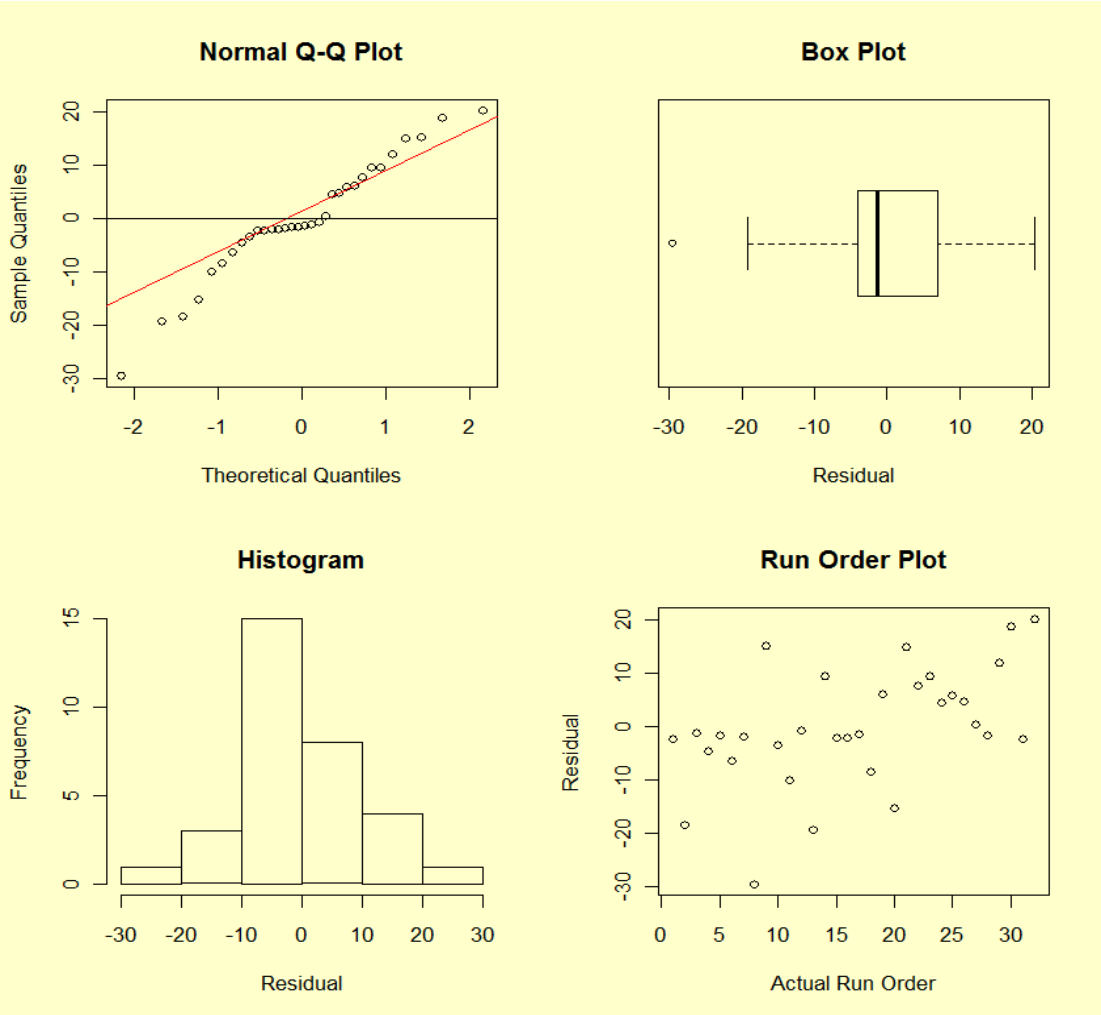
This model has a very large R^2 and adjusted R^2. The residual plots (shown below) are quite a bit better behaved than before

Residual plots from model with transformed response

The plot of the residuals versus the predicted values indicates that the transformation has resolved the problem of increasing variace with increasing strength.



The normal probability plot, box plot, and the histogram of the residuals do not indicate any serious violations of the model assumptions. The run sequence plot of the residuals does not indicate any time dependent patterns.



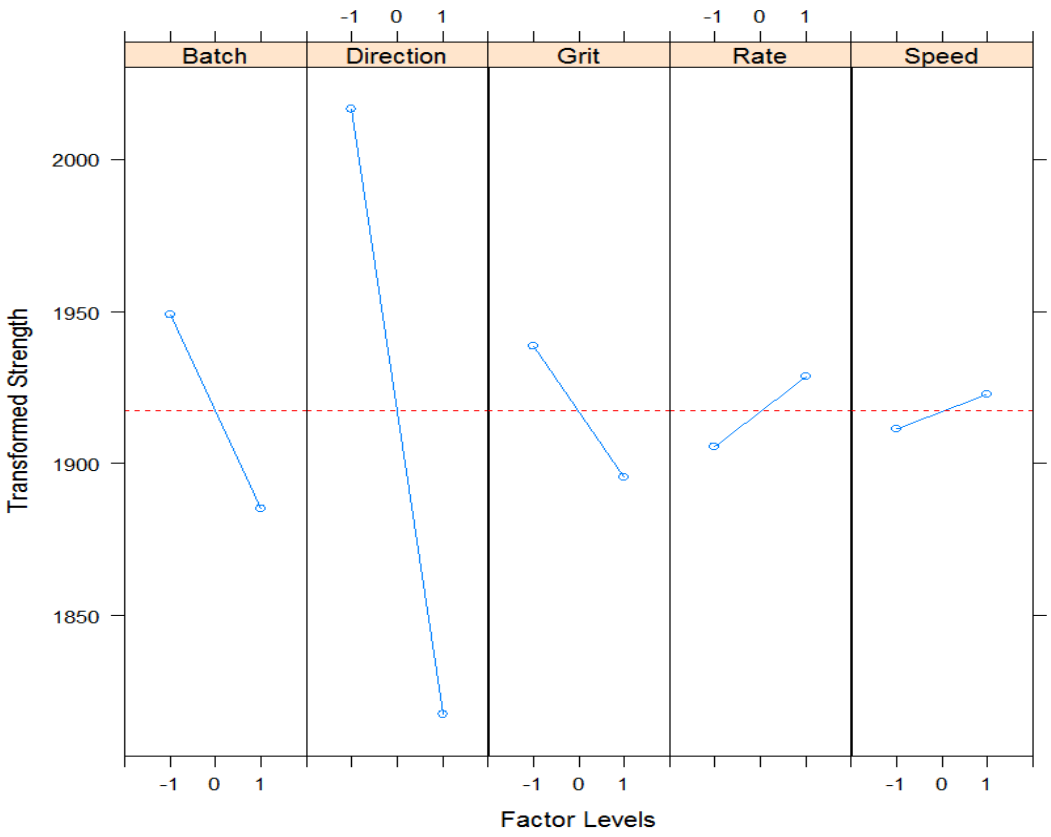
Step 5. Answer the questions in your experimental objectives

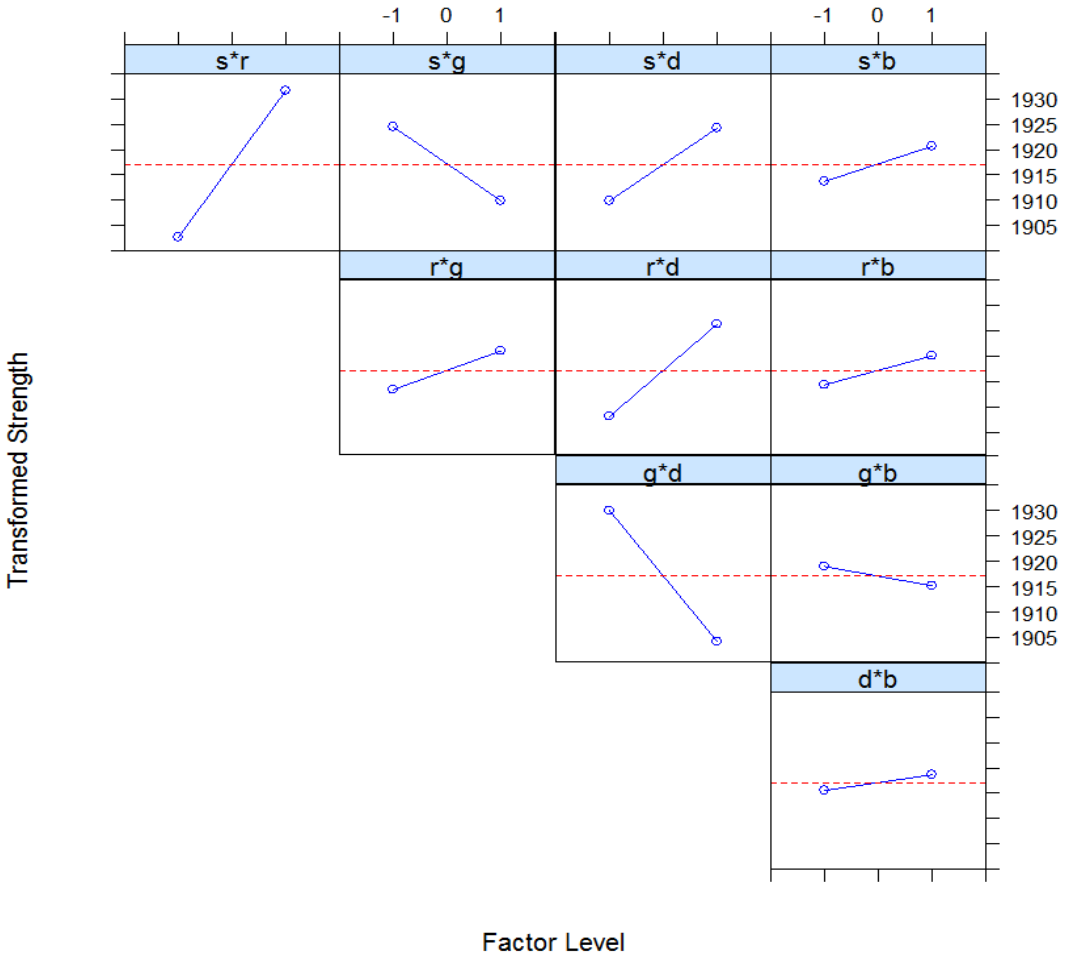
Important main effects and interaction effects

The magnitudes of the effect estimates show that "Direction" is by far the most important factor. "Batch" plays the next most critical role, followed by "Wheel Grit". Then, there are several important interactions followed by "Feed Rate". "Table Speed" plays a role in almost every significant interaction term, but is the least important main effect on its own. Note that large interactions can obscure main effects.

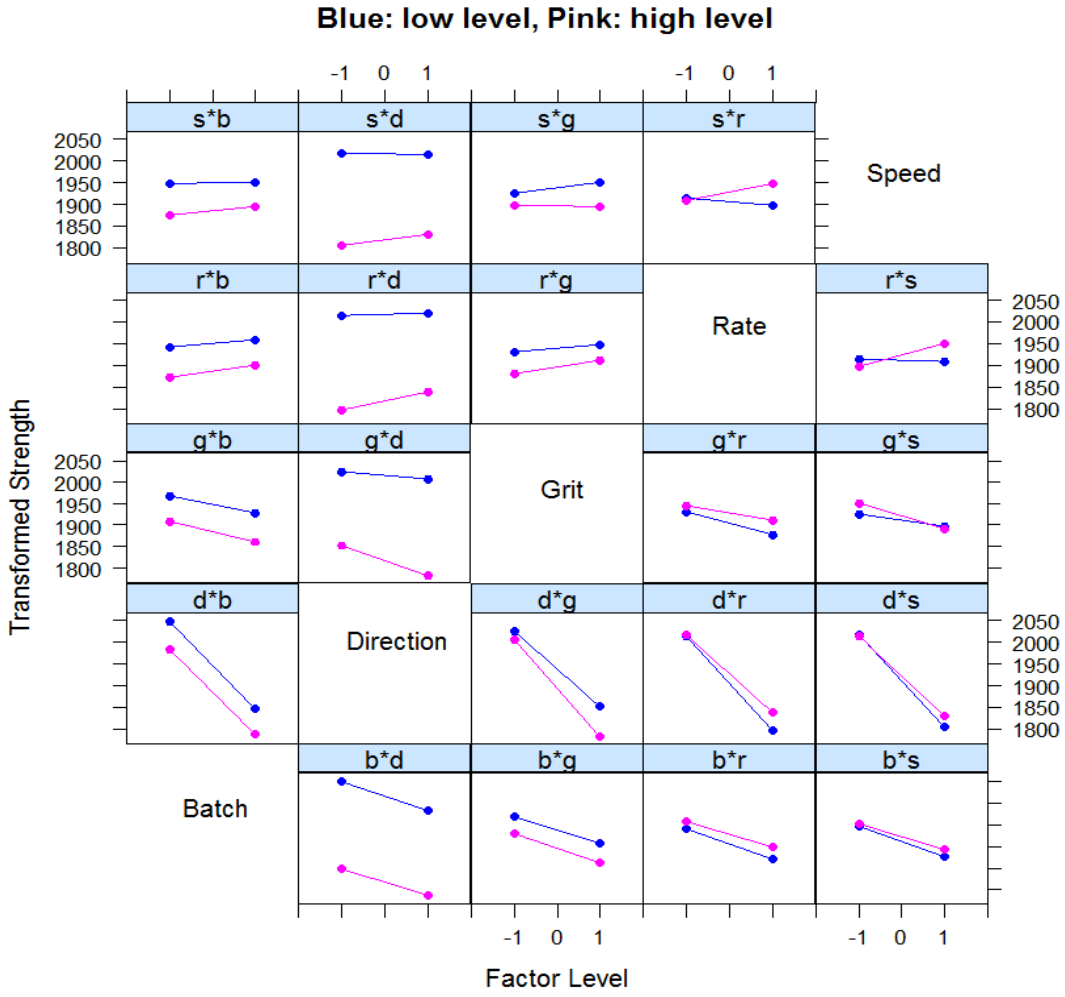
Plots of the main effects and significant 2-way interactions

Plots of the main effects and the significant 2-way interactions are shown below.


Batch
Direction
Grit
Rate
Speed
-1 0 1
2000
1950
1900
1850
Transformed Strength
Factor Levels



Next, we plot 2-way interaction plot showing means for all combinations of levels for the two factors.



The labels located in the diagonal spaces of the plot grid have two purposes. First, the label indicates the factor associated with the x-axis for all plots in the same row. Second, the label indicates the factor defining the two lines for plots in the same column.

For example, the plot labeled r*s contains averages for low and high levels of the rate variable (x-axis) for both levels of speed. The blue line represents the low level of speed and the pink line represents the high level of speed. The two lines in the r*s plot cross, indicating that there is interaction between rate and speed. Parallel lines indicate that there is no interaction between the two factors.

Optimal Settings

Based on the analyses, we can select factor settings that maximize ceramic strength. Translating from "-1" and "+1" back to the actual factor settings, we have: Table speed at "1" or .125m/s; Down Feed Rate at "1" or .125 mm; Wheel Grit at "-1" or 140/170; and Direction at "-1" or longitudinal.

Unfortunately, "Batch" is also a very significant factor, with the first batch giving higher strengths than the second. Unless it is possible to learn what worked well with this batch, and how to repeat it, not much can be done about this factor.

Comments

Analyses with value of Direction fixed

1. One might ask what an analysis of just the 2^4 factorial with "Direction" kept at -1 (i.e., longitudinal) would yield. This analysis turns out to have a very simple model; only "Wheel Grit" and "Batch" are significant main effects and no interactions are

indicates complex model is needed only for transverse cut

significant.

If, on the other hand, we do an analysis of the 2^4 factorial with "Direction" kept at +1 (i.e., transverse), then we obtain a 7-parameter model with all the main effects and interactions we saw in the 2^5 analysis, except, of course, any terms involving "Direction".

So it appears that the complex model of the full analysis came from the physical properties of a transverse cut, and these complexities are not present for longitudinal cuts.

Half fraction design

2. If we had assumed that three-factor and higher interactions were negligible before experimenting, a 2_V^{5-1} half fraction design might have been chosen. In hindsight, we would have obtained valid estimates for all main effects and two-factor interactions except for *X*3 and *X*5, which would have been aliased with *X*1**X*2**X*4 in that half fraction.

Natural log transformation

3. Finally, we note that many analysts might prefer to adopt a natural logarithm transformation (i.e., use ln *Y*) as the response instead of using a Box-Cox transformation with an exponent of 0.2. The natural logarithm transformation corresponds to an exponent of $\lambda = 0$ in the Box-Cox graph.

5.4.7.2. Fractional factorial example

A "Catapult" Fractional Factorial Experiment

A step-by-step analysis of a fractional factorial "catapult" experiment

This experiment was conducted by a team of students on a *catapult* – a table-top wooden device used to teach design of experiments and statistical process control. The catapult has several controllable factors and a response easily measured in a classroom setting. It has been used for over 10 years in hundreds of classes.

Catapult

Description of Experiment: Response and Factors

The experiment has five factors that might affect the distance the golf ball travels

Purpose: To determine the significant factors that affect the distance the ball is thrown by the catapult, and to determine the settings required to reach three different distances (30, 60 and 90 inches).

Response Variable: The distance in inches from the front of the catapult to the spot where the ball lands. The ball is a plastic golf ball.

Number of observations: 20 (a 2^{5-1} resolution V design with 4 center points).

Variables:

1. Response Variable Y = distance
2. Factor 1 = band height (height of the pivot point for the rubber bands – levels were 2.25 and 4.75 inches with a centerpoint level of 3.5)
3. Factor 2 = start angle (location of the arm when the operator releases– starts the forward motion of the arm – levels were 0 and 20 degrees with a centerpoint level of 10 degrees)
4. Factor 3 = rubber bands (number of rubber bands used on the catapult– levels were 1 and 2 bands)
5. Factor 4 = arm length (distance the arm is extended – levels were 0 and 4 inches with a centerpoint level of 2 inches)

6. Factor 5 = stop angle (location of the arm where the forward motion of the arm is stopped and the ball starts flying – levels were 45 and 80 degrees with a centerpoint level of 62 degrees)

Design matrix and responses (in run order)

The design matrix appears below in (randomized) run order.

distance	height	start	bands	length	stop	order
28.00	3.25	0	1	0	80	1
99.00	4	10	2	2	62	2
126.50	4.75	20	2	4	80	3
126.50	4.75	0	2	4	45	4
45.00	3.25	20	2	4	45	5
35.00	4.75	0	1	0	45	6
45.00	4	10	1	2	62	7
28.25	4.75	20	1	0	80	8
85.00	4.75	0	1	4	80	9
8.00	3.25	20	1	0	45	10
36.50	4.75	20	1	4	45	11
33.00	3.25	0	1	4	45	12
84.50	4	10	2	2	62	13
28.50	4.75	20	2	0	45	14
33.50	3.25	0	2	0	45	15
36.00	3.25	20	2	0	80	16
84.00	4.75	0	2	0	80	17
45.00	3.25	20	1	4	80	18
37.50	4	10	1	2	62	19
106.00	3.25	0	2	4	80	20

One discrete factor

Note that four of the factors are *continuous*, and one – number of rubber bands – is *discrete*. Due to the presence of this discrete factor, we actually have two different centerpoints, each with two runs. Runs 7 and 19 are with one rubber band, and the center of the other factors, while runs 2 and 13 are with two rubber bands and the center of the other factors.

Five confirmatory runs

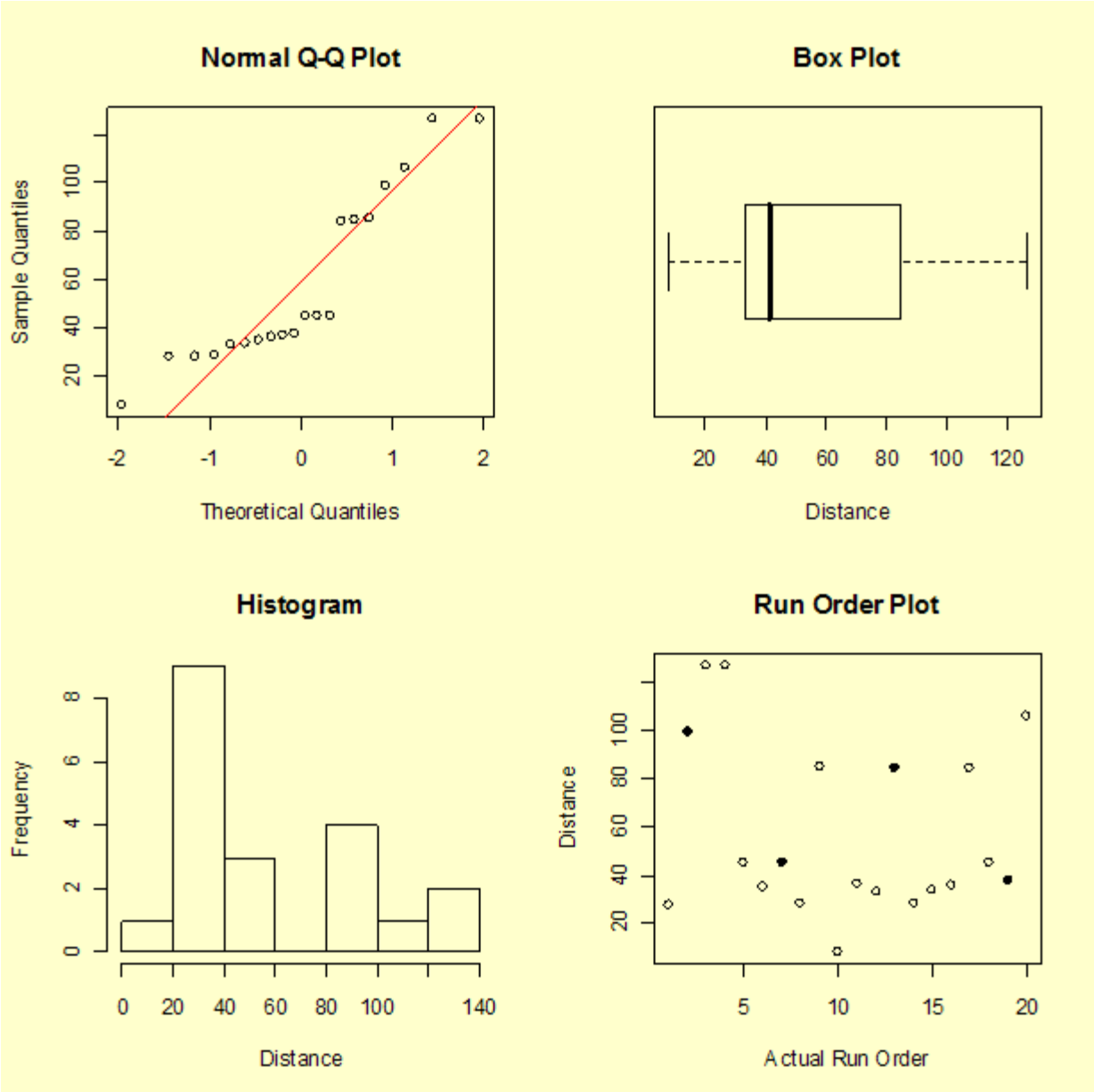
After analyzing the 20 runs and determining factor settings needed to achieve predicted distances of 30, 60 and 90 inches, the team was asked to conduct five confirmatory runs at each of the derived settings.

Analysis of the Experiment

Step 1: Look at the data

Histogram, box plot, normal probability plot, and run order plot of the response

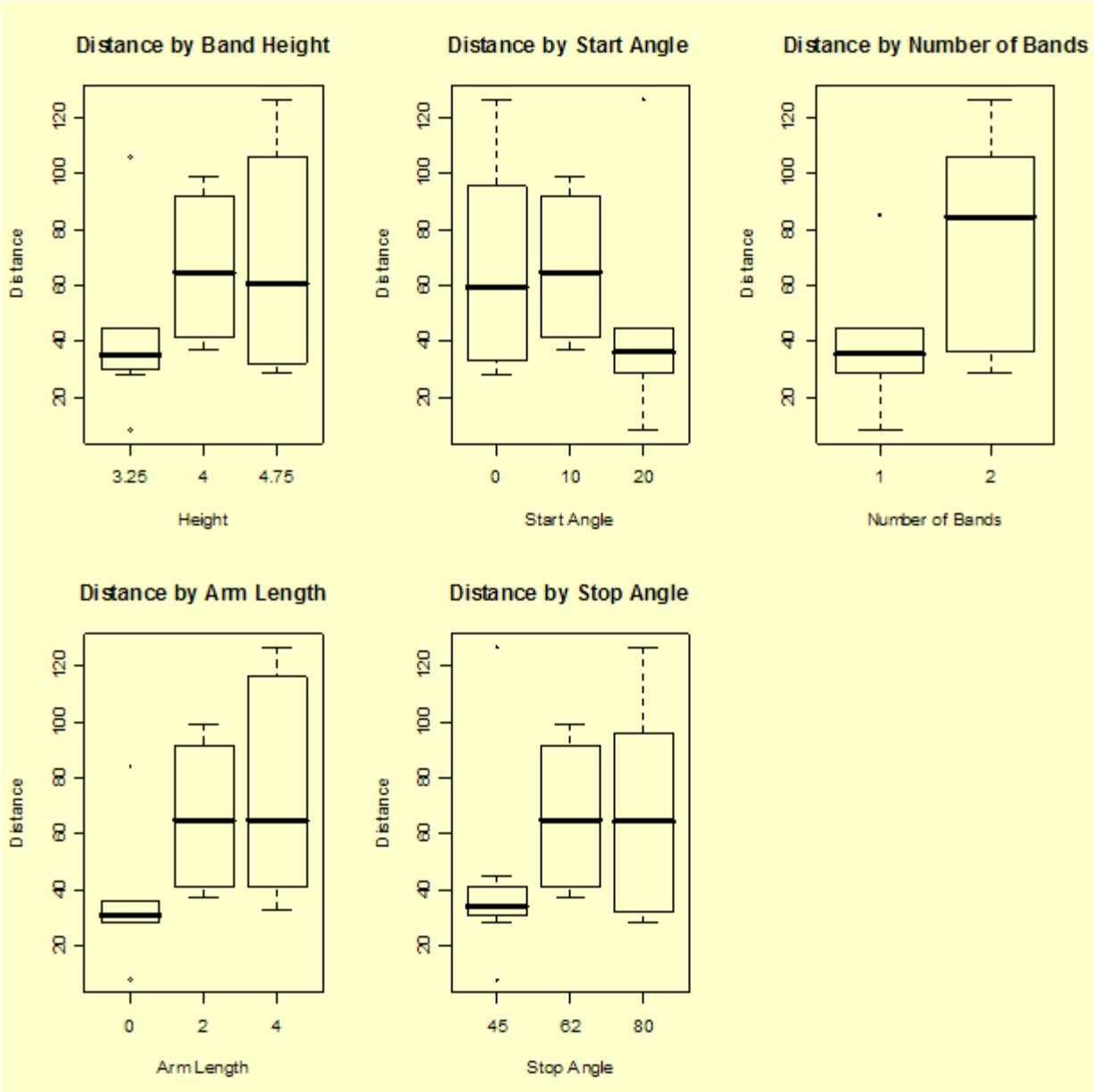
We start by plotting the data several ways to see if any trends or anomalies appear that would not be accounted for by the models.



We can see the large spread of the data and a pattern to the data that should be explained by the analysis. The run order plot does not indicate an obvious time sequence. The four highlighted points in the run order plot are the center points in the design. Recall that runs 2 and 13 had two rubber bands and runs 7 and 19 had one rubber band. There may be a slight aging of the rubber bands in that the second center point resulted in a distance that was a little shorter than the first for each pair.

Plots of responses versus factor columns

Next look at the plots of responses sorted by factor columns.



Several factors appear to change the average response level and most have large spread at each of the levels.

Step 2: Create the theoretical model

The resolution V design can estimate main effects and all two-factor interactions

With a resolution V design we are able to estimate all the main effects and all two-factor interactions – without worrying about confounding. Therefore, the initial model will have 16 terms – the intercept term, the 5 main effects, and the 10 two-factor interactions.

Step 3: Create the actual model from the data

Variable coding

Note we have used the orthogonally coded columns for the analysis, and have abbreviated the factor names as follows:

Height (h) = band height

Start (s) = start angle
Bands (b) = number of rubber bands
Stop (e) = stop angle
Length (l) = arm length.

Trial model with all main factors and two-factor interactions

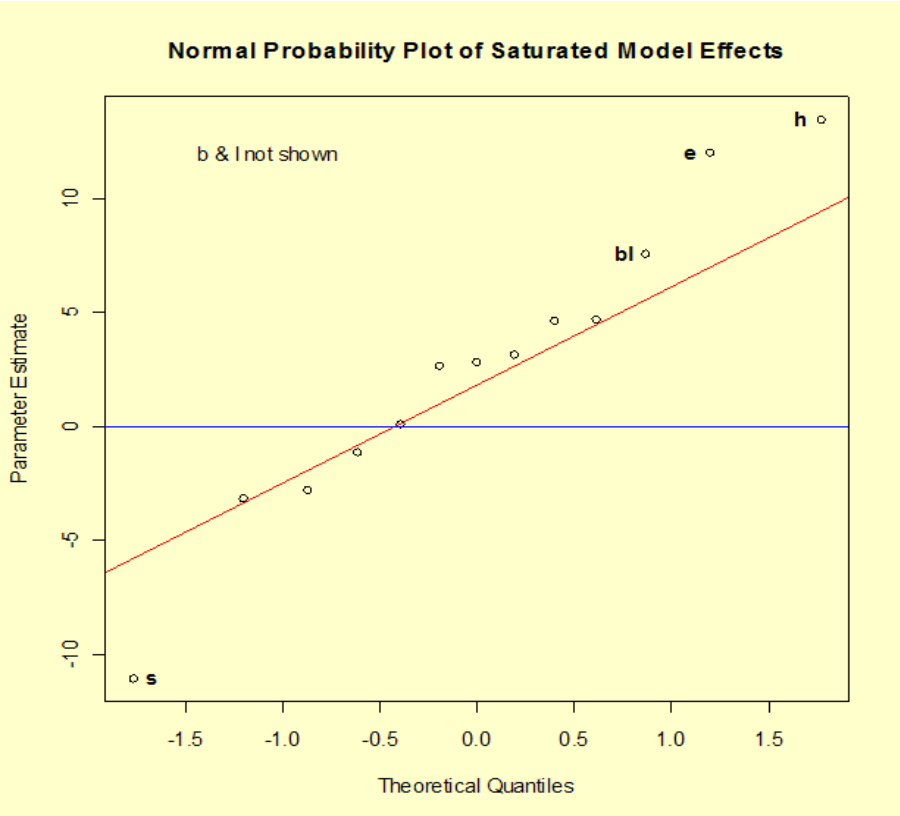
The results of fitting the trial model that includes all main factors and two-factor interactions follow.

```
Source      Estimate  Std. Error  t value  Pr(>|t|)
---------   --------  ----------  -------  --------
Intercept    57.5375      2.9691   19.378  4.18e-05 ***
h            13.4844      3.3196    4.062   0.01532 *
s           -11.0781      3.3196   -3.337   0.02891 *
b            19.4125      2.9691    6.538   0.00283 **
l            20.1406      3.3196    6.067   0.00373 **
e            12.0469      3.3196    3.629   0.02218 *
h*s          -2.7656      3.3196   -0.833   0.45163
h*b           4.6406      3.3196    1.398   0.23467
h*l           4.7031      3.3196    1.417   0.22950
h*e           0.1094      3.3196    0.033   0.97529
s*b          -3.1719      3.3196   -0.955   0.39343
s*l          -1.1094      3.3196   -0.334   0.75502
s*e           2.6719      3.3196    0.805   0.46601
b*l           7.6094      3.3196    2.292   0.08365 .
b*e           2.8281      3.3196    0.852   0.44225
l*e           3.1406      3.3196    0.946   0.39768

Significance codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:  13.28, based on 4 degrees of freedom
Multiple R-squared:  0.9709
Adjusted R-squared:  0.8619
F-statistic:  8.905, based on 15 and 4 degrees of freedom
p-value: 0.02375
```

Use p-values and a normal probability plot to help select significant effects

The model has a good R^2 value, but the fact that R^2 adjusted is considerably smaller indicates that we undoubtedly have some terms in our model that are not significant. Scanning the column of *p*-values (labeled Pr(>|t|)) for small values shows five significant effects at the 0.05 level and another one at the 0.10 level.

A normal probability plot of effects is a useful graphical tool to determine significant effects. The graph below shows that there are nine terms in the model that can be assumed to be noise. That would leave six terms to be included in the model. Whereas the output above shows a *p*-value of 0.0836 for the interaction of Bands (b) and Length (l), the normal plot suggests we treat this interaction as significant.



Refit using just the effects that appear to matter

Remove the non-significant terms from the model and refit to produce the following analysis of variance table.

Source	Df	Sum of Sq	Mean Sq	F value	Pr(>F)
Model	6	22148.55	3691.6		
Total error	13	2106.99	162.1	22.77	3.5e-06
Lack-of-fit	11	1973.74	179.4		
Pure error	2	133.25	66.6	2.69	0.3018

```
Residual standard error:  12.73 based on 13 degrees of freedom
Multiple R-squared:  0.9131
Adjusted R-squared:  0.873
p-value:
```

R^2 is OK and there is no significant model "lack of fit"

The R^2 and R^2 adjusted values are acceptable. The ANOVA table shows us that the model is significant, and the lack-of-fit test is not significant. Parameter estimates are below.

Source	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	57.537	2.847	20.212	3.33e-11	***
h	13.484	3.183	4.237	0.00097	***
s	-11.078	3.183	-3.481	0.00406	**
b	19.412	2.847	6.819	1.23e-05	***
l	20.141	3.183	6.328	2.62e-05	***
e	12.047	3.183	3.785	0.00227	**
b*l	7.609	3.183	2.391	0.03264	*

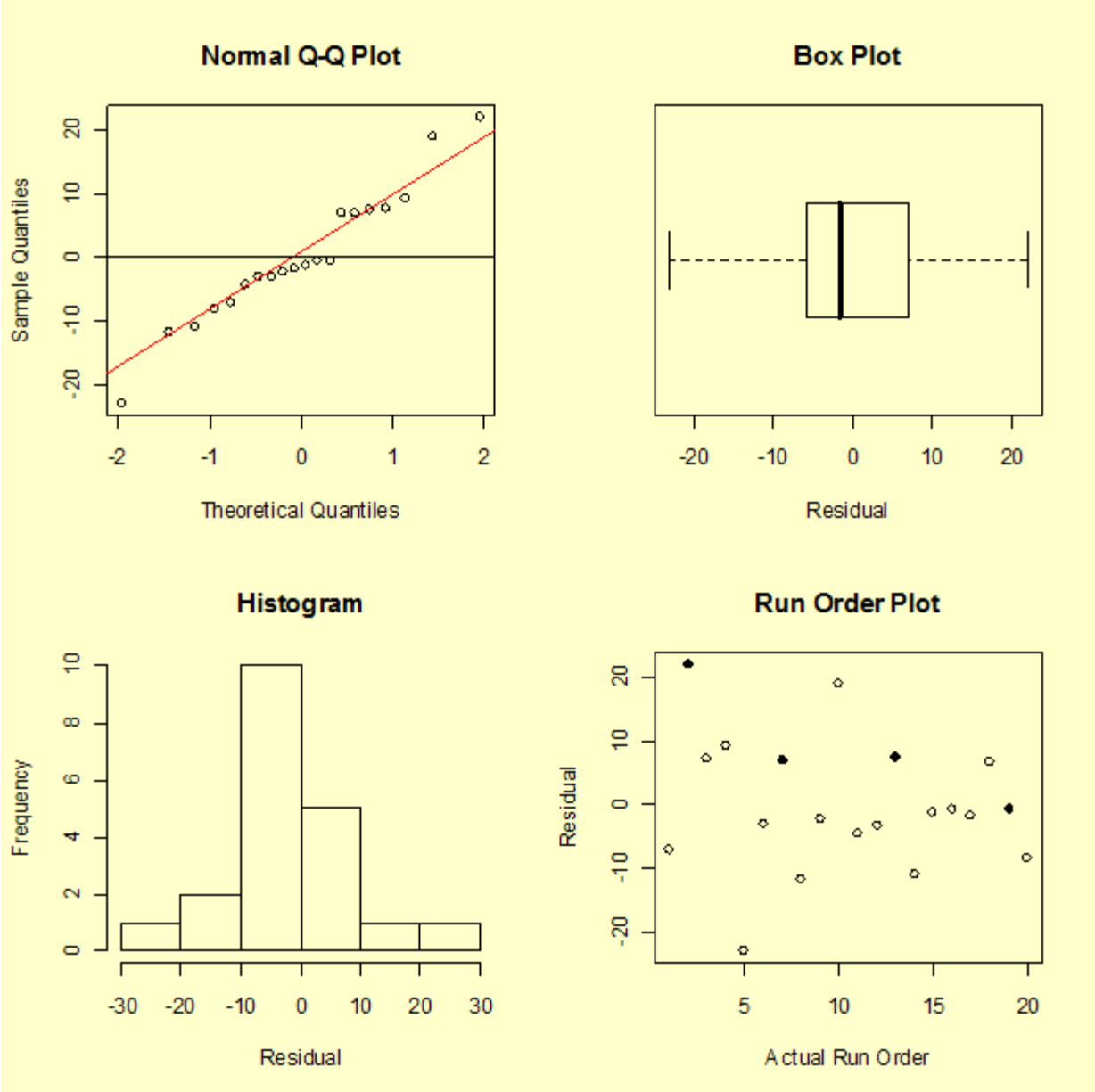
```
Significance codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 4: Test the model assumptions using residual graphs (adjust and simplify as needed)

Diagnostic residual plots

To examine the assumption that the residuals are approximately normally distributed, are independent, and have equal variances, we generate four plots of the residuals: a normal probability plot, box plot, histogram, and a run-order plot of the residuals. In the run-order

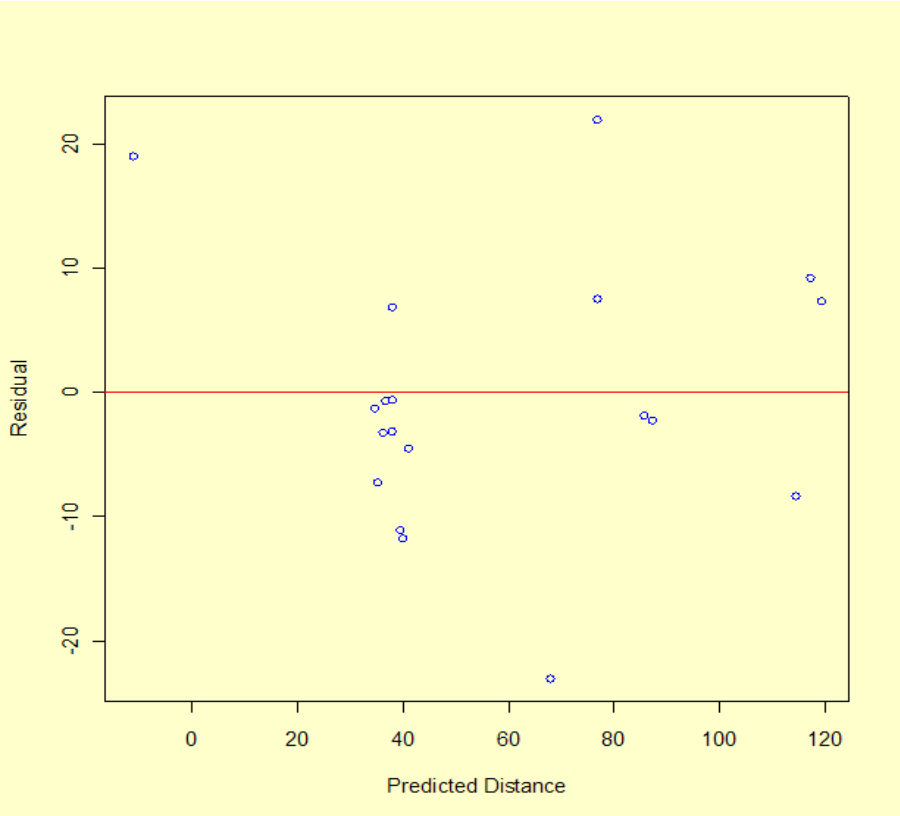
plot, the highlighted points are the centerpoint values. Recall that run numbers 2 and 13 had two rubber bands while run numbers 7 and 19 had only one rubber band.



The residuals do appear to have, at least approximately, a normal distributed.

Plot of residuals versus predicted values

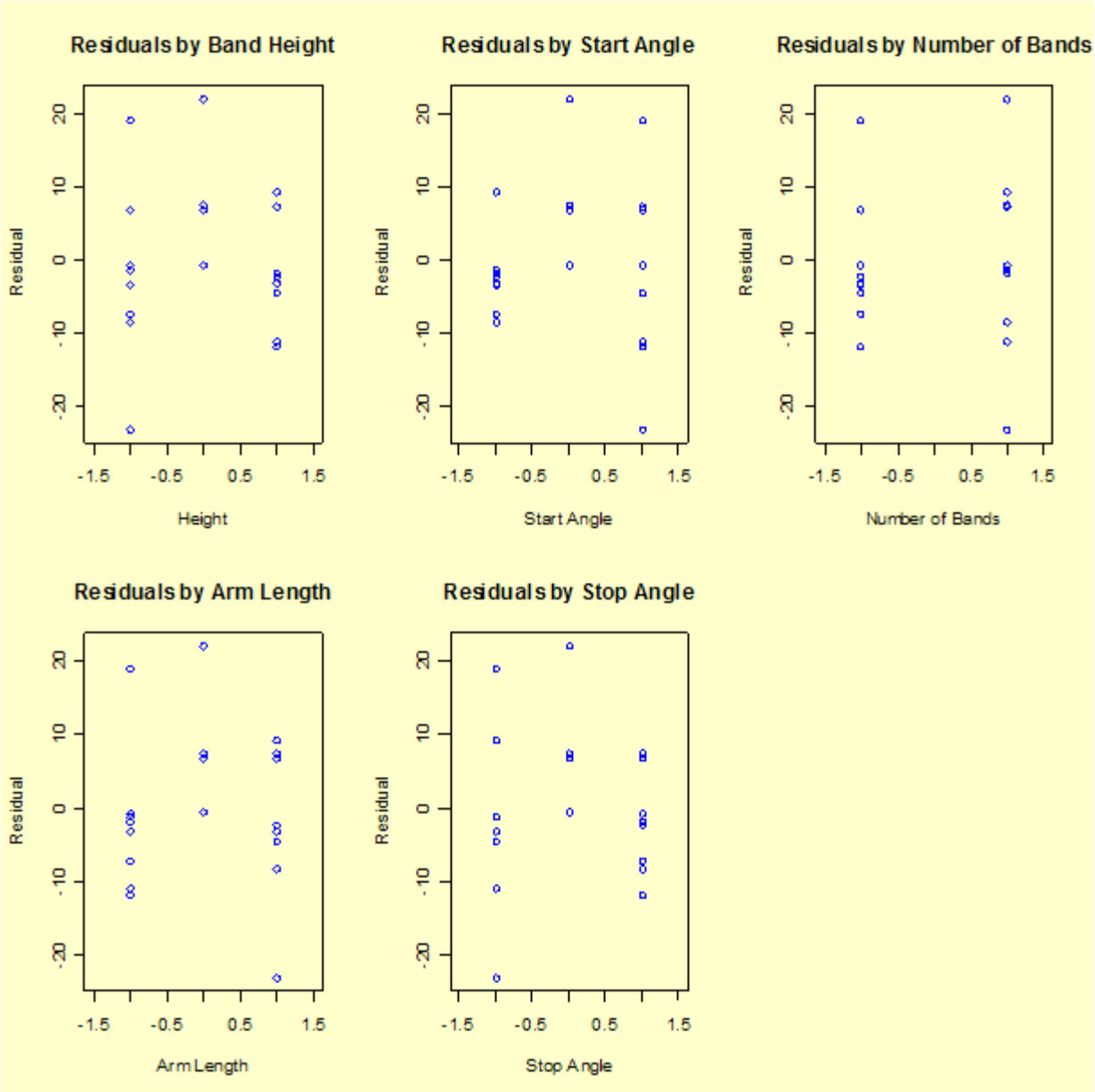
Next we plot the residuals versus the predicted values.



There does not appear to be a pattern to the residuals. One observation about the graph, from a single point, is that the model performs poorly in predicting a short distance. In fact, run number 10 had a measured distance of 8 inches, but the model predicts -11 inches, giving a residual of 19 inches. The fact that the model predicts an impossible negative distance is an obvious shortcoming of the model. We may not be successful at predicting the catapult settings required to hit a distance less than 25 inches. This is not surprising since there is only one data value less than 28 inches. Recall that the objective is to achieve distances of 30, 60, and 90 inches.

Plots of residuals versus the factor variables

Next we look at the residual values versus each of the factors.



The residual graphs are not ideal, although the model passes the lack-of-fit test

Most of the residual graphs versus the factors appear to have a slight "frown" on the graph (higher residuals in the center). This may indicate a lack of fit, or sign of curvature at the centerpoint values. The lack- of-fit test, however, indicates that the lack of fit is not significant.

Consider a transformation of the response variable to see if we can obtain a better model

At this point, since there are several unsatisfactory features of the model we have fit and the resultant residuals, we should consider whether a simple transformation of the response variable (Y = "Distance") might improve the situation.

There are at least two good reasons to suspect that using the logarithm of distance as the response might lead to a better model.

1. A linear model fit to ln(Y) will always predict a positive distance when converted back to the original scale for any possible combination of X factor values.
2. Physical considerations suggest that a realistic model for distance might require quadratic terms since gravity plays a key role - taking logarithms often reduces the impact of non-linear terms.

To see whether using ln(*Y*) as the response leads to a more satisfactory model, we return to step 3.

Step 3a: Fit the full model using ln(Y) as the response

First a main effects and two-factor interaction model is fit to the log distance responses

Proceeding as before, using the coded values of the factor levels and the natural logarithm of distance as the response, we obtain the following parameter estimates.

```
Source      Estimate  Std. Error  t value  Pr(>|t|)
---------   --------  ----------  -------  --------
(Intercept)  3.85702     0.06865   56.186  6.01e-07 ***
h            0.25735     0.07675    3.353   0.02849 *
s           -0.24174     0.07675   -3.150   0.03452 *
b            0.34880     0.06865    5.081   0.00708 **
l            0.39437     0.07675    5.138   0.00680 **
e            0.26273     0.07675    3.423   0.02670 *
h*s         -0.02582     0.07675   -0.336   0.75348
h*b         -0.02035     0.07675   -0.265   0.80403
h*l         -0.01396     0.07675   -0.182   0.86457
h*e         -0.04873     0.07675   -0.635   0.55999
s*b          0.00853     0.07675    0.111   0.91686
s*l          0.06775     0.07675    0.883   0.42724
s*e          0.07955     0.07675    1.036   0.35855
b*l          0.01499     0.07675    0.195   0.85472
b*e         -0.01152     0.07675   -0.150   0.88794
l*e         -0.01120     0.07675   -0.146   0.89108

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:  0.307 based on 4 degrees of freedom
Multiple R-squared:  0.9564
Adjusted R-squared:  0.7927
F-statistic:  5.845 based on 15 and 4 degrees of freedom
p-value:  0.0502
```

A simpler model with just main effects has a satisfactory fit

Examining the *p*-values of the 16 model coefficients, only the intercept and the 5 main effect terms appear significant. Refitting the model with just these terms yields the following results.

```
Source           Df Sum of Sq  Mean Sq   F value    Pr(>F)
-----------      -- ---------  -------   -------    ------
Model             5   8.02079   1.60416   36.285   1.6e-07
Total error      14   0.61896   0.04421

Lack-of-fit      12   0.58980   0.04915
Pure error        2   0.02916   0.01458    3.371    0.2514


Source      Estimate  Std. Error  t value  Pr(>|t|)
---------   --------  ----------  -------  --------
Intercept    3.85702     0.04702   82.035   < 2e-16 ***
h            0.25735     0.05257    4.896  0.000236 ***
s           -0.24174     0.05257   -4.599  0.000413 ***
b            0.34880     0.04702    7.419  3.26e-06 ***
l            0.39437     0.05257    7.502  2.87e-06 ***
e            0.26273     0.05257    4.998  0.000195 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:  0.2103 based on 14 degrees of freedom
Multiple R-squared:  0.9284
Adjusted R-squared:  0.9028
```

This is a simpler model than previously obtained in Step 3 (no interaction term). All the terms are highly significant and there is no indication of a significant lack of fit.

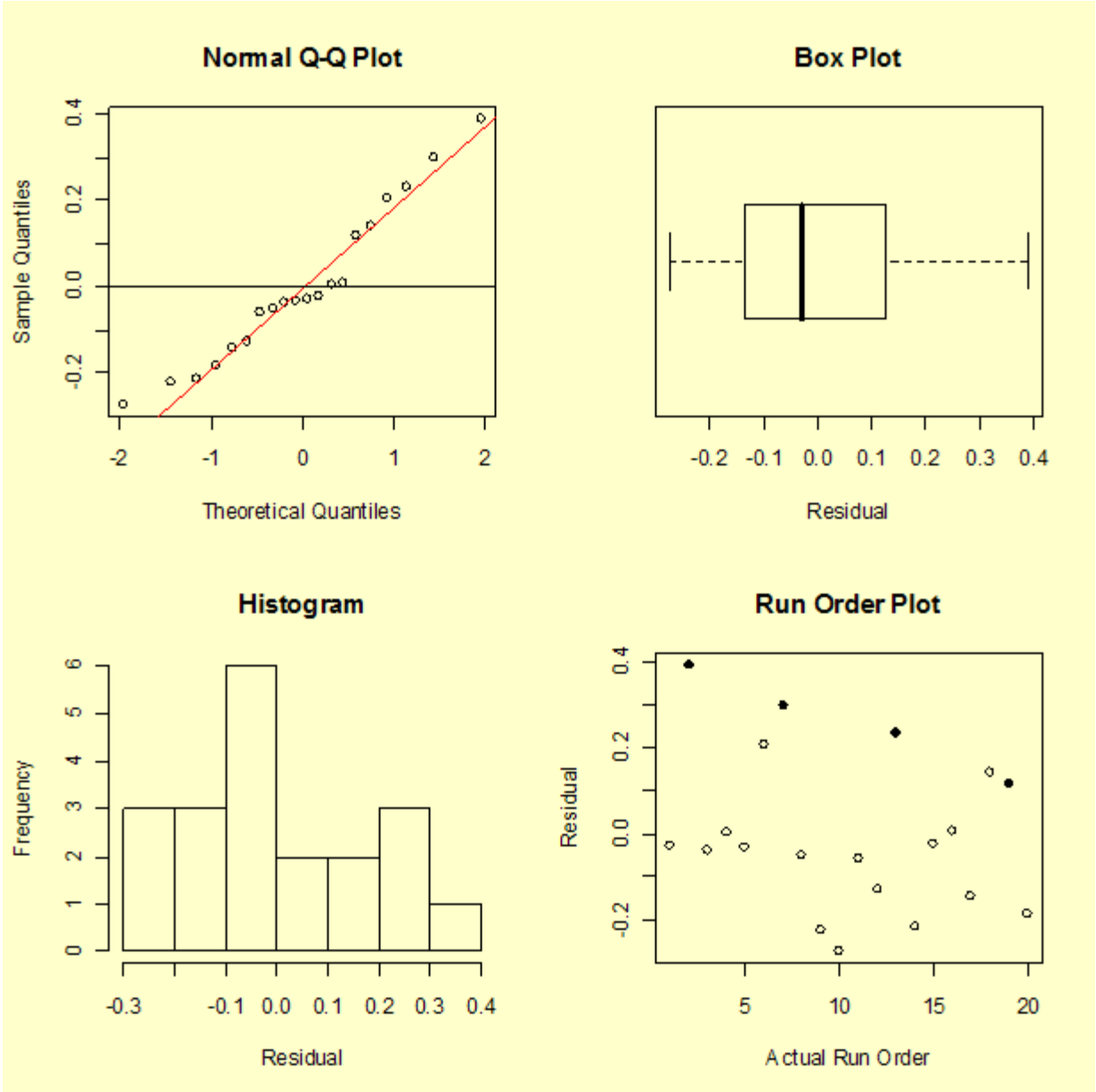
We next look at the residuals for this new model fit.

Step 4a: Test the (new) model assumptions using residual graphs (adjust and simplify as

needed)

Normal probability plot, box plot, histogram, and run-order plot of the residuals

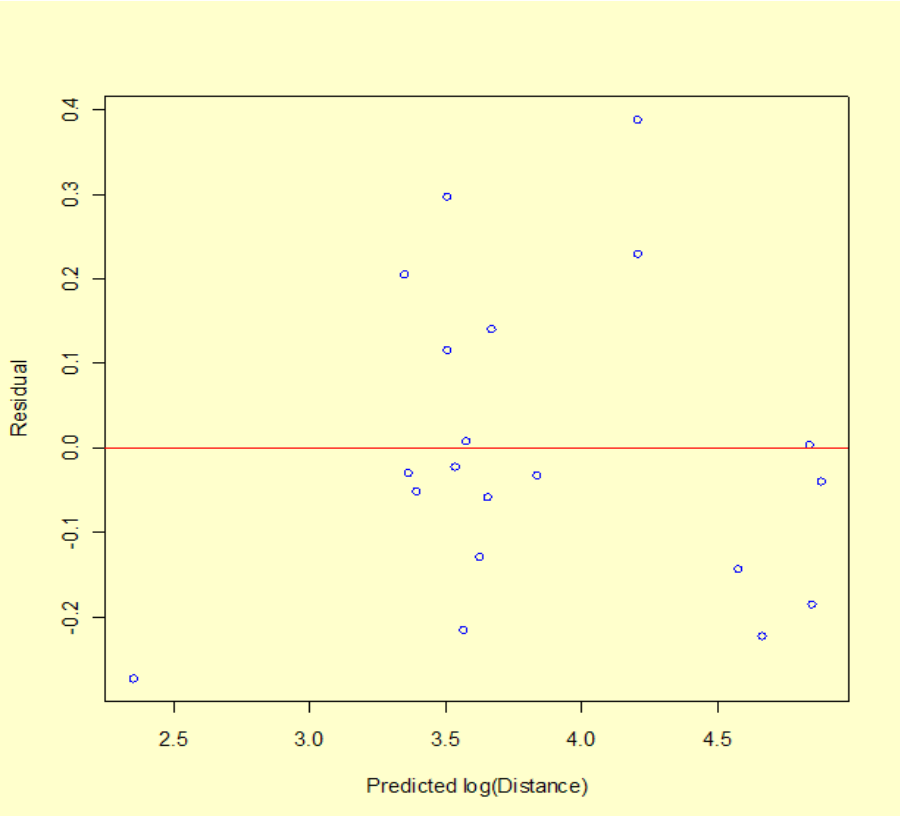
The following normal plot, box plot, histogram and run-order plot of the residuals shows no problems.



Residuals plotted versus run order again show a possible slight decreasing trend (rubber band fatigue?).

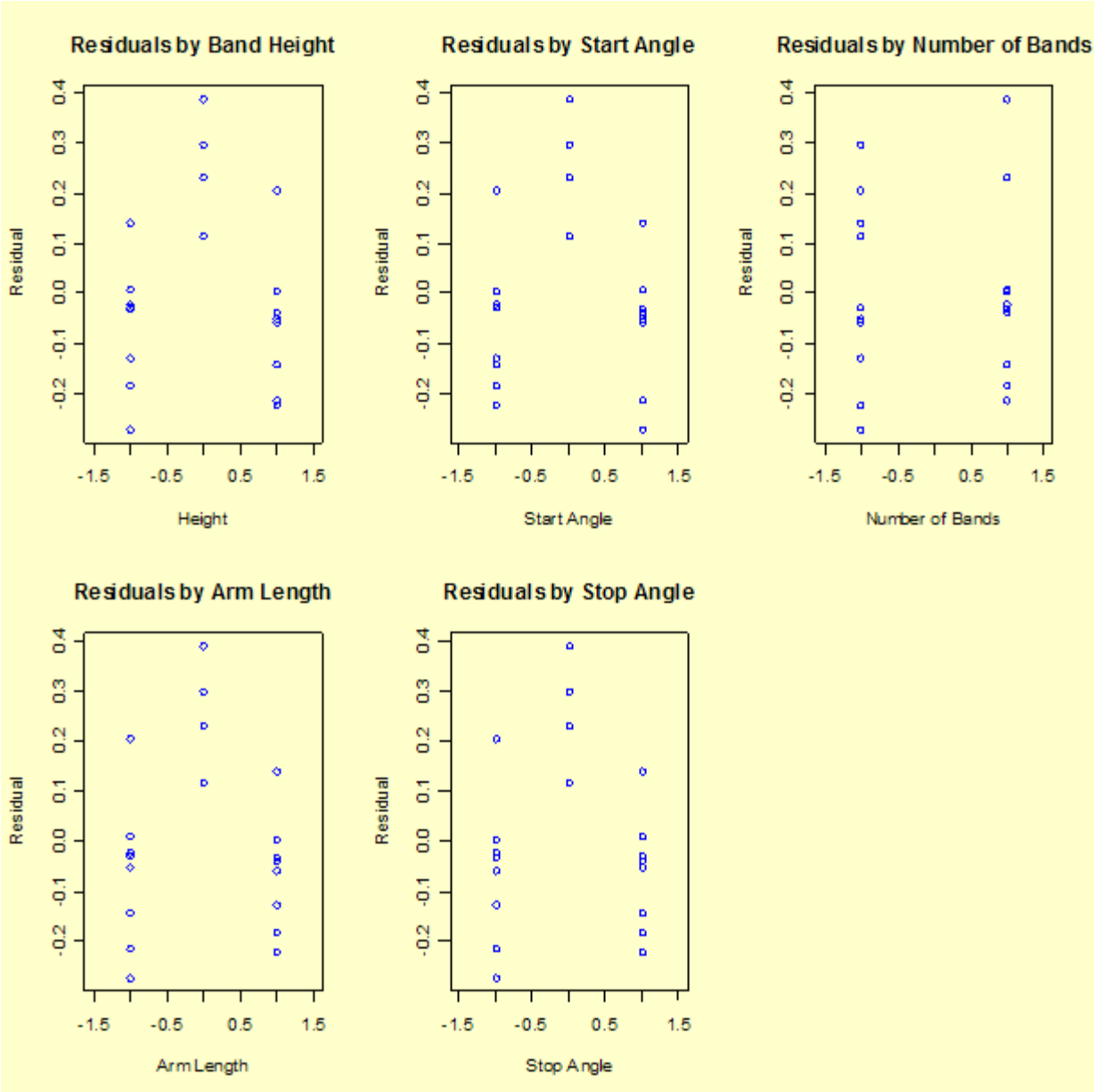
Plot of residuals versus predicted ln(Y) values

A plot of the residuals versus the predicted ln(*Y*) values looks reasonable, although there might be a tendency for the model to overestimate slightly for high predicted values.



Plot of residuals versus the factor variables

Next we look at the residual values versus each of the factors.



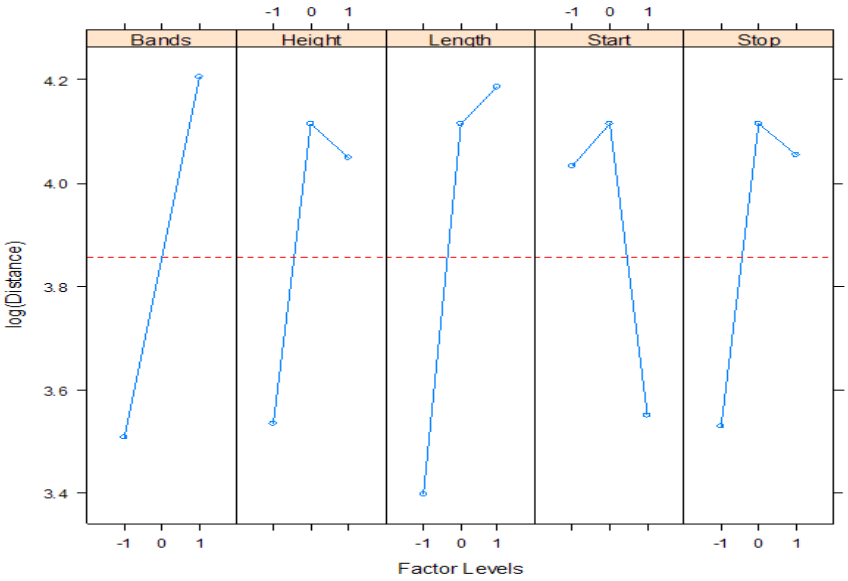
The residuals for the main effects model (fit to natural log of distance) are reasonably well behaved

These plots still appear to have a slight "frown" on the graph (higher residuals in the center). However, the model is generally an improvement over the previous model and will be accepted as possibly the best that can be done without conducting a new experiment designed to fit a quadratic model.

Step 5: Use the results to answer the questions in your experimental objectives

Final step: Predict the settings that should be used to obtain desired distances

Based on the analyses and plots, we can select factor settings that maximize the log-transaformed distance. Translating from "-1", "0", and "+1" back to the actual factor settings, we have: band height at "0" or 3.5 inches; start angle at "0" or 10 degrees; number of rubber bands at "1" or 2 bands; arm length at "1" or 4 inches , and stop angle at "0" or 80 degrees.



"Confirmation" runs were successful

In the confirmatory runs that followed the experiment, the team was successful at hitting all three targets, but did not hit them all five times. The model discovery and fitting process, as illustrated in this analysis, is often an iterative process.

5.4.7.3. Response surface model example

Data Source

A CCD with two responses

This example uses experimental data published in Czitrom and Spagon, (1997), *Statistical Case Studies for Industrial Process Improvement.* The material is copyrighted by the American Statistical Association and the Society for Industrial and Applied Mathematics, and is used with their permission. Specifically, Chapter 15, titled "Elimination of TiN Peeling During Exposure to CVD Tungsten Deposition Process Using Designed Experiments", describes a semiconductor wafer processing experiment (labeled *Experiment 2*).

Goal, response variables, and factor variables

The goal of this experiment was to fit response surface models to the two responses, *deposition layer Uniformity* and *deposition layer Stress*, as a function of two particular controllable factors of the chemical vapor deposition (CVD) reactor process. These factors were *Pressure* (measured in torr) and the ratio of the gaseous reactants H_2 and WF_6 (called H_2/WF_6).

The experiment also included an important third (categorical) response - the presence or absence of titanium nitride (TiN) peeling. The third response has been omitted in this example in order to focus on the response surface aspects of the experiment.

To summarize, the goal is to obtain a response surface model for two responses, Uniformity and Stress. The factors are: Pressure and H_2/WF_6.

Experiment Description

The design is a 13-run CCI design with 3 centerpoints

The minimum and maximum values chosen for Pressure were 4 torr and 80 torr (0.5333 kPa and 10.6658 kPa). Although the international system of units indicates that the standard unit for pressure is Pascal, or 1 N/m^2, we use torr to be consistent with the analysis appearing in the paper by Czitrom and Spagon.

The minimum and maximum H_2/WF_6 ratios were chosen to be 2 and 10. Since response curvature, especially for Uniformity, was a distinct possibility, an experimental design that allowed estimating a second order (quadratic) model was needed. The experimenters decided to use a central composite inscribed (CCI) design. For two factors, this design is typically recommended to have 13 runs with 5 centerpoint runs. However, the experimenters, perhaps to conserve a limited supply of wafer resources, chose to include only 3 centerpoint runs. The design is still rotatable, but the uniform precision property has been sacrificed.

Table containing the CCI design points and experimental responses

The table below shows the CCI design and experimental responses, in the order in which they were run (presumably randomized). The last two columns show coded values of the factors.

Run	Pressure	H_2/WF_6	Uniformity	Stress	Coded Pressure	Coded H_2/WF_6
1	80	6	4.6	8.04	1	0
2	42	6	6.2	7.78	0	0
3	68.87	3.17	3.4	7.58	0.71	-0.71
4	15.13	8.83	6.9	7.27	-0.71	0.71
5	4	6	7.3	6.49	-1	0
6	42	6	6.4	7.69	0	0
7	15.13	3.17	8.6	6.66	-0.71	-0.71
8	42	2	6.3	7.16	0	-1
9	68.87	8.83	5.1	8.33	0.71	0.71
10	42	10	5.4	8.19	0	1
11	42	6	5.0	7.90	0	0

Low values of both responses are better than high

Uniformity is calculated from four-point probe sheet resistance measurements made at 49 different locations across a wafer. The value in the table is the standard deviation of the 49 measurements divided by their mean, expressed as a percentage. So a smaller value of Uniformity indicates a more uniform layer - hence, lower values are desirable. The Stress calculation is based on an optical measurement of wafer bow, and again lower values are more desirable.

Analysis of DOE Data

Steps for fitting a response surface model

The steps for fitting a response surface (second-order or quadratic) model are as follows:

1. Fit the full model to the first response.
2. Use stepwise regression, forward selection, or backward elimination to identify important variables.
3. When selecting variables for inclusion in the model, follow the hierarchy principle and keep all main effects that are part of significant higher-order terms or interactions, even if the main effect *p*-value is larger than you would like (note that not all analysts agree with this principle).
4. Generate diagnostic residual plots (histograms, box plots, normal plots, etc.) for the model selected.
5. Examine the fitted model plot, interaction plots, and ANOVA statistics (R^2, adjusted R^2, lack-of-fit test, etc.). Use all these plots and statistics to determine whether the model fit is satisfactory.
6. Use contour plots of the response surface to explore the effect of changing factor levels on the response.
7. Repeat all the above steps for the second response variable.
8. After satisfactory models have been fit to both responses, you can overlay the surface contours for both responses.
9. Find optimal factor settings.

Fitting a Model to the Uniformity Response, Simplifying the Model and Checking Residuals

Fit full quadratic model to Uniformity response

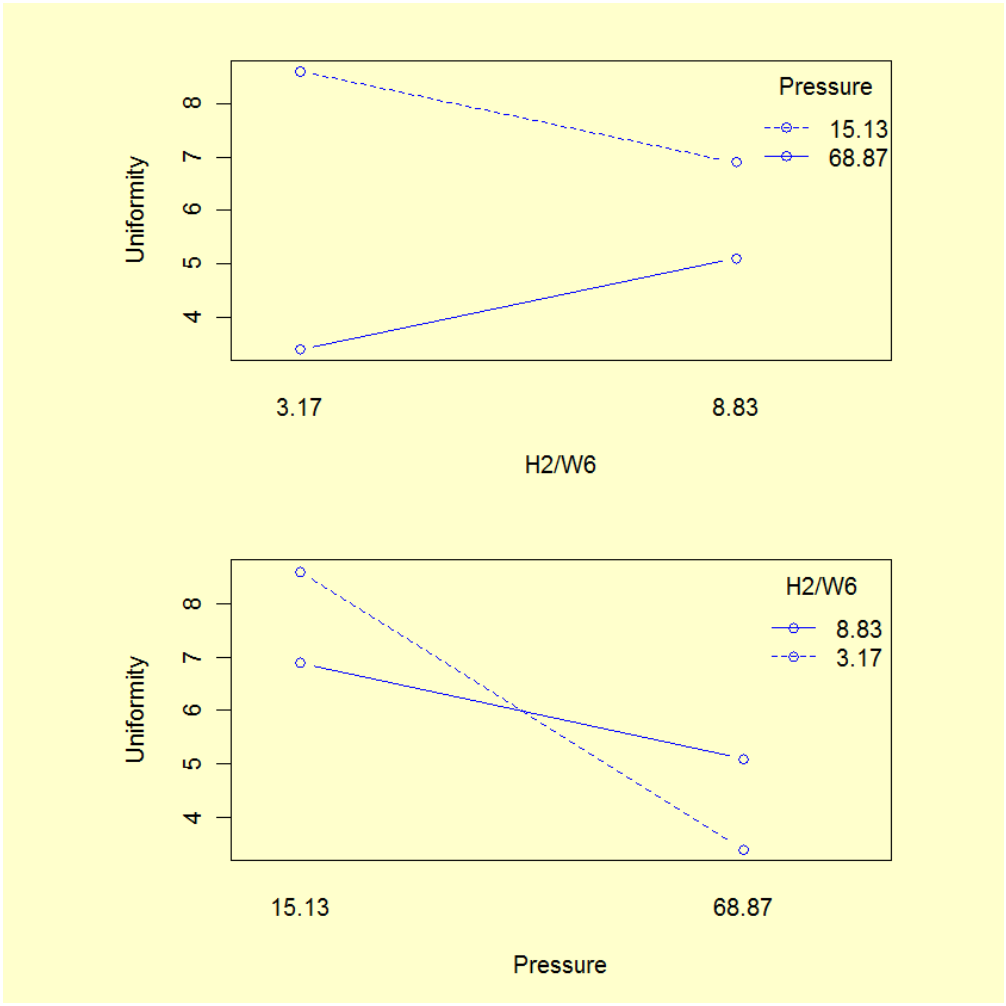
Source	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	5.86613	0.41773	14.043	3.29e-05
Pressure	-1.90967	0.36103	-5.289	0.00322
H2/WF6	-0.22408	0.36103	-0.621	0.56201
Pressure*H2/WF6	1.68617	0.71766	2.350	0.06560
Pressure^2	0.13373	0.60733	0.220	0.83442
H2/WF6^2	0.03373	0.60733	0.056	0.95786

```
Residual standard error:  0.7235 based on 5 degrees of freedom
Multiple R-squared:  0.8716
Adjusted R-squared:  0.7431
F-statistic:  6.787 based on 5 and 5 degrees of freedom
p-value:  0.0278
```

Stepwise regression for Uniformity

```
Start:   AIC=-3.79
Model:   Uniformity ~ Pressure + H2/WF6 + Pressure*H2/WF6 +
Pressure^2 + H2/WF6^2

Step 1:  Remove H2/WF6^2, AIC=-5.79
Model:   Uniformity ~ Pressure + H2/WF6 + Pressure*H2/WF6 +
Pressure^2

Step 2:  Remove Pressure^2,  AIC=-7.69
Model:   Uniformity ~ Pressure + H2/WF6 + Pressure*H2/WF6

Step 3:  Remove H2/WF6,  AIC=-8.88
Model:   Uniformity ~ Pressure + Pressure*H2/WF6
```

The stepwise routine selects a model containing the intercept, Pressure, and the interaction term. However, many statisticians do not think an interaction term should be included in a model unless both main effects are also included. Thus, we will use the model from Step 2 that included Pressure, H2/WF6, and the interaction term. Interaction plots confirm the need for an interaction term in the model.



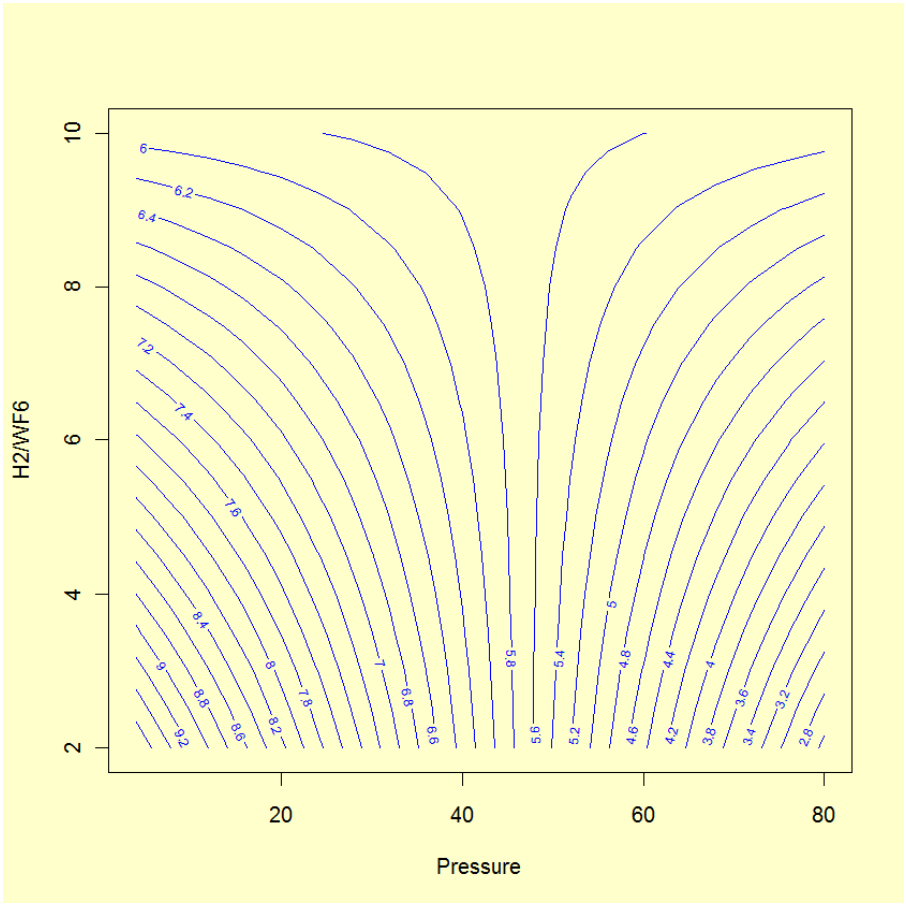
Analysis of model selected by stepwise regression for Uniformity

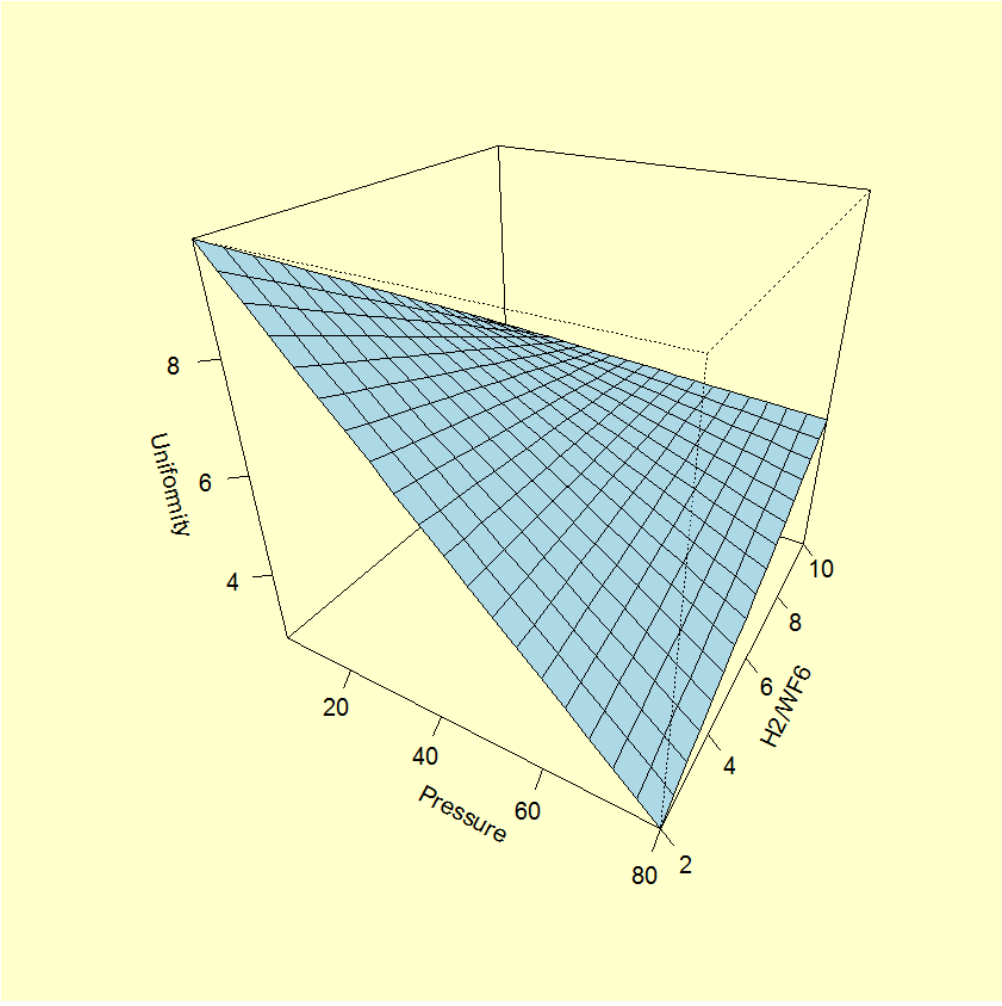
Source	DF	Sum of Sq	Mean Sq	F value	Pr(>F)
Model	3	17.739	5.9130	15.66	0.0017
Total error	7	2.643	0.3776		
Lack-of-fit	5	1.4963	0.2993	0.52	0.7588
Pure error	2	1.1467	0.5734		

```
Residual standard error:  0.6145 based on 7 degrees of freedom
Multiple R-squared:  0.8703
Adjusted R-squared:  0.8148
```

Source	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	5.9273	0.1853	31.993	7.54e-09
Pressure	-1.9097	0.3066	-6.228	0.000433
H2/WF6	-0.2241	0.3066	-0.731	0.488607
Pressure*H2/WF6	1.6862	0.6095	2.767	0.027829

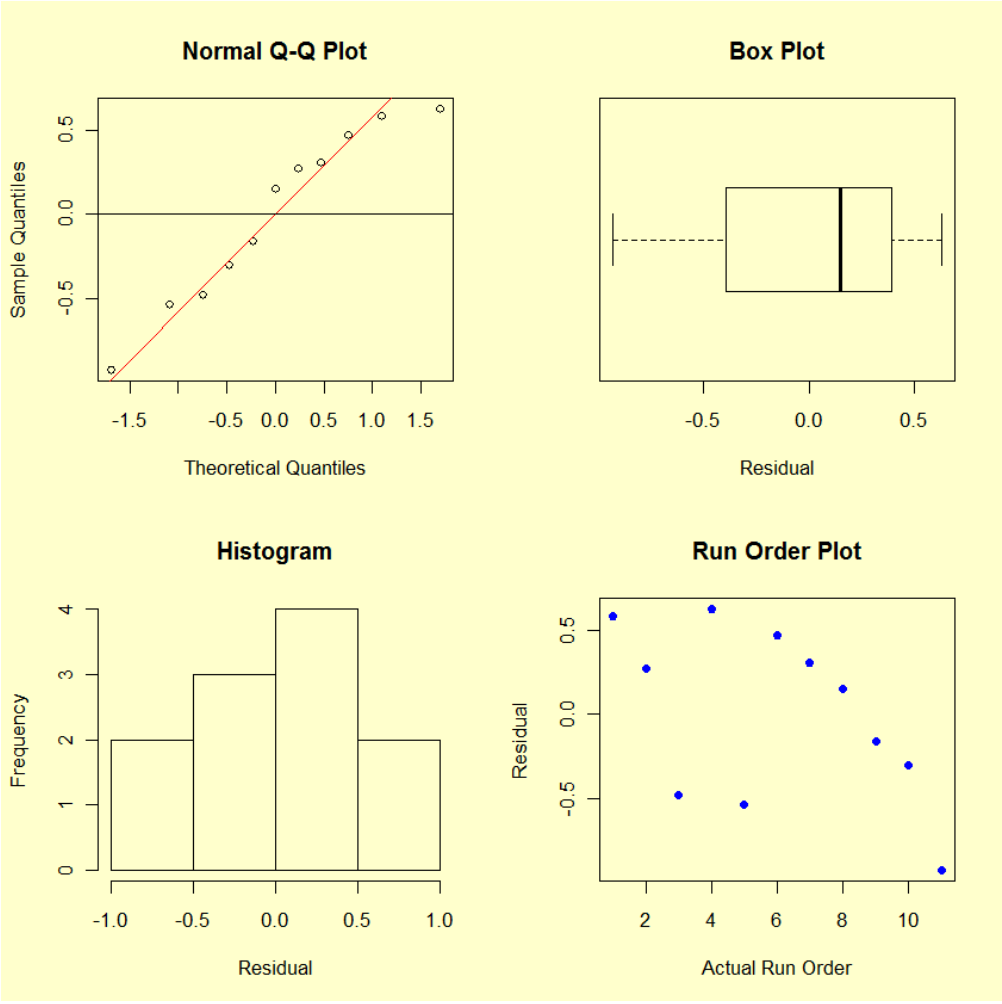
A contour plot and perspective plot of Uniformity provide a visual display of the response surface.


H2/WF6
Pressure
2
4
6
8
10
20
40
60
80
6
6.2
6.4
7.2
7.4
7.6
8.4
9
8.8
8.6
9.2
8
7.8
8.2
7
6.8
6.6
5.8
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
3.8
3.6
3.4
3.2
2.8


Uniformity
Pressure
H2/WF6
4
6
8
20
40
60
80
2
4
6
8
10

Residual plots

We perform a residuals analysis to validate the model assumptions. We generate a normal plot, a box plot, a histogram and a run-order plot of the residuals.



The residual plots do not indicate problems with the underlying assumptions.

Conclusions from the analysis

From the above output, we make the following conclusions.

- The R^2 is reasonable for fitting Uniformity (well known to be a difficult response to model).
- The lack-of-fit test is not significant (very small "Prob > F " would indicate a lack of fit).
- The residual plots do not reveal any major violations of the underlying assumptions.
- The interaction plot shows why an interaction term is needed (parallel lines would suggest no interaction).

Fitting a Model to the Stress Response, Simplifying the Model and Checking Residuals

Fit full quadratic model to Stress

Source	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	8.056791	0.179561	44.869	1.04e-07
Pressure	0.735933	0.038524	19.103	7.25e-06
H2/WF6	0.852099	0.198192	4.299	0.00772
Pressure*H2/WF6	0.069431	0.076578	0.907	0.40616

response

```
Pressure^2        -0.528848    0.064839   -8.156   0.00045
H2/WF6^2          -0.007414    0.004057   -1.827   0.12722


Residual standard error:  0.07721 based on 5 degrees of freedom
Multiple R-squared:  0.9917
Adjusted R-squared:  0.9834
F-statistic:  119.8 based on 5 and 5 degrees of freedom
p-value:  3.358e-05
```

Stepwise regression for Stress

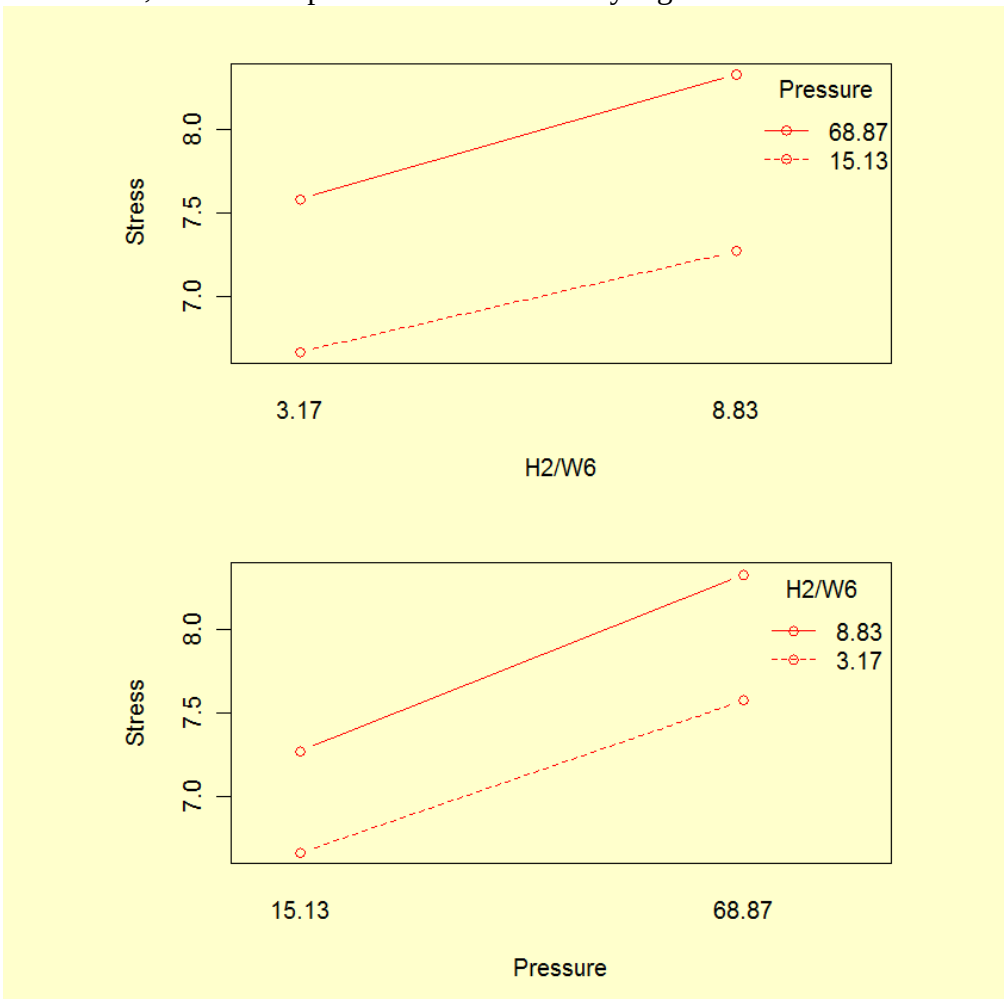
```
Start:   AIC=-53.02
Model:   Stress ~ Pressure + H2/WF6 + Pressure*H2/WF6 +
Pressure^2 + H2/WF6^2

Step 1:  AIC=-53.35
Model:   Stress ~ Pressure + H2/WF6 + Pressure^2 + H2/WF6^2
```

The stepwise routine identifies a model containing the intercept, the main effects, and both squared terms. However, the fit of the full quadratic model indicates that neither the H2/WF6 squared term nor the interaction term are significant. A comparison of the full model and the model containing just the main effects and squared pressure terms indicates that there is no significant difference between the two models.

```
Model 1: Stress ~ Pressure + H2/WF6 + Pressure^2
Model 2: Stress ~ Pressure + H2/WF6 + Pressure^2 +
Pressure*H2/WF6 + H2/WF6^2
```

Source	DF	Sum of Sq	Mean Sq	F value	Pr(>F)
Model 1	2	0.024802	0.01240	2.08	0.22
Model 2	5	0.029804	0.00596		

In addition, interaction plots do not indicate any significant interaction.



Thus, we will proceed with the model containing main effects and the squared pressure term.

The fact that the stepwise procedure selected a model for Stress containing a term that was not significant indicates that all output generated by statistical software should be carefully examined. In this case, the stepwise procedure identified the model with the lowest AIC (Akaike information criterion), but did not take into account contributions by individual terms. Other software using a different criteria may identify a different model, so it is important to understand the algorithms being used.

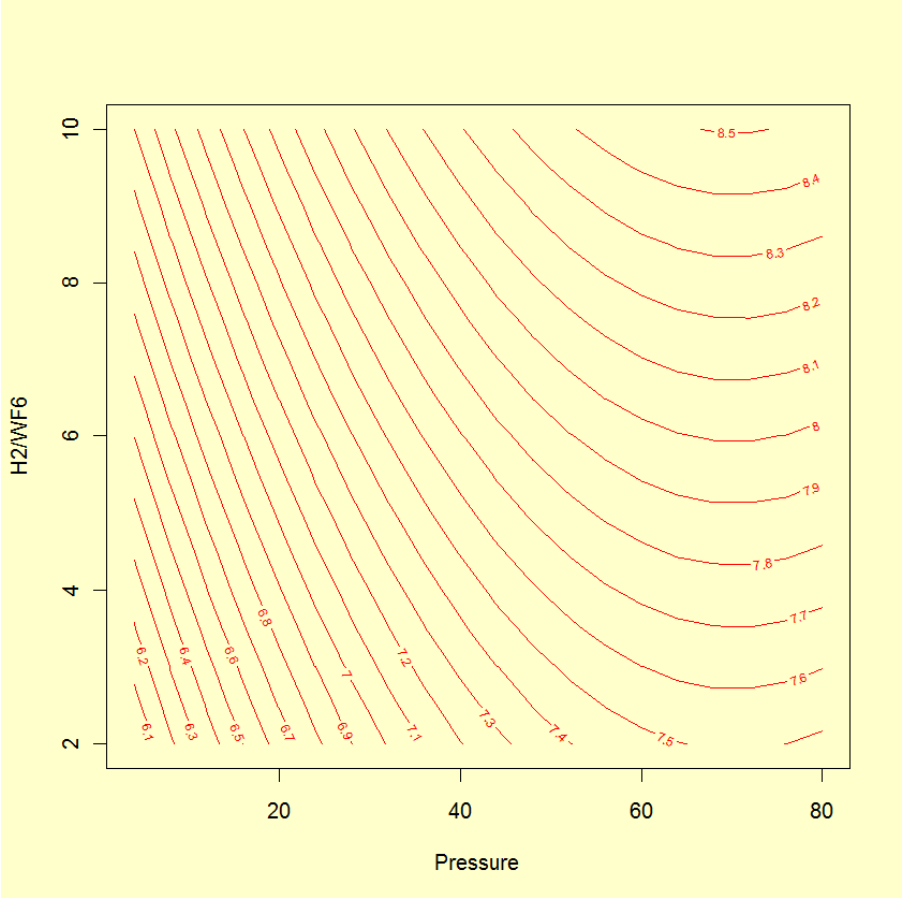
Analysis of reduced model for Stress

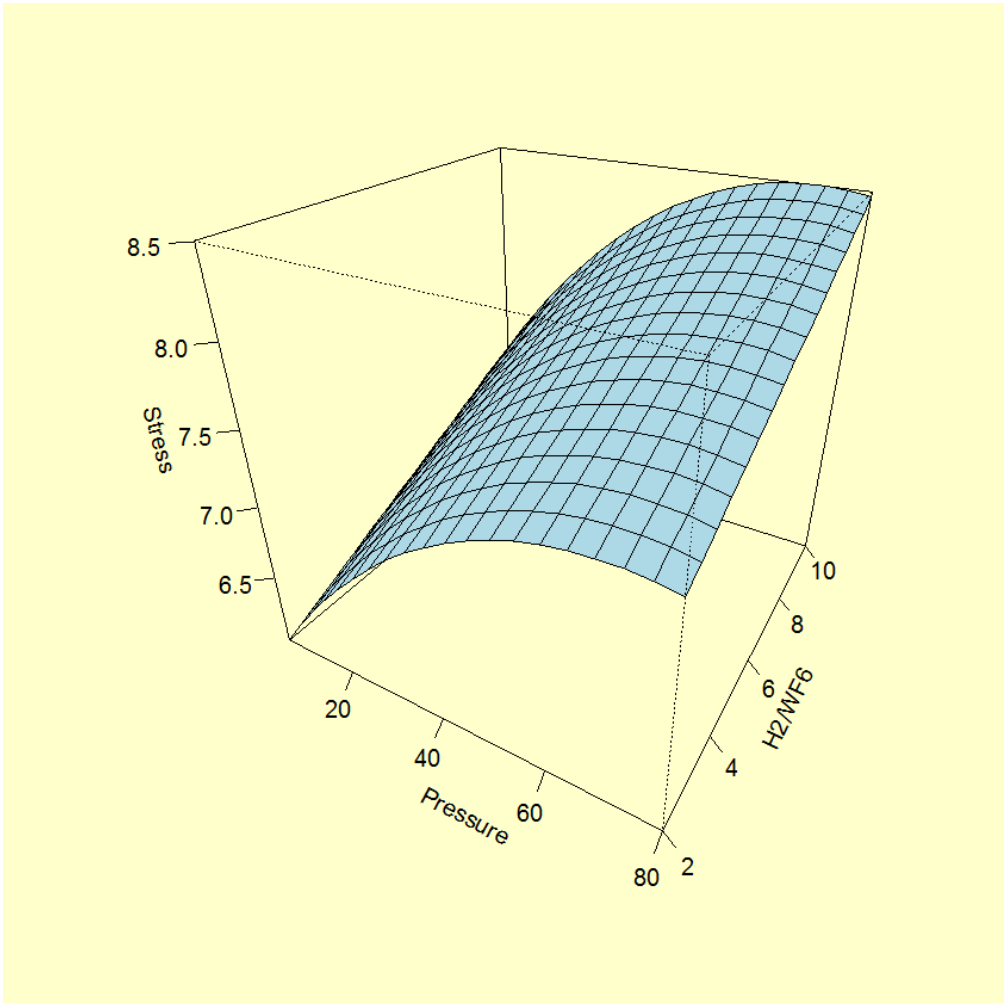
Source	DF	Sum of Sq	Mean Sq	F value	Pr(>F)
Model	3	3.5454	1.1818	151.5	9.9e-07
Total error	7	0.0546	0.0078		
Lack-of-fit	5	0.032405	0.00065	0.58	0.73
Pure error	2	0.022200	0.01110		

```
Residual standard error:  0.0883 based on 7 degrees of freedom
Multiple R-squared:  0.9848
Adjusted R-squared:  0.9783
```

Source	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	7.73410	0.03715	208.185	1.56e-14
Pressure	0.73593	0.04407	16.699	6.75e-07
H2/WF6	0.49686	0.04407	11.274	9.65e-06
Pressure^2	-0.49426	0.07094	-6.967	0.000218

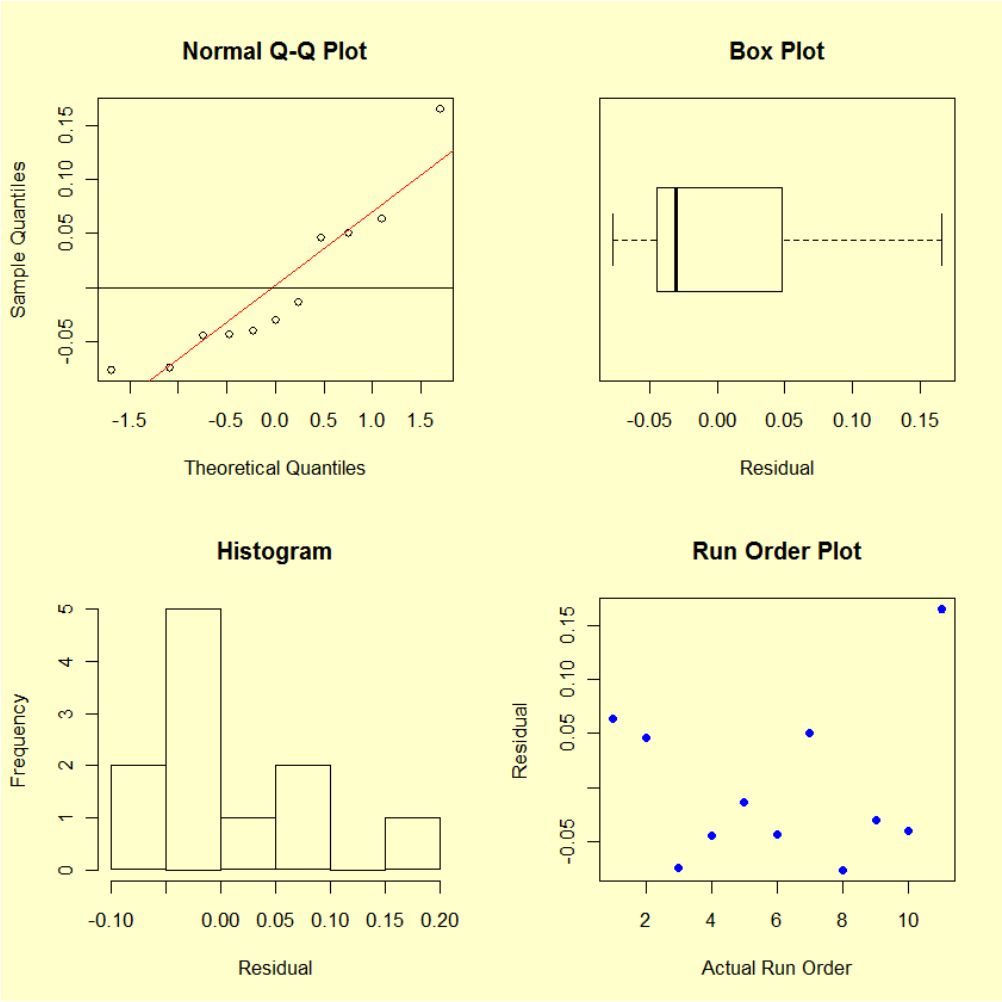
A contour plot and perspective plot of Stress provide a visual representation of the response surface.





Residual plots

We perform a residuals analysis to validate the model by generating a run-order plot, box plot, histogram, and normal probability plot of the residuals.



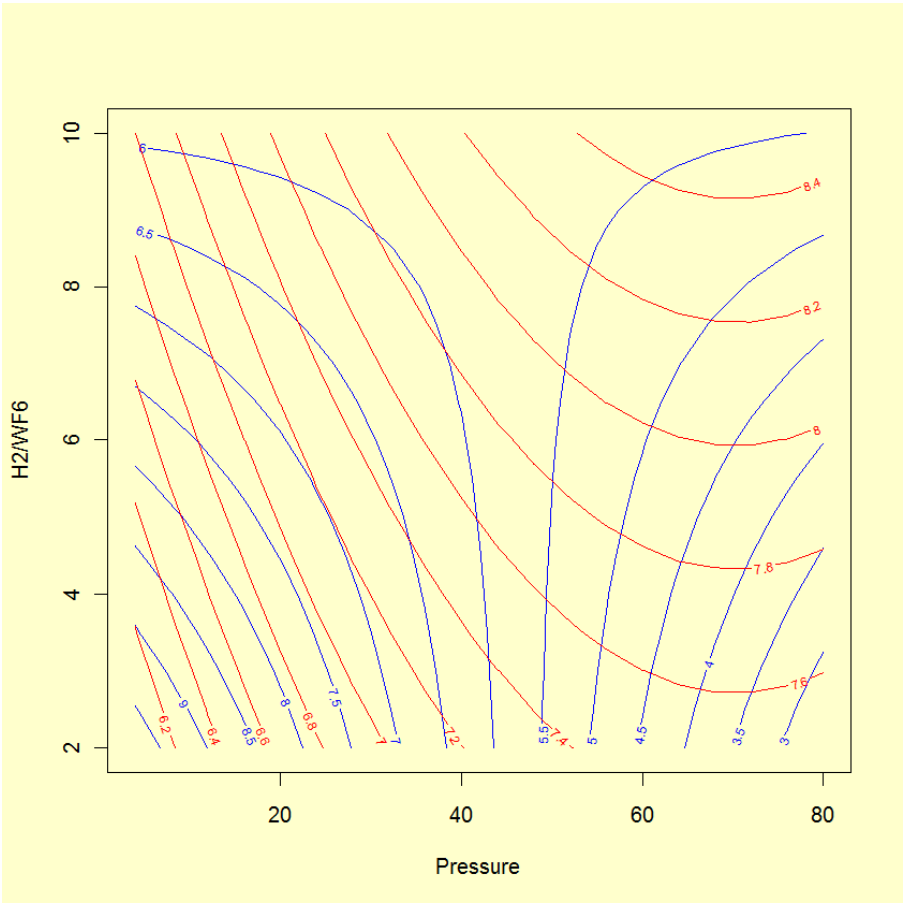
The residual plots do not indicate any major violations of the underlying assumptions.

Conclusions From the above output, we make the following conclusions.

- The R^2 is very good for fitting Stress.
- The lack-of-fit test is not significant (very small "Prob > F " would indicate a lack of fit).
- The residual plots do not reveal any major violations of the underlying assumptions.
- The nearly parallel lines in the interaction plots show why an interaction term is not needed.

Response Surface Contours for Both Responses

Overlay contour plots We overlay the contour plots for the two responses to visually compare the surfaces over the region of interest.



Summary

Final response surface models

The response surface models fit to (coded) Uniformity and Stress were:

Uniformity = 5.93 - 1.91*Pressure - 0.22*H_2/WF_6 + 1.70*Pressure*H_2/WF_6

Stress = 7.73 + 0.74*Pressure + 0.50*H_2/WF_6 - 0.49*Pressure^2

Trade-offs are often needed for multiple responses

The models and the corresponding contour plots show that trade-offs have to be made when trying to achieve low values for both Uniformity and Stress since a high value of Pressure is good for Uniformity while a low value of Pressure is good for Stress. While low values of H_2/WF_6 are good for both responses, the situation is further complicated by the fact that the Peeling response (not considered in this analysis) was unacceptable for values of H_2/WF_6 below approximately 5.

Uniformity was chosen as more important

In this case, the experimenters chose to focus on optimizing Uniformity while keeping H_2/WF_6 at 5. That meant setting Pressure at 80 torr.

Confirmation runs validated the model

A set of 16 verification runs at the chosen conditions confirmed that all goals, except those for the Stress response, were met by this set of process settings.

projections

5.5. Advanced topics

Contents of "Advanced Topics" section

This section builds on the basics of DOE described in the preceding sections by adding brief or survey descriptions of a selection of useful techniques. Subjects covered are:

1. When classical designs don't work
2. Computer-aided designs
    1. D-Optimal designs
    2. Repairing a design
3. Optimizing a Process
    1. Single response case
        1. Path of steepest ascent
        2. Confidence region for search path
        3. Choosing the step length
        4. Optimization when there is adequate quadratic fit
        5. Effect of sampling error on optimal solution
        6. Optimization subject to experimental region constraints
    2. Multiple response case
        1. Path of steepest ascent
        2. Desirability function approach
        3. Mathematical programming approach
4. Mixture designs
    1. Mixture screening designs
    2. Simplex-lattice designs
    3. Simplex-Centroid designs
    4. Constrained mixture designs
    5. Treating mixture and process variables together
5. Nested variation
6. Taguchi designs
7. John's 3/4 fractional factorial designs
8. Small composite designs
9. An EDA approach to experimental design
    1. Ordered data plot
    2. DOE scatter plot
    3. DOE mean plot
    4. Interaction effects matrix plot
    5. Block plot
    6. DOE Youden plot
    7. |Effects| plot
    8. Half-normal probability plot
    9. Cumulative residual standard deviation plot
    10. DOE contour plot

5.5.1. What if classical designs don't work?

Reasons designs don't work

Most experimental situations call for standard designs that can be constructed with many statistical software packages. Standard designs have assured degrees of precision, orthogonality, and other optimal properties that are important for the exploratory nature of most experiments. In some situations, however, standard designs are not appropriate or are impractical. These may include situations where

1. The required blocking structure or blocking size of the experimental situation does not fit into a standard blocked design
2. Not all combinations of the factor settings are feasible, or for some other reason the region of experimentation is constrained or irregularly shaped.
3. A classical design needs to be 'repaired'. This can happen due to improper planning with the original design treatment combinations containing forbidden or unreachable combinations that were not considered before the design was generated.
4. A nonlinear model is appropriate.
5. A quadratic or response surface design is required in the presence of qualitative factors.
6. The factors in the experiment include both components of a mixture and other process variables.
7. There are multiple sources of variation leading to nested or hierarchical data structures and restrictions on what can be randomized.
8. A standard fractional factorial design requires too many treatment combinations for the given amount of time and/or resources.

Computer-aided designs

When situations such as the above exist, computer-aided designs are a useful option. In some situations, computer-aided designs are the only option an experimenter has.

5.5.2. What is a computer-aided design?

Computer-aided designs are generated by a computer algorithm and constructed to be optimal for certain models according to one of many types of optimality criteria	Designs generated from a computer algorithm are referred to as computer-aided designs. Computer-aided designs are experimental designs that are generated based on a particular optimality criterion and are generally 'optimal' only for a specified model. As a result, they are sometimes referred to as optimal designs and generally do not satisfy the desirable properties such as independence among the estimators that standard classical designs do. The design treatment runs that are generated by the algorithms are chosen from an overall candidate set of possible treatment combinations. The candidate set consists of all the possible treatment combinations that one wishes to consider in an experiment.
Optimality critieria	There are various forms of optimality criteria that are used to select the points for a design.
D-Optimality	One popular criterion is *D-optimality*, which seeks to maximize |X'X|, the determinant of the *information matrix* X'X of the design. This criterion results in minimizing the generalized variance of the parameter estimates based on a pre-specified model.
A-Optimality	Another criterion is *A-optimality*, which seeks to minimize the trace of the inverse of the information matrix. This criterion results in minimizing the average variance of the parameter estimates based on a pre-specified model.
G-Optimality	A third criterion is *G-optimality*, which seeks to minimize the maximum prediction variance, i.e., minimize max. $[d=x'(X'X)^{-1}x]$, over a specified set of design points.
V-Optimality	A fourth criterion is *V-optimality*, which seeks to minimize the average prediction variance over a specified set of design points.

Optimality of a given design is model dependent

Since the optimality criterion of most computer-aided designs is based on some function of the information matrix, the 'optimality' of a given design is model dependent. That is, the experimenter must specify a model for the design and the final number of design points desired before the 'optimal' design' can be generated. The design generated by the computer algorithm is 'optimal' only for that model.

5.5.2.1. D-Optimal designs

D-optimal designs are often used when classical designs do not apply

D-optimal designs are one form of design provided by a computer algorithm. These types of computer-aided designs are particularly useful when classical designs do not apply.

Unlike standard classical designs such as factorials and fractional factorials, D-optimal design matrices are usually not orthogonal and effect estimates are correlated.

These types of designs are always an option regardless of the type of model the experimenter wishes to fit (for example, first order, first order plus some interactions, full quadratic, cubic, etc.) or the objective specified for the experiment (for example, screening, response surface, etc.). D-optimal designs are straight optimizations based on a chosen optimality criterion and the model that will be fit. The optimality criterion used in generating D-optimal designs is one of maximizing |X'X|, the determinant of the information matrix X'X.

This optimality criterion results in minimizing the generalized variance of the parameter estimates for a pre-specified model. As a result, the 'optimality' of a given D-optimal design is model dependent. That is, the experimenter must specify a model for the design before a computer can generate the specific treatment combinations. Given the total number of treatment runs for an experiment and a specified model, the computer algorithm chooses the optimal set of design runs from a *candidate set* of possible design treatment runs. This candidate set of treatment runs usually consists of all possible combinations of various factor levels that one wishes to use in the experiment.

In other words, the candidate set is a collection of treatment combinations from which the D-optimal algorithm chooses the treatment combinations to include in the design. The computer algorithm generally uses a stepping and exchanging process to select the set of treatment runs.

Note: There is no guarantee that the design the computer generates is actually D-optimal.

D-optimal

The reasons for using D-optimal designs instead of standard

designs are useful when resources are limited or there are constraints on factor settings

classical designs generally fall into two categories:

1. standard factorial or fractional factorial designs require too many runs for the amount of resources or time allowed for the experiment
2. the design space is constrained (the process space contains factor settings that are not feasible or are impossible to run).

Example

Suppose an industrial process has three design variables (k = 3), and engineering judgment specifies the following model as an appropriate representation of the process.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} {X_1}^2 + \epsilon$$

The levels being considered by the researcher are (coded)

X1: 5 levels (-1, -0.5, 0, 0.5, 1)
X2: 2 levels (-1, 1)
X3: 2 levels (-1, 1)

Due to resource limitations, only n = 12 data points can be collected.

Create the candidate set

Given the experimental specifications, the first step in generating the design is to create a candidate set of points. The candidate set is a data table with a row for each point (run) to be considered for the design, often a full factorial. For our problem, the candidate set is a full factorial in all factors containing 5*2*2 = 20 possible design runs.

Table containing the candidate set

TABLE 5.1 Candidate Set for Variables X1, X2, X3

X1	X2	X3
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
-0.5	-1	-1
-0.5	-1	+1
-0.5	+1	-1
-0.5	+1	+1
0	-1	-1
0	-1	+1
0	+1	-1
0	+1	+1
0.5	-1	-1
0.5	-1	+1

0.5	+1	-1
0.5	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

Generating a D-optimal design

D-optimal designs maximize the D-efficiency, which is a volume criterion on the generalized variance of the parameter estimates. The D-efficiency of the standard fractional factorial is 100 %, but it is not possible to achieve 100 % D-efficiency when pure quadratic terms such as $(X1)^2$ are included in the model.

The D-efficiency values are a function of the number of points in the design, the number of independent variables in the model, and the maximum standard error for prediction over the design points. The best design is the one with the highest D-efficiency. Other reported efficiencies (e.g. A, G, I) help choose an optimal design when various models produce similar D-efficiencies.

D-optimal design

The D-optimal design (D=0.6825575, A=2.2, G=1, I=4.6625) using 12 runs is shown in Table 5.2 in standard order. The standard error of prediction is also shown. The design runs should be randomized before the treatment combinations are executed.

TABLE 5.2 Final D-optimal Design

X1	X2	X3	OptStdPred
-1	-1	-1	0.645497
-1	-1	+1	0.645497
-1	+1	-1	0.645497
-1	+1	+1	0.645497
0	-1	-1	0.645497
0	-1	+1	0.645497
0	+1	-1	0.645497
0	+1	+1	0.645497
+1	-1	-1	0.645497
+1	-1	+1	0.645497
+1	+1	-1	0.645497
+1	+1	+1	0.645497

Software note

Software packages may have different procedures for generating D-optimal designs, so the final design may be different depending on the software packaged being used.

5.5.2.2. Repairing a design

Repair or augment classical designs	Computer-aided designs are helpful in either repairing or augmenting a current experimental design. They can be used to repair a 'broken' standard classical design.
Original design matrix may contain runs that were lost or impossible to acieve	There may be situations in which, due to improper planning or other issues, the original design matrix contains forbidden or unreachable combinations of the factor settings. A computer-aided design (for example a D-optimal design) can be used to 'replace' those runs from the original design that were unattainable. The runs from the original design that are attainable are labeled as 'inclusion' runs and will be included in the final computer-aided design.
Computer-aided design can generate additional attainable runs	Given a pre-specified model, the computer-aided design can generate the additional attainable runs that are necessary in order to estimate the model of interest. As a result, the computer-aided design is just replacing those runs in the original design that were unattainable with a new set of runs that are attainable, and which still allows the experimenter to obtain information regarding the factors from the experiment.
Properties of this final design may not compare with those of the original design	The properties of this final design will probably not compare with those of the original design and there may exist some correlation among the estimates. However, instead of not being able to use any of the data for analysis, generating the replacement runs from a computer-aided design, a D-optimal design for example, allows one to analyze the data. Furthermore, computer-aided designs can be used to augment a classical design with treatment combinations that will break alias chains among the terms in the model or permit the estimation of curvilinear effects.

5.5.3. How do you optimize a process?

How do you determine the optimal region to run a process?

Often the primary DOE goal is to find the operating conditions that maximize (or minimize) the system responses

The optimal region to run a process is usually determined after a sequence of experiments has been conducted and a series of empirical models obtained. In many engineering and science applications, experiments are conducted and empirical models are developed with the objective of improving the responses of interest. From a mathematical point of view, the objective is to find the operating conditions (or factor levels) X_1, X_2, ..., X_k that maximize or minimize the r system response variables Y_1, Y_2, ..., Y_r. In experimental optimization, different optimization techniques are applied to the *fitted* response equations $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_r$.

Provided that the fitted equations approximate adequately the true (unknown) system responses, the optimal operating conditions of the model will be "close" to the optimal operating conditions of the true system.

The DOE approach to optimization

The experimental optimization of response surface models differs from classical optimization techniques in at least three ways:

Find approximate (good) models and iteratively search for (near) optimal operating conditions

1. Experimental optimization is an iterative process; that is, experiments conducted in one set of experiments result in fitted models that indicate where to search for improved operating conditions in the next set of experiments. Thus, the coefficients in the fitted equations (or the form of the fitted equations) may change during the optimization process. This is in contrast to classical optimization in which the functions to optimize are supposed to be fixed and given.

Randomness (sampling variability) affects the final

2. The response models are fit from experimental data that usually contain random variability due to uncontrollable or unknown causes. This implies that an experiment, if repeated, will result in a different fitted response surface model that might lead to

answers and should be taken into account	different optimal operating conditions. Therefore, sampling variability should be considered in experimental optimization. In contrast, in classical optimization techniques the functions are deterministic and given.
Optimization process requires input of the experimenter	3. The fitted responses are local approximations, implying that the optimization process requires the input of the experimenter (a person familiar with the process). This is in contrast with classical optimization which is always automated in the form of some computer algorithm.

5.5.3.1. Single response case

Optimizing of a single response usually starts with line searches in the direction of maximum improvement

The experimental optimization of a single response is usually conducted in two phases or steps, following the advice of Box and Wilson. The first phase consists of a sequence of line searches in the direction of maximum improvement. Each search in the sequence is continued until there is evidence that the direction chosen does not result in further improvements. The sequence of line searches is performed as long as there is no evidence of lack of fit for a simple first-order model of the form

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

If there is lack of fit for linear models, quadratic models are tried next

The second phase is performed when there is lack of linear fit in Phase I, and instead, a second-order or quadratic polynomial regression model of the general form

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + b_{11} X_1^2 + b_{22} X_2^2 + \ldots + b_{kk} X_k^2 + b_{12} X_1 X_2 + b_{13} X_1 X_3 + \ldots + b_{1k} X_1 X_k + b_{23} X_2 X_3 + \ldots + b_{2k} X_2 X_k + \ldots + b_{k-1,k} X_{k-1} X_k$$

is fit. Not all responses will require quadratic fit, and in such cases Phase I is stopped when the response of interest cannot be improved any further. Each phase is explained and illustrated in the next few sections.

"Flowchart" for two phases of experimental optimization

The following is a flow chart showing the two phases of experimental optimization.

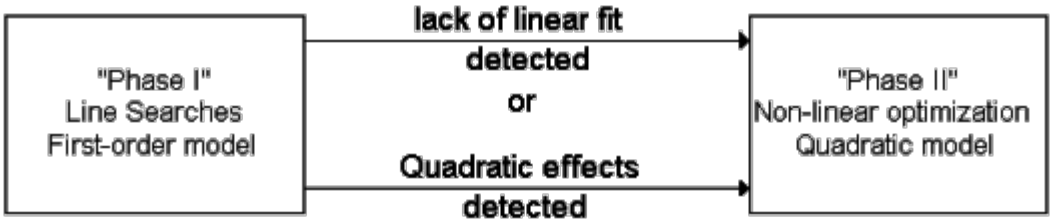


FIGURE 5.1: The Two Phases of Experimental Optimization

5. Process Improvement
5.5. Advanced topics
5.5.3. How do you optimize a process?
5.5.3.1. Single response case

5.5.3.1.1. Single response: Path of steepest ascent

Starting at the current operating conditions, fit a linear model

If experimentation is initially performed in a new, poorly understood production process, chances are that the initial operating conditions X_1, X_2, ...,X_k are located far from the region where the factors achieve a maximum or minimum for the response of interest, **Y**. A first-order model will serve as a good local approximation in a small region close to the initial operating conditions and far from where the process exhibits curvature. Therefore, it makes sense to fit a simple first-order (or linear polynomial) model of the form:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

Experimental strategies for fitting this type of model were discussed earlier. Usually, a 2^{k-p} fractional factorial experiment is conducted with repeated runs at the current operating conditions (which serve as the origin of coordinates in orthogonally coded factors).

Determine the directions of steepest ascent and continue experimenting until no further improvement occurs - then iterate the process

The idea behind "Phase I" is to keep experimenting along the direction of steepest ascent (or descent, as required) until there is no further improvement in the response. At that point, a new fractional factorial experiment with center runs is conducted to determine a new search direction. This process is repeated until at some point significant curvature in $\hat{Y}$ is detected. This implies that the operating conditions X_1, X_2, ...,X_k are close to where the maximum (or minimum, as required) of **Y** occurs. When significant curvature, or lack of fit, is detected, the experimenter should proceed with "Phase II". Figure 5.2 illustrates a sequence of line searches when seeking a region where curvature exists in a problem with 2 factors (i.e., $k=2$).

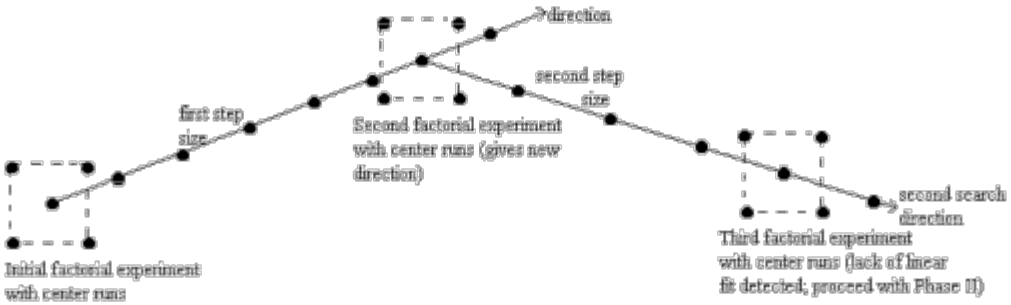


FIGURE 5.2: A Sequence of Line Searches for a 2-Factor Optimization Problem

Two main decisions: search direction and length of step

There are two main decisions an engineer must make in Phase I:

1. determine the search direction;
2. determine the length of the step to move from the current operating conditions.

Figure 5.3 shows a flow diagram of the different iterative tasks required in Phase I. This diagram is intended as a guideline and should not be automated in such a way that the experimenter has no input in the optimization process.

Flow chart of iterative search process

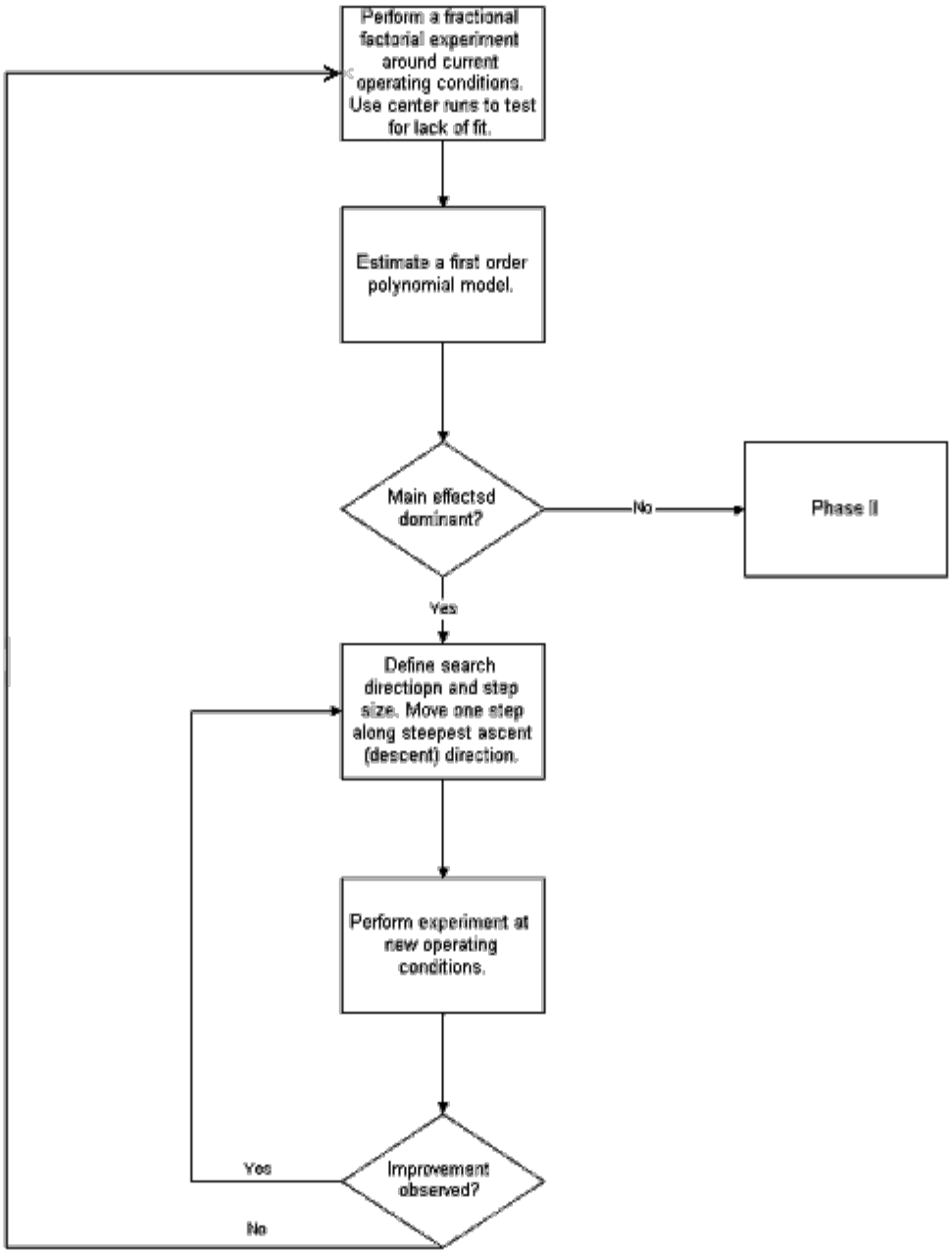


FIGURE 5.3: Flow Chart for the First Phase of the Experimental Optimization Procedure

Procedure for Finding the Direction of Maximum Improvement

The direction of steepest ascent is

Suppose a first-order model (like above) has been fit and provides a useful approximation. As long as lack of fit (due to pure quadratic curvature and interactions) is very small compared to the main effects, steepest ascent can be

determined by the gradient of the fitted model

attempted. To determine the direction of maximum improvement we use

1. the estimated direction of steepest ascent, given by the gradient of $\hat{Y}$, if the objective is to *maximize* **Y**;
2. the estimated direction of steepest descent, given by the negative of the gradient of $\hat{Y}$, if the objective is to *minimize* **Y**.

The direction of steepest ascent depends on the scaling convention - equal variance scaling is recommended

The direction of the gradient, $\boldsymbol{g}$, is given by the values of the parameter estimates, that is, $\boldsymbol{g' = (b_1, b_2, ..., b_k)}$. Since the parameter estimates $\boldsymbol{b_1, b_2, ..., b_k}$ depend on the scaling convention for the factors, the steepest ascent (descent) direction is also scale dependent. That is, two experimenters using different scaling conventions will follow different paths for process improvement. This does not diminish the general validity of the method since the region of the search, as given by the signs of the parameter estimates, does not change with scale. An equal variance scaling convention, however, is recommended. The coded factors $\boldsymbol{x_i}$, in terms of the factors in the original units of measurement, $\boldsymbol{X_i}$, are obtained from the relation

$$x_i = \frac{X_i - (X_{low} + X_{high})/2}{(X_{high} - X_{low})/2} \qquad i = 1, 2, ..., k$$

This coding convention is recommended since it provides parameter estimates that are scale independent, generally leading to a more reliable search direction. The coordinates of the factor settings in the direction of steepest ascent, positioned a distance ρ from the origin, are given by:

$$\text{maximize} \quad b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

$$\text{subject to:} \quad \sum_{i=1}^{k} x_i^2 \leq \rho^2$$

Solution is a simple equation

This problem can be solved with the aid of an optimization solver (e.g., like the solver option of a spreadsheet). However, in this case this is not really needed, as the solution is a simple equation that yields the coordinates

$$x_i^* = \rho \frac{b_i}{\sqrt{\sum_{j=1}^{k} b_j^2}} \qquad i = 1, 2, ..., k.$$

Equation can be computed for increasing values of ρ

An engineer can compute this equation for different increasing values of ρ and obtain different factor settings, all on the steepest ascent direction.

To see the details that explain this equation, see Technical Appendix 5A.

Example: Optimization of a Chemical Process

Optimization by search example

It has been concluded (perhaps after a factor screening experiment) that the yield (**Y**, in %) of a chemical process is mainly affected by the temperature ($\boldsymbol{X_1}$, in ^{o}C) and by the reaction time ($\boldsymbol{X_2}$, in minutes). Due to safety reasons, the region of

operation is limited to

$$50 \le X_1 \le 250$$
$$150 \le X_2 \le 500$$

Factor levels

The process is currently run at a temperature of 200 °C and a reaction time of 200 minutes. A process engineer decides to run a 2^2 full factorial experiment with factor levels at

factor	low	center	high
X_1	170	200	230
X_2	150	200	250

Orthogonally coded factors

Five repeated runs at the center levels are conducted to assess lack of fit. The orthogonally coded factors are

$$x_1 = \frac{X_1 - 200}{30} \quad \text{and} \quad x_2 = \frac{X_2 - 200}{50}.$$

Experimental results

The experimental results were:

x_1	x_2	X_1	X_2	Y (= yield)
-1	-1	170	150	32.79
+1	-1	230	150	24.07
-1	+1	170	250	48.94
+1	+1	230	250	52.49
0	0	200	200	38.89
0	0	200	200	48.29
0	0	200	200	29.68
0	0	200	200	46.50
0	0	200	200	44.15

ANOVA table

The corresponding ANOVA table for a first-order polynomial model is

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F VALUE	PROB>F
MODEL	503.3035	2	251.6517	4.7972	0.0687
CURVATURE	8.2733	1	8.2733	0.1577	0.7077
RESIDUAL	262.2893	5	52.4579		
LACK OF FIT	37.6382	1	37.6382	0.6702	0.4590
PURE ERROR	224.6511	4	56.1628		
COR TOTAL	773.8660	8			

Resulting model

It can be seen from the ANOVA table that there is no significant lack of linear fit due to an interaction term and there is no evidence of curvature. Furthermore,

there is evidence that the first-order model is significant. The resulting model (in the coded variables) is

$$\hat{Y} = 40.644 - 1.2925x_1 + 11.14x_2$$

Diagnostic checks

The usual diagnostic checks show conformance to the regression assumptions, although the R^2 value is not very high: $R^2 = 0.6504$.

Determine level of factors for next run using direction of steepest ascent

To *maximize* $\hat{Y}$, we use the direction of steepest ascent. The engineer selects $\rho = 1$ since a point on the steepest ascent direction one unit (in the coded units) from the origin is desired. Then from the equation above for the predicted *Y* response, the coordinates of the factor levels for the next run are given by:

$$x_1^* = \frac{\rho b_1}{\sqrt{\sum_{j=1}^{2} b_j^2}} = \frac{(1)(-1.2925)}{\sqrt{(-1.2925)^2 + (11.14)^2}} = -0.1152$$

and

$$x_2^* = \frac{\rho b_2}{\sqrt{\sum_{j=1}^{2} b_j^2}} = \frac{(1)(11.14)}{\sqrt{(-1.2925)^2 + (11.14)^2}} = 0.9933$$

This means that to improve the process, for every (-0.1152)(30) = -3.456 °C that temperature is varied (decreased), the reaction time should be varied by (0.9933)(50) = 49.66 minutes.

==========

Technical Appendix 5A: finding the factor settings on the steepest ascent direction a specified distance from the origin

Details of how to determine the path of steepest ascent

The problem of finding the factor settings on the steepest ascent/descent direction that are located a distance ρ from the origin is given by the optimization problem,

$$\text{maximize} \quad b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

$$\text{subject to:} \quad \sum_{i=1}^{k} x_i^2 \leq \rho^2$$

Solve using a Lagrange multiplier approach

To solve it, use a Lagrange multiplier approach. First, add a penalty λ for solutions not satisfying the constraint (since we want a direction of steepest ascent, we maximize, and therefore the penalty is negative). For steepest descent we minimize and the penalty term is added instead.

$$\text{maximize} \quad L = b'x - \lambda(x'x - \rho^2)$$

Compute the partials and equate them to zero

$$\frac{\partial L}{\partial x} = b - 2\lambda x = 0$$

$$\frac{\partial L}{\partial \lambda} = -(x'x - \rho^2) = 0$$

Solve two equations in two unknowns

These two equations have two unknowns (the vector x and the scalar λ) and thus can be solved yielding the desired solution:

$$x^* = \rho \frac{b}{||b||}$$

or, in non-vector notation:

$$x_i^* = \rho \frac{b_i}{\sqrt{\sum_{j=1}^{k} b_j^2}} \qquad i = 1, 2, ..., k$$

Multiples of the direction of the gradient

From this equation we can see that any multiple ρ of the direction of the gradient (given by $b/||b||$) will lead to points on the steepest ascent direction. For steepest descent, use instead $-b_i$ in the numerator of the equation above.

5. Process Improvement
5.5. Advanced topics
5.5.3. How do you optimize a process?
5.5.3.1. Single response case

5.5.3.1.2. Single response: Confidence region for search path

"Randomness" means that the steepest ascent direction is just an estimate and it is possible to construct a confidence "cone" around this direction estimate

The direction given by the gradient $\boldsymbol{g}' = (b_0, b_2, \ldots, b_k)$ constitutes only a single (point) estimate based on a sample of **N** runs. If a different set of **N** runs were conducted, these would provide different parameter estimates, which in turn would give a different gradient. To account for this sampling variability, Box and Draper gave a formula for constructing a "cone" around the direction of steepest ascent that with certain probability contains the true (unknown) system gradient given by $(\beta_1, \beta_2, \ldots, \beta_k)$. The width of the confidence cone is useful to assess how reliable an estimated search direction is.

Figure 5.4 shows such a cone for the steepest ascent direction in an experiment with two factors. If the cone is so wide that almost every possible direction is inside the cone, an experimenter should be very careful in moving too far from the current operating conditions along the path of steepest ascent or descent. Usually this will happen when the linear fit is quite poor (i.e., when the R^2 value is low). Thus, plotting the confidence cone is not so important as computing its width.

If you are interested in the details on how to compute such a cone (and its width), see Technical Appendix 5B.

Graph of a confidence cone for the steepest ascent direction

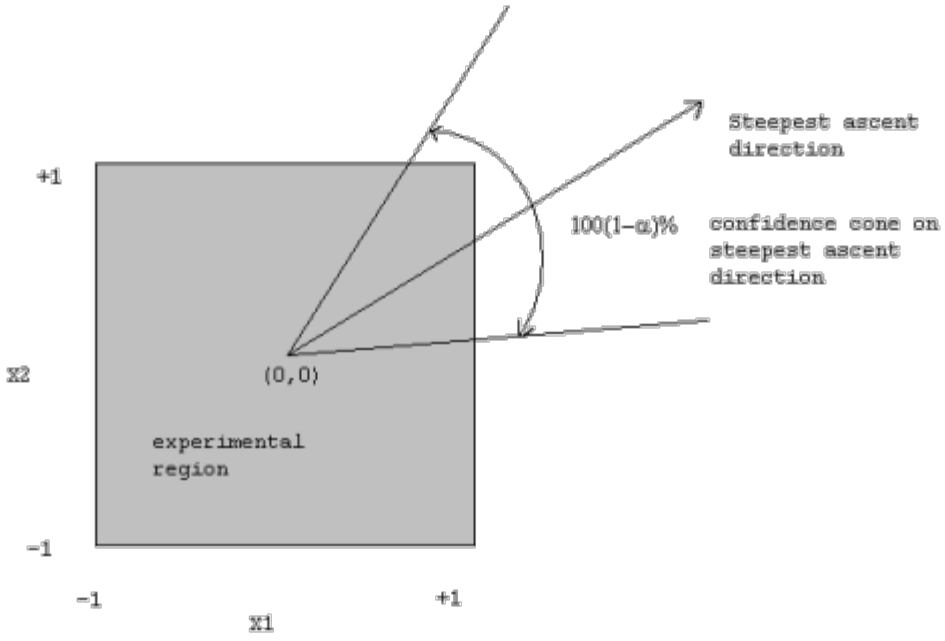


FIGURE 5.4: A Confidence Cone for the Steepest Ascent Direction in an Experiment with 2 Factors

==========

Technical Appendix 5B: Computing a Confidence Cone on the Direction of Steepest Ascent

Details of how to construct a confidence cone for the direction of steepest ascent

Suppose the response of interest is adequately described by a first-order polynomial model. Consider the inequality

$$\sum_{i=1}^{k} b_i^2 - \frac{\left(\sum_{i=1}^{k} b_i x_i\right)^2}{\sum_{i=1}^{k} x_i^2} \leq (k-1) s_b^2 F_{\alpha,k-1,n-p}$$

with

$$s_b^2 = SS_{error} \frac{C_{jj}}{n-p}$$

C_{jj} is the j-th diagonal element of the matrix $(X'X)^{-1}$ (for $j = 1, ..., k$ these values are all equal if the experimental design is a 2^{k-p} factorial of at least Resolution III), and X is the model matrix of the experiment (including columns for the intercept and second-order terms, if any). Any operating condition with coordinates $x' = (x_1, x_2, ..., x_k)$ that satisfies this inequality generates a direction that lies within the 100(1- α) % confidence cone of steepest *ascent* if

$$\sum_{i=1}^{k} b_i x_i > 0$$

or inside the 100(1- α) % confidence cone of steepest *descent* if

$$\sum_{i=1}^{k} b_i x_i < 0.$$

Inequality defines a cone

The inequality defines a cone with the apex at the origin and center line located along the gradient of $\hat{Y}$.

A measure of goodnes of fit: θ_α

A measure of "goodness" of a search direction is given by the fraction of directions *excluded* by the 100(1- α) % confidence cone around the steepest ascent/descent direction (see Box and Draper, 1987) which is given by:

$$\theta_\alpha = 1 - \phi_\alpha = 1 - T_{k-1}\left(\frac{\sum_{i=1}^{k} b_i^2}{s_b^2 F_{\alpha,k-1,n-p}} - (k-1)\right)^{1/2}$$

with $T_{k-1}()$ denoting the complement of the Student's *t* distribution function with $k-1$ degrees of freedom (that is, $T_{k-1}(x) = P(t_{k-1} \geq x)$) and $F_{\alpha, k-1, n-p}$ denotes an α percentage point of the *F* distribution with $k-1$ and $n-p$ degrees of freedom, with $n-p$ denoting the error degrees of freedom. The value of ϕ_α represents the fraction of directions included by the confidence cone. The smaller θ_α is, the wider the cone is, with $0 \leq \theta_\alpha \leq 1$. Note that the inequality equation and the "goodness measure" equation are valid when operating conditions are given in coded units.

Example: Computing θ_α

Compute s_b^2 from ANOVA table and C_{jj}

From the ANOVA table in the chemical experiment discussed earlier

$$s_b^2 = \frac{1}{4}(52.4579) = 13.1145$$

since $C_{jj} = 1/4$ (j=2,3) for a 2^2 factorial. The fraction of directions excluded by a 95 % confidence cone in the direction of steepest ascent is:

Compute θ_α

$$\theta_{0.05} = 1 - T_1\left[\frac{(-1.2925)^2 + (11.14)^2}{(13.1145)(5.99)} - 1\right]^{0.5}$$

$$\theta_{0.05} = 1 - 0.29 = 0.71$$

Conclusions for this example

since $F_{0.05,1,6} = 5.99$. Thus 71 % of the possible directions from the current operating point are excluded with 95 % confidence. This is useful information that

can be used to select a step length. The smaller θ_α is, the shorter the step should be, as the steepest ascent direction is less reliable. In this example, with high confidence, the true steepest ascent direction is within this cone of 29 % of possible directions. For k=2, 29 % of 360^o = 104.4^o, so we are 95 % confident that our estimated steepest ascent path is within plus or minus 52.2^o of the true steepest path. In this case, we should not use a large step along the estimated steepest ascent path.

5. Process Improvement
5.5. Advanced topics
5.5.3. How do you optimize a process?
5.5.3.1. Single response case

5.5.3.1.3. Single response: Choosing the step length

A procedure for choosing how far along the direction of steepest ascent to go for the next trial run

Once the search direction is determined, the second decision needed in Phase I relates to how far in that direction the process should be "moved". The most common procedure for selecting a step length is based on choosing a step size in one factor and then computing step lengths in the other factors proportional to their parameter estimates. This provides a point on the direction of maximum improvement. The procedure is given below. A similar approach is obtained by choosing increasing values of ρ in

$$x_i^* = \rho \frac{b_i}{\sqrt{\sum_{j=1}^{k} b_j^2}} \quad i = 1, 2, \ldots, k.$$

However, the procedure below considers the original units of measurement which are easier to deal with than the coded "distance" ρ.

Procedure: selection of step length

Procedure for selecting the step length

The following is the procedure for selecting the step length.

1. Choose a step length ΔX_j (in natural units of measurement) for some factor j. Usually, factor j is chosen to be the one engineers feel more comfortable varying, or the one with the largest $|b_j|$. The value of ΔX_j can be based on the width of the confidence cone around the steepest ascent/descent direction. Very wide cones indicate that the estimated steepest ascent/descent direction is not reliable, and thus ΔX_j should be small. This usually occurs when the R^2 value is low. In such a case, additional experiments can be conducted in the current experimental region to obtain a better model fit and a better search direction.

2. Transform to coded units:

$$\Delta x_j = \frac{\Delta X_j}{s_j}$$

with s_j denoting the scale factor used for factor j (e.g., s_j = **range**$_j$/**2**).

3. Set $\Delta x_i = \frac{b_i}{b_j}\Delta x_j$ for all other factors i.

4. Transform all the Δx_i's to natural units: $\Delta X_i = (\Delta x_i)(s_i)$.

Example: Step Length Selection.

An example of step length selection

The following is an example of the step length selection procedure.

- For the chemical process experiment described previously, the process engineer selected $\Delta X_2 = 50$ minutes. This was based on process engineering considerations. It was also felt that $\Delta X_2 = 50$ does not move the process too far away from the current region of experimentation. This was desired since the R^2 value of 0.6580 for the fitted model is quite low, providing a not very reliable steepest ascent direction (and a wide confidence cone, see Technical Appendix 5B).
- $\Delta x_2 = \frac{50}{50} = 1.0$
- $\Delta x_1 = \frac{-1.2925}{11.14} = -0.1160$
- $\Delta X_2 = (-0.1160)(30) = -3.48^{o}C$.

Thus the step size is $\Delta X' = (-3.48^{o}C, 50 \text{ minutes})$.

Procedure: Conducting Experiments Along the Direction of Maximum Improvement

Procedure for conducting experiments along the direction of maximum improvement

The following is the procedure for conducting experiments along the direction of maximum improvement.

1. Given current operating conditions $X_0' = (X_1, X_2, ..., X_k)$ and a step size $\Delta X' = (\Delta X_1, \Delta X_2, ..., \Delta X_k)$, perform experiments at factor levels $X_0 + \Delta X$, $X_0 + 2\Delta X$, $X_0 + 3\Delta X$, ... as long as improvement in the response Y (decrease or increase, as desired) is

observed.

2. Once a point has been reached where there is no further improvement, a new first-order experiment (e.g., a 2^{k-p} fractional factorial) should be performed with repeated center runs to assess lack of fit. If there is no significant evidence of lack of fit, the new first-order model will provide a new search direction, and another iteration is performed as indicated in Figure 5.3. Otherwise (there is evidence of lack of fit), the experimental design is augmented and a second-order model should be fitted. That is, the experimenter should proceed to "Phase II".

Example: Experimenting Along the Direction of Maximum Improvement

Step 1: increase factor levels by Δ

Step 1:

Given $\boldsymbol{X_0}$ = (200°C, 200 minutes) and $\Delta\boldsymbol{X}$ = (-3.48°C, 50 minutes), the next experiments were performed as follows (the step size in temperature was rounded to -3.5°C for practical reasons):

	X_1	X_2	x_1	x_2	Y (= yield)
X_0	200	200	0	0	
$X_0 + \Delta X$	196.5	250	-0.1160	1	56.2
$X_0 + 2\Delta X$	193.0	300	-0.2320	2	71.49
$X_0 + 3\Delta X$	189.5	350	-0.3480	3	75.63
$X_0 + 4\Delta X$	186.0	400	-0.4640	4	72.31
$X_0 + 5\Delta X$	182.5	450	-0.5800	5	72.10

Since the goal is to maximize $\boldsymbol{Y}$, the point of maximum observed response is $\boldsymbol{X_1}$ = 189.5°C, $\boldsymbol{X_2}$ = 350 minutes. Notice that the search was stopped after 2 consecutive drops in response, to assure that we have passed by the "peak" of the "hill".

Step 2: new factorial experiment

Step 2:

A new 2^2 factorial experiment is performed with $\boldsymbol{X}$' = **(189.5, 350)** as the origin. Using the same scaling factors as before, the new scaled controllable factors are:

$$x_1 = \frac{X_1 - 189.5}{30} \quad \text{and} \quad x_2 = \frac{X_2 - 350}{50}$$

Five center runs (at $X_1 = 189.5, X_2 = 350$) were repeated to assess lack of fit. The experimental results were:

x_1	x_2	X_1	X_2	Y (= yield)
-1	-1	159.5	300	64.33
+1	-1	219.5	300	51.78
-1	+1	159.5	400	77.30
+1	+1	219.5	400	45.37
0	0	189.5	350	62.08
0	0	189.5	350	79.36
0	0	189.5	350	75.29
0	0	189.5	350	73.81
0	0	189.5	350	69.45

The corresponding ANOVA table for a linear model is

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F VALUE	PROB > F
MODEL	505.376	2	252.688	4.731	0.0703
CURVATURE	336.364	1	336.364	6.297	0.0539
RESIDUAL	267.075	5	53.415		
LACK OF FIT	93.896	1	93.896	2.168	0.2149
PURE ERROR	173.179	4	43.295		
COR TOTAL	1108.815	8			

From the table, the linear effects (model) are significant and there is no evidence of lack of fit. However, there is a significant curvature effect (at the 5.4 % significance level), which implies that the optimization should proceed with Phase II; that is, the fit and optimization of a second-order model.

5.5.3.1.4. Single response: Optimization when there is adequate quadratic fit

Regions where quadratic models or even cubic models are needed occur in many instances in industry

After a few steepest ascent (or descent) searches, a first-order model will eventually lead to no further improvement or it will exhibit lack of fit. The latter case typically occurs when operating conditions have been changed to a region where there are quadratic (second-order) effects present in the response. A second-order polynomial can be used as a local approximation of the response in a small region where, hopefully, optimal operating conditions exist. However, while a quadratic fit is appropriate in most of the cases in industry, there will be a few times when a quadratic fit will not be sufficiently flexible to explain a given response. In such cases, the analyst generally does one of the following:

1. Uses a transformation of Y or the X_is to improve the fit.
2. Limits use of the model to a smaller region in which the model does fit.
3. Adds other terms to the model.

Procedure: obtaining the estimated optimal operating conditions

Second-order polynomial model

Once a linear model exhibits lack of fit or when significant curvature is detected, the experimental design used in Phase I (recall that a 2^{k-p} factorial experiment might be used) should be augmented with axial runs on each factor to form what is called a *central composite design*. This experimental design allows estimation of a second-order polynomial of the form

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum_{i<j}^{k} \sum_{j=1}^{k} b_{ij} x_i x_j$$

Steps to find optimal operating conditions

If the corresponding analysis of variance table indicates no lack of fit for this model, the engineer can proceed to determine the estimated optimal operating conditions.

1. Using some graphics software, obtain a contour plot of the fitted response. If the number of factors (k) is greater than 2, then plot contours in all planes corresponding to all the possible pairs of factors. For k greater than, say, 5, this could be too cumbersome (unless the graphic software plots all pairs automatically). In such a case, a "canonical analysis" of the surface is recommended (see Technical Appendix 5D).

2. Use an optimization solver to maximize or minimize (as desired) the estimated

response $\hat{Y}$.

3. Perform a confirmation experiment at the estimated optimal operating conditions given by the solver in step 2.

Chemical experiment example

We illustrate these steps using the chemical experiment discussed previously. For a technical description of a formula that provides the coordinates of the stationary point of the surface, see Technical Appendix 5C.

Example: Second Phase Optimization of Chemical Process

Experimental results for axial runs

Recall that in the chemical experiment, the ANOVA table, obtained from using an experiment run around the coordinates $X_1 = 189.5$, $X_2 = 350$, indicated significant curvature effects. Augmenting the 2^2 factorial experiment with axial runs at $\pm\alpha = \pm\sqrt{2}$ to achieve a rotatable central composite experimental design, the following experimental results were obtained:

x_1	x_2	X_1	X_2	Y (= yield)
-1.414	0	147.08	350	72.58
+1.414	0	231.92	350	37.42
0	-1.414	189.5	279.3	54.63
0	+1.414	189.5	420.7	54.18

ANOVA table

The ANOVA table corresponding to a cubic model with an interaction term (contained in the quadratic sum-of-squares partition) is

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F VALUE	PROB > F
MEAN	51418.2	1	51418.2		
Linear	1113.7	2	556.8	5.56	0.024
Quadratic	768.1	3	256.0	7.69	0.013
Cubic	9.9	2	5.0	0.11	0.897
RESIDUAL	223.1	5	44.6		
TOTAL	53533.0	13			

Lack-of-fit tests and auxillary diagnostic statistics

From the ANOVA table, the linear and quadratic effects are significant. The lack-of-fit tests and auxiliary diagnostic statistics for linear, quadratic, and cubic models are:

MODEL	SUM OF SQUARES	DF	MEAN SQUARE	F VALUE	PROB > F
Linear	827.9	6	138.0	3.19	0.141
Quadratic	59.9	3	20.0	0.46	0.725
Cubic	49.9	1	49.9	1.15	0.343
PURE ERROR	173.2	4	43.3		

MODEL	ROOT MSE	R-SQR	ADJ R-SQR	PRED R-SQR	PRESS
Linear	10.01	0.5266	0.4319	0.2425	1602.02
Quadratic	5.77	0.8898	0.8111	0.6708	696.25
Cubic	6.68	0.8945	0.7468	-0.6393	3466.71

The quadratic model has a larger *p*-value for the lack of fit test, higher adjusted R^2, and a

lower PRESS statistic; thus it should provide a reliable model. The fitted quadratic equation, in coded units, is

$$\hat{Y} = 72.0 - 11.78x_1 + 0.74x_2 - 7.25x_1^2 - 7.55x_2^2 - 4.85x_1x_2$$

Step 1:

Contour plot of the fitted response function

A contour plot of this function (Figure 5.5) shows that it appears to have a single optimum point in the region of the experiment (this optimum is calculated below to be (-0.9285, 0.3472), in coded x_1, x_2 units, with a predicted response value of 77.59).

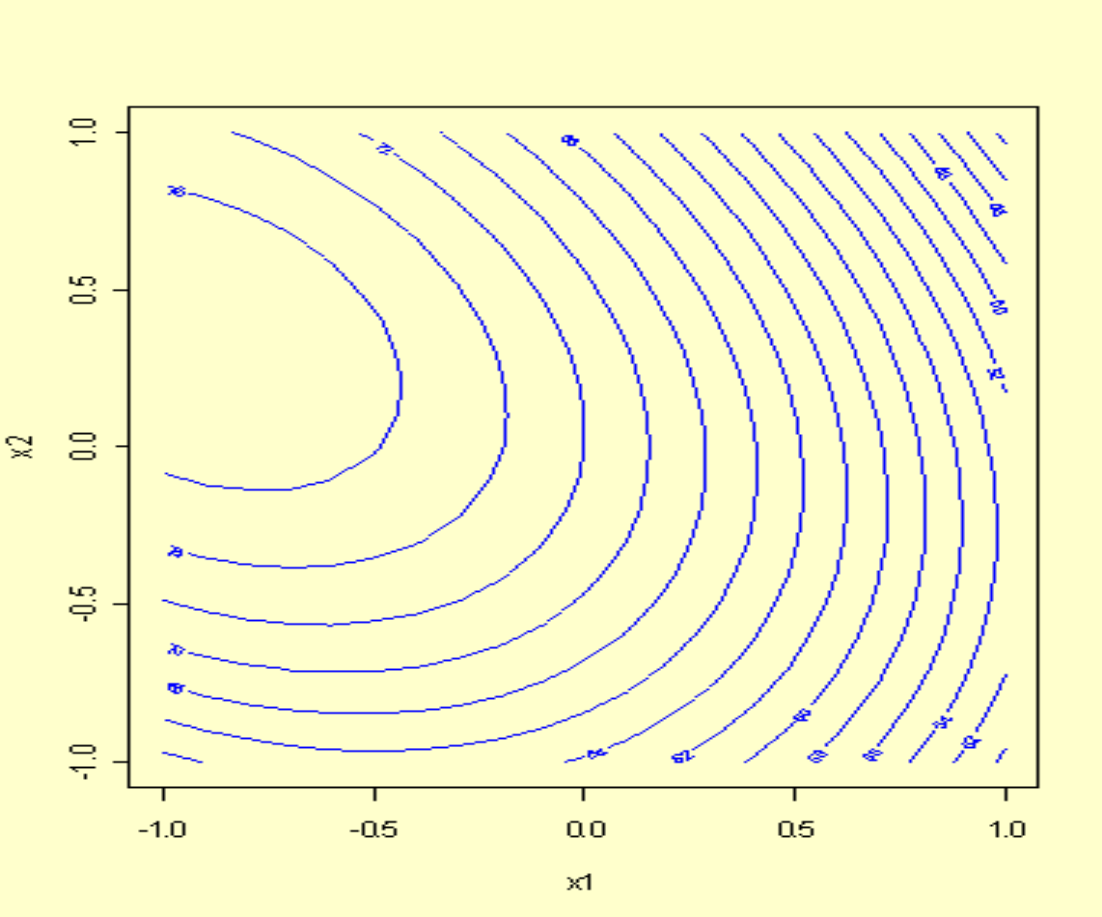


FIGURE 5.5: Contour Plot of the Fitted Response in the Example

3D plot of the fitted response function

Since there are only two factors in this example, we can also obtain a 3D plot of the fitted response against the two factors (Figure 5.6).

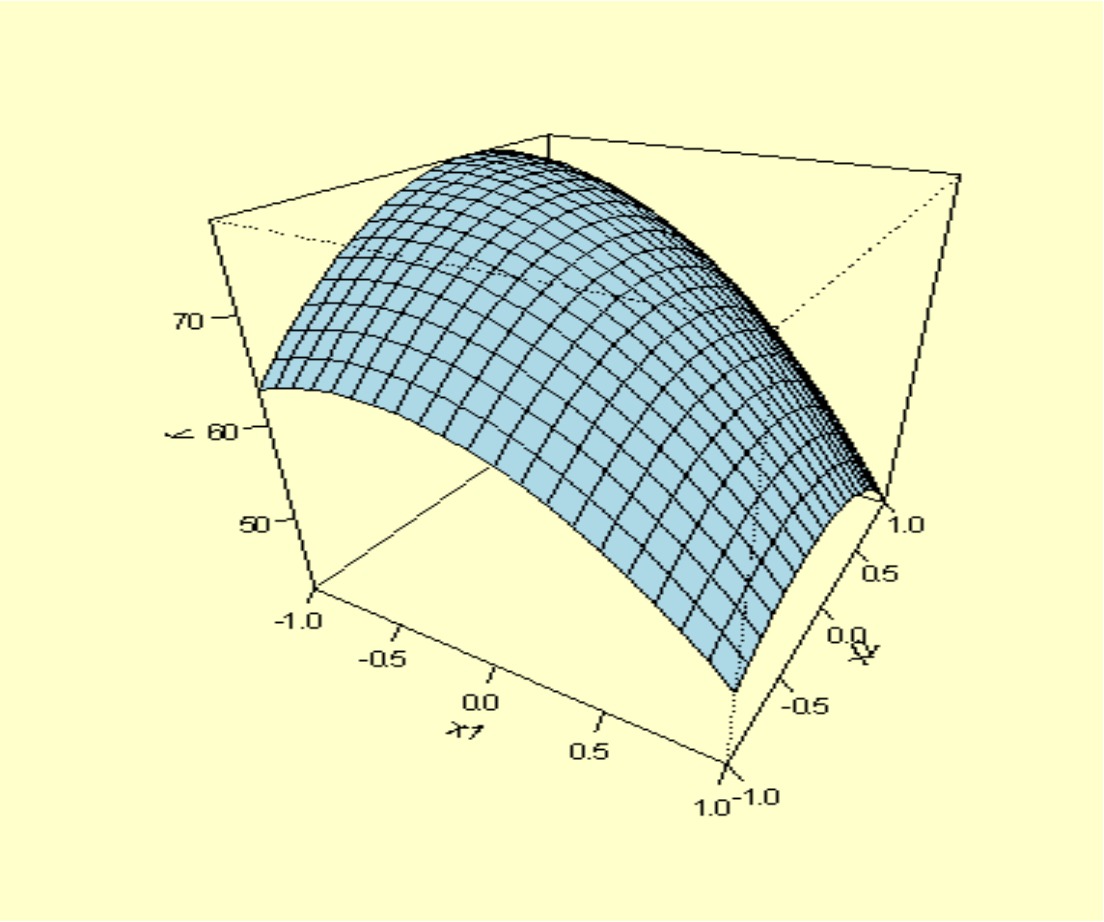


FIGURE 5.6: 3D Plot of the Fitted Response in the Example

Step 2:

Optimization point

An optimization routine was used to maximize $\hat{Y}$. The results are $X_1^* = 161.64^oC$, $X_2^* = 367.32$ minutes. The estimated yield at the optimal point is $\hat{Y}(X^*) = 77.59\ \%$.

Step 3:

Confirmation experiment

A confirmation experiment was conducted by the process engineer at settings $X_1 = 161.64$, $X_2 = 367.32$. The observed response was $\hat{Y}(X^*) = 76.5\ \%$, which is satisfactorily close to the estimated optimum.

==========

Technical Appendix 5C: Finding the Factor Settings for the Stationary Point of a Quadratic Response

How to find the maximum or minimum point for a quadratic response

1. Rewrite the fitted equation using matrix notation as

$$\hat{Y}(\mathbf{x}) = b_o + \mathbf{b}'\mathbf{x} + \mathbf{x}'\mathbf{B}\mathbf{x}$$

where $\mathbf{b}' = (b_1, b_2, ..., b_k)$ denotes a vector of first-order parameter estimates,

$$B = \begin{pmatrix} b_{11} & b_{12}/2 & \cdots & b_{ik}/2 \\ & b_{22} & & \\ & & \ddots & \vdots \\ \text{symmetric} & & & b_{kk} \end{pmatrix}$$

is a matrix of second-order parameter estimates and $x' = (x_1, x_2, ..., x_k)$ is the vector of controllable factors. Notice that the off-diagonal elements of B are equal to half the two-factor interaction coefficients.

2. Equating the partial derivatives of $\hat{Y}$ with respect to x to zeroes and solving the resulting system of equations, the coordinates of the stationary point of the response are given by

$$x^* = -\frac{1}{2}B^{-1}b$$

Nature of the stationary point is determined by B

The nature of the stationary point (whether it is a point of maximum response, minimum response, or a saddle point) is determined by the matrix B. The two-factor interactions do not, in general, let us "see" what type of point x^* is. One thing that can be said is that if the diagonal elements of B (b_{ii}) have mixed signs, x^* is a saddle point. Otherwise, it is necessary to look at the characteristic roots or eigenvalues of B to see whether B is "positive definite" (so x^* is a point of minimum response) or "negative definite" (the case in which x^* is a point of maximum response). This task is easier if the two-factor interactions are "eliminated" from the fitted equation as is described in Technical Appendix 5D.

Example: computing the stationary point, Chemical Process experiment

Example of computing the stationary point

The fitted quadratic equation in the chemical experiment discussed in Section 5.5.3.1.1 is, in coded units,

$$\hat{Y} = 72.0 - 11.78x_1 + 0.74x_2 - 7.25x_1^2 - 7.55x_2^2 - 4.85x_1x_2$$

from which we obtain $b' = (-11.78, 0.74)$,

$$B = \begin{pmatrix} -7.25 & -2.2425 \\ -2.425 & -7.55 \end{pmatrix} \; ; \; B^{-1} = \begin{pmatrix} -0.1545 & 0.0496 \\ 0.0496 & -0.1483 \end{pmatrix}$$

and

$$x^* = -\frac{1}{2}\begin{pmatrix} -0.1545 & 0.0496 \\ 0.0496 & -0.1483 \end{pmatrix}\begin{pmatrix} -11.78 \\ 0.74 \end{pmatrix} = \begin{pmatrix} -0.9285 \\ 0.3472 \end{pmatrix}$$

Transforming back to the original units of measurement, the coordinates of the stationary point are

$$X^* = \begin{pmatrix} 161.64°C \\ 367.36 \text{ minutes} \end{pmatrix}.$$

The predicted response at the stationary point is $\hat{Y}(X^*) = 77.59$ %.

Technical Appendix 5D: "Canonical Analysis" of Quadratic Responses

Case for a single controllable response

Whether the stationary point X^* represents a point of maximum or minimum response, or is just a saddle point, is determined by the matrix of second-order coefficients, B. In the simpler case of just a single controllable factor ($k=1$), B is a scalar proportional to the second derivative of $\hat{Y}(x)$ with respect to x. If $d^2\hat{Y}/dx^2$ is positive, recall from calculus that the function $\hat{Y}(x)$ is convex ("bowl shaped") and x^* is a point of minimum response.

Case for multiple controllable responses not so easy

Unfortunately, the multiple factor case ($k>1$) is not so easy since the two-factor interactions (the off-diagonal elements of B) obscure the picture of what is going on. A recommended procedure for analyzing whether B is "positive definite" (we have a minimum) or "negative definite" (we have a maximum) is to rotate the axes $x_1, x_2, ..., x_k$ so that the two-factor interactions disappear. It is also customary (Box and Draper, 1987; Khuri and Cornell, 1987; Myers and Montgomery, 1995) to translate the origin of coordinates to the stationary point so that the intercept term is eliminated from the equation of $\hat{Y}(x)$. This procedure is called the canonical analysis of $\hat{Y}(x)$.

Procedure: Canonical Analysis

Steps for performing the canonical analysis

1. Define a new axis $z = x - x^*$ (translation step). The fitted equation becomes

$$\hat{Y}(z) = \hat{Y}(x^*) + z'Bz.$$

2. Define a new axis $w = E'z$, with $E'BE = D$ and D a diagonal matrix to be defined (rotation step). The fitted equation becomes

$$\hat{Y}(w) = \hat{Y}(x^*) + w'Dw.$$

   This is the so-called canonical form of the model. The elements on the diagonal of D, λ_i $(i = 1, 2, ..., k)$ are the eigenvalues of B. The columns of E', e_i, are the *orthonormal eigenvectors* of B, which means that the e_i satisfy $(B - \lambda_i)e_i = 0$, $e_i'e_j = 0$ for $i \neq j$, and $e_i'e_i = 1.0$.

3. If all the λ_i are negative, x^* is a point of maximum response. If all the λ_i are positive, x^* is a point of minimum response. Finally, if the λ_i are of mixed signs, the response is a saddle function and x^* is the saddle point.

Eigenvalues that are approximately zero

If some $\lambda_i \approx 0$, the fitted ellipsoid

$$\hat{Y}(w) = \hat{Y}(x^*) + \sum_{i=1}^{k} \lambda_i w_i^2$$

is elongated (i.e., it is flat) along the direction of the w_i axis. Points along the w_i axis will have an estimated response close to optimal; thus the process engineer has flexibility in

choosing "good" operating conditions. If two eigenvalues (say λ_i and λ_j) are close to zero, a plane in the (w_i, w_j) coordinates will have close to optimal operating conditions, etc.

Canonical analysis typically performed by software

Software is available to compute the eigenvalues λ_i and the orthonormal eigenvectors e_i; thus there is no need to do a canonical analysis by hand.

Example: Canonical Analysis of Yield Response in Chemical Experiment

***B* matrix for this example**

Let us return to the chemical experiment example to illustrate the method. Keep in mind that when the number of factors is small (e.g., $k=2$ as in this example) canonical analysis is not recommended in practice since simple contour plotting will provide sufficient information. The fitted equation of the model yields

$$B = \begin{pmatrix} -7.25 & -2.2425 \\ -2.425 & -7.55 \end{pmatrix}$$

Compute the eigenvalues and find the orthonormal eigenvectors

To compute the eigenvalues λ_i, we have to find all roots of the expression that results from equating the determinant of $B - \lambda_i I$ to zero. Since B is symmetric and has real coefficients, there will be k real roots λ_i, $i = 1, 2, ..., k$. To find the orthonormal eigenvectors, solve the simultaneous equations $(B - \lambda_i I)e_i = 0$ and $e_i' e_i = 1$.

Canonical analysis results

The results of the canonical analysis are as follows:

Eigenvalues	Eigenvectors X1	Eigenvectors X2
-4.973187	0.728460	-0.685089
-9.827317	0.685089	0.728460

Notice that the eigenvalues are the two roots of

$$\det(B - \lambda I) = (-7.25 - \lambda)(-7.55 - \lambda) - (2.425(-2.2425)) = 0$$

As mentioned previously, the stationary point is $(x^*)' = (-0.9278, 0.3468)$, which corresponds to $X^{*\prime} = (161.64, 367.36)$. Since both eigenvalues are negative, x^* is a point of maximum response. To obtain the directions of the axis of the fitted ellipsoid, compute

$$w_1 = 0.7285(x_1 + 0.9278) - 0.6851(x_2 - 0.3468) = 0.9143 + 0.7285x_1 - 0.6851x_2$$

and

$$w_2 = 0.6851(x_1 + 0.9278) - 0.7285(x_2 - 0.3468) = 0.8830 + 0.6851x_1 + 0.7285x_2$$

Since $|\lambda_1| < |\lambda_2|$, there is somewhat more elongation in the w_i direction. However, since both eigenvalues are quite far from zero, there is not much flexibility in choosing operating conditions. It can be seen from Figure 5.5 that the fitted ellipses do not have a great elongation in the w_1 direction, the direction of the major axis. It is important to emphasize that confirmation experiments at x^* should be performed to check the validity of the

estimated optimal solution.

5.5.3.1.5. Single response: Effect of sampling error on optimal solution

Experimental error means all derived optimal operating conditions are just estimates - confidence regions that are likely to contain the optimal points can be derived

Process engineers should be aware that the estimated optimal operating conditions $\boldsymbol{x}^*$ represent a single estimate of the true (unknown) system optimal point. That is, due to sampling (experimental) error, if the experiment is repeated, a different quadratic function will be fitted which will yield a different stationary point $\boldsymbol{x}^*$. Some authors (Box and Hunter, 1954; Myers and Montgomery, 1995) provide a procedure that allows one to compute a region in the factor space that, with a specified probability, contains the system stationary point. This region is useful information for a process engineer in that it provides a measure of how "good" the point estimate $\boldsymbol{x}^*$ is. In general, the larger this region is, the less reliable the point estimate $\boldsymbol{x}^*$ is. When the number of factors, **k**, is greater than 3, these confidence regions are difficult to visualize.

Confirmation runs are very important

Awareness of experimental error should make a process engineer realize the importance of performing confirmation runs at $\boldsymbol{x}^*$, the estimated optimal operating conditions.

5. Process Improvement
5.5. Advanced topics
5.5.3. How do you optimize a process?
5.5.3.1. Single response case

5.5.3.1.6. Single response: Optimization subject to experimental region constraints

Optimal operating conditions may fall outside region where experiment conducted

Sometimes the optimal operating conditions x^* simply fall outside the region where the experiment was conducted. In these cases, constrained optimization techniques can be used to find the solution x^* that optimizes $\hat{Y}(x)$ without leaving the region in the factor space where the experiment took place.

Ridge analysis is a method for finding optimal factor settings that satisfy certain constraints

"Ridge Analysis", as developed by Hoerl (1959), Hoerl (1964) and Draper (1963), is an optimization technique that finds factor settings x^* such that they

optimize $\hat{Y}(x) = b_0 + b'x + x'Bx$

subject to: $x'x = \rho^2$

The solution x^* to this problem provides operating conditions that yield an estimated absolute maximum or minimum response on a sphere of radius ρ. Different solutions can be obtained by trying different values of ρ.

Solve with non-linear programming software

The original formulation of Ridge Analysis was based on the eigenvalues of a stationarity system. With the wide availability of non-linear programming codes, Ridge Analysis problems can be solved without recourse to eigenvalue analysis.

5.5.3.2. Multiple response case

When there are multiple responses, it is often impossible to simultaneously optimize each one - trade-offs must be made

In the multiple response case, finding process operating conditions that simultaneously maximize (or minimize, as desired) all the responses is quite difficult, and often impossible. Almost inevitably, the process engineer must make some trade-offs in order to find process operating conditions that are satisfactory for most (and hopefully all) the responses. In this subsection, we examine some effective ways to make these trade-offs.

- Path of steepest ascent
- The desirability function approach
- The mathematical programming approach
  - Dual response systems
  - More than 2 responses

5.5.3.2.1. Multiple responses: Path of steepest ascent

Objective: consider and balance the individual paths of maximum improvement

When the responses exhibit adequate linear fit (i.e., the response models are all linear), the objective is to find a direction or path that simultaneously considers the individual paths of maximum improvement and balances them in some way. This case is addressed next.

When there is a mix of linear and higher-order responses, or when all empirical response models are of higher-order, see sections 5.5.3.2.2 and 5.5.3.2.3. The desirability method (section 5.5.3.2.2) can also be used when all response models are linear.

Procedure: Path of Steepest Ascent, Multiple Responses.

A weighted priority strategy is described using the path of steepest ascent for each response

The following is a weighted priority strategy using the path of steepest ascent for each response.

1. Compute the gradients $\boldsymbol{g_i}$ (i = 1, 2, . . ., k) of all responses as explained in section 5.5.3.1.1. If one of the responses is clearly of primary interest compared to the others, use only the gradient of this response and follow the procedure of section 5.5.3.1.1. Otherwise, continue with step 2.

2. Determine relative priorities π_i for each of the **k** responses. Then, the weighted gradient for the search direction is given by

$$g = \frac{\pi_1 g_1 + \pi_2 g_2 + \ldots + \pi_k g_k}{\sum_{i=1}^{k} \pi_i}$$

and the weighted direction is

$$d = \frac{g}{||g||}$$

Weighting factors based on R^2

The confidence cone for the direction of maximum improvement explained in section 5.5.3.1.2 can be used to weight down "poor" response models that provide very wide cones and unreliable directions. Since the width of the cone is proportional to **(1 - R^2)**, we can use

$$\pi_j = \frac{R_j^2}{\sum_{i=1}^{k} R_i^2} \quad j = 1, 2, \ldots, k$$

Single response steepest ascent procedure

Given a weighted direction of maximum improvement, we can follow the single response steepest ascent procedure as in section 5.5.3.1.1 by selecting points with coordinates $x^* = \rho d_i$, $i = 1, 2, ..., k$. These and related issues are explained more fully in Del Castillo (1996).

Example: Path of Steepest Ascent, Multiple Response Case

An example using the weighted priority method

Suppose the response model:

$$\hat{y}_1 = 711.0 + 50.9x_1 + 154.8x_2$$

with $R_1^2 = 0.8968$ represents the average yield of a production process obtained from a replicated factorial experiment in the two controllable factors (in coded units). From the same experiment, a second response model for the process standard deviation of the yield is obtained and given by

$$\hat{y}_2 = 19.26 + 6.31x_1 + 6.28x_2$$

with $R_2^2 = 0.5977$. We wish to maximize the mean yield while minimizing the standard deviation of the yield.

Step 1: compute the gradients:

Compute the gradients

We compute the gradients as follows.

$$g_1' = \left(\frac{50.9}{\sqrt{50.9^2 + 154.8^2}}, \frac{154.8}{\sqrt{50.9^2 + 154.8^2}} \right) = (0.3124, 0.9500)$$

$$g_2' = \left(\frac{-6.31}{\sqrt{6.31^2 + 6.28^2}}, \frac{-6.28}{\sqrt{6.31^2 + 6.28^2}} \right) = (-0.7088, -0.7054)$$

(recall we wish to *minimize* y_2).

Step 2: find relative priorities:

Find relative priorities

Since there are no clear priorities, we use the quality of fit as the priority:

$$\pi_1 = \frac{0.8968}{0.8968 + 0.5977} = 0.6$$

$$\pi_2 = \frac{0.5977}{0.8968 + 0.5977} = 0.4$$

Then, the weighted gradient is

g' = (0.6(0.3124) + 0.4(-0.7088), 0.6(0.95) + 0.4(-0.7054)) = (-0.096, 0.2878)

which, after scaling it (by dividing each coordinate by $\sqrt{(-0.096)^2 + 0.2878^2}$), gives the weighted direction ***d'*** **= (-.03164, 0.9486)**.

Therefore, if we want to move ρ **= 1** coded units along the path of maximum improvement, we will set x_1 **= (1)(-0.3164) = -0.3164,** x_2 **= (1)(0.9486) = 0.9486** in the next run or experiment.

5.5.3.2.2. Multiple responses: The desirability approach

The desirability approach is a popular method that assigns a "score" to a set of responses and chooses factor settings that maximize that score

The desirability function approach is one of the most widely used methods in industry for the optimization of multiple response processes. It is based on the idea that the "quality" of a product or process that has multiple quality characteristics, with one of them outside of some "desired" limits, is completely unacceptable. The method finds operating conditions $\boldsymbol{x}$ that provide the "most desirable" response values.

For each response $\boldsymbol{Y_i(x)}$, a desirability function $\boldsymbol{d_i(Y_i)}$ assigns numbers between 0 and 1 to the possible values of $\boldsymbol{Y_i}$, with $\boldsymbol{d_i(Y_i) = 0}$ representing a completely undesirable value of $\boldsymbol{Y_i}$ and $\boldsymbol{d_i(Y_i) = 1}$ representing a completely desirable or ideal response value. The individual desirabilities are then combined using the geometric mean, which gives the *overall desirability* $\boldsymbol{D}$:

$$D = (d_1(Y_1) \times d_2(Y_2) \times \ldots \times d_k(Y_k))^{1/k}$$

with **k** denoting the number of responses. Notice that if any response $\boldsymbol{Y_i}$ is completely undesirable $\boldsymbol{(d_i(Y_i) = 0)}$, then the overall desirability is zero. In practice, fitted response values $\hat{Y}_i$ are used in place of the $\boldsymbol{Y_i}$.

Desirability functions of Derringer and Suich

Depending on whether a particular response $\boldsymbol{Y_i}$ is to be maximized, minimized, or assigned a target value, different desirability functions $\boldsymbol{d_i(Y_i)}$ can be used. A useful class of desirability functions was proposed by Derringer and Suich (1980). Let $\boldsymbol{L_i}$, $\boldsymbol{U_i}$ and $\boldsymbol{T_i}$ be the lower, upper, and target values, respectively, that are desired for response $\boldsymbol{Y_i}$, with $\boldsymbol{L_i \leq T_i \leq U_i}$.

Desirability function for "target is best"

If a response is of the "target is best" kind, then its individual desirability function is

$$d_i(\hat{Y}_i) = \begin{cases} 0 & \text{if } \hat{Y}_i(\boldsymbol{x}) < L_i \\ \left(\frac{\hat{Y}_i(\boldsymbol{x}) - L_i}{T_i - L_i}\right)^s & \text{if } L_i \leq \hat{Y}_i(\boldsymbol{x}) \leq T_i \\ \left(\frac{\hat{Y}_i(\boldsymbol{x}) - U_i}{T_i - U_i}\right)^t & \text{if } T_i \leq \hat{Y}_i(\boldsymbol{x}) \leq U_i \\ 0 & \text{if } \hat{Y}_i(\boldsymbol{x}) > U_i \end{cases}$$

with the exponents $\boldsymbol{s}$ and $\boldsymbol{t}$ determining how important it is to hit the target value. For $\boldsymbol{s = t = 1}$, the desirability function increases linearly towards $\boldsymbol{T_i}$; for $\boldsymbol{s < 1, t < 1}$,

the function is convex, and for $s > 1, t > 1$, the function is concave (see the example below for an illustration).

Desirability function for maximizing a response

If a response is to be maximized instead, the individual desirability is defined as

$$d_i(\hat{Y}_i) = \begin{cases} 0 & \text{if } \hat{Y}_i(x) < L_i \\ \left(\frac{\hat{Y}_i(x) - L_i}{T_i - L_i}\right)^s & \text{if } L_i \le \hat{Y}_i(x) \le T_i \\ 1.0 & \text{if } \hat{Y}_i(x) > T_i \end{cases}$$

with $\mathbf{T_i}$ in this case interpreted as a large enough value for the response.

Desirability function for minimizing a response

Finally, if we want to minimize a response, we could use

$$d_i(\hat{Y}_i) = \begin{cases} 1.0 & \text{if } \hat{Y}_i(x) < T_i \\ \left(\frac{\hat{Y}_i(x) - U_i}{T_i - U_i}\right)^s & \text{if } T_i \le \hat{Y}_i(x) \le U_i \\ 0 & \text{if } \hat{Y}_i(x) > U_i \end{cases}$$

with $\mathbf{T_i}$ denoting a small enough value for the response.

Desirability approach steps

The desirability approach consists of the following steps:

1. Conduct experiments and fit response models for all **k** responses;
2. Define individual desirability functions for each response;
3. *Maximize* the overall desirability ***D*** with respect to the controllable factors.

Example:

An example using the desirability approach

Derringer and Suich (1980) present the following multiple response experiment arising in the development of a tire tread compound. The controllable factors are: x_1, hydrated silica level, x_2, silane coupling agent level, and x_3, sulfur level. The four responses to be optimized and their desired ranges are:

Factor and response variables

Source	Desired range
PICO Abrasion index, Y_1	$120 < Y_1$
200% modulus, Y_2	$1000 < Y_2$
Elongation at break, Y_3	$400 < Y_3 < 600$
Hardness, Y_4	$60 < Y_4 < 75$

The first two responses are to be maximized, and the value **s=1** was chosen for their desirability functions. The last two responses are "target is best" with $T_3 = 500$ and $T_4 = 67.5$. The values **s=t=1** were chosen in both cases.

Experimental

The following experiments were conducted using a central composite design.

runs from a central composite design

Run Number	x_1	x_2	x_3	Y_1	Y_2	Y_3	Y_4
1	-1.00	-1.00	-1.00	102	900	470	67.5
2	+1.00	-1.00	-1.00	120	860	410	65.0
3	-1.00	+1.00	-1.00	117	800	570	77.5
4	+1.00	+1.00	-1.00	198	2294	240	74.5
5	-1.00	-1.00	+1.00	103	490	640	62.5
6	+1.00	-1.00	+1.00	132	1289	270	67.0
7	-1.00	+1.00	+1.00	132	1270	410	78.0
8	+1.00	+1.00	+1.00	139	1090	380	70.0
9	-1.63	0.00	0.00	102	770	590	76.0
10	+1.63	0.00	0.00	154	1690	260	70.0
11	0.00	-1.63	0.00	96	700	520	63.0
12	0.00	+1.63	0.00	163	1540	380	75.0
13	0.00	0.00	-1.63	116	2184	520	65.0
14	0.00	0.00	+1.63	153	1784	290	71.0
15	0.00	0.00	0.00	133	1300	380	70.0
16	0.00	0.00	0.00	133	1300	380	68.5
17	0.00	0.00	0.00	140	1145	430	68.0
18	0.00	0.00	0.00	142	1090	430	68.0
19	0.00	0.00	0.00	145	1260	390	69.0
20	0.00	0.00	0.00	142	1344	390	70.0

Fitted response

Using ordinary least squares and standard diagnostics, the fitted responses are:

$$\begin{aligned}\hat{Y}_1 &= 139.12 + 16.49x_1 + 17.88x_2 + 2.21x_3 \\ &\quad -4.01x_1^2 - 3.45x_2^2 - 1.57x_3^2 \\ &\quad +5.12x_1x_2 - 7.88x_1x_3 - 7.13x_2x_3\end{aligned}$$

(**R^2 = 0.8369** and adjusted **R^2 = 0.6903**);

$$\begin{aligned}\hat{Y}_2 &= 1261.13 + 268.15x_1 + 246.5x_2 - 102.63x_3 \\ &\quad -83.5x_1^2 - 124.82x_2^2 + 199.2x_3^2 \\ &\quad +69.3x_1x_2 - 104.38x_1x_3 - 94.13x_2x_3\end{aligned}$$

(**R^2 = 0.7137** and adjusted **R^2 = 0.4562**);

$$\hat{Y}_3 = 417.5 - 99.67x_1 - 31.4x_2 - 27.42x_3$$

(**R^2 = 0.682** and adjusted **R^2 = 0.6224**);

$$\begin{aligned}\hat{Y}_4 &= 68.91 - 1.41x_1 + 4.32x_2 + 0.21x_3 \\ &\quad +1.56x_1^2 + 0.058x_2^2 - 0.32x_3^2 \\ &\quad -1.62x_1x_2 + 0.25x_1x_3 - 0.12x_2x_3\end{aligned}$$

(R^2 = **0.8667** and adjusted R^2 = **0.7466**).

Note that no interactions were significant for response 3 and that the fit for response 2 is quite poor.

Best Solution

The best solution is $(x^*)'$ = **(-0.10, 0.15, -1.0)** and results in:

$d_1(\hat{Y}_1)$ = **0.34** $(\hat{Y}_1(x^*)$ = **136.4)**

$d_2(\hat{Y}_2)$ = **1.0** $(\hat{Y}_2(x^*)$ = **1571.05)**

$d_3(\hat{Y}_3)$ = **0.49** $(\hat{Y}_3(x^*)$ = **450.56)**

$d_4(\hat{Y}_4)$ = **0.76** $(\hat{Y}_4(x^*)$ = **69.26)**

The overall desirability for this solution is 0.596. All responses are predicted to be within the desired limits.

3D plot of the overall desirability function

Figure 5.8 shows a 3D plot of the overall desirability function $D(x)$ for the (x_2, x_3) plane when x_1 is fixed at -0.10. The function $D(x)$ is quite "flat" in the vicinity of the optimal solution, indicating that small variations around x^* are predicted to not change the overall desirability drastically. However, the importance of performing confirmatory runs at the estimated optimal operating conditions should be emphasized. This is particularly true in this example given the poor fit of the response models (e.g., $\hat{Y}_2$).

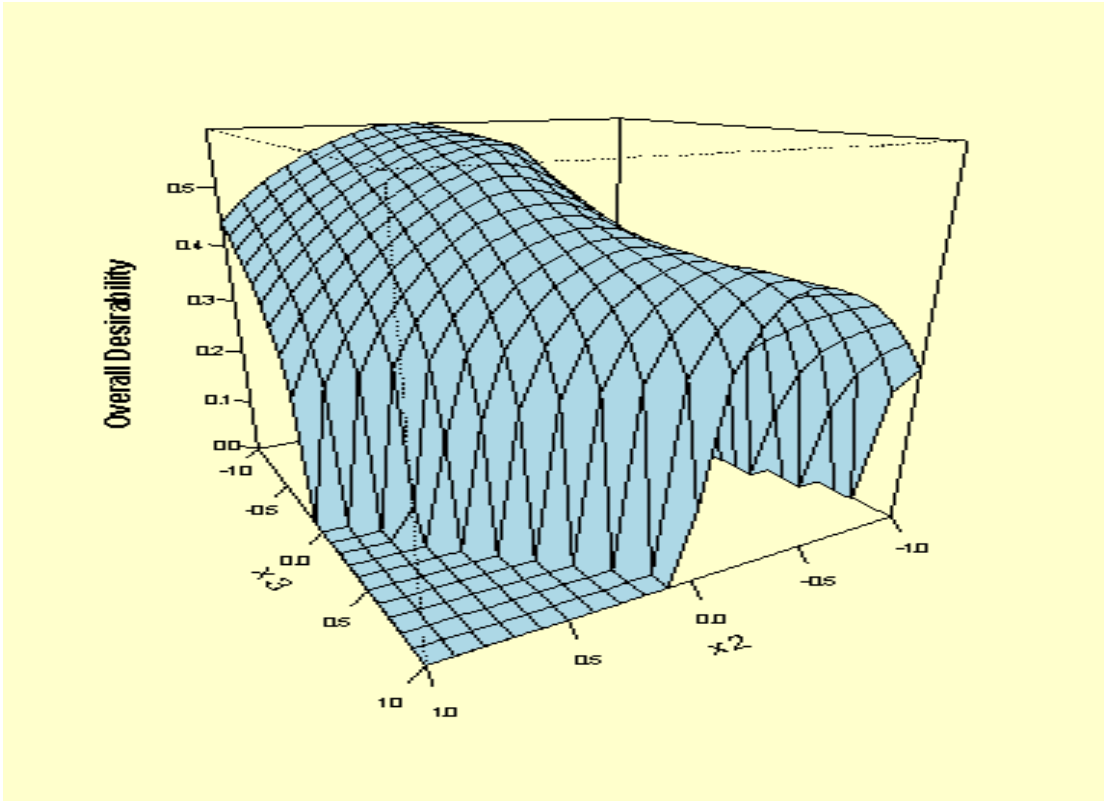


FIGURE 5.8 Overall Desirability Function for Example Problem

5. Process Improvement
5.5. Advanced topics
5.5.3. How do you optimize a process?
5.5.3.2. Multiple response case

5.5.3.2.3. Multiple responses: The mathematical programming approach

The mathematical programming approach maximizes or minimizes a primary response, subject to appropriate constraints on all other responses

The analysis of multiple response systems usually involves some type of optimization problem. When one response can be chosen as the "primary", or most important response, and bounds or targets can be defined on all other responses, a mathematical programming approach can be taken. If this is not possible, the desirability approach should be used instead.

In the mathematical programming approach, the primary response is maximized or minimized, as desired, subject to appropriate constraints on all other responses. The case of two responses ("dual" responses) has been studied in detail by some authors and is presented first. Then, the case of more than 2 responses is illustrated.

- Dual response systems
- More than 2 responses

Dual response systems

Optimization of dual response systems

The optimization of dual response systems (DRS) consists of finding operating conditions ***x*** that

$$\text{optimize} \quad \hat{Y}_p(x)$$

$$\text{subject to:} \quad \hat{Y}_s(x) = T$$
$$x'x \leq \rho^2$$

with **T** denoting the target value for the secondary response, ***p*** the number of primary responses (i.e., responses to be optimized), ***s*** the number of secondary responses (i.e., responses to be constrained), and ρ is the radius of a spherical constraint that limits the region in the controllable factor space where the search should be undertaken. The value of ρ should be chosen with the purpose of avoiding solutions that extrapolate too far outside the region where the experimental data were obtained. For example, if the experimental design is a central composite design, choosing $\rho = \alpha$ (axial distance) is a logical choice. Bounds of the form

$L \leq x_i \leq U$ can be used instead if a cubical experimental region were used (e.g., when using a factorial experiment). Note that a Ridge Analysis problem is related to a DRS problem when the secondary constraint is absent. Thus, any algorithm or solver for DRS's will also work for the Ridge Analysis of single response systems.

Nonlinear programming software required for DRS

In a DRS, the response models $\hat{Y}_p$ and $\hat{Y}_s$ can be linear, quadratic or even cubic polynomials. A nonlinear programming algorithm has to be used for the optimization of a DRS. For the particular case of quadratic responses, an equality constraint for the secondary response, and a spherical region of experimentation, specialized optimization algorithms exist that guarantee global optimal solutions. In such a case, the algorithm DRSALG can be used (download from http://www.stat.cmu.edu/jqt/29-3), but a Fortran compiler is necessary.

More general case

In the more general case of inequality constraints or a cubical region of experimentation, a general purpose nonlinear solver must be used and several starting points should be tried to avoid local optima. This is illustrated in the next section.

Example for more than 2 responses

Example: problem setup

The values of three components (x_1, x_2, x_3) of a propellant need to be selected to maximize a primary response, burning rate (Y_1), subject to satisfactory levels of two secondary reponses; namely, the variance of the burning rate (Y_2) and the cost (Y_3). The three components must add to 100% of the mixture. The fitted models are:

$$\begin{aligned}\hat{Y}_1 &= 35.4x_1 + 42.77x_2 + 70.36x_3 + 16.02x_1x_2 \\ &\quad +36.33x_1x_3 + 136.8x_2x_3 + 854.9x_1x_2x_3\end{aligned}$$

$$\begin{aligned}\hat{Y}_2 &= 3.88x_1 + 9.03x_2 + 13.63x_3 - 0.1904x_1x_2 \\ &\quad -16.61x_1x_3 - 27.67x_2x_3\end{aligned}$$

$$\hat{Y}_3 = 23.13x_1 + 19.73x_2 + 14.73x_3$$

The optimization problem

The optimization problem is therefore:

maximize $\hat{Y}_1(x)$

subject to: $\hat{Y}_2(x) \leq 4.5$

$\hat{Y}_3(x) \leq 20$

$x_1 + x_2 + x_3 = 1.0$

$0 \leq x_1 \leq 1$

$$0 \leq x_2 \leq 1$$
$$0 \leq x_3 \leq 1$$

Solution

The solution is $(x^*)' = (0.212, 0.343, 0.443)$ which provides $\hat{Y}_1 = 106.62$, $\hat{Y}_2 = 4.17$, and $\hat{Y}_3 = 18.23$. Therefore, both secondary responses are below the specified upper bounds. The optimization should be implemented using a variety of starting points to avoid local optima. Once again, confirmatory experiments should be conducted at the estimated optimal operating conditions.

The solution to the optimization problem can be obtained using R code.

5.5.4. What is a mixture design?

When the factors are proportions of a blend, you need to use a mixture design

In a mixture experiment, the independent factors are proportions of different components of a blend. For example, if you want to optimize the tensile strength of stainless steel, the factors of interest might be the proportions of iron, copper, nickel, and chromium in the alloy. The fact that the proportions of the different factors must sum to 100% complicates the design as well as the analysis of mixture experiments.

Standard mixture designs and constrained mixture designs

When the mixture components are subject to the constraint that they must sum to one, there are standard mixture designs for fitting standard models, such as *Simplex-Lattice* designs and *Simplex-Centroid* designs. When mixture components are subject to additional constraints, such as a maximum and/or minimum value for each component, designs other than the standard mixture designs, referred to as constrained mixture designs or *Extreme-Vertices* designs, are appropriate.

Measured response assumed to depend only on relative proportions

In mixture experiments, the measured response is assumed to depend only on the relative proportions of the ingredients or components in the mixture and not on the amount of the mixture. The amount of the mixture could also be studied as an additional factor in the experiment; however, this would be an example of mixture and process variables being treated together.

Proportions of each variable must sum to 1

The main distinction between mixture experiments and independent variable experiments is that with the former, the input variables or components are non-negative proportionate amounts of the mixture, and if expressed as fractions of the mixture, they must sum to one. If for some reason, the sum of the component proportions is less than one, the variable proportions can be rewritten as scaled fractions so that the scaled fractions sum to one.

Purpose of a mixture design

In mixture problems, the purpose of the experiment is to model the blending surface with some form of mathematical equation so that:

1. Predictions of the response for any mixture or combination of the ingredients can be made empirically, or

2. Some measure of the influence on the response of each component singly and in combination with other components can be obtained.

Assumptions for mixture experiments

The usual assumptions made for factorial experiments are also made for mixture experiments. In particular, it is assumed that the errors are independent and identically distributed with zero mean and common variance. Another assumption that is made, as with factorial designs, is that the true underlying response surface is continuous over the region being studied.

Steps in planning a mixture experiment

Planning a mixture experiment typically involves the following steps (Cornell and Piepel, 1994):

1. Define the objectives of the experiment.
2. Select the mixture components and any other factors to be studied. Other factors may include process variables or the total amount of the mixture.
3. Identify any constraints on the mixture components or other factors in order to specify the experimental region.
4. Identify the response variable(s) to be measured.
5. Propose an appropriate model for modeling the response data as functions of the mixture components and other factors selected for the experiment.
6. Select an experimental design that is sufficient not only to fit the proposed model, but which allows a test of model adequacy as well.

5.5.4.1. Mixture screening designs

Screening experiments can be used to identify the important mixture factors

In some areas of mixture experiments, for example, certain chemical industries, there is often a large number, *q*, of potentially important components that can be considered candidates in an experiment. The objective of these types of experiments is to screen the components to identify the ones that are most important. In this type of situation, the experimenter should consider a screening experiment to reduce the number of possible components.

A first order mixture model

The construction of screening designs and their corresponding models often begins with the first-order or first-degree mixture model

$$E(Y) = \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_q x_q$$

for which the beta coefficients are non-negative and sum to one.

Choices of types of screening designs depend on constraints

If the experimental region is a simplex, it is generally a good idea to make the ranges of the components as similar as possible. Then the relative effects of the components can be assessed by ranking the ratios of the parameter estimates (i.e., the estimates of the β_i), relative to their standard errors.

Simplex screening designs are recommended when it is possible to experiment over the total simplex region. Constrained mixture designs are suggested when the proportions of some or all of the components are restricted by upper and lower bounds. If these designs are not feasible in this situation, then D-optimal designs for a linear model are always an option.

5.5.4.2. Simplex-lattice designs

Definition of simplex-lattice points

A {q, m} simplex-lattice design for q components consists of points defined by the following coordinate settings: the proportions assumed by each component take the $m+1$ equally spaced values from 0 to 1,

$$x_i = 0, 1/m, 2/m, \ldots , 1 \text{ for } i = 1, 2, \ldots , q$$

and all possible combinations (mixtures) of the proportions from this equation are used.

Except for the center, all design points are on the simplex boundaries

Note that the standard Simplex-Lattice and the Simplex-Centroid designs (described later) are boundary-point designs; that is, with the exception of the overall centroid, all the design points are on the boundaries of the simplex. When one is interested in prediction in the interior, it is highly desirable to augment the simplex-type designs with interior design points.

Example of a three-component simplex lattice design

Consider a three-component mixture for which the number of equally spaced levels for each component is four (i.e., $x_i = 0$, 0.333, 0.667, 1). In this example $q = 3$ and $m = 3$. If one uses all possible blends of the three components with these proportions, the {3, 3} simplex-lattice then contains the 10 blending coordinates listed in the table below. The experimental region and the distribution of design runs over the simplex region are shown in the figure below. There are 10 design runs for the {3, 3} simplex-lattice design.

Design table

TABLE 5.3 Simplex Lattice Design

*X*1	*X*2	*X*3
0	0	1
0	0.667	0.333
0	1	0
0.333	0	0.667
0.333	0.333	0.333
0.333	0.6667	0
0.667	0	0.333
0.667	0.333	0

1 0 0

Diagram showing configuration of design runs

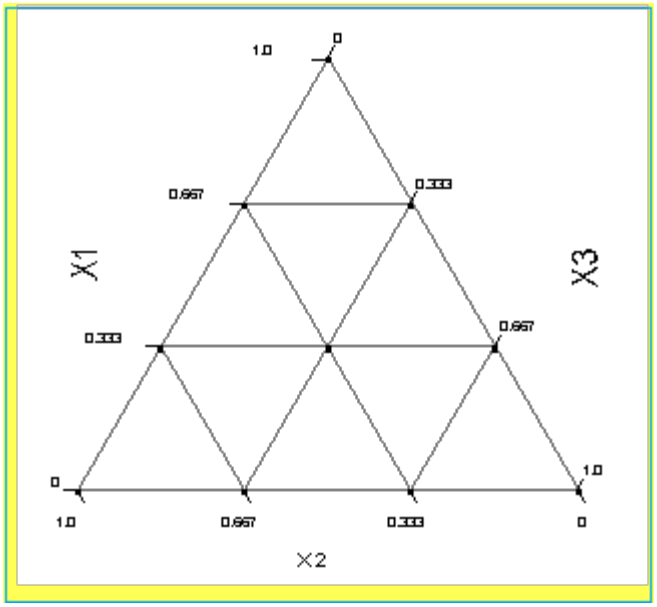


FIGURE 5.9 Configuration of Design Runs for a {3,3} Simplex-Lattice Design

The number of design points in the simplex-lattice is $(q+m-1)!/(m!(q-1)!)$.

Definition of canonical polynomial model used in mixture experiments

Now consider the form of the polynomial model that one might fit to the data from a mixture experiment. Due to the restriction $x_1 + x_2 + ... + x_q = 1$, the form of the regression function that is fit to the data from a mixture experiment is somewhat different from the traditional polynomial fit and is often referred to as the canonical polynomial. Its form is derived using the general form of the regression function that can be fit to data collected at the points of a $\{q, m\}$ simplex-lattice design and substituting into this function the dependence relationship among the x_i terms. The number of terms in the $\{q, m\}$ polynomial is $(q+m-1)!/(m!(q-1)!)$, as stated previously. This is equal to the number of points that make up the associated $\{q, m\}$ simplex-lattice design.

Example for a {q, m=1} simplex-lattice design

For example, the equation that can be fit to the points from a $\{q, m=1\}$ simplex-lattice design is

$$E(Y) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

Multiplying β_0 by $(x_1 + x_2 + ... + x_q = 1)$, the resulting equation is

$$E(Y) = \beta_1^* x_1 + ... + \beta_q^* x_q$$

with $\beta_i^* = \beta_0 + \beta_i$ for all $i = 1, ..., q$.

First-

This is called the canonical form of the first-order mixture model.

order canonical form

In general, the canonical forms of the mixture models (with the asterisks removed from the parameters) are as follows:

Summary of canonical mixture models

Linear $$E(Y) = \sum_{i=1}^{q} \beta_i x_i$$

Quadratic $$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j} \beta_{ij} x_i x_j$$

Cubic $$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j} \beta_{ij} x_i x_j + \sum_{j=1}^{q} \sum_{i<j} \delta_{ij} x_i x_j (x_i - x_j) + \sum_{k=1}^{q} \sum_{j<k} \sum_{i<j} \beta_{ijk} x_i x_j x_k$$

Special Cubic $$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j} \beta_{ij} x_i x_j + \sum_{k=1}^{q} \sum_{j<k} \sum_{i<j} \beta_{ijk} x_i x_j x_k$$

Linear blending portion

The terms in the canonical mixture polynomials have simple interpretations. Geometrically, the parameter β_i in the above equations represents the expected response to the pure mixture x_i=1, x_j=0, $i \neq j$, and is the height of the mixture surface at the vertex x_i=1. The portion of each of the above polynomials given by

$$\sum_{i=1}^{q} \beta_i x_i$$

is called the linear blending portion. When blending is strictly additive, then the linear model form above is an appropriate model.

Three-component mixture example

The following example is from Cornell (1990) and consists of a three-component mixture problem. The three components are Polyethylene (X1), polystyrene (X2), and polypropylene (X3), which are blended together to form fiber that will be spun into yarn. The product developers are only interested in the pure and binary blends of these three materials. The response variable of interest is yarn elongation in kilograms of force applied. A {3,2} simplex-lattice design is used to study the blending process. The simplex region and the six design runs are shown in the figure below. The design and the observed responses are listed in Table 5.4. There were two replicate observations run at each of the pure blends. There were three replicate observations run at the binary blends. There are 15 observations with six unique design runs.

Diagram showing the

designs runs for this example

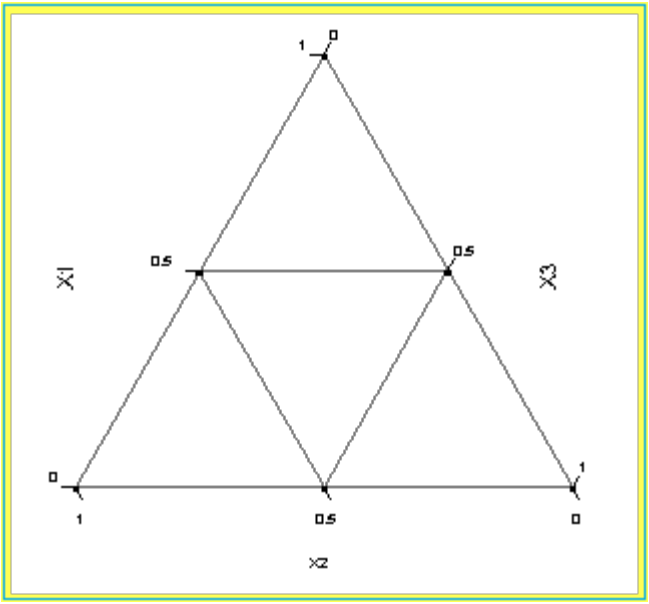


FIGURE 5.10 Design Runs for the {3,2} Simplex-Lattice Yarn Elongation Problem

Table showing the simplex-lattice design and observed responses

TABLE 5.4 Simplex-Lattice Design for Yarn Elongation Problem

*X*1	*X*2	*X*3	Observed Elongation Values
0.0	0.0	1.0	16.8, 16.0
0.0	0.5	0.5	10.0, 9.7, 11.8
0.0	1.0	0.0	8.8, 10.0
0.5	0.0	0.5	17.7, 16.4, 16.6
0.5	0.5	0.0	15.0, 14.8, 16.1
1.0	0.0	0.0	11.0, 12.4

Fit a quadratic mixture model

The design runs listed in the above table are in standard order. The actual order of the 15 treatment runs was completely randomized. Since there are three levels of each of the three mixture components, a quadratic mixture model can be fit to the data. The results of the model fit are shown below. Note that there was no intercept in the model.

Summary of Fit

```
RSquare                        0.951356
RSquare Adj                    0.924331
Root Mean Square Error         0.85375
Mean of Response              13.54
Observations (or Sum Wgts)    15
```

Analysis of Variance

```
Source   DF  Sum of Squares  Mean Square  F Ratio
Prob > F
Model     5     2878.27        479.7117   658.141
1.55e-13
Error     9        6.56          0.7289
C Total  14     2884.83
```

Parameter Estimates

```
Term    Estimate  Std Error   t Ratio  Prob>|t|
```

X1	11.7	0.603692	19.38	<.0001
X2	9.4	0.603692	15.57	<.0001
X3	16.4	0.603692	27.17	<.0001
X2*X1	19	2.608249	7.28	<.0001
X3*X1	11.4	2.608249	4.37	0.0018
X3*X2	-9.6	2.608249	-3.68	0.0051

Interpretation of results

Under the parameter estimates section of the output are the individual t-tests for each of the parameters in the model. The three cross product terms are significant (X1*X2, X3*X1, X3*X2), indicating a significant quadratic fit.

The fitted quadratic model

The fitted quadratic mixture model is

$$\hat{y} = 11.7x_1 + 9.4x_2 + 16.4x_3 + 19.0x_1x_2 + 11.4x_1x_3 - 9.6x_2x_3$$

Conclusions from the fitted quadratic model

Since $b_3 > b_1 > b_2$, one can conclude that component 3 (polypropylene) produces yarn with the highest elongation. Additionally, since b_{12} and b_{13} are positive, blending components 1 and 2 or components 1 and 3 produces higher elongation values than would be expected just by averaging the elongations of the pure blends. This is an example of 'synergistic' blending effects. Components 2 and 3 have antagonistic blending effects because b_{23} is negative.

Contour plot of the predicted elongation values

The figure below is the contour plot of the elongation values. From the plot it can be seen that if maximum elongation is desired, a blend of components 1 and 3 should be chosen consisting of about 75% - 80% component 3 and 20% - 25% component 1.

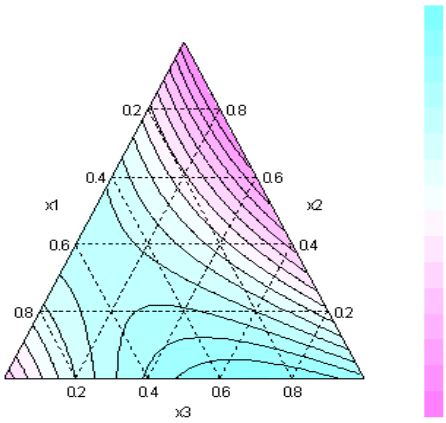


FIGURE 5.11 Contour Plot of Predicted Elongation Values from {3,2} Simplex-Lattice Design

The analyses in this page can be obtained using R code.

5. Process Improvement
5.5. Advanced topics
5.5.4. What is a mixture design?

5.5.4.3. Simplex-centroid designs

Definition of simplex-centroid designs

A second type of mixture design is the simplex-centroid design. In the q-component simplex-centroid design, the number of distinct points is $2^q - 1$. These points correspond to q permutations of (1, 0, 0, ..., 0) or q single component blends, the $\binom{q}{2}$ permutations of (.5, .5, 0, ..., 0) or all binary mixtures, the $\binom{q}{3}$ permutations of (1/3, 1/3, 1/3, 0, ..., 0), ..., and so on, with finally the overall centroid point ($1/q$, $1/q$, ..., $1/q$) or q-nary mixture.

The design points in the Simplex-Centroid design will support the polynomial

Model supported by simplex-centroid designs

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{i<j} \beta_{ij} x_i x_j + \sum_{k=1}^{q} \sum_{j<k} \sum_{i<j} \beta_{ijk} x_i x_j x_k + \ldots + \beta_{12\ldots q} x_i x_j \ldots x_q$$

which is the qth-order mixture polynomial. For q = 2, this is the quadratic model. For q = 3, this is the special cubic model.

Example of runs for three and four components

For example, the fifteen runs for a four component (q = 4) simplex-centroid design are:

(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), (.5,.5,0,0), (.5,0,.5,0) ..., (0,0,.5,.5), (1/3,1/3,1/3,0), ...,(0,1/3,1/3,1/3), (1/4,1/4,1/4,1/4).

The runs for a three component simplex-centroid design of degree 2 are

(1,0,0), (0,1,0), (0,0,1), (.5,.5,0), (.5,0,.5), (0,.5,.5), (1/3, 1/3, 1/3).

However, in order to fit a first-order model with q =4, only the five runs with a "1" and all "1/4's" would be needed. To fit a second-order model, add the six runs with a ".5" (this also fits a saturated third-order model, with no degrees of freedom left for error).

5.5.4.4. Constrained mixture designs

Upper and/or lower bound constraints may be present

In mixture designs when there are constraints on the component proportions, these are often upper and/or lower bound constraints of the form $L_i \leq x_i \leq U_i$, i = 1, 2,..., q, where L_i is the lower bound for the i-th component and U_i the upper bound for the i-th component. The general form of the constrained mixture problem is

Typical additional constraints

$$x_1 + x_2 + ... + x_q = 1$$

$$L_i \leq x_i \leq U_i, \quad \text{for } i = 1, 2,..., q$$

with $L_i \geq 0$ and $U_i \leq 1$.

Example using only lower bounds

Consider the following case in which only the lower bounds in the above equation are imposed, so that the constrained mixture problem becomes

$$x_1 + x_2 + ... + x_q = 1$$

$$L_i \leq x_i \leq 1, \quad \text{for } i = 1, 2,..., q$$

Assume we have a three-component mixture problem with constraints

$$0.3 \leq x_1 \quad 0.4 \leq x_2 \quad 0.1 \leq x_3$$

Feasible mixture region

The feasible mixture space is shown in the figure below. Note that the existence of lower bounds does not affect the shape of the mixture region, it is still a simplex region. In general, this will always be the case if only lower bounds are imposed on any of the component proportions.

Diagram showing the feasible mixture space

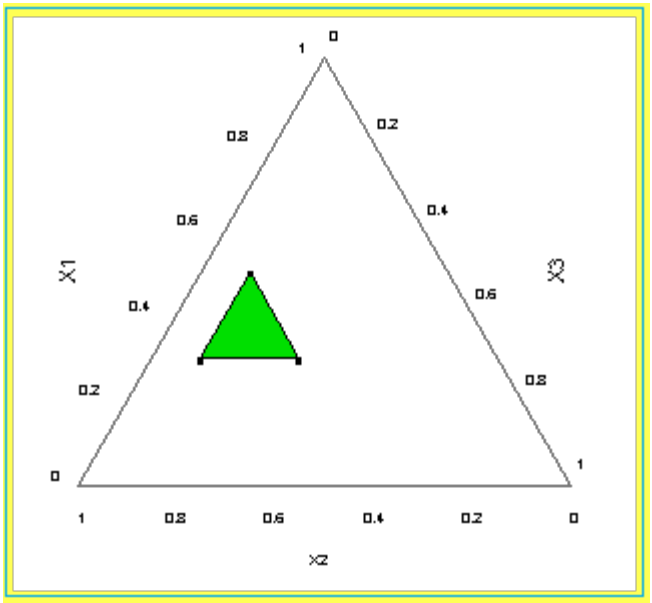


FIGURE 5.12 The Feasible Mixture Space (Shaded Region) for Three Components with Lower Bounds

A simple transformation helps in design construction and analysis

Since the new region of the experiment is still a simplex, it is possible to define a new set of components that take on the values from 0 to 1 over the feasible region. This will make the design construction and the model fitting easier over the constrained region of interest. These new components (x_i^*) are called pseudo components and are defined using the following formula

Formula for pseudo components

$$x_i^* = \frac{x_i - L_i}{1 - L}$$

with

$$L = \sum_{i=1}^{q} L_i < 1$$

denoting the sum of all the lower bounds.

Computation of the pseudo components for the example

In the three component example above, the pseudo components are

$$x_1^* = \frac{x_1 - 0.3}{0.2} \qquad x_2^* = \frac{x_2 - 0.4}{0.2} \qquad x_3^* = \frac{x_3 - 0.1}{0.2}$$

Constructing the design in the pseudo components

Constructing a design in the pseudo components is accomplished by specifying the design points in terms of the x_i^* and then converting them to the original component settings using

$$x_i = L_i + (1 - L)x_i^*$$

Select appropriate design

In terms of the pseudo components, the experimenter has the choice of selecting a Simplex-Lattice or a Simplex-Centroid design, depending on the objectives of the experiment.

Simplex-centroid design example (after transformation)

Suppose, we decided to use a Simplex-centroid design for the three-component experiment. The table below shows the design points in the pseudo components, along with the corresponding setting for the original components.

Table showing the design points in both the pseudo components and the original components

TABLE 5.5 Pseudo Component Settings and Original Component Settings, Three-Component Simplex-Centroid Design

Pseudo Components			Original Components		
X_1	X_2	X_3	x_1^*	x_2^*	x_3^*
1	0	0	0.5	0.4	0.1
0	1	0	0.3	0.6	0.1
0	0	1	0.3	0.4	0.3
0.5	0.5	0	0.4	0.5	0.1
0.5	0	0.5	0.4	0.4	0.2
0	0.5	0.5	0.3	0.5	0.2
0.3333	0.3333	0.3333	0.3667	0.4667	0.1666

Use of pseudo components (after transformation) is recommended

It is recommended that the pseudo components be used to fit the mixture model. This is due to the fact that the constrained design space will usually have relatively high levels of multicollinearity among the predictors. Once the final predictive model for the pseudo components has been determined, the equation in terms of the original components can be determined by substituting the relationship between x_i and x_i^*.

D-optimal designs can also be used

Computer-aided designs (D-optimal, for example) can be used to select points for a mixture design in a constrained region. See Myers and Montgomery (1995) for more details on using D-optimal designs in mixture experiments.

Extreme vertice designs anre another option

Note: There are other mixture designs that cover only a sub-portion or smaller space within the simplex. These types of mixture designs (not covered here) are referred to as *extreme vertices designs*. (See chapter 11 of Myers and Montgomery (1995) or Cornell (1990).

5.5.4.5. Treating mixture and process variables together

Options for setting up experiments for processes that have both standard process variables and mixture variables

Consider a mixture experiment consisting of q mixture components and k process variables. First consider the case in which each of the process variables to be studied has only two levels. Orthogonally scaled factor settings for the process variables will be used (i.e., -1 is the low level, 1 is the high level, and 0 is the center point). Also assume that each of the components x_i can range from 0 to 1. The region of interest then for the process variables is a k-dimensional hypercube.

The region of interest for the mixture components is the (q-1)-dimensional simplex. The combined region of interest for both the process variables and the mixture components is of dimensionality q - 1 + k.

Example of three mixture components and three process variables

For example, consider three mixture components (x_1, x_2, x_3) with three process variables (z_1, z_2, z_3). The dimensionality of the region is 5. The combined region of interest for the three mixture components and three process variables is shown in the two figures below. The complete space of the design can be viewed in either of two ways. The first diagram shows the idea of a full factorial at each vertex of the three-component simplex region. The second diagram shows the idea of a three-component simplex region at each point in the full factorial. In either case, the same overall process space is being investigated.

Diagram showing simplex region of a 3-component mixture with a 2^3 full factorial at each pure mixture run

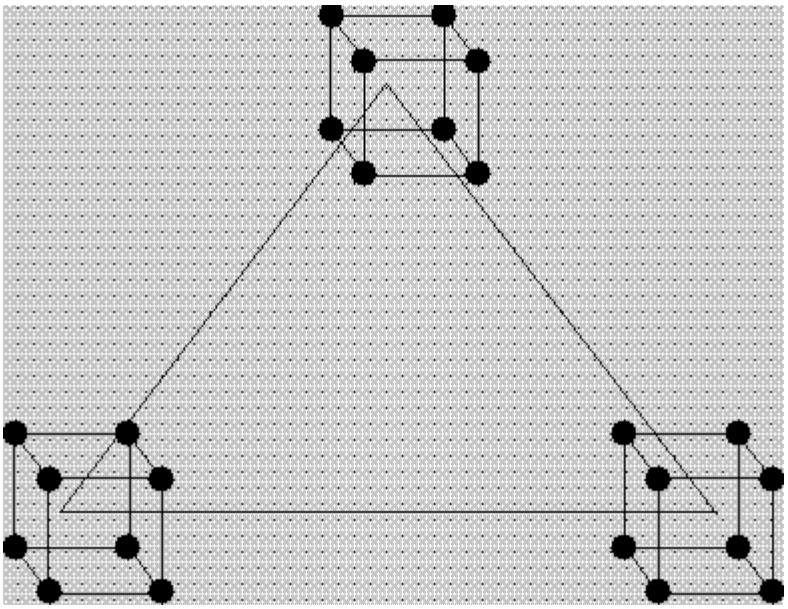


FIGURE 5.13 Simplex Region of a Three Component Mixture with a 2^3 Full Factorial at Each Pure Mixture Run

Diagram showing process space of a 2^3 full factorial with the 3-component simplex region at each point of the full factorial

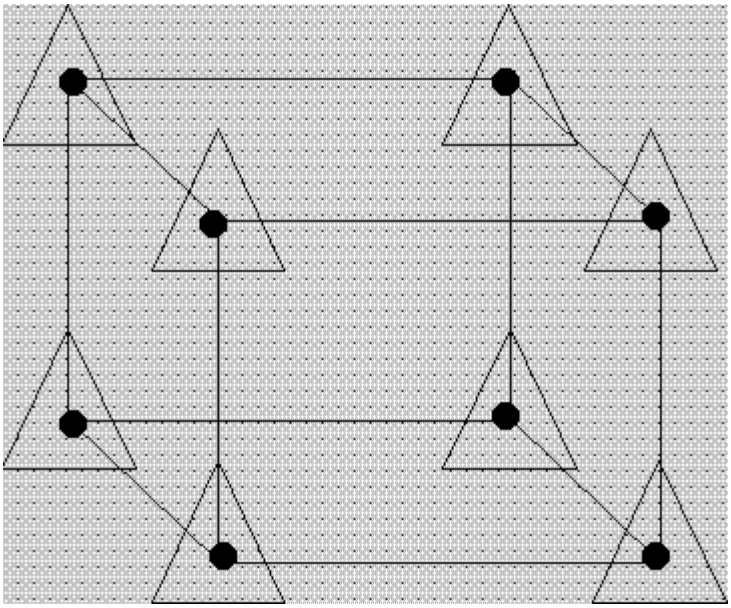


FIGURE 5.14 Process Space of a 2^3 Full Factorial with the Three Component Simplex Region at Each Point of the Full Factorial

Additional options available

As can be seen from the above diagrams, setting up the design configurations in the process variables and mixture components involves setting up either a mixture design at each point of a configuration in the process variables, or similarly, creating a factorial arrangement in the process variables at each point of composition in the mixture components. For the example depicted in the above two diagrams, this is not the only design available for this number of mixture components with the specified number of process variables. Another option might be to run a fractional factorial design at each vertex or point of the mixture design,

with the same fraction run at each mixture design point. Still another option might be to run a fractional factorial design at each vertex or point of the mixture design, with a different fraction run at each mixture design point.

5.5.5. How can I account for nested variation (restricted randomization)?

Nested data structures are common and lead to many sources of variability

Many processes have more than one source of variation in them. In order to reduce variation in processes, these multiple sources must be understood, and that often leads to the concept of nested or hierarchical data structures. For example, in the semiconductor industry, a batch process may operate on several wafers at a time (wafers are said to be *nested* within batch). Understanding the input variables that control variation among those wafers, as well as understanding the variation across each wafer in a run, is an important part of the strategy for minimizing the total variation in the system.

Example of nested data

Figure 5.15 below represents a batch process that uses 7 monitor wafers in each run. The plan further calls for measuring response on each wafer at each of 9 sites. The organization of the sampling plan has a hierarchical or nested structure: the batch run is the topmost level, the second level is an individual wafer, and the third level is the site on the wafer.

The total amount of data generated per batch run will be 7*9 = 63 data points. One approach to analyzing these data would be to compute the mean of all these points as well as their standard deviation and use those results as responses for each run.

Diagram illustrating the example

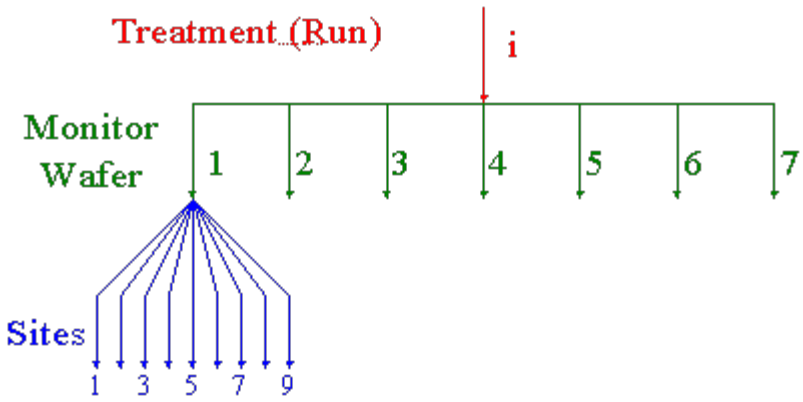


FIGURE 5.15 Hierarchical Data Structure Example

Sites nested within wafers and wafers are nested within runs

Analyzing the data as suggested above is not absolutely incorrect, but doing so loses information that one might otherwise obtain. For example, site 1 on wafer 1 is physically different from site 1 on wafer 2 or on any other wafer. The same is true for any of the sites on any of the wafers. Similarly, wafer 1 in run 1 is physically different from wafer 1 in run 2, and so on. To describe this situation one says that *sites are nested within wafers* while *wafers are nested within runs*.

Nesting places restrictions on the randomization

As a consequence of this nesting, there are restrictions on the randomization that can occur in the experiment. This kind of restricted randomization always produces nested sources of variation. Examples of nested variation or restricted randomization discussed on this page are *split-plot* and *strip-plot* designs.

Wafer-to-wafer and site-to-site variations are often "noise factors" in an experiment

The objective of an experiment with the type of sampling plan described in Figure 5.15 is generally to reduce the variability due to sites on the wafers and wafers within runs (or batches) in the process. The sites on the wafers and the wafers within a batch become sources of unwanted variation and an investigator seeks to make the system robust to those sources -- in other words, one could treat wafers and sites as noise factors in such an experiment.

Treating wafers and sites as random effects allows calculation of variance estimates

Because the wafers and the sites represent unwanted sources of variation and because one of the objectives is to reduce the process sensitivity to these sources of variation, treating wafers and sites as random effects in the analysis of the data is a reasonable approach. In other words, nested variation is often another way of saying nested random effects or nested sources of noise. If the factors "wafers" and "sites", are treated as random effects, then it is possible to estimate a variance component due to each source of variation through analysis of variance techniques. Once estimates of the variance components have been obtained, an investigator is then able to determine the largest source of variation in the process under experimentation, and also determine the magnitudes of the other sources of variation in relation to the largest source.

Nested random effects same as nested variation

If an experiment or process has nested variation, the experiment or process has multiple sources of random error that affect its output. Having nested random effects in a model is the same thing as having nested variation in a model.

Split-Plot Designs

Split-plot designs often arise when

Split-plot designs result when a particular type of restricted randomization has occurred during the experiment. A simple factorial experiment can result in a split-plot type of design

because of the way the experiment was actually executed.

some factors are "hard to vary" or when batch processes are run

In many industrial experiments, three situations often occur:

1. some of the factors of interest may be 'hard to vary' while the remaining factors are easy to vary. As a result, the order in which the treatment combinations for the experiment are run is determined by the ordering of these 'hard-to-vary' factors
2. experimental units are processed together as a batch for one or more of the factors in a particular treatment combination
3. experimental units are processed individually, one right after the other, for the same treatment combination without resetting the factor settings for that treatment combination.

A split-plot experiment example

An experiment run under one of the above three situations usually results in a split-plot type of design. Consider an experiment to examine electroplating of aluminum (non-aqueous) on copper strips. The three factors of interest are: current (A); solution temperature (T); and the solution concentration of the plating agent (S). Plating rate is the measured response. There are a total of 16 copper strips available for the experiment. The treatment combinations to be run (orthogonally scaled) are listed below in standard order (i.e., they have not been randomized):

Table showing the design matrix

TABLE 5.6 Orthogonally Scaled Treatment Combinations from a 2^3 Full Factorial

Current	Temperature	Concentration
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

Concentration is hard to vary, so minimize the number of times it is changed

Consider running the experiment under the first condition listed above, with the factor solution concentration of the plating agent (S) being hard to vary. Since this factor is hard to vary, the experimenter would like to randomize the treatment combinations so that the solution concentration factor has a minimal number of changes. In other words, the randomization of the treatment runs is restricted somewhat by the level of the solution concentration factor.

Randomize so that all runs for one level of concentration are run first

As a result, the treatment combinations might be randomized such that those treatment runs corresponding to one level of the concentration (-1) are run first. Each copper strip is individually plated, meaning only one strip at a time is placed in the solution for a given treatment combination. Once the four runs at the low level of solution concentration have been completed, the solution is changed to the high level of concentration (1), and the remaining four runs of the experiment are performed (where again, each strip is individually plated).

Performing replications

Once one complete replicate of the experiment has been completed, a second replicate is performed with a set of four copper strips processed for a given level of solution concentration before changing the concentration and processing the remaining four strips. Note that the levels for the remaining two factors can still be randomized. In addition, the level of concentration that is run first in the replication runs can also be randomized.

Whole plot and subplot factors

Running the experiment in this way results in a split-plot design. Solution concentration is known as the *whole plot factor* and the *subplot factors* are the current and the solution temperature.

Definition of experimental units and whole plot and subplot factors for this experiment

A split-plot design has more than one size experimental unit. In this experiment, one size experimental unit is an individual copper strip. The treatments or factors that were applied to the individual strips are solution temperature and current (these factors were changed each time a new strip was placed in the solution). The other or larger size experimental unit is a set of four copper strips. The treatment or factor that was applied to a set of four strips is solution concentration (this factor was changed after four strips were processed). The smaller size experimental unit is referred to as the subplot experimental unit, while the larger experimental unit is referred to as the whole plot unit.

Each size of experimental unit leads to an error term in the model for the experiment

There are 16 subplot experimental units for this experiment. Solution temperature and current are the subplot factors in this experiment. There are four whole-plot experimental units in this experiment. Solution concentration is the whole-plot factor in this experiment. Since there are two sizes of experimental units, there are two error terms in the model, one that corresponds to the whole-plot error or whole-plot experimental unit and one that corresponds to the subplot error or subplot experimental unit.

Partial ANOVA table

The ANOVA table for this experiment would look, in part, as follows:

Source	DF
Replication	1

```
Concentration                                1
Error (Whole plot) = Rep*Conc                1
Temperature                                  1
Rep*Temp                                     1
Current                                      1
Rep*Current                                  1
Temp*Conc                                    1
Rep*Temp*Conc                                1
Temp*Current                                 1
Rep*Temp*Current                             1
Current*Conc                                 1
Rep*Current*Conc                             1
Temp*Current*Conc                            1
Error (Subplot) =Rep*Temp*Current*Conc       1
```

The first three sources are from the whole-plot level, while the next 12 are from the subplot portion. A normal probability plot of the 12 subplot term estimates could be used to look for significant terms.

A batch process leads to a different experiment - also a strip-plot

Consider running the experiment under the second condition listed above (i.e., a batch process) for which four copper strips are placed in the solution at one time. A specified level of current can be applied to an individual strip within the solution. The same 16 treatment combinations (a replicated 2^3 factorial) are run as were run under the first scenario. However, the way in which the experiment is performed would be different. There are four treatment combinations of solution temperature and solution concentration: (-1, -1), (-1, 1), (1, -1), (1, 1). The experimenter randomly chooses one of these four treatments to set up first. Four copper strips are placed in the solution. Two of the four strips are randomly assigned to the low current level. The remaining two strips are assigned to the high current level. The plating is performed and the response is measured. A second treatment combination of temperature and concentration is chosen and the same procedure is followed. This is done for all four temperature / concentration combinations.

This also a split-plot design

Running the experiment in this way also results in a split-plot design in which the whole-plot factors are now solution concentration and solution temperature, and the subplot factor is current.

Defining experimental units

In this experiment, one size experimental unit is again an individual copper strip. The treatment or factor that was applied to the individual strips is current (this factor was changed each time for a different strip within the solution). The other or larger size experimental unit is again a set of four copper strips. The treatments or factors that were applied to a set of four strips are solution concentration and solution temperature (these factors were changed after four strips were processed).

Subplot experimental unit

The smaller size experimental unit is again referred to as the subplot experimental unit. There are 16 subplot experimental units for this experiment. Current is the subplot factor in this experiment.

Whole-plot experimental unit

The larger-size experimental unit is the whole-plot experimental unit. There are four whole plot experimental units in this experiment and solution concentration and solution temperature are the whole plot factors in this experiment.

Two error terms in the model

There are two sizes of experimental units and there are two error terms in the model: one that corresponds to the whole-plot error or whole-plot experimental unit, and one that corresponds to the subplot error or subplot experimental unit.

Partial ANOVA table

The ANOVA for this experiment looks, in part, as follows:

Source	DF
Concentration	1
Temperature	1
Error (Whole plot) = Conc*Temp	1
Current	1
Conc*Current	1
Temp*Current	1
Conc*Temp*Current	1
Error (Subplot)	8

The first three sources come from the whole-plot level and the next 5 come from the subplot level. Since there are 8 degrees of freedom for the subplot error term, this MSE can be used to test each effect that involves current.

Running the experiment under the third scenario

Consider running the experiment under the third scenario listed above. There is only one copper strip in the solution at one time. However, two strips, one at the low current and one at the high current, are processed one right after the other under the same temperature and concentration setting. Once two strips have been processed, the concentration is changed and the temperature is reset to another combination. Two strips are again processed, one after the other, under this temperature and concentration setting. This process is continued until all 16 copper strips have been processed.

This also a split-plot design

Running the experiment in this way also results in a split-plot design in which the whole-plot factors are again solution concentration and solution temperature and the subplot factor is current. In this experiment, one size experimental unit is an individual copper strip. The treatment or factor that was applied to the individual strips is current (this factor was changed each time for a different strip within the solution). The other or larger-size experimental unit is a set of two copper strips. The treatments or factors that were applied to a pair of two strips are solution concentration and solution temperature (these factors were changed after two strips were processed). The smaller size experimental unit is referred to as the subplot experimental unit.

Current is the

There are 16 subplot experimental units for this experiment.

subplot factor and temperature and concentration are the whole plot factors

Current is the subplot factor in the experiment. There are eight whole-plot experimental units in this experiment. Solution concentration and solution temperature are the whole plot factors. There are two error terms in the model, one that corresponds to the whole-plot error or whole-plot experimental unit, and one that corresponds to the subplot error or subplot experimental unit.

Partial ANOVA table

The ANOVA for this (third) approach is, in part, as follows:

Source	DF
Concentration	1
Temperature	1
Conc*Temp	1
Error (Whole plot)	4
Current	1
Conc*Current	1
Temp*Current	1
Conc*Temp*Current	1
Error (Subplot)	4

The first four terms come from the whole-plot analysis and the next 5 terms come from the subplot analysis. Note that we have separate error terms for both the whole plot and the subplot effects, each based on 4 degrees of freedom.

Primary distinction of split-plot designs is that they have more than one experimental unit size (and therefore more than one error term)

As can be seen from these three scenarios, one of the major differences in split-plot designs versus simple factorial designs is the number of different sizes of experimental units in the experiment. Split-plot designs have more than one size experimental unit, i.e., more than one error term. Since these designs involve different sizes of experimental units and different variances, the standard errors of the various mean comparisons involve one or more of the variances. Specifying the appropriate model for a split-plot design involves being able to identify each size of experimental unit. The way an experimental unit is defined relative to the design structure (for example, a completely randomized design versus a randomized complete block design) and the treatment structure (for example, a full 2^3 factorial, a resolution V half fraction, a two-way treatment structure with a control group, etc.). As a result of having greater than one size experimental unit, the appropriate model used to analyze split-plot designs is a mixed model.

Using wrong model can lead to invalid conclusions

If the data from an experiment are analyzed with only one error term used in the model, misleading and invalid conclusions can be drawn from the results. For a more detailed discussion of these designs and the appropriate analysis procedures, see Milliken, Analysis of Messy Data, Vol. 1.

Strip-Plot Designs

Strip-plot

Similar to a split-plot design, a strip-plot design can result

desgins often result from experiments that are conducted over two or more process steps

when some type of restricted randomization has occurred during the experiment. A simple factorial design can result in a strip-plot design depending on how the experiment was conducted. Strip-plot designs often result from experiments that are conducted over two or more process steps in which each process step is a batch process, i.e., completing each treatment combination of the experiment requires more than one processing step with experimental units processed together at each process step. As in the split-plot design, strip-plot designs result when the randomization in the experiment has been restricted in some way. As a result of the restricted randomization that occurs in strip-plot designs, there are multiple sizes of experimental units. Therefore, there are different error terms or different error variances that are used to test the factors of interest in the design. A traditional strip-plot design has three sizes of experimental units.

Example with two steps and three factor variables

Consider the following example from the semiconductor industry. An experiment requires an implant step and an anneal step. At both the anneal and the implant steps there are three factors to test. The implant process accommodates 12 wafers in a batch, and implanting a single wafer under a specified set of conditions is not practical nor does doing so represent economical use of the implanter. The anneal furnace can handle up to 100 wafers.

Explanation of the diagram that illustrates the design structure of the example

The figure below shows the design structure for how the experiment was run. The rectangles at the top of the diagram represent the settings for a two-level factorial design for the three factors in the implant step (A, B, C). Similarly, the rectangles at the lower left of the diagram represent a two-level factorial design for the three factors in the anneal step (D, E, F).

The arrows connecting each set of rectangles to the grid in the center of the diagram represent a randomization of trials in the experiment. The horizontal elements in the grid represent the experimental units for the anneal factors. The vertical elements in the grid represent the experimental units for the implant factors. The intersection of the vertical and horizontal elements represents the experimental units for the interaction effects between the implant factors and the anneal factors. Therefore, this experiment contains three sizes of experimental units, each of which has a unique error term for estimating the significance of effects.

Diagram of the split-plot design

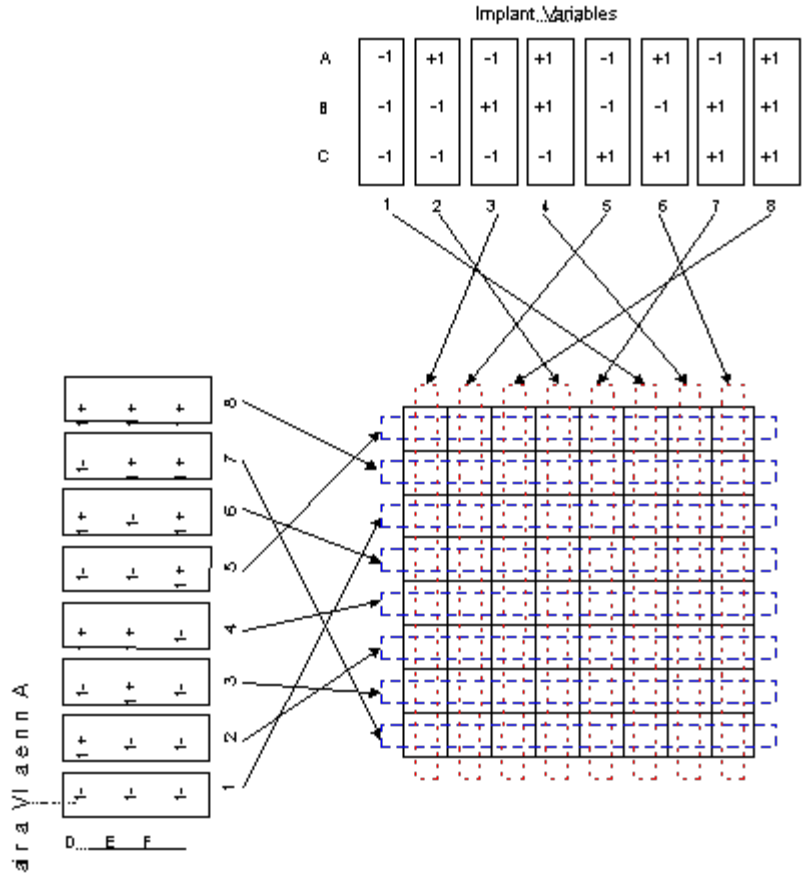


FIGURE 5.16 Diagram of a strip-plot design involving two process steps with three factors in each step

Physical meaning of the experimental units

To put actual physical meaning to each of the experimental units in the above example, consider each cell in the grid as an individual wafer. A batch of eight wafers goes through the implant step first. According to the figure, treatment combination #3 in factors A, B, and C is the first implant treatment run. This implant treatment is applied to all eight wafers at once. Once the first implant treatment is finished, another set of eight wafers is implanted with treatment combination #5 of factors A, B, and C. This continues until the last batch of eight wafers is implanted with treatment combination #6 of factors A, B, and C. Once all of the eight treatment combinations of the implant factors have been run, the anneal step starts. The first anneal treatment combination to be run is treatment combination #5 of factors D, E, and F. This anneal treatment combination is applied to a set of eight wafers, with each of these eight wafers coming from one of the eight implant treatment combinations. After this first batch of wafers has been annealed, the second anneal treatment is applied to a second batch of eight wafers, with these eight wafers coming from one each of the eight implant treatment combinations. This is continued until the last batch of eight wafers has been implanted with a particular combination of factors D, E, and F.

Three sizes of

Running the experiment in this way results in a strip-plot

experimental units	design with three sizes of experimental units. A set of eight wafers that are implanted together is the experimental unit for the implant factors A, B, and C and for all of their interactions. There are eight experimental units for the implant factors. A different set of eight wafers are annealed together. This different set of eight wafers is the second size experimental unit and is the experimental unit for the anneal factors D, E, and F and for all of their interactions. The third size experimental unit is a single wafer. This is the experimental unit for all of the interaction effects between the implant factors and the anneal factors.
Replication	Actually, the above figure of the strip-plot design represents one block or one replicate of this experiment. If the experiment contains no replication and the model for the implant contains only the main effects and two-factor interactions, the three-factor interaction term A*B*C (1 degree of freedom) provides the error term for the estimation of effects within the implant experimental unit. Invoking a similar model for the anneal experimental unit produces the three-factor interaction term D*E*F for the error term (1 degree of freedom) for effects within the anneal experimental unit.
Further information	For more details about strip-plot designs, see Milliken and Johnson (1987) or Miller (1997).

5.5.6. What are Taguchi designs?

Taguchi designs are related to fractional factorial designs - many of which are large screening designs

Genichi Taguchi, a Japanese engineer, proposed several approaches to experimental designs that are sometimes called "Taguchi Methods." These methods utilize two-, three-, and mixed-level fractional factorial designs. Large screening designs seem to be particularly favored by Taguchi adherents.

Taguchi refers to experimental design as "off-line quality control" because it is a method of ensuring good performance in the design stage of products or processes. Some experimental designs, however, such as when used in evolutionary operation, can be used on-line while the process is running. He has also published a booklet of design nomograms ("Orthogonal Arrays and Linear Graphs," 1987, American Supplier Institute) which may be used as a design guide, similar to the table of fractional factorial designs given previously in Section 5.3. Some of the well-known Taguchi orthogonal arrays (L9, L18, L27 and L36) were given earlier when three-level, mixed-level and fractional factorial designs were discussed.

If these were the only aspects of "Taguchi Designs," there would be little additional reason to consider them over and above our previous discussion on factorials. "Taguchi" designs are similar to our familiar fractional factorial designs. However, Taguchi has introduced several noteworthy new ways of conceptualizing an experiment that are very valuable, especially in product development and industrial engineering, and we will look at two of his main ideas, namely Parameter Design and Tolerance Design.

Parameter Design

Taguchi advocated using inner and outer array designs to take into account noise factors (outer) and design factors (inner)

The aim here is to make a product or process less variable (more robust) in the face of variation over which we have little or no control. A simple fictitious example might be that of the starter motor of an automobile that has to perform reliably in the face of variation in ambient temperature and varying states of battery weakness. The engineer has control over, say, number of armature turns, gauge of armature wire, and ferric content of magnet alloy.

Conventionally, one can view this as an experiment in five factors. Taguchi has pointed out the usefulness of viewing it as a set-up of three inner array factors (turns, gauge, ferric %) over which we have design control, plus an outer array of factors over which we have control only in the laboratory (temperature, battery voltage).

Pictorial representation of Taguchi designs

Pictorially, we can view this design as being a conventional design in the inner array factors (compare Figure 3.1) with the addition of a "small" outer array factorial design at each corner of the "inner array" box.

Let I1 = "turns," I2 = "gauge," I3 = "ferric %," E1 = "temperature," and E2 = "voltage." Then we construct a 2^3 design "box" for the I's, and at each of the eight corners so constructed, we place a 2^2 design "box" for the E's, as is shown in Figure 5.17.

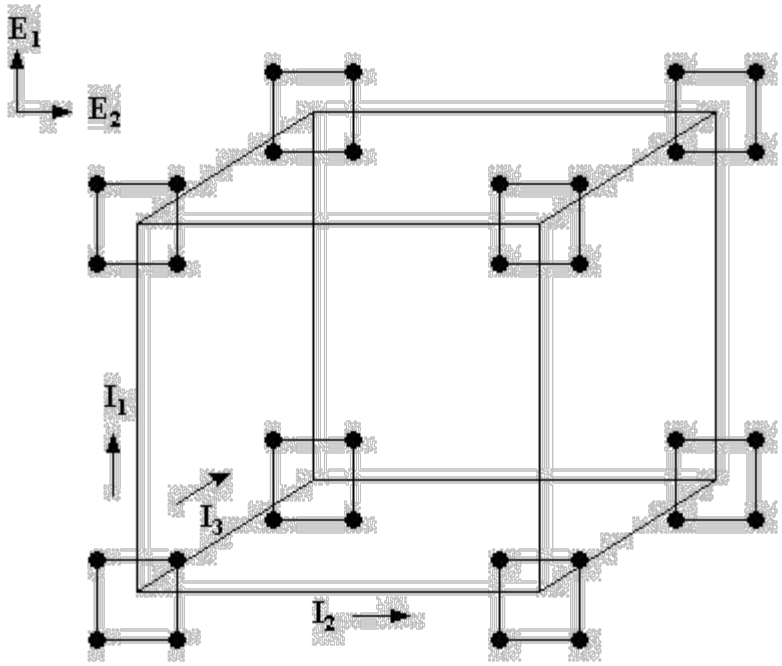


FIGURE 5.17 Inner 2^3 and outer 2^2 arrays for robust design with `I' the inner array, `E' the outer array.

An example of an inner and outer array designed experiment

We now have a total of 8x4 = 32 experimental settings, or runs. These are set out in Table 5.7, in which the 2^3 design in the I's is given in standard order on the left of the table and the 2^2 design in the E's is written out sideways along the top. Note that the experiment would not be run in the standard order but should, as always, have its runs randomized. The output measured is the percent of (theoretical) maximum torque.

Table showing the Taguchi design and the responses from the experiment

TABLE 5.7 Design table, in standard order(s) for the parameter design of Figure 5.9

Run Number				1	2	3	4		
				E1	**-1**	**+1**	**-1**	**+1**	**Output**	**Output**
I1	**I2**	**I3**		**E2**	**-1**	**-1**	**+1**	**+1**	**MEAN**	**STD. DEV**

1	**-1**	**-1**	**-1**	75	86	67	98	**81.5**	**13.5**
2	**+1**	**-1**	**-1**	87	78	56	91	**78.0**	**15.6**
3	**-1**	**+1**	**-1**	77	89	78	8	**63.0**	**37.1**
4	**+1**	**+1**	**-1**	95	65	77	95	**83.0**	**14.7**
5	**-1**	**-1**	**+1**	78	78	59	94	**77.3**	**14.3**
6	**+1**	**-1**	**+1**	56	79	67	94	**74.0**	**16.3**
7	**-1**	**+1**	**+1**	79	80	66	85	**77.5**	**8.1**
8	**+1**	**+1**	**+1**	71	80	73	95	**79.8**	**10.9**

Interpretation of the table

Note that there are four outputs measured on each row. These correspond to the four `outer array' design points at each corner of the `outer array' box. As there are eight corners of the outer array box, there are eight rows in all.

Each row yields a mean and standard deviation % of maximum torque. Ideally there would be one row that had both the highest average torque and the lowest standard deviation (variability). Row 4 has the highest torque and row 7 has the lowest variability, so we are forced to compromise. We can't simply `pick the winner.'

Use contour plots to see inside the box

One might also observe that all the outcomes occur at the corners of the design `box', which means that we cannot see `inside' the box. An optimum point might occur within the box, and we can search for such a point using contour plots. Contour plots were illustrated in the example of response surface design analysis given in Section 4.

Fractional factorials

Note that we could have used fractional factorials for either the inner or outer array designs, or for both.

Tolerance Design

Taguchi also advocated tolerance studies to determine, based on a loss or cost function, which variables have critical tolerances that need to be tightened

This section deals with the problem of how, and when, to specify tightened tolerances for a product or a process so that quality and performance/productivity are enhanced. Every product or process has a number—perhaps a large number—of components. We explain here how to identify the critical components to target when tolerances have to be tightened.

It is a natural impulse to believe that the quality and performance of any item can easily be improved by merely tightening up on some or all of its tolerance requirements. By this we mean that if the old version of the item specified, say, machining to ± 1 micron, we naturally believe that we can obtain better performance by specifying machining to ± ½ micron.

This can become expensive, however, and is often not a guarantee of much better performance. One has merely to witness the high initial and maintenance costs of such tight-tolerance-level items as space vehicles, expensive automobiles, etc. to realize that tolerance design—the selection of critical tolerances and the re-specification of those critical

tolerances—is not a task to be undertaken without careful thought. In fact, it is recommended that *only after extensive parameter design studies have been completed* should tolerance design be performed *as a last resort* to improve quality and productivity.

Example

Example: measurement of electronic component made up of two components

Customers for an electronic component complained to their supplier that the measurement reported by the supplier on the as-delivered items appeared to be imprecise. The supplier undertook to investigate the matter.

The supplier's engineers reported that the measurement in question was made up of two components, which we label x and y, and the final measurement M was reported according to the standard formula

$$\boldsymbol{M = K\,x/y}$$

with `K' a known physical constant. Components x and y were measured separately in the laboratory using two different techniques, and the results combined by software to produce M. Buying new measurement devices for both components would be prohibitively expensive, and it was not even known by how much the x or y component tolerances should be improved to produce the desired improvement in the precision of M.

Taylor series expansion

Assume that in a measurement of a standard item the `true' value of x is x_0 and for y it is y_0. Let f(x, y) = M; then the Taylor Series expansion for f(x, y) is

$$\begin{aligned} f(x,y) \;=\; & f(x_o,y_o) + (x-x_o)\tfrac{df}{dx} + (y-y_o)\tfrac{df}{dy} + (x-x_o)^2\tfrac{d^2f}{dx^2} \\ & +(y-y_o)^2\tfrac{d^2y}{dy^2} + (x-x_o)(y-y_o)\tfrac{d^2f}{dxdy} + \\ & (\text{higher-order terms}) \end{aligned}$$

with all the partial derivatives, `df/dx', etc., evaluated at (x_0, y_0).

Apply formula to M

Applying this formula to $M(x, y)$ = Kx/y, we obtain

$$\begin{aligned} M(x,y) \;=\; & K\tfrac{x_o}{y_o} + (x-x_o)\tfrac{K}{y_o} - (y-y_o)\tfrac{Kx_o}{y_o^2} - 2(y-y_o)^2\tfrac{K}{y_o^3} \\ & -(x-x_o)(y-y_o)\tfrac{K}{y_o^2} + (\text{higher-order terms}) \end{aligned}$$

It is assumed known from experience that the measurements of x show a distribution with an average value x_0, and with a standard deviation $\boldsymbol{\sigma}_x$ = 0.003 x-units.

Assume distribution of x is normal

In addition, we assume that the distribution of x is normal. Since 99.74% of a normal distribution's range is covered by 6$\boldsymbol{\sigma}$, we take 3$\boldsymbol{\sigma}_x$ = 0.009 x-units to be the existing tolerance T_x for measurements on x. That is, T_x = ± 0.009 x-units is the `play' around x_0 that we expect from the existing measurement system.

Assume distribution of y is normal

It is also assumed known that the y measurements show a normal distribution around y_0, with standard deviation σ_y = 0.004 y-units. Thus $T_y = \pm 3\sigma_y = \pm 0.012$.

Worst case values

Now $\pm T_x$ and $\pm T_y$ may be thought of as `worst case' values for $(x\text{-}x_0)$ and $(y\text{-}y_0)$. Substituting T_x for $(x\text{-}x_0)$ and T_y for $(y\text{-}y_0)$ in the expanded formula for $M(x, y)$, we have

$$M_T = K\frac{x_o}{y_o} + T_x\frac{K}{y_o} - T_y\frac{Kx_o}{y_o^2} - 2T_y^2\frac{K}{y_o^3} - T_xT_y\frac{K}{y_o^2} + (\text{higher-order terms})$$

Drop some terms

The T_y^2 and T_xT_y terms, and all terms of higher order, are going to be at least an order of magnitude smaller than terms in T_x and in T_y, and for this reason we drop them, so that

$$M_T = K\frac{x_o}{y_o} + T_x\frac{K}{y_o} - T_y\frac{Kx_o}{y_o^2}$$

Worst case Euclidean distance

Thus, a `worst case' Euclidean distance δ of $M(x, y)$ from its ideal value Kx_0/y_0 is (approximately)

$$\Delta = \sqrt{(T_x\frac{K}{y_o})^2 + (T_y\frac{Kx_o}{y_o^2})^2} = \sqrt{(0.009\frac{K}{y_o})^2 + (0.012\frac{Kx_o}{y_o^2})^2}$$

This shows the relative contributions of the components to the variation in the measurement.

Economic decision

As y_0 is a known quantity and reduction in T_x and in T_y each carries its own price tag, it becomes an economic decision whether one should spend resources to reduce T_x or T_y, or both.

Simulation an alternative to Taylor series approximation

In this example, we have used a Taylor series approximation to obtain a simple expression that highlights the benefit of T_x and T_y. Alternatively, one might simulate values of $M = K*x/y$, given a specified (T_x,T_y) and (x_0,y_0), and then summarize the results with a model for the variability of M as a function of (T_x,T_y).

Functional form may not be available

In other applications, no functional form is available and one must use experimentation to empirically determine the optimal tolerance design. See Bisgaard and Steinberg (1997).

5.5.7. What are John's 3/4 fractional factorial designs?

John's designs require only 3/4 of the number of runs a full 2^n factorial would require

Three-quarter (¾) designs are two-level factorial designs that require only three-quarters of the number of runs of the `original' design. For example, instead of making all of the sixteen runs required for a $2^4$ fractional factorial design, we need only run 12 of them. Such designs were invented by Professor Peter John of the University of Texas, and are sometimes called`John's ¾ designs.'

Three-quarter fractional factorial designs can be used to save on resources in two different contexts. In one scenario, we may wish to perform additional runs after having completed a fractional factorial, so as to de-alias certain specific interaction patterns. Second , we may wish to use a ¾ design to begin with and thus save on 25% of the run requirement of a regular design.

Semifolding Example

Four experimental factors

We have four experimental factors to investigate, namely X1, X2, X3, and X4, and we have designed and run a 2^{4-1} fractional factorial design. Such a design has eight runs, or rows, if we don't count center point runs (or replications).

Resolution IV design

The 2^{4-1} design is of resolution IV, which means that main effects are confounded with, at worst, three-factor interactions, and two-factor interactions are confounded with other two factor interactions.

Design matrix

The design matrix, in standard order, is shown in Table 5.8 along with all the two-factor interaction columns. Note that the column for X4 is constructed by multiplying columns for X1, X2, and X3 together (i.e., 4=123).

Table 5.8 The 2^{4-1} design plus 2-factor interaction columns shown in standard order. Note that 4=123.

Run					Two-Factor Interaction Columns					
Number	X1	X2	X3	X4	X1*X2	X1*X3	X1*X4	X2*X3	X2*X4	X3*X4
1	-1	-1	-1	-1	+1	+1	+1	+1	+1	+1
2	+1	-1	-1	+1	-1	-1	+1	+1	-1	-1
3	-1	+1	-1	+1	-1	+1	-1	-1	+1	-1
4	+1	+1	-1	-1	+1	-1	-1	-1	-1	+1
5	-1	-1	+1	+1	+1	-1	-1	-1	-1	+1

6	+1	-1	+1	-1	-1	+1	-1	-1	+1	-1
7	-1	+1	+1	-1	-1	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

Confounding of two-factor interactions

Note also that 12=34, 13=24, and 14=23. These follow from the generating relationship 4=123 and tells us that we cannot estimate any two-factor interaction that is free of some other two-factor alias.

Estimating two-factor interactions free of confounding

Suppose that we became interested in estimating some or all of the two-factor interactions that involved factor X1; that is, we want to estimate one or more of the interactions 12, 13, and 14 free of two-factor confounding.

One way of doing this is to run the `other half' of the design—an additional eight rows formed from the relationship 4 = -123. Putting these two `halves' together—the original one and the new one, we'd obtain a 2^4 design in sixteen runs. Eight of these runs would already have been run, so all we'd need to do is run the remaining half.

Alternative method requiring fewer runs

There is a way, however, to obtain what we want while adding only four more runs. These runs are selected in the following manner: take the four rows of Table 5.8 that have `-1' in the `X1' column and switch the `-' sign under X1 to `+' to obtain the four-row table of Table 5.9. This is called a foldover on X1, choosing the subset of runs with X1 = -1. Note that this choice of 4 runs is not unique, and that if the initial design suggested that X1 = -1 were a desirable level, we would have chosen to experiment at the other four treatment combinations that were omitted from the initial design.

Table of the additional design points

TABLE 5.9 Foldover on `X1' of the 2^{4-1} design of Table 5.5

Run Number	X1	X2	X3	X4
9	+1	-1	-1	-1
10	+1	+1	-1	+1
11	+1	-1	+1	+1
12	+1	+1	+1	-1

Table with new design points added to the original design points

Add this new block of rows to the bottom of Table 5.8 to obtain a design in twelve rows. We show this in Table 5.10 and also add in the two-factor interactions as well for illustration (not needed when we do the runs).

TABLE 5.10 A twelve-run design based on the 2^{4-1} also showing all two-factor interaction columns

Run					Two-Factor Interaction Columns					
Number	X1	X2	X3	X4	X1*X2	X1*X3	X1*X4	X2*X3	X2*X4	X3*X4

1	-1	-1	-1	-1	+1	+1	+1	+1	+1	+1
2	+1	-1	-1	+1	-1	-1	+1	+1	-1	-1
3	-1	+1	-1	+1	-1	+1	-1	-1	+1	-1
4	+1	+1	-1	-1	+1	-1	-1	-1	-1	+1
5	-1	-1	+1	+1	+1	-1	-1	-1	-1	+1
6	+1	-1	+1	-1	-1	+1	-1	-1	+1	-1
7	-1	+1	+1	-1	-1	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
1	+1	-1	-1	-1	-1	-1	-1	+1	+1	+1
10	+1	+1	-1	+1	+1	-1	+1	-1	+1	-1
11	+1	-1	+1	+1	-1	+1	+1	-1	-1	+1
12	+1	+1	+1	-1	+1	+1	-1	+1	-1	-1

Design is resolution V

Examine the two-factor interaction columns and convince yourself that no two are alike. This means that no two-factor interaction involving X1 is aliased with any other two-factor interaction. Thus, the design is resolution V, which is not always the case when constructing these types of ¾ foldover designs.

Estimating X1 two-factor interactions

What we now have is a design with 12 runs, with which we can estimate all the two-factor interactions involving X1 free of aliasing with any other two-factor interaction. It is called a ¾ design because it has ¾ the number of rows of the next regular factorial design (a 2^4).

Standard errors of effect estimates

If one fits a model with an intercept, a block effect, the four main effects and the six two-factor interactions, then each coefficient has a standard error of $\sigma/8^{1/2}$ - instead of $\sigma/12^{1/2}$ - because the design is not orthogonal and each estimate is correlated with two other estimates. Note that no degrees of freedom exists for estimating σ. Instead, one should plot the 10 effect estimates using a normal (or half-normal) effects plot to judge which effects to declare significant.

Further information

For more details on ¾ fractions obtained by adding a follow-up design that is half the size of the original design, see Mee and Peralta (2000).

Next we consider an example in which a ¾ fraction arises when the (¾) 2^{k-p} design is planned from the start because it is an efficient design that allows estimation of a sufficient number of effects.

A 48-Run 3/4 Design Example

Estimate all main effects and two-factor interactions

Suppose we wish to run an experiment for k=8 factors, with which we want to estimate all main effects and two-factor interactions. We could use the 2^{8-2}_V design described in the summary table of fractional factorial designs, but this would require a 64-run experiment to estimate the 1 + 8 + 28 = 37 desired coefficients. In this context, and especially for larger

for 8 factors

resolution V designs, ¾ of the design points will generally suffice.

Construction of the 48-run design

The 48 run-design is constructed as follows: start by creating the full 2_V^{8-2} design using the generators 7 = 1234 and 8 = 1256. The defining relation is I = 12347 = 12568 = 345678 (see the summary table details for this design).

Next, arrange these 64 treatment combinations into four blocks of size 16, blocking on the interactions 135 and 246 (i.e., block 1 has 135 = 246 = -1 runs, block 2 has 135 = -1, 246 = +1, block 3 has 135 = +1, 246 = -1 and block 4 has 135 = 246 = +1). If we exclude the first block in which 135 = 246 = -1, we have the desired ¾ design reproduced below (the reader can verify that these are the runs described in the summary table, excluding the runs numbered 1, 6, 11, 16, 18, 21, 28, 31, 35, 40, 41,46, 52, 55, 58 and 61).

Table containing the design matrix

X1	X2	X3	X4	X5	X6	X7	X8
+1	-1	-1	-1	-1	-1	-1	-1
-1	+1	-1	-1	-1	-1	-1	-1
+1	+1	-1	-1	-1	-1	+1	+1
-1	-1	+1	-1	-1	-1	-1	+1
-1	+1	+1	-1	-1	-1	+1	-1
+1	+1	+1	-1	-1	-1	-1	+1
-1	-1	-1	+1	-1	-1	-1	+1
+1	-1	-1	+1	-1	-1	+1	-1
+1	+1	-1	+1	-1	-1	-1	+1
-1	-1	+1	+1	-1	-1	+1	+1
+1	-1	+1	+1	-1	-1	-1	-1
-1	+1	+1	+1	-1	-1	-1	-1
-1	-1	-1	-1	+1	-1	+1	-1
-1	+1	-1	-1	+1	-1	-1	+1
+1	+1	-1	-1	+1	-1	+1	-1
+1	-1	+1	-1	+1	-1	+1	+1
-1	+1	+1	-1	+1	-1	+1	+1
+1	+1	+1	-1	+1	-1	-1	-1
-1	-1	-1	+1	+1	-1	-1	-1
+1	-1	-1	+1	+1	-1	+1	+1
-1	+1	-1	+1	+1	-1	+1	+1
-1	-1	+1	+1	+1	-1	+1	-1
+1	-1	+1	+1	+1	-1	-1	+1
+1	+1	+1	+1	+1	-1	+1	-1
-1	-1	-1	-1	-1	+1	+1	-1
+1	-1	-1	-1	-1	+1	-1	+1
+1	+1	-1	-1	-1	+1	+1	-1

-1	-1	+1	-1	-1	+1	-1	-1
+1	-1	+1	-1	-1	+1	+1	+1
-1	+1	+1	-1	-1	+1	+1	+1
+1	-1	-1	+1	-1	+1	+1	+1
-1	+1	-1	+1	-1	+1	+1	+1
+1	+1	-1	+1	-1	+1	-1	-1
-1	-1	+1	+1	-1	+1	+1	-1
-1	+1	+1	+1	-1	+1	-1	+1
+1	+1	+1	+1	-1	+1	+1	-1
-1	-1	-1	-1	+1	+1	+1	+1
+1	-1	-1	-1	+1	+1	-1	-1
-1	+1	-1	-1	+1	+1	-1	-1
-1	-1	+1	-1	+1	+1	-1	+1
+1	-1	+1	-1	+1	+1	+1	-1
+1	+1	+1	-1	+1	+1	-1	+1
-1	-1	-1	+1	+1	+1	-1	+1
-1	+1	-1	+1	+1	+1	+1	-1
+1	+1	-1	+1	+1	+1	-1	+1
+1	-1	+1	+1	+1	+1	-1	-1
-1	+1	+1	+1	+1	+1	-1	-1
+1	+1	+1	+1	+1	+1	+1	+1

Good precision for coefficient estimates

This design provides 11 degrees of freedom for error and also provides good precision for coefficient estimates (some of the coefficients have a standard error of $\sigma/\sqrt{32}$ and some have a standard error of $\sigma/\sqrt{42.55}$).

Further information

More about John's ¾ designs can be found in John (1971) or Diamond (1989).

5.5.8. What are small composite designs?

Small composite designs save runs, compared to Resolution V response surface designs, by adding star points to a Resolution III design

Response surface designs (RSD) were described earlier. A typical RSD requires about 13 runs for 2 factors, 20 runs for 3 factors, 31 runs for 4 factors, and 32 runs for 5 factors. It is obvious that, once you have four or more factors you wish to include in a RSD, you will need more than one lot (i.e., batch) of experimental units for your basic design. This is what most statistical software today will give you. However, there is a way to cut down on the number of runs, as suggested by H.O. Hartley in his paper '*Smallest Composite Designs for Quadratic Response Surfaces*', published in Biometrics, December 1959.

This method addresses the theory that using a Resolution V design as the smallest fractional design to create a RSD is unnecessary. The method adds star points to designs of Resolution III and uses the star points to clear the main effects of aliasing with the two-factor interactions. The resulting design allows estimation of the higher-order interactions. It also provides poor interaction coefficient estimates and should not be used unless the error variability is negligible compared to the systematic effects of the factors.

Useful for 4 or 5 factors

This could be particularly useful when you have a design containing four or five factors and you wish to only use the experimental units from one lot (i.e., batch).

Table containing design matrix for four factors

The following is a design for four factors. You would want to randomize these runs before implementing them; -1 and +1 represent the low and high settings, respectively, of each factor.

TABLE 5.11 Four factors: Factorial design section is based on a generator of I = X1*X2*X3, Resolution III; $-\alpha$ and $+\alpha$ are the star points, calculated beyond the factorial range; 0 represents the midpoint of the factor range.

Row	X1	X2	X3	X4
1	+1	-1	-1	-1
2	-1	+1	-1	-1

3	-1	-1	+1	-1
4	+1	+1	+1	-1
5	+1	-1	-1	+1
6	-1	+1	-1	+1
7	-1	-1	+1	+1
8	+1	+1	+1	+1
9	$-\alpha$	0	0	0
10	α	0	0	0
11	0	$-\alpha$	0	0
12	0	α	0	0
13	0	0	$-\alpha$	0
14	0	0	α	0
15	0	0	0	$-\alpha$
16	0	0	0	α
17	0	0	0	0
18	0	0	0	0
19	0	0	0	0
20	0	0	0	0

Determining α in Small Composite Designs

α based on number of treatment combinations in the factorial portion

To maintain rotatability for usual CCD's, the value of α is determined by the number of treatment combinations in the factorial portion of the central composite design:

$$\alpha = \left[number\ of\ factorial\ runs \right]^{1/4}$$

Small composite designs not rotatable

However, small composite designs are not rotatable, regardless of the choice of α. For small composite designs, α should not be smaller than [number of factorial runs]$^{1/4}$ nor larger than $k^{1/2}$.

5.5.9. An EDA approach to experimental design

Introduction

This section presents an exploratory data analysis (EDA) approach to analyzing the data from a designed experiment. This material is meant to complement, not replace, the more model-based approach for analyzing experiment designs given in section 4 of this chapter.

Choosing an appropriate design is discussed in detail in section 3 of this chapter.

Starting point

Problem category

The problem category we will address is the screening problem. Two characteristics of screening problems are:

1. There are many factors to consider.
2. Each of these factors may be either continuous or discrete.

Desired output

The desired output from the analysis of a screening problem is:

- A ranked list (by order of importance) of factors.
- The best settings for each of the factors.
- A good model.
- Insight.

Problem essentials

The essentials of the screening problem are:

- There are k factors with n observations.
- The generic model is:

$$Y = f(X_1, X_2, ..., X_k) + \varepsilon$$

Design type

In particular, the EDA approach is applied to 2^k full factorial and 2^{k-p} fractional factorial designs.

An EDA approach is particularly applicable to screening designs because we are in the preliminary stages of understanding our process.

EDA philosophy

EDA is not a single technique. It is an approach to analyzing data.

- EDA is data-driven. That is, we do not assume an initial model. Rather, we attempt to let the data speak for themselves.
- EDA is question-based. That is, we select a technique to answer one or more questions.
- EDA utilizes multiple techniques rather than depending on a single technique. Different plots have a different basis, focus, and sensitivities, and therefore may bring out different aspects of the data. When multiple techniques give us a redundancy of conclusions, this increases our confidence that our conclusions are valid. When they give conflicting conclusions, this may be giving us a clue as to the nature of our data.
- EDA tools are often graphical. The primary objective is to provide insight into the data, which graphical techniques often provide more readily than quantitative techniques.

10-Step process

The following is a 10-step EDA process for analyzing the data from 2^k full factorial and 2^{k-p} fractional factorial designs.

1. Ordered data plot
2. DOE scatter plot
3. DOE mean plot
4. Interaction effects matrix plot
5. Block plot
6. DOE Youden plot
7. |Effects| plot
8. Half-normal probability plot
9. Cumulative residual standard deviation plot
10. DOE contour plot

Each of these plots will be presented with the following format:

- Purpose of the plot
- Output of the plot
- Definition of the plot
- Motivation for the plot
- An example of the plot using the defective springs data
- A discussion of how to interpret the plot
- Conclusions we can draw from the plot for the defective springs data

Data set

Defective springs data

The plots presented in this section are demonstrated with a data set from Box and Bisgaard (1987).

These data are from a 2^3 full factorial data set that contains the following variables:

1. Response variable Y = percentage of springs without cracks
2. Factor 1 = oven temperature (2 levels: 1450 and 1600 F)
3. Factor 2 = carbon concentration (2 levels: 0.5% and 0.7%)
4. Factor 3 = quench temperature (2 levels: 70 and 120 F)

Y Percent Acceptable Temperature	X1 Oven Temperature	X2 Carbon Concentration	X3 Quench
67	-1	-1	-1
79	+1	-1	-1
61	-1	+1	-1
75	+1	+1	-1
59	-1	-1	+1
90	+1	-1	+1
52	-1	+1	+1
87	+1	+1	+1

5.5.9.1. Ordered data plot

Purpose The ordered data plot answers the following two questions:

1. What is the best setting (based on the data) for each of the *k* factors?
2. What is the most important factor?

In the above two questions, the terms "best" and "important" need more precise definitions.

Settings may be declared as "best" in three different ways:

1. "best" with respect to the data;
2. "best" on average;
3. "best" with respect to predicted values from an adequate model.

In the worst case, each of the above three criteria may yield different "best settings". If that occurs, then the three answers must be consolidated at the end of the 10-step process.

The ordered data plot will yield best settings based on the first criteria (data). That is, this technique yields those settings that correspond to the best response value, with the best value dependent upon the project goals:

1. maximization of the response;
2. minimization of the response;
3. hitting a target for the response.

This, in turn, trivially yields the best response value:

1. maximization: the observed maximum data point;
2. minimization: the observed minimum data point;
3. target: the observed data value closest to the specified target.

With respect to the most "important" factor, this by default refers to the single factor which causes the greatest change in the value of the response variable as we proceed from the "-" setting to the "+" setting of the factor. In practice, if a factor has one setting for the best and near-best response values and the opposite setting for the worst and near-worst response values, then that factor is usually the most important factor.

Output The output from the ordered data plot is:

1. Primary: Best setting for each of the *k* factors.
2. Secondary: The name of the most important factor.

Definition

An ordered data plot is formed by:

- Vertical Axis: The ordered (smallest to largest) raw response value for each of the *n* runs in the experiment.
- Horizontal Axis: The corresponding dummy run index (1 to *n*) with (at each run) a designation of the corresponding settings (- or +) for each of the *k* factors.

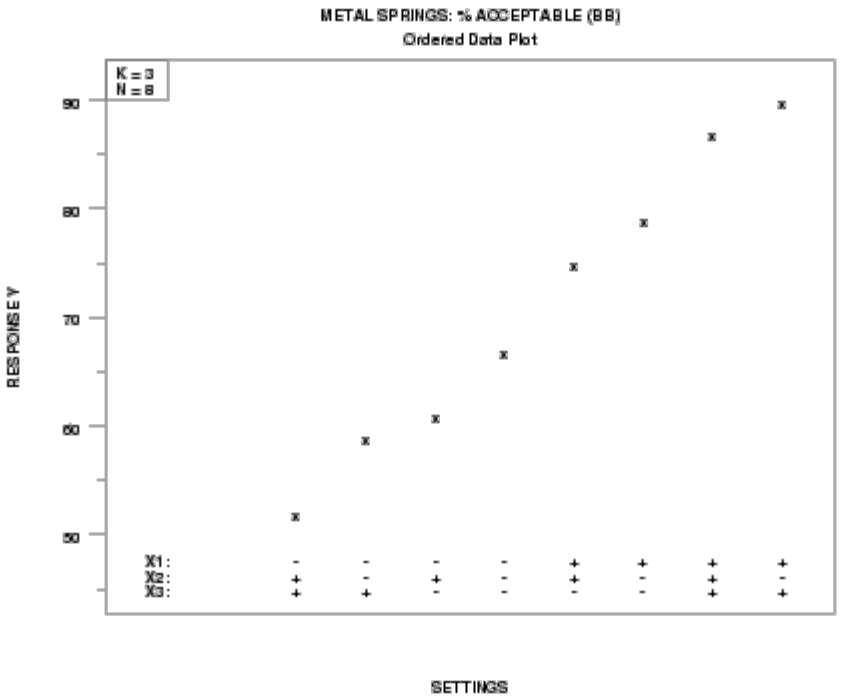
In essence, the ordered data plot may be viewed as a scatter plot of the ordered data versus a single *n*-treatment consolidation factor.

Motivation

To determine the best setting, an obvious place to start is the best response value. What constitutes "best"? Are we trying to maximize the response, minimize the response, or hit a specific target value? This non-statistical question must be addressed and answered by the analyst. For example, if the project goal is ultimately to achieve a large response, then the desired experimental goal is maximization. In such a case, the analyst would note from the plot the largest response value and the corresponding combination of the *k*-factor settings that yielded that best response.

Plot for defective springs data

Applying the ordered response plot for the defective springs data set yields the following plot.



How to interpret

From the ordered data plot, we look for the following:

1. best settings;
2. most important factor.

Best Settings (Based on the Data):

At the best (highest or lowest or target) response value, what are the corresponding settings for each of the *k* factors? This defines the best setting based on the raw data.

Most Important Factor:

For the best response point and for the nearby neighborhood of near-best response points, which (if any) of the *k* factors has consistent settings? That is, for the subset of response values that is best or near-best, do all of these values emanate from an identical level of some factor?

Alternatively, for the best half of the data, does this half happen to result from some factor with a common setting? If yes, then the factor that displays such consistency is an excellent candidate for being declared the "most important factor". For a balanced experimental design, when all of the best/near-best response values come from one setting, it follows that all of the worst/near-worst response values will come from the other setting of that factor. Hence that factor becomes "most important".

At the bottom of the plot, step though each of the *k* factors and determine which factor, if any, exhibits such behavior. This defines the "most important" factor.

Conclusions for the defective springs data

The application of the ordered data plot to the defective springs data set results in the following conclusions:

1. Best Settings (Based on the Data):

   $(X_1, X_2, X_3) = (+, -, +) = (+1, -1, +1)$ is the best setting since

   1. the project goal is maximization of the percent acceptable springs;
   2. $Y = 90$ is the largest observed response value; and
   3. $(X_1, X_2, X_3) = (+, -, +)$ at $Y = 90$.

2. Most important factor:

   X_1 is the most important factor since the four largest response values (90, 87, 79, and 75) have factor X_1 at +1, and the four smallest response values (52, 59, 61, and 67) have factor X_1 at -1.

5.5.9.2. DOE scatter plot

Purpose The DOE (design of experiments) scatter plot answers the following three questions:

1. What are the most important factors?
2. What is the best setting for each of these important factors?
3. What data points are outliers?

In the above questions, the terms "important", "best", and "outliers" need clarification and specificity:

Important

A factor can be "important" if it leads to a significant shift in either the location or the variation of the response variable as we go from the "-" setting to the "+" setting of the factor. Both definitions are relevant and acceptable. The default definition of "important" in engineering/scientific applications is a shift in location. Unless specified otherwise, when a factor is claimed to be important, the implication is that the factor caused a large location shift in the response.

Best

A factor setting is "best" if it results in a typical response that is closest, in location, to the desired project goal (maximization, minimization, target). This desired project goal is an engineering, not a statistical, question, and so the desired optimization goal must be specified by the engineer.

Outlier

A data point is an "outlier" if it comes from a different probability distribution or from a different deterministic model than the remainder of the data. A single outlier in a data set can affect all effect estimates and so in turn can potentially invalidate the factor rankings in terms of importance.

Given the above definitions, the DOE scatter plot is a useful early-step tool for determining the important factors, best settings, and outliers. An alternate name for the DOE scatter plot is "main effects plot".

Output The output for the DOE scatter plot is:

1. Primary: Identification of the important factors.
2. Secondary: Best setting for these factors and identification of outliers.

Definition The DOE scatter plot is formed by

- Vertical Axis: The response (= the raw data) for a given setting (- or +) of a factor for each of the *k* factors.
- Horizontal Axis: The *k* factors, and the two settings (- and +) within each factor.

Motivation

The scatter plot is the primary data analysis tool for determining if and how a response relates to another factor. Determining if such a relationship exists is a necessary first step in converting statistical association to possible engineering cause-and-effect. Looking at how the raw data change as a function of the different levels of a factor is a fundamental step which, it may be argued, should never be skipped in any data analysis.

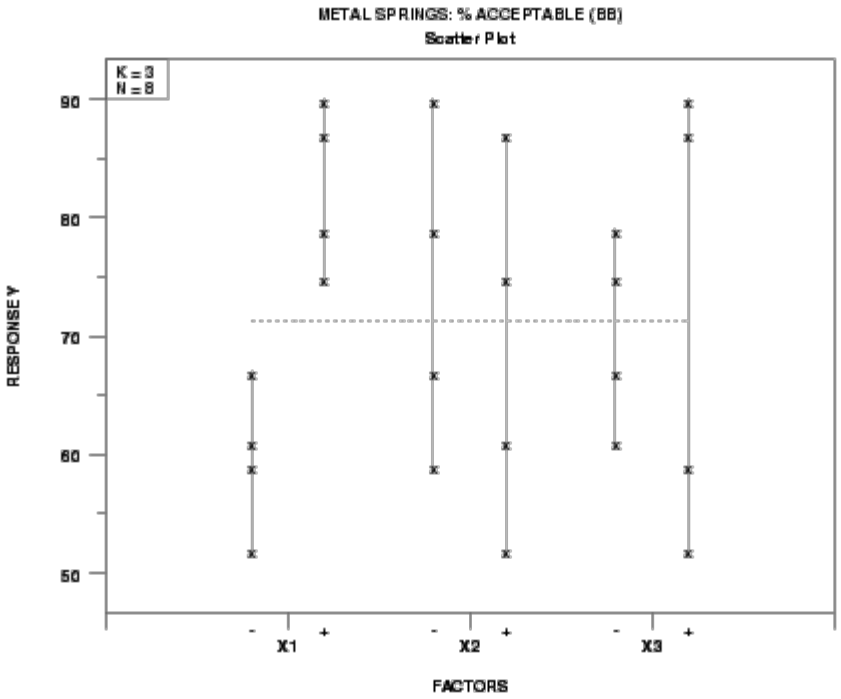
From such a foundational plot, the analyst invariably extracts information dealing with location shifts, variation shifts, and outliers. Such information may easily be washed out by other "more advanced" quantitative or graphical procedures (even computing and plotting means!). Hence there is motivation for the DOE scatter plot.

If we were interested in assessing the importance of a single factor, and since "important" by default means shift in location, then the simple scatter plot is an ideal tool. A large shift (with little data overlap) in the body of the data from the "-" setting to the "+" setting of a given factor would imply that the factor is important. A small shift (with much overlap) would imply the factor is not important.

The DOE scatter plot is actually a sequence of *k* such scatter plots with one scatter plot for each factor.

Plot for defective springs data

The DOE scatter plot for the defective springs data set is as follows.



How to interpret

As discussed previously, the DOE scatter plot is used to look for the following:

1. Most Important Factors;
2. Best Settings of the Most Important Factors;
3. Outliers.

Each of these will be discussed in turn.

Most Important Factors:

For each of the *k* factors, as we go from the "-" setting to the "+" setting within the factor, is there a location shift in the body of the data? If yes, then

1. Which factor has the biggest such data location shift (that is, has least data overlap)? This defines the "most important factor".

2. Which factor has the next biggest shift (that is, has next least data overlap)? This defines the "second most important factor".

3. Continue for the remaining factors.

In practice, the DOE scatter plot will typically only be able to discriminate the most important factor (largest shift) and perhaps the second most important factor (next largest shift). The degree of overlap in remaining factors is frequently too large to ascertain with certainty the ranking for other factors.

Best Settings for the Most Important Factors:

For each of the most important factors, which setting ("-" or "+") yields the "best" response?

In order to answer this question, the engineer must first define "best". This is done with respect to the overall project goal in conjunction with the specific response variable under study. For some experiments (e.g., maximizing the speed of a chip), "best" means we are trying to maximize the response (speed). For other experiments (e.g., semiconductor chip scrap), "best" means we are trying to minimize the response (scrap). For yet other experiments (e.g., designing a resistor) "best" means we are trying to hit a specific target (the specified resistance). Thus the definition of "best" is an engineering precursor to the determination of best settings.

Suppose the analyst is attempting to maximize the response. In such a case, the analyst would proceed as follows:

1. For factor 1, for what setting (- or +) is the body of the data higher?
2. For factor 2, for what setting (- or +) is the body of the data higher?
3. Continue for the remaining factors.

The resulting *k*-vector of best settings:

(x1best, x2best, ..., xkbest)

is thus theoretically obtained by looking at each factor individually in the DOE scatter plot and choosing the setting (- or +) that has the body of data closest to the desired optimal (maximal, minimal, target) response.

As indicated earlier, the DOE scatter plot will typically be able to estimate best settings for only the first few important factors. Again, the degree of data overlap precludes ascertaining best settings for the remaining factors. Other tools, such as the DOE mean plot, will do a better job of determining such settings.

Outliers:

Do any data points stand apart from the bulk of the data? If so, then such values are candidates for further investigation as outliers. For multiple outliers, it is of interest to note if all such anomalous data cluster at the same setting for any of the various factors. If so, then such settings become candidates for avoidance or inclusion, depending on the nature (bad or good), of the outliers.

Conclusions for the defective springs data

The application of the DOE scatter plot to the defective springs data set results in the following conclusions:

1. Most Important Factors:
    1. X_1 (most important);
    2. X_2 (of lesser importance);
    3. X_3 (of least importance).

    that is,

    - factor 1 definitely looks important;
    - factor 2 is a distant second;
    - factor 3 has too much overlap to be important with respect to location, but is flagged for further investigation due to potential differences in variation.

2. Best Settings:

    $(X_1, X_2, X_3) = (+, -, -) = (+1, -1, -1)$

3. Outliers: None detected.

5.5.9.3. DOE mean plot

Purpose The DOE (design of experiments) mean plot answers the following two questions:

1. What is the ranked list of factors (not including the interactions)? The ranking is from the most important factor to least important factor.
2. What is the best setting for each of the *k* factors?

In the above two questions, the terms "important" and "best" need clarification and specificity.

A factor can be important if it leads to a significant shift in the location of the response variable as we go from the "-" setting of the factor to the "+" setting of the factor. Alternatively, a factor can be important if it leads to a significant change in variation (spread) as we go from the "-" to the "+" settings. Both definitions are relevant and acceptable. The default definition of "important" in engineering/scientific applications is the former (shift in location). Unless specified to the contrary, when a factor is claimed to be important, the implication is that the factor caused a large location shift in the response.

In this context, a factor setting is best if it results in a typical response that is closest (in location) to the desired project goal (that is, a maximization, minimization, or hitting a target). This desired project goal is an engineering, not a statistical, question, and so the desired optimization goal must be overtly specified by the engineer.

Given the above two definitions of important and best, the DOE mean plot is a useful tool for determining the important factors and for determining the best settings.

An alternate name for the DOE mean plot is the "main effects plot".

Output The output from the DOE mean plot is:

1. Primary: A ranked list of the factors (not including interactions) from most important to least important.
2. Secondary: The best setting for each of the *k* factors.

Definition The DOE mean plot is formed by:

- Vertical Axis: The mean response for a given setting ("-" or "+") of a factor, for each of the *k* factors.
- Horizontal Axis: The *k* factors and the two settings ("-" and "+") within each

factor.

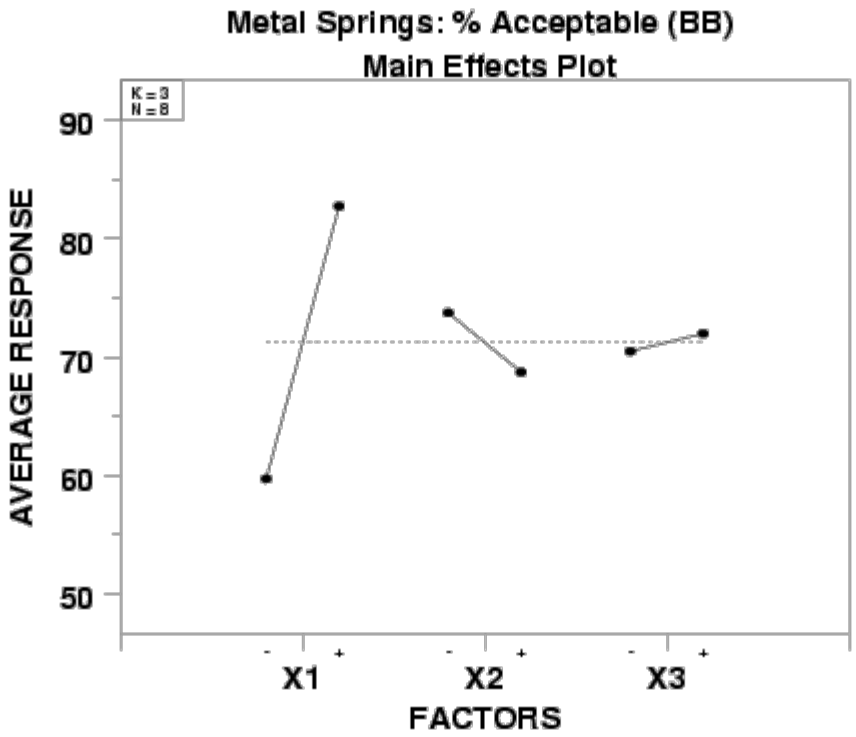
Motivation

If we were interested in assessing the importance of a single factor, and since important, by default, means shift in location, and the average is the simplest location estimator, a reasonable graphics tool to assess a single factor's importance would be a simple mean plot. The vertical axis of such a plot would be the mean response for each setting of the factor and the horizontal axis is the two settings of the factor: "-" and "+" (-1 and +1). A large difference in the two means would imply the factor is important while a small difference would imply the factor is not important.

The DOE mean plot is actually a sequence of *k* such plots, with one mean plot for each factor. To assist in comparability and relative importance, all of the mean plots are on the same scale.

Plot for defective springs data

Applying the DOE mean plot to the defective springs data yields the following plot.



How to interpret

From the DOE mean plot, we look for the following:

1. A ranked list of factors from most important to least important.
2. The best settings for each factor (on average).

Ranked List of Factors--Most Important to Least Important:

For each of the *k* factors, as we go from the "-" setting to the "+" setting for the factor, is there a shift in location of the average response?

If yes, we would like to identify the factor with the biggest shift (the "most

important factor"), the next biggest shift (the "second most important factor"), and so on until all factors are accounted for.

Since we are only plotting the means and each factor has identical (-,+) = (-1,+1) coded factor settings, the above simplifies to

1. What factor has the steepest line? This is the most important factor.
2. The next steepest line? This is the second most important factor.
3. Continue for the remaining factors.

This ranking of factors based on local means is the most important step in building the definitive ranked list of factors as required in screening experiments.

Best Settings (on Average):

For each of the *k* factors, which setting (- or +) yields the "best" response?

In order to answer this, the engineer must first define "best". This is done with respect to the overall project goal in conjunction with the specific response variable under study. For some experiments, "best" means we are trying to maximize the response (e.g., maximizing the speed of a chip). For other experiments, "best" means we are trying to minimize the response (e.g., semiconductor chip scrap). For yet other experiments, "best" means we are trying to hit a specific target (e.g., designing a resistor to match a specified resistance). Thus the definition of "best" is a precursor to the determination of best settings.

For example, suppose the analyst is attempting to maximize the response. In that case, the analyst would proceed as follows:

1. For factor 1, what setting (- or +) has the largest average response?
2. For factor 2, what setting (- or +) has the largest average response?
3. Continue for the remaining factors.

The resulting *k*-vector of best settings:

(x1best, x2best, ..., xkbest)

is in general obtained by looking at each factor individually in the DOE mean plot and choosing that setting (- or +) that has an average response closest to the desired optimal (maximal, minimal, target) response.

This candidate for best settings is based on the averages. This *k*-vector of best settings should be similar to that obtained from the DOE scatter plot, though the DOE mean plot is easier to interpret.

Conclusions for the defective springs data

The application of the DOE mean plot to the defective springs data set results in the following conclusions:

1. Ranked list of factors (excluding interactions):
    1. X_1 (most important). Qualitatively, this factor looks definitely important.
    2. X_2 (of lesser importantance). Qualitatively, this factor is a distant second to X_1.

3. X_3 (unimportant). Qualitatively, this factor appears to be unimportant.

2. Best settings (on average):

$(X_1, X_2, X_3) = (+, -, +) = (+1, -1, +1)$

5.5.9.4. Interaction effects matrix plot

Purpose

The interaction effects matrix plot is an extension of the DOE mean plot to include both main effects and 2-factor interactions (the DOE mean plot focuses on main effects only). The interaction effects matrix plot answers the following two questions:

1. What is the ranked list of factors (including 2-factor interactions), ranked from most important to least important; and
2. What is the best setting for each of the *k* factors?

For a *k*-factor experiment, the effect on the response could be due to main effects and various interactions all the way up to *k*-term interactions. As the number of factors, *k*, increases, the total number of interactions increases exponentially. The total number of possible interactions of all orders = 2^k - 1 - *k*. Thus for *k* = 3, the total number of possible interactions = 4, but for *k* = 7 the total number of possible interactions = 120.

In practice, the most important interactions are likely to be 2-factor interactions. The total number of possible 2-factor interactions is

$$\left(\begin{array}{c} k \\ 2 \end{array} \right) = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$$

Thus for *k* = 3, the number of 2-factor interactions = 3, while for *k* = 7, the number of 2-factor interactions = 21.

It is important to distinguish between the number of interactions that are active in a given experiment versus the number of interactions that the analyst is capable of making definitive conclusions about. The former depends only on the physics and engineering of the problem. The latter depends on the number of factors, *k*, the choice of the *k* factors, the constraints on the number of runs, *n*, and ultimately on the experimental design that the analyst chooses to use. In short, the number of possible interactions is **not** necessarily identical to the number of interactions that we can detect.

Note that

1. with full factorial designs, we can uniquely estimate interactions of all orders;
2. with fractional factorial designs, we can uniquely estimate only some (or at times no) interactions; the more fractionated the design, the fewer interactions that we can estimate.

Output The output for the interaction effects matrix plot is

1. Primary: Ranked list of the factors (including 2-factor interactions) with the factors are ranked from important to unimportant.
2. Secondary: Best setting for each of the *k* factors.

Definition The interaction effects matrix plot is an upper right-triangular matrix of mean plots consisting of *k* main effects plots on the diagonal and *k**(*k*-1)/2 2-factor interaction effects plots on the off-diagonal.

In general, interactions are **not** the same as the usual (multiplicative) cross-products. However, for the special case of 2-level designs coded as (-, +) = (-1, +1), the interactions **are** identical to cross-products. By way of contrast, if the 2-level designs are coded otherwise (e.g., the (1, 2) notation espoused by Taguchi and others), then this equivalance is **not** true. Mathematically,

{-1, +1} x {-1, +1} => {-1, +1}

but

{1, 2} x {1, 2} => {1, 2, 4}

Thus, coding does make a difference. We recommend the use of the (-, +) coding.

It is remarkable that with the - and + coding, the 2-factor interactions are dealt with, interpreted, and compared in the same way that the *k* main effects are handled. It is thus natural to include both 2-factor interactions and main effects within the same matrix plot for ease of comparison.

For the off-diagonal terms, the first construction step is to form the horizontal axis values, which will be the derived values (also - and +) of the cross-product. For example, the settings for the X_1*X_2 interaction are derived by simple multiplication from the data as shown below.

X_1	X_2	X_1*X_2
-	-	+
+	-	-
-	+	-
+	+	+

Thus X_1, X_2, and X_1*X_2 all form a closed (-, +) system. The advantage of the closed system is that graphically interactions can be interpreted in the exact same fashion as the *k* main effects.

After the entire X_1*X_2 vector of settings has been formed in this way, the vertical axis of the X_1*X_2 interaction plot is formed:

1. the plot point above X_1*$X2$ = "-" is simply the mean of all response values for which X_1*X_2 = "-"
2. the plot point above X_1*X_2 = "+" is simply the mean of all response values for

which X_1*X_2 = "+".

We form the plots for the remaining 2-factor interactions in a similar fashion.

All the mean plots, for both main effects and 2-factor interactions, have a common scale to facilitate comparisons. Each mean plot has

1. Vertical Axis: The mean response for a given setting (- or +) of a given factor or a given 2-factor interaction.
2. Horizontal Axis: The 2 settings (- and +) within each factor, or within each 2-factor interaction.
3. Legend:
    1. A tag (1, 2, ..., k, 12, 13, etc.), with 1 = X_1, 2 = X_2, ..., k = X_k, 12 = X_1*X_2, 13 = X_1*X_3, 35 = X_3*X_5, 123 = X_1*X_2*X_3, etc.) which identifies the particular mean plot; and
    2. The least squares estimate of the factor (or 2-factor interaction) effect. These effect estimates are large in magnitude for important factors and near-zero in magnitude for unimportant factors.

In a later section, we discuss in detail the models associated with full and fractional factorial 2-level designs. One such model representation is

$$Y = \mu + \beta_1*X_1 + \beta_2*X_2 + \beta_{12}*X_1*X_2 + ... + \varepsilon$$

For factor variables coded with + and - settings, the β_i coefficient is one half of the effect estimate due to factor X_i. Thus, if we multiply the least-squares coefficients by two, due to orthogonality, we obtain the simple difference of means at the + setting and the - setting. This is true for the k main factors. It is also true for all two-factor and multi-factor interactions.

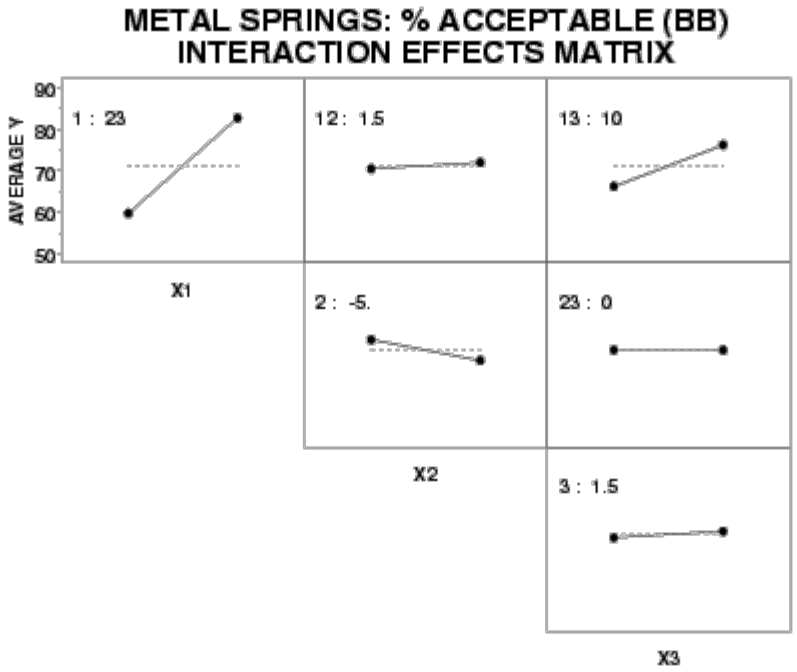
Thus, visually, the difference in the mean values on the plot is identically the least squares estimate for the effect. Large differences (steep lines) imply important factors while small differences (flat lines) imply unimportant factors.

Motivation

As discussed in detail above, the next logical step beyond main effects is displaying 2-factor interactions, and this plot matrix provides a convenient graphical tool for examining the relative importance of main effects and 2-factor interactions in concert. To do so, we make use of the striking aspect that in the context of 2-level designs, the 2-factor interactions are identical to cross-products and the 2-factor interaction effects can be interpreted and compared the same way as main effects.

Plot for defective springs data

Constructing the interaction effects matrix plot for the defective springs data set yields the following plot.



How to interpret

From the interaction effects matrix, we can draw three important conclusions:

1. Important Factors (including 2-factor interactions);
2. Best Settings;
3. Confounding Structure (for fractional factorial designs).

We discuss each of these in turn.

1. **Important factors (including 2-factor interactions):**

   Jointly compare the *k* main factors and the *k**(*k*-1)/2 2-factor interactions. For each of these subplots, as we go from the "-" setting to the "+" setting within a subplot, is there a shift in location of the average data (yes/no)? Since all subplots have a common (-1, +1) horizontal axis, questions involving shifts in location translate into questions involving steepness of the mean lines (large shifts imply steep mean lines while no shifts imply flat mean lines).

   1. Identify the factor or 2-factor interaction that has the largest shift (based on averages). This defines the "most important factor". The largest shift is determined by the steepest line.

   2. Identify the factor or 2-factor interaction that has the next largest shift (based on averages). This defines the "second most important factor". This shift is determined by the next steepest line.

   3. Continue for the remaining factors.

   This ranking of factors and 2-factor interactions based on local means is a major step in building the definitive list of ranked factors as required for screening experiments.

2. **Best settings:**

   For each factor (of the *k* main factors along the diagonal), which setting (- or +) yields the "best" (highest/lowest) average response?

   Note that the experimenter has the ability to change settings for only the *k* main factors, not for any 2-factor interactions. Although a setting of some 2-factor interaction may yield a better average response than the alternative setting for that same 2-factor interaction, the experimenter is unable to set a 2-factor interaction setting in practice. That is to say, there is no "knob" on the machine that controls 2-factor interactions; the "knobs" only control the settings of the *k* main factors.

   How then does this matrix of subplots serve as an improvement over the *k* best settings that one would obtain from the DOE mean plot? There are two common possibilities:

   1. Steep Line:

      For those main factors along the diagonal that have steep lines (that is, are important), choose the best setting directly from the subplot. This will be the same as the best setting derived from the DOE mean plot.

   2. Flat line:

      For those main factors along the diagonal that have flat lines (that is, are unimportant), the naive conclusion to use either setting, perhaps giving preference to the cheaper setting or the easier-to-implement setting, may be unwittingly incorrect. In such a case, the use of the off-diagonal 2-factor interaction information from the interaction effects matrix is critical for deducing the better setting for this nominally "unimportant" factor.

      To illustrate this, consider the following example:

      - Suppose the factor X_1 subplot is steep (important) with the best setting for X_1 at "+".
      - Suppose the factor X_2 subplot is flat (unimportant) with both settings yielding about the same mean response.

      Then what setting should be used for X_2? To answer this, consider the following two cases:

      1. Case 1. If the X_1*X_2 interaction plot happens also to be flat (unimportant), then choose either setting for X_2 based on cost or ease.
      2. Case 2. On the other hand, if the X_1*X_2 interaction plot is steep (important), then this dictates a prefered setting for X_2 **not** based on cost or ease.

      To be specific for case 2, if X_1*X_2 is important, with X_1*X_2 = "+" being the better setting, and if X_1 is important, with X_1 = "+" being the better

setting, then this implies that the best setting for X_2 must be "+" (to assure that X_1*$X2$ (= +*+) will also be "+"). The reason for this is that since we are already locked into X_1 = "+", and since X_1*X_2 = "+" is better, then the only way we can obtain X_1*X_2 = "+" with X_1 = "+" is for X_2 to be "+" (if X_2 were "-", then X_1*X_2 with X_1 = "+" would yield X_1*X_2 = "-").

In general, if X_1 is important, X_1*X_2 is important, and X_2 is not important, then there are four distinct cases for deciding what the best setting is for X_2:

X_1	X_1*X_2	=> X_2
+	+	+
+	-	-
-	+	-
-	-	+

By similar reasoning, examining each factor and pair of factors, we thus arrive at a resulting vector of the *k* best settings:

(x1best, x2best, ..., xkbest)

This average-based *k*-vector should be compared with best settings *k*-vectors obtained from previous steps (in particular, from step 1 in which the best settings were drawn from the best data value).

When the average-based best settings and the data-based best settings agree, we benefit from the increased confidence given our conclusions.

When the average-based best settings and the data-based best settings disagree, then what settings should the analyst finally choose? Note that in general the average-based settings and the data-based settings will invariably be identical for all "important" factors. Factors that do differ are virtually always "unimportant". Given such disagreement, the analyst has three options:

1. Use the average-based settings for minor factors. This has the advantage of a broader (average) base of support.
2. Use the data-based settings for minor factors. This has the advantage of demonstrated local optimality.
3. Use the cheaper or more convenient settings for the local factor. This has the advantage of practicality.

Thus the interaction effects matrix yields important information not only about the ranked list of factors, but also about the best settings for each of the *k* main factors. This matrix of subplots is one of the most important tools for the experimenter in the analysis of 2-level screening designs.

3. **Confounding Structure (for Fractional Factorial Designs)**

When the interaction effects matrix is used to analyze 2-level fractional (as opposed to full) factorial designs, important additional information can be extracted from the matrix regarding confounding structure.

It is well-known that all fractional factorial designs have confounding, a property whereby every estimated main effect is confounded/contaminated/biased by some high-order interactions. The practical effect of this is that the analyst is unsure of how much of the estimated main effect is due to the main factor itself and how much is due to some confounding interaction. Such contamination is the price that is paid by examining k factors with a sample size n that is less than a full factorial $n = 2^k$ runs.

It is a "fundamental theorem" of the discipline of experimental design that for a given number of factors k and a given number of runs n, some fractional factorial designs are better than others. "Better" in this case means that the intrinsic confounding that **must** exist in all fractional factorial designs has been minimized by the choice of design. This minimization is done by constructing the design so that the main effect confounding is pushed to as high an order interaction as possible.

The rationale behind this is that in physical science and engineering systems it has been found that the "likelihood" of high-order interactions being significant is small (compared to the likelihood of main effects and 2-factor interactions being significant). Given this, we would prefer that such inescapable main effect confounding be with the highest order interaction possible, and hence the bias to the estimated main effect be as small as possible.

The worst designs are those in which the main effect confounding is with 2-factor interactions. This may be dangerous because in physical/engineering systems, it is quite common for Nature to have some real (and large) 2-factor interactions. In such a case, the 2-factor interaction effect will be inseparably entangled with some estimated main effect, and so the experiment will be flawed in that

1. ambiguous estimated main effects and
2. an ambiguous list of ranked factors

will result.

If the number of factors, k, is large and the number of runs, n, is constrained to be small, then confounding of main effects with 2-factor interactions is unavoidable. For example, if we have $k = 7$ factors and can afford only $n = 8$ runs, then the corresponding 2-level fractional factorial design is a 2^{7-4} which necessarily will have main effects confounded with (3) 2-factor interactions. This cannot be avoided.

On the other hand, situations arise in which 2-factor interaction confounding with main effects results not from constraints on k or n, but on poor design construction. For example, if we have $k = 7$ factors and can afford $n = 16$ runs, a poorly constructed design might have main effects counfounded with 2-factor interactions, but a well-constructed design with the same $k = 7$, $n = 16$ would have main effects confounded with 3-factor interactions but **no** 2-factor interactions. Clearly, this latter design is preferable in terms of minimizing main effect confounding/contamination/bias.

For those cases in which we do have main effects confounded with 2-factor interactions, an important question arises:

> For a particular main effect of interest, how do we know which 2-factor interaction(s) confound/contaminate that main effect?

The usual answer to this question is by means of generator theory, confounding tables, or alias charts. An alternate complementary approach is given by the interaction effects matrix. In particular, if we are examining a 2-level fractional factorial design and

1. if we are not sure that the design has main effects confounded with 2-factor interactions, or
2. if we are sure that we have such 2-factor interaction confounding but are not sure what effects are confounded,

then how can the interaction effects matrix be of assistance? The answer to this question is that the confounding structure can be read **directly** from the interaction effects matrix.

For example, for a 7-factor experiment, if, say, the factor X_3 is confounded with the 2-factor interaction X_2*X_5, then

1. the appearance of the factor X_3 subplot and the appearance of the 2-factor interaction X_2*X_5 subplot will necessarily be identical, and
2. the value of the estimated main effect for X_3 (as given in the legend of the main effect subplot) and the value of the estimated 2-factor interaction effect for X_2*X_5 (as given in the legend of the 2-factor interaction subplot) will also necessarily be identical.

The above conditions are necessary, but not sufficient for the effects to be confounded.

Hence, in the abscence of tabular descriptions (from your statistical software program) of the confounding structure, the interaction effect matrix offers the following graphical alternative for deducing confounding structure in fractional factorial designs:

1. scan the main factors along the diagonal subplots and choose the subset of factors that are "important".
2. For each of the "important" factors, scan all of the 2-factor interactions and compare the main factor subplot and estimated effect with each 2-factor interaction subplot and estimated effect.
3. If there is no match, this implies that the main effect is **not** confounded with any 2-factor interaction.
4. If there is a match, this implies that the main effect **may** be confounded with that 2-factor interaction.
5. If none of the main effects are confounded with any 2-factor interactions, we can have high confidence in the integrity (non-contamination) of our estimated main effects.
6. In practice, for highly-fractionated designs, each main effect may be confounded with several 2-factor interactions. For example, for a 2^{7-4} fractional factorial design, each main effect will be confounded with

three 2-factor interactions. These 1 + 3 = 4 identical subplots will be blatantly obvious in the interaction effects matrix.

Finally, what happens in the case in which the design the main effects are **not** confounded with 2-factor interactions (**no** diagonal subplot matches any off-diagonal subplot). In such a case, does the interaction effects matrix offer any useful further insight and information?

The answer to this question is yes because even though such designs have main effects unconfounded with 2-factor interactions, it is fairly common for such designs to have 2-factor interactions confounded with one another, and on occasion it may be of interest to the analyst to understand that confounding. A specific example of such a design is a 2^{4-1} design formed with X_4 settings = X_1*X_2*X_3. In this case, the 2-factor-interaction confounding structure may be deduced by comparing all of the 2-factor interaction subplots (and effect estimates) with one another. Identical subplots and effect estimates hint strongly that the two 2-factor interactions are confounded. As before, such comparisons provide necessary (but not sufficient) conditions for confounding. Most statistical software for analyzing fractional factorial experiments will explicitly list the confounding structure.

Conclusions for the defective springs data

The application of the interaction effects matrix plot to the defective springs data set results in the following conclusions:

1. Ranked list of factors (including 2-factor interactions):
    1. X_1 (estimated effect = 23.0)
    2. X_1*X_3 (estimated effect = 10.0)
    3. X_2 (estimated effect = -5.0)
    4. X_3 (estimated effect = 1.5)
    5. X_1*X_2 (estimated effect = 1.5)
    6. X_2*X_3 (estimated effect = 0.0)

    Factor 1 definitely looks important. The X_1*X_3 interaction looks important. Factor 2 is of lesser importance. All other factors and 2-factor interactions appear to be unimportant.

2. Best Settings (on the average):

    (X_1, X_2, X_3) = (+, -, +) = (+1, -1, +1)

5.5.9.5. Block plot

Purpose The block plot answers the following two general questions:

1. What are the important factors (including interactions)?
2. What are the best settings for these important factors?

The basic (single) block plot is a multifactor EDA technique to determine if a factor is important and to ascertain if that importance is unconditional (robust) over all settings of all other factors in the system. In an experimental design context, the block plot is actually a sequence of block plots with one plot for each of the k factors.

Due to the ability of the block plot to determine whether a factor is important over all settings of all other factors, the block plot is also referred to as a DOE robustness plot.

Output The block plot provides specific information on

1. Important factors (of the k factors and the $\binom{k}{2}$ 2-factor interactions); and
2. Best settings of the important factors.

Definition The block plot is a series of k basic block plots with each basic block plot for a main effect. Each basic block plot asks the question as to whether that particular factor is important:

1. The first block plot asks the question: "Is factor X_1 important?
2. The second block plot asks the question: "Is factor X_2 important?
3. Continue for the remaining factors.

The i-th basic block plot, which targets factor i and asks whether factor X_i is important, is formed by:

- Vertical Axis: Response
- Horizontal Axis: All 2^{k-1} possible combinations of the (k-1) non-target factors (that is, "robustness" factors). For example, for the block plot focusing on factor X_1 from a 2^3 full factorial experiment, the horizontal axis will consist of all $2^{3-1} = 4$ distinct combinations of factors X_2 and X_3. We create this robustness factors axis because we are interested in determining if X_1 is

important robustly. That is, we are interested in whether X_1 is important not only in a general/summary kind of way, but also whether the importance of X is universally and consistently valid over each of the $2^{3-1} = 4$ combinations of factors X_2 and X_3. These four combinations are (X_2, X_3) = (+, +), (+, -), (-, +), and (-, -). The robustness factors on the horizontal axis change from one block plot to the next. For example, for the $k = 3$ factor case:

1. the block plot targeting X_1 will have robustness factors X_2 and X_3;
2. the block plot targeting X_2 will have robustness factors X_1 and X_3;
3. the block plot targeting X_3 will have robustness factors X_1 and X_2.

- Plot Character: The setting (- or +) for the target factor X_i. Each point in a block plot has an associated setting for the target factor X_i. If X_i = "-", the corresponding plot point will be "-"; if X_i = "+", the corresponding plot point will be "+".

For a particular combination of robustness factor settings (horizontally), there will be two points plotted above it (vertically):

1. one plot point for X_i = "-"; and
2. the other plot point for X_i = "+".

In a block plot, these two plot points are surrounded by a box (a block) to focus the eye on the internal within-block differences as opposed to the distraction of the external block-to-block differences. Internal block differences reflect on the importance of the target factor (as desired). External block-to-block differences reflect on the importance of various robustness factors, which is not of primary interest.

Large within-block differences (that is, tall blocks) indicate a large local effect on the response which, since all robustness factors are fixed for a given block, can only be attributed to the target factor. This identifies an "important" target factor. Small within-block differences (small blocks) indicate that the target factor X_i is unimportant.

For a given block plot, the specific question of interest is thus

> Is the target factor X_i important? That is, as we move within a block from the target factor setting of "-" to the target factor setting of "+", does the response variable value change by a large amount?

The height of the block reflects the "local" (that is, for that particular combination of robustness factor settings) effect on the response due to a change in the target factor settings. The "localized" estimate for the target factor effect for X_i is in fact identical to the difference in the response between the target factor X_i at the "+" setting and at the "-" setting. Each block height of a robustness plot is thus a localized estimate of the target factor effect.

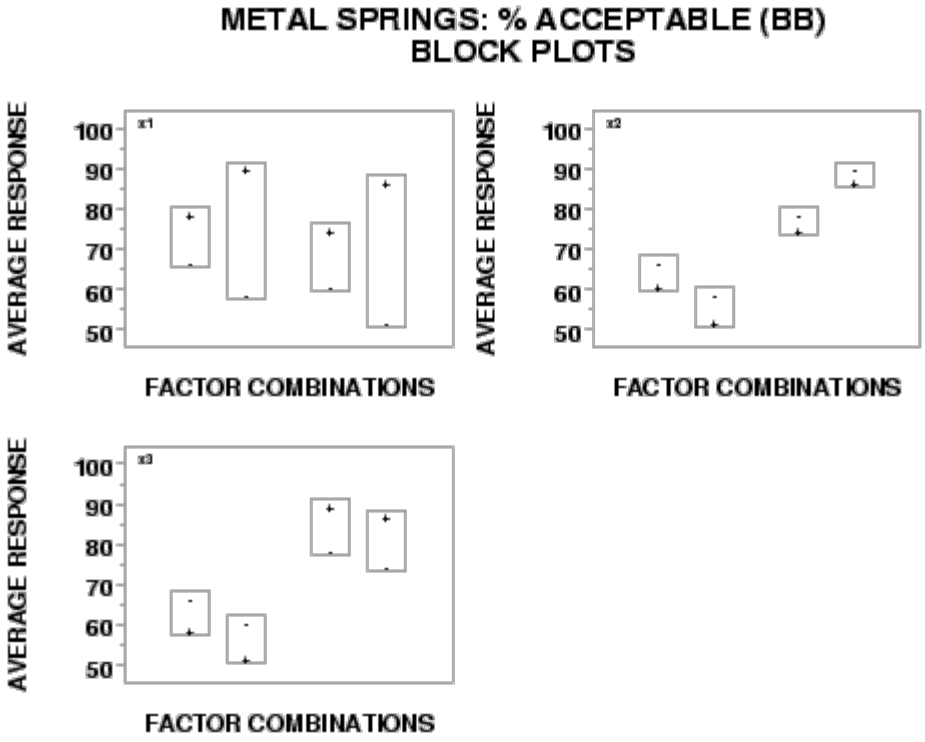
In summary, important factors will have both

1. consistently large block heights; and
2. consistent +/- sign arrangements

where the "consistency" is over all settings of robustness factors. Less important factors will have only one of these two properties. Unimportant factors will have neither property.

Plot for defective springs data

Applying the ordered response plot to the defective springs data set yields the following plot.



How to interpret

From the block plot, we are looking for the following:

1. Important factors (including 2-factor interactions);
2. Best settings for these factors.

We will discuss each of these in turn.

Important factors (including 2-factor interactions):

Look at each of the *k* block plots. Within a given block plot,

> Are the corresponding block heights consistently large as we scan across the within-plot robustness factor settings--yes/no; and are the within-block sign patterns (+ above -, or - above +) consistent across all robustness factors settings--yes/no?

To facilitate intercomparisons, all block plots have the same vertical axis scale. Across such block plots,

1. Which plot has the consistently largest block heights, along with consistent arrangement of within-block +'s and -'s? This defines the "most important factor".

2. Which plot has the consistently next-largest block heights, along with consistent arrangement of within-block +'s and -'s? This defines the "second most important factor".

3. Continue for the remaining factors.

This scanning and comparing of the *k* block plots easily leads to the identification of the most important factors. This identification has the additional virtue over previous steps in that it is robust. For a given important factor, the consistency of block heights and sign arrangement across robustness factors gives additional credence to the robust importance of that factor. The factor is important (the change in the response will be large) irrespective of what settings the robustness factors have. Having such information is both important and comforting.

Important Special Case; Large but Inconsistent:

What happens if the block heights are large but **not** consistent? Suppose, for example, a 2^3 factorial experiment is being analyzed and the block plot focusing on factor X_1 is being examined and interpreted so as to address the usual question of whether factor X_1 is important.

Let us consider in some detail how such a block plot might appear. This X_1 block plot will have $2^{3-1} = 4$ combinations of the robustness factors X_2 and X_3 along the horizontal axis in the following order:

$$(X_2, X_3) = (+, +); (X_2, X_3) = (+, -); (X_2, X_3) = (-, +); (X_2, X_3) = (-, -).$$

If the block heights are consistently large (with "+" above "-" in each block) over the four combinations of settings for X_2 and X_3, as in

(X_2, X_3)	block height (= local X_1 effect)
(+, +)	30
(+, -)	29
(-, +)	29
(-, -)	31

then from binomial considerations there is one chance in $2^{4-1} = 1/8 \approx 12.5$ % of the the four local X_1 effects having the same sign (i.e., all positive or all negative). The usual statistical cutoff of 5 % has not been achieved here, but the 12.5 % is suggestive. Further, the consistency of the four X_1 effects (all near 30) is evidence of a robustness of the X effect over the settings of the other two factors. In summary, the above suggests:

1. Factor 1 is probably important (the issue of how large the effect has to be in order to be considered important will be discussed in more detail in a later section); and
2. The estimated factor 1 effect is about 30 units.

On the other hand, suppose the 4 block heights for factor 1 vary in the following cyclic way:

(X_2, X_3)	block height (= local X_1 effect)
(+, +)	30
(+, -)	20
(-, +)	30
(-, -)	20

then how is this to be interpreted?

The key here to such interpretation is that the block plot is telling us that the estimated X_1 effect is in fact at least 20 units, but **not** consistent. The effect is changing, but it is changing in a structured way. The "trick" is to scan the X_2 and X_3 settings and deduce what that substructure is. Doing so from the above table, we see that the estimated X_1 effect is 30

- for point 1 (X_2, X_3) = (+, +) and
- for point 3 (X_2, X_3) = (-, +)

and then the estimated X_1 effect drops 10 units to 20

- for point 2 (X_2, X_3) = (+, -) and
- for point 4 (X_2, X_3) = (-, -)

We thus deduce that the estimated X_1 effect is

1. 30 whenever X_3 = "+"
2. 20 whenever X_3 = "-"

When the factor X_1 effect is not consistent, but in fact changes depending on the setting of factor X_3, then definitionally that is said to be an "X_1*X_3 interaction". That is precisely the case here, and so our conclusions would be:

1. factor X_1 is probably important;
2. the estimated factor X_1 effect is 25 (the average of 30, 20, 30, and 20);
3. the X_1*X_3 interaction is probably important;
4. the estimated X_1*X_3 interaction is about 10 (the change in the factor X_1 effect as X_3 changes = 30 - 20 = 10);
5. hence the X_1*X_3 interaction is less important than the X_1 effect.

Note that we are using the term important in a qualitative sense here. More precise determinations of importance in terms of statistical or engineering significance are discussed in later sections.

The block plot gives us the structure and the detail to allow such conclusions to be drawn and to be understood. It is a valuable adjunct to the previous analysis steps.

Best settings:

After identifying important factors, it is also of use to determine the best settings for

these factors. As usual, best settings are determined for main effects only (since main effects are all that the engineer can control). Best settings for interactions are not done because the engineer has no direct way of controlling them.

In the block plot context, this determination of best factor settings is done simply by noting which factor setting (+ or -) within each block is closest to that which the engineer is ultimately trying to achieve. In the defective springs case, since the response variable is percent acceptable springs, we are clearly trying to maximize (as opposed to minimize, or hit a target) the response and the ideal optimum point is 100 %. Given this, we would look at the block plot of a given important factor and note within each block which factor setting (+ or -) yields a data value closest to 100 % and then select that setting as the best for that factor.

From the defective springs block plots, we would thus conclude that

1. the best setting for factor 1 is +;
2. the best setting for factor 2 is -;
3. the best setting for factor 3 cannot be easily determined.

Conclusions for the defective springs data

In summary, applying the block plot to the defective springs data set results in the following conclusions:

1. Unranked list of important factors (including interactions):
    - X_1 is important;
    - X_2 is important;
    - X_1*X_3 is important.
2. Best Settings:

    $(X_1, X_2, X_3) = (+, -, ?) = (+1, -1, ?)$

5.5.9.6. DOE Youden plot

Purpose The DOE (design of experiments) Youden plot answers the following question:

What are the important factors (including interactions)?

In its original interlab rendition, the Youden plot was a graphical technique developed in the 1960's by Jack Youden of NIST for assessing between-lab biases and within-lab variation problems in the context of interlab experimentation. In particular, it was appropriate for the analysis of round-robin data when exactly two materials, batches, etc. were used in the design.

In a design of experiments context, we borrow this duality emphasis and apply it to 2-level designs. The 2-component emphasis of the Youden plot makes it a natural to be applied to such designs.

Output The DOE Youden plot provides specific information on

1. Ranked list of factors (including interactions); and
2. Separation of factors into two categories: important and unimportant.

The primary output from a DOE Youden plot is the ranked list of factors (out of the *k* factors and interactions). For full factorial designs, interactions include the full complement of interactions at all orders; for fractional factorial designs, interactions include only some, and occasionally none, of the actual interactions. Further, the DOE Youden plot yields information identifying which factors/interactions are important and which are unimportant.

Definition The DOE Youden plot consists of the following:

- Vertical Axis: Mean response at the "+" setting for each factor and each interaction. For a given factor or interaction, *n*/2 response values will go into computing the "+" mean.
- Horizontal Axis: Mean response at the "-" setting for each factor and each interaction. For a given factor or interaction, *n*/2 response values will go into computing the "-" mean.
- Plot Character: Factor/interaction identification for which

  1 indicates factor X_1;
  2 indicates factor X_2;
  ...
  12 indicates the 2-factor X_1*X_2 interaction

123 indicates the 3-factor $X_1*X_2*X_3$ interaction
etc.

In essence, the DOE Youden plot is a scatter plot of the "+" average responses versus the "-" average responses. The plot will consist of n - 1 points with one point for each factor and one point for each (available) interaction. Each point on the plot is annotated to identify which factor or interaction is being represented.

Motivation

Definitionally, if a factor is **unimportant**, the "+" average will be approximately the same as the "-" average, and if a factor is **important**, the "+" average will be considerably different from the "-" average. Hence a plot that compares the "+" averages with the "-" averages directly seems potentially informative.

From the definition above, the DOE Youden plot is a scatter plot with the "+" averages on the vertical axis and the "-" averages on the horizontal axis. Thus, unimportant factors will tend to cluster in the middle of the plot and important factors will tend to be far removed from the middle.

Because of an arithmetic identity which requires that the average of any corresponding "+" and "-" means must equal the grand mean, all points on a DOE Youden plot will lie on a -45 degree diagonal line. Or to put it another way, for each factor

average (+) + average (-) = constant (with constant = grand mean)

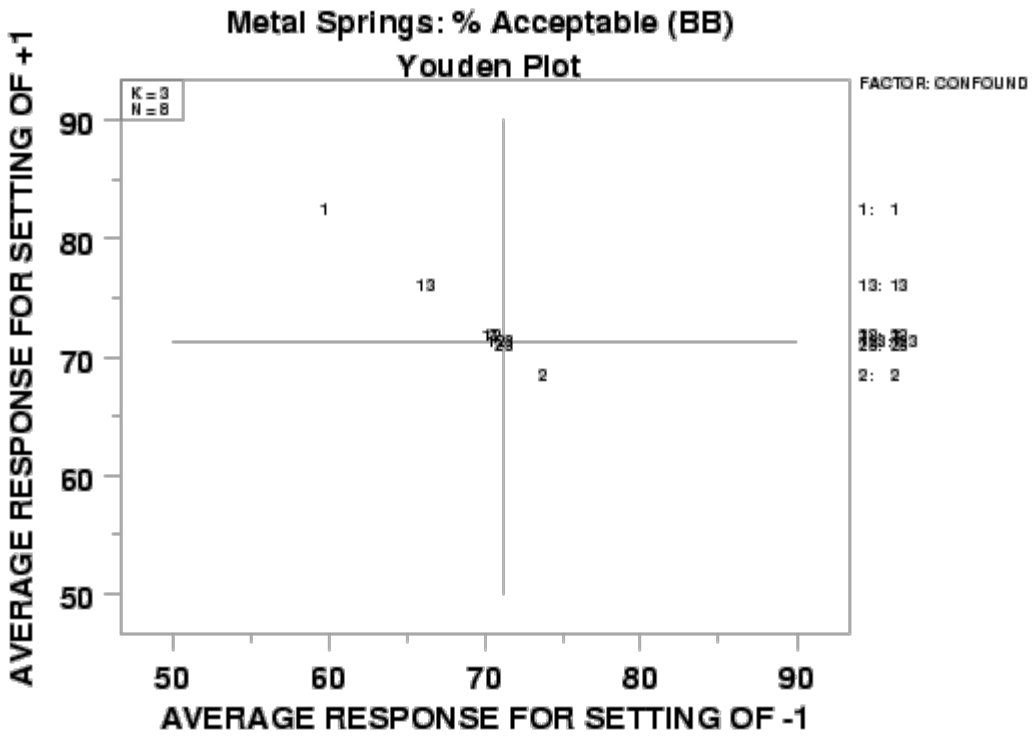
So

average (+) = constant - average (-)

Therefore, the slope of the line is -1 and all points lie on the line. Important factors will plot well-removed from the center because average (+) = average (-) at the center.

Plot for defective springs data

Applying the DOE Youden plot for the defective springs data set yields the following plot.



How to interpret

In the DOE Youden plot, we look for the following:

1. A ranked list of factors (including interactions). The intersecting dotted lines at the center of the plot are the value of the grand mean on both the vertical and horizontal axes. Scan the points along the negative-slope diagonal line and note as to whether such points are clustered around the grand mean or are displaced up or down the diagonal line.

    1. Which point is farthest away from the center? This defines the "most important" factor.
    2. Which point is next farthest away from the center? This defines the "second most important" factor.
    3. Continue in a similar manner for the remaining points. The points closest to the center define the "least important" factors.

2. Separation of factors into important/unimportant categories. Interpretationally, if a factor is **unimportant**, the "+" average will be about the same as the "-" average, so the plot of "+" vertically and "-" horizontally will be near the grand mean of all n - 1 data points.

    Conversely, if a factor is important, the "+" average will differ greatly from the "-" average, and so the plot of "+" vertically and "-" horizontally will be considerably displaced up into the top left quadrant or down into the bottom right quadrant.

    The separation of factors into important/unimportant categories is thus done by answering the question:

    > Which points visually form a cluster around the center? (these define the "unimportant factors"--all remaining factors are "important").

This ranked list of important factors derived from the DOE Youden plot is to be compared with the ranked lists obtained from previous steps. Invariably, there will be a large degree of consistency exhibited across all/most of the techniques.

Conclusions for the defective springs data

The application of the DOE Youden plot to the defective springs data set results in the following conclusions:

1. Ranked list of factors (including interactions):
    1. X_1 (most important)
    2. X_1*X_3 (next most important)
    3. X_2
    4. other factors are of lesser importance
2. Separation of factors into important/unimportant categories:
    - "Important": X_1, X_1*X_3, and X_2
    - "Unimportant": the remainder

5.5.9.7. |Effects| plot

Purpose The |effects| plot answers the question:

What are the important factors (including interactions)?

Quantitatively, the question as to what is the estimated effect of a given factor or interaction and what is its rank relative to other factors and interactions is answered via the least squares estimation criterion (that is, forming effect estimates that minimize the sum of the squared differences between the raw data and the fitted values from such estimates). Based on such an estimation criterion, one could then construct a tabular list of the factors and interactions ordered by the effect magnitude.

The |effects| plot provides a graphical representation of these ordered estimates, Pareto-style from largest to smallest.

The |effects| plot, as presented here, yields both of the above: the plot itself, and the ranked list table. Further, the plot also presents auxiliary confounding information, which is necessary in forming valid conclusions for fractional factorial designs.

Output The output of the |effects| plot is:

1. Primary: A ranked list of important effects (and interactions). For full factorial designs, interactions include the full complement of interactions at all orders; for fractional factorial designs, interactions include only some, and occasionally none, of the actual interactions.
2. Secondary: Grouping of factors (and interactions) into two categories: important and unimportant.

Definition The |effects| plot is formed by:

- Vertical Axis: Ordered (largest to smallest) absolute value of the estimated effects for the main factors and for (available) interactions. For n data points (no replication), typically (n-1) effects will be estimated and the (n-1) |effects| will be plotted.
- Horizontal Axis : Factor/interaction identification:
  1 indicates factor X_1;
  2 indicates factor X_2;
  ...
  12 indicates the 2-factor X_1*X_2 interaction

123 indicates the 3-factor $X_1*X_2*X_3$ interaction,
etc.

- Far right margin : Factor/interaction identification (built-in redundancy):
  1 indicates factor X_1;
  2 indicates factor X_2;
  ...
  12 indicates the 2-factor X_1*X_2 interaction
  123 indicates the 3-factor $X_1*X_2*X_3$ interaction,
  etc.
  If the design is a fractional factorial,the confounding structure is provided for main factors and 2-factor interactions.

- Upper right table: Ranked (largest to smallest by magnitude) list of the least squares estimates for the main effects and for (available) interactions.

  As before, if the design is a fractional factorial, the confounding structure is provided for main factors and 2-factor interactions.

The estimated effects that form the basis for the vertical axis are optimal in the least squares sense. No other estimators exist that will yield a smaller sum of squared deviations between the raw data and the fitted values based on these estimates.

For both the 2^k full factorial designs and 2^{k-p} fractional factorial designs, the form for the least squares estimate of the factor *i* effect, the 2-factor interaction effect, and the multi-factor interaction effect has the following simple form:

factor *i* effect = $\bar{Y}$(+) - $\bar{Y}$(-)
2-factor interaction effect = $\bar{Y}$(+) - $\bar{Y}$(-)
multi-factor interaction effect = $\bar{Y}$(+) - $\bar{Y}$(-)

with $\bar{Y}$(+) denoting the average of all response values for which factor *i* (or the 2-factor or multi-factor interaction) takes on a "+" value, and $\bar{Y}$(-) denoting the average of all response values for which factor *i* (or the 2-factor or multi-factor interaction) takes on a "-" value.

The essence of the above simplification is that the 2-level full and fractional factorial designs are all orthogonal in nature, and so all off-diagonal terms in the least squares *X'X* matrix vanish.

Motivation

Because of the difference-of-means definition of the least squares estimates, and because of the fact that all factors (and interactions) are standardized by taking on values of -1 and +1 (simplified to - and +), the resulting estimates are all on the same scale. Therefore, comparing and ranking the estimates based on magnitude makes eminently good sense.

Moreover, since the sign of each estimate is completely arbitrary and will reverse depending on how the initial assignments were made (e.g., we could assign "-" to treatment A and "+" to treatment B or just as easily assign "+" to treatment A and "-" to treatment B), forming a ranking based on magnitudes (as opposed to signed effects) is preferred.

Given that, the ultimate and definitive ranking of factor and interaction effects will

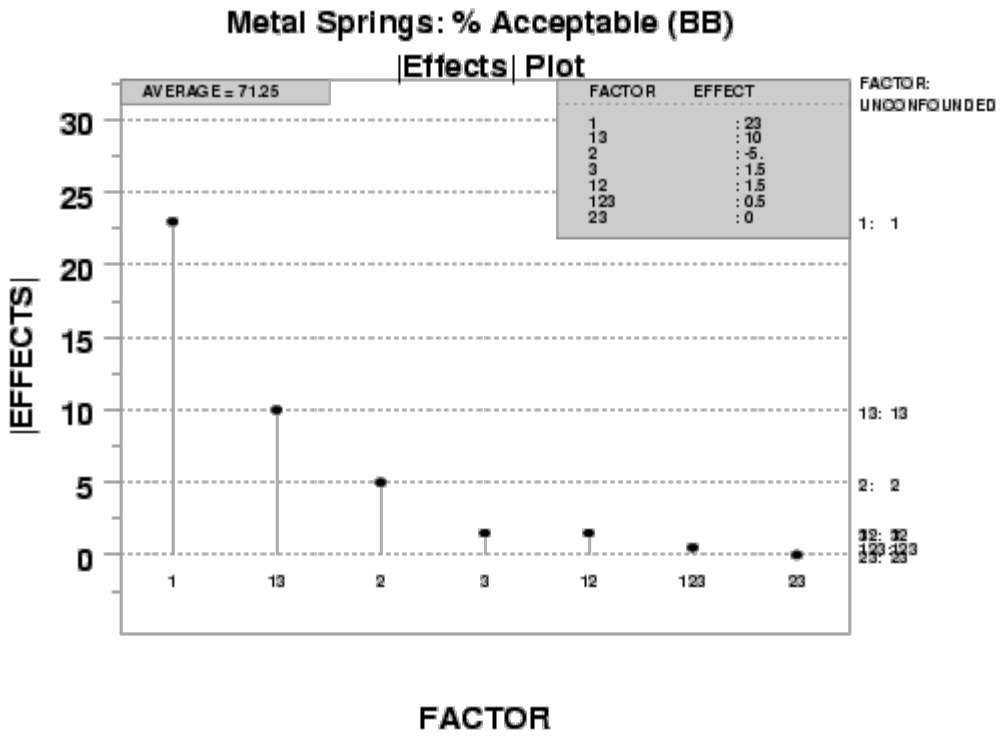
be made based on the ranked (magnitude) list of such least squares estimates. Such rankings are given graphically, Pareto-style, within the plot; the rankings are given quantitatively by the tableau in the upper right region of the plot. For the case when we have fractional (versus full) factorial designs, the upper right tableau also gives the confounding structure for whatever design was used.

If a factor is important, the "+" average will be considerably different from the "-" average, and so the absolute value of the difference will be large. Conversely, unimportant factors have small differences in the averages, and so the absolute value will be small.

We choose to form a Pareto chart of such |effects|. In the Pareto chart, the largest effects (most important factors) will be presented first (to the left) and then progress down to the smallest effects (least important) factors to the right.

Plot for defective springs data

Applying the |effects| plot to the defective springs data yields the following plot.



How to interpret

From the |effects| plot, we look for the following:

1. The ranked list of factors (including interactions) is given by the left-to-right order of the spikes. These spikes should be of decreasing height as we move from left to right. Note the factor identifier associated with each of these bars.

2. Identify the important factors. Forming the ranked list of factors is important, but is only half of the analysis. The second part of the analysis is to take the ranking and "draw the (horizontal) line" in the list and on the graph so that factors above the line are deemed "important while factors below the line are deemed unimportant.

Since factor effects are frequently a continuum ranging from the very large through the moderate and down to the very small, the separation of all such factors into two groups (important and unimportant) may seem arbitrary and severe. However, in practice, from both a research funding and a modeling point of view, such a bifurcation is both common and necessary.

From an engineering research-funding point of view, one must frequently focus on a subset of factors for future research, attention, and money, and thereby necessarily set aside other factors from any further consideration. From a model-building point of view, a final model either has a term in it or it does not--there is no middle ground. Parsimonious models require in-or-out decisions. It goes without saying that as soon as we have identified the important factors, these are the factors that will comprise our (parsimonious) good model, and those that are declared as unimportant will not be in the model.

Given that, where does such a bifurcation line go?

There are four ways, each discussed in turn, to draw such a line:

1. Statistical significance;
2. Engineering significance;
3. Numerical significance; and
4. Pattern significance.

The ranked list and segregation of factors derived from the |effects| plot are to be compared with the ranked list of factors obtained in previous steps. Invariably, there will be a considerable degree of consistency exhibited across all of the techniques.

Conclusions for the defective springs data

The application of the |effects| plot to the defective springs data set results in the following conclusions:

1. Ranked list of factors (including interactions):
    1. X_1 (most important)
    2. X_1*X_3 (next most important)
    3. X_2
    4. other factors are of lesser importance
2. Separation of factors into important/unimportant categories:
    - Important: X_1, X_1*X_3, and X_2
    - Unimportant: the remainder

5.5.9.7.1. Statistical significance

Formal statistical methods

Formal statistical methods to answer the question of statistical significance commonly involve the use of

- ANOVA (analysis of variance); and
- *t*-based confidence intervals for the effects.

ANOVA

The virtue of ANOVA is that it is a powerful, flexible tool with many applications. The drawback of ANOVA is that

- it is heavily quantitative and non-intuitive;
- it must have an assumed underlying model; and
- its validity depends on assumptions of a constant error variance and normality of the errors.

t confidence intervals

T confidence intervals for the effects, using the *t*-distribution, are also heavily used for determining factor significance. As part of the *t* approach, one first needs to determine *sd*(*effect*), the standard deviation of an effect. For 2-level full and fractional factorial designs, such a standard deviation is related to σ, the standard deviation of an observation under fixed conditions, via the formula:

$$sd(effect) = \frac{2\sigma}{\sqrt{n}}$$

which in turn leads to forming 95% confidence intervals for an effect via

c * *sd*(*effect*)

for an appropriate multiple *c* (from the *t* distribution). Thus in the context of the |effects| plot, "drawing the line" at *c* * *sd*(*effect*) would serve to separate, as desired, the list of effects into 2 domains:

- significant (that is, important); and
- not significant (that is, unimportant).

Estimating sd(effect)

The key in the above approach is to determine an estimate for *sd*(*effect*). Three statistical approaches are common:

1. Prior knowledge about σ:

   If σ is known, we can compute *sd*(*effect*) from the above expression and make use of a conservative (normal-based) 95% confidence interval by drawing the line at

   $$2sd(effect) = 2(\frac{2\sigma}{\sqrt{n}})$$

   This method is rarely used in practice because σ is rarely known.

2. Replication in the experimental design:

   Replication will allow σ to be estimated from the data without depending on the correctness of a deterministic model. This is a real benefit. On the other hand, the downside of such replication is that it increases the number of runs, time, and expense of the experiment. If replication can be afforded, this method should be used. In such a case, the analyst separates important from unimportant terms by drawing the line at

   $$t * sd(effect) = t * (\frac{2\hat{\sigma}}{\sqrt{n}})$$

   with *t* denoting the 97.5 percent point from the appropriate Student's-*t* distribution.

3. Assume 3-factor interactions and higher are zero:

   This approach "assumes away" all 3-factor interactions and higher and uses the data pertaining to these interactions to estimate σ. Specifically,

   $$\hat{\sigma} = \sqrt{\frac{SSQ}{h}}$$

   with *h* denoting the number of 3-factor interactions and higher, and SSQ is the sum of squares for these higher-order effects. The analyst separates important from unimportant effects by drawing the line at

   $$t * sd(effect) = t * (\frac{2\hat{\sigma}}{\sqrt{n}})$$

   with *t* denoting the 97.5 percent point from the appropriate (with *h* degrees of freedom) Student's-*t* distribution.

This method warrants caution:

- it involves an untestable assumption (that such interactions = 0);
- it can result in an estimate for *sd*(*effect*) based on few terms (even a single term); and
- it is virtually unusable for highly-fractionated designs (since high-order interactions are not directly estimable).

Non-statistical considerations

The above statistical methods can and should be used. Additionally, the non-statistical considerations discussed in the next few sections are frequently insightful in practice and have their place in the EDA approach as advocated here.

5.5.9.7.2. Engineering significance

Engineering cutoff

Draw the horizontal line on the chart at that value which you as an engineer have declared beforehand as the engineering cutoff. Any effect larger than this cutoff will be considered as significant from an engineering point of view.

Specifying a cutoff value requires non-statistical thinking, but is frequently useful

This approach requires preliminary, data-free thinking on the part of the analyst as to how big (= what number?) an effect (any effect) must be before the analyst would "care" as an engineer/scientist? In other words, in the units of the response variable, how much would the response variable have to change consistently before the analyst would say "that's a big enough change for me from an engineering point of view"? An engineering number, a cutoff value, is needed here. This value is non-statistical; thie value must emanate from the engineer's head.

If upon reflection the analyst does **not** have such a value in mind, this "engineering significance" approach would be set aside. From experience, it has been found that the engineering soul-searching that goes into evoking such a cutoff value is frequently useful and should be part of the decision process, independent of statistical considerations, of separating the effects into important/unimportant categories.

A rough engineering cutoff

In the absence of a known engineering cutoff, a rough cutoff value is commonly 5 % or 10 % of the **average** (or current) production response for the system. Thus, if a chemical reaction production process is yielding a reaction rate of about 70, then
5 % of 70 = 3. The engineer may declare any future effect that causes an average change of 3 or more units in the response (that is, any estimated effect whose magnitude exceeds 3) to be "engineering significant". In the context of the |effects| plot, the engineer would draw the line at a height of 3 on the plot, and all effects that are above the line are delared as significant and all below the line are declared not significant.

5.5.9.7.3. Numerical significance

10 % of the largest effect	Note the height of the largest bar (= the magnitude of the largest effect). Declare as "significant" any effect that exceeds 10 % of the **largest** effect. The 10 % is arbitrary and has no statistical (or engineering) basis, but it does have a "numeric" basis in that it results in keeping the largest effect and any effects that are within 90 % of the largest effect.
Apply with caution	As with any rule-of-thumb, some caution should be used in applying this critierion. Specifically, if the largest effect is in fact not very large, this rule-of-thumb may not be useful.

5.5.9.7.4. Pattern significance

Look for L-shaped pattern

The |effects| plot has a characteristic horizontally-elongated *L*-shaped pattern. The vertical arm of the *L* consists of important factors. The horizontal arm is comprised of unimportant factors. If a factor is important, the bar height will be large and succeeding bar heights may drop off considerably (perhaps by 50 %)--such factors make up the left arm of the *L*. On the other hand, if a factor is not important, its bar height will tend to be small and near-zero--such factors make up the bottom arm of the *L*. It is of interest to note where the kink is in the *L*. Factors to the left of that kink are arguably declared important while factors at the kink point and to the right of it are declared unimportant.

Factor labels

As a consequence of this "kinking", note the labels on the far right margin of the plot. Factors to the left and above the kink point tend to have far-right labels distinct and isolated. Factors at, to the right, and below the kink point tend to have far right labels that are overstruck and hard to read. A (rough) rule-of-thumb would then be to declare as important those factors/interactions whose far-right labels are easy to distinguish, and to declare as unimportant those factors/interactions whose far-right labels are overwritten and hard to distinguish.

5.5.9.8. Half-normal probability plot

Purpose The half-normal probability plot answers the question:

What are the important factors (including interactions)?

Quantitatively, the estimated effect of a given main effect or interaction and its rank relative to other main effects and interactions is given via least squares estimation (that is, forming effect estimates that minimize the sum of the squared differences between raw data and the fitted values from such estimates). Having such estimates in hand, one could then construct a list of the main effects and interactions ordered by the effect magnitude.

The half-normal probability plot is a graphical tool that uses these ordered estimated effects to help assess which factors are important and which are unimportant.

A half-normal distribution is the distribution of the $|X|$ with X having a normal distribution.

Output The outputs from the half-normal probablity plot are

1. Primary: Grouping of factors and interactions into two categories: important and unimportant. For full factorial designs, interactions include the full complement of interactions of all orders; for fractional factorial designs, interactions include only some, and occasionally none, of the actual interactions (when they aren't estimable).
2. Secondary: Ranked list of factors and interactions from most important down to least important.

Definition A half-normal probability plot is formed by

- Vertical Axis: Ordered (largest to smallest) absolute value of the estimated effects for the main factors and available interactions. If n data points (no replication) have been collected, then typically (n-1) effects will be estimated and the (n-1) |effects| will be plotted.
- Horizontal Axis: (n-1) theoretical order statistic medians from a half-normal distribution. These (n-1) values are not data-dependent. They depend only on the half-normal distribution and the number of items plotted (= n-1). The theoretical medians represent an "ideal" typical ordered data set that would have been obtained from a random drawing of (n-1) samples from a half-normal distribution.

- Far right margin : Factor/interaction identification:
  1 indicates factor X_1;
  2 indicates factor X_2;
  ...
  12 indicates the 2-factor X_1*X_2 interaction
  123 indicates the 3-factor $X_1*X_2*X_3$ interaction,
  etc.

  If the design is a fractional factorial, the confounding structure is provided for main effects and 2-factor interactions.

Motivation

To provide a rationale for the half-normal probability plot, we first dicuss the motivation for the normal probability plot (which also finds frequent use in these 2-level designs).

The basis for the normal probability plot is the mathematical form for each (and all) of the estimated effects. As discussed for the |effects| plot, the estimated effects are the optimal least squares estimates. Because of the orthogonality of the 2^k full factorial and the 2^{k-p} fractional factorial designs, all least squares estimators for main effects and interactions simplify to the form:

estimated effect = $\bar{Y}(+) - \bar{Y}(-)$

with $\bar{Y}(+)$ the average of all response values for which the factor or interaction takes on a "+" value, and where $\bar{Y}(-)$ is the average of all response values for which the factor or interaction takes on a "-" value.

Under rather general conditions, the Central Limit Thereom allows that the difference-of-sums form for the estimated effects tends to follow a normal distribution (for a large enough sample size n) a normal distribution.

The question arises as to what normal distribution; that is, a normal distribution with what mean and what standard deviation? Since all estimators have an identical form (a difference of averages), the standard deviations, though unknown, will in fact be the same under the assumption of constant σ. This is good in that it simplifies the normality analysis.

As for the means, however, there will be differences from one effect to the next, and these differences depend on whether a factor is unimportant or important. **Unimportant** factors are those that have near-zero effects and **important** factors are those whose effects are considerably removed from zero. Thus, **unimportant** effects tend to have a normal distribution centered near zero while **important** effects tend to have a normal distribution centered at their respective true large (but unknown) effect values.

In the simplest experimental case, if the experiment were such that no factors were important (that is, all effects were near zero), the (n-1) estimated effects would behave like random drawings from a normal distribution centered at zero. We can test for such normality (and hence test for a null-effect experiment) by using the normal probability plot. Normal probability plots are easy to interpret. In simplest terms:

if linear, then normal

If the normal probability plot of the (n-1) estimated effects is linear, this implies that all of the true (unknown) effects are zero or near-zero. That is, no factor is important.

On the other hand, if the truth behind the experiment is that there is exactly one factor that was important (that is, significantly non-zero), and all remaining factors are unimportant (that is, near-zero), then the normal probability plot of all (n-1) effects is near-linear for the (n-2) unimportant factors and the remaining single important factor would stand well off the line.

Similarly, if the experiment were such that some subset of factors were important and all remaining factors were unimportant, then the normal probability plot of all (n-1) effects would be near-linear for all unimportant factors with the remaining important factors all well off the line.

In real life, with the number of important factors unknown, this suggests that one could form a normal probability plot of the (n-1) estimated effects and draw a line through those (unimportant) effects in the vicinity of zero. This identifies and extracts all remaining effects off the line and declares them as important.

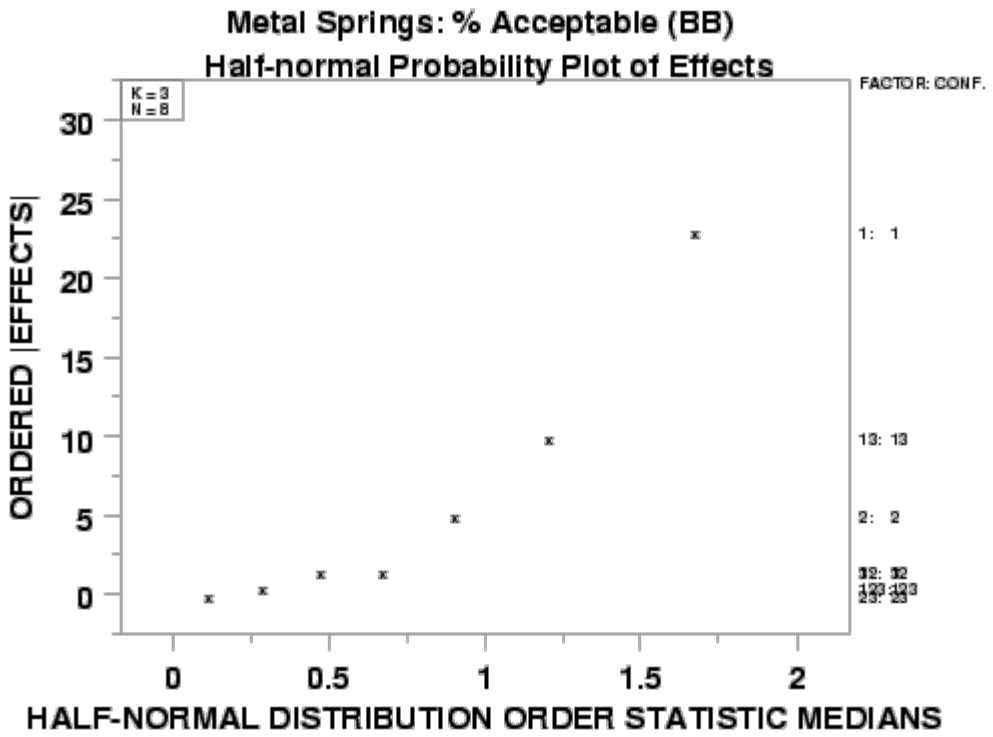
The above rationale and methodology works well in practice, with the net effect that the normal probability plot of the effects is an important, commonly used and successfully employed tool for identifying important factors in 2-level full and factorial experiments. Following the lead of Cuthbert Daniel (1976), we augment the methodology and arrive at a further improvement. Specifically, the sign of each estimate is completely arbitrary and will reverse depending on how the initial assignments were made (e.g., we could assign "-" to treatment A and "+" to treatment B or just as easily assign "+" to treatment A and "-" to treatment B).

This arbitrariness is addressed by dealing with the effect magnitudes rather than the signed effects. If the signed effects follow a normal distribution, the absolute values of the effects follow a half-normal distribution.

In this new context, one tests for important versus unimportant factors by generating a half-normal probability plot of the absolute value of the effects. As before, linearity implies half-normality, which in turn implies all factors are unimportant. More typically, however, the half-normal probability plot will be only partially linear. Unimportant (that is, near-zero) effects manifest themselves as being near zero and on a line while important (that is, large) effects manifest themselves by being off the line and well-displaced from zero.

Plot for defective springs data

The half-normal probability plot of the effects for the defectice springs data set is as follows.



How to interpret

From the half-normal probability plot, we look for the following:

1. Identifying Important Factors:

   Determining the subset of important factors is the most important task of the half-normal probability plot of |effects|. As discussed above, the estimated |effect| of an **un**important factor will typically be on or close to a near-zero line, while the estimated |effect| of an important factor will typically be displaced well off the line.

   The separation of factors into important/unimportant categories is thus done by answering the question:

   > Which points on the half-normal probability plot of |effects| are large and well-off the linear collection of points drawn in the vicinity of the origin?

   This line of unimportant factors typically encompasses the majority of the points on the plot. The procedure consists, therefore, of the following:

   1. identifying this line of near-zero (unimportant) factors; then
   2. declaring the remaining off-line factors as important.

   Note that the half-normal probability plot of |effects| and the |effects| plot have the same vertical axis; namely, the ordered |effects|, so the following discussion about right-margin factor identifiers is relevant to both plots. As a consequence of the natural on-line/off-line segregation of the |effects| in half-normal probability plots, factors off-line tend to have far-right labels that are distinct and isolated while factors near the line tend to have far-right labels that are overstruck and hard to read. The rough rule-of-thumb would then be

to declare as important those factors/interactions whose far-right labels are easy to distinguish and to declare as unimportant those factors/interactions whose far-right labels are overwritten and hard to distinguish.

2. Ranked List of Factors (including interactions):

   This is a minor objective of the half-normal probability plot (it is better done via the |effects| plot). To determine the ranked list of factors from a half-normal probability plot, simply scan the vertical axis |effects|

   1. Which |effect| is largest? Note the factor identifier associated with this largest |effect| (this is the "most important factor").

   2. Which |effect| is next in size? Note the factor identifier associated with this next largest |effect| (this is the "second most important factor").

   3. Continue for the remaining factors. In practice, the bottom end of the ranked list (the unimportant factors) will be hard to extract because of overstriking, but the top end of the ranked list (the important factors) will be easy to determine.

In summary, it should be noted that since the signs of the estimated effects are arbitrary, we recommend the use of the half-normal probability plot of |effects| technique over the normal probability plot of the |effects|. These probability plots are among the most commonly-employed EDA procedure for identification of important factors in 2-level full and factorial designs. The half-normal probability plot enjoys widespread usage across both "classical" and Taguchi camps. It deservedly plays an important role in our recommended 10-step graphical procedure for the analysis of 2-level designed experiments.

Conclusions for the defective springs data

The application of the half-normal probability plot to the defective springs data set results in the following conclusions:

1. Ranked list of factors (including interactions):
   1. X_1 (most important)
   2. X_1*X_3 (next most important)
   3. X_2
   4. other factors are of lesser importance

2. Separation of factors into important/unimportant categories:
   Important: X_1, X_1*X_3, and X_2
   Unimportant: the remainder

5.5.9.9. Cumulative residual standard deviation plot

Purpose The cumulative residual sd (standard deviation) plot answers the question:

What is a good model for the data?

The prior 8 steps in this analysis sequence addressed the two important goals:

1. Factors: determining the most important factors that affect the response, and
2. Settings: determining the best settings for these factors.

In addition to the above, a third goal is of interest:

3. Model: determining a model (that is, a prediction equation) that functionally relates the observed response *Y* with the various main effects and interactions.

Such a function makes particular sense when all of the individual factors are continuous and ordinal (such as temperature, pressure, humidity, concentration, etc.) as opposed to any of the factors being discrete and non-ordinal (such as plant, operator, catalyst, supplier).

In the continuous-factor case, the analyst could use such a function for the following purposes.

1. Reproduction/Smoothing: predict the response at the observed design points.
2. Interpolation: predict what the response would be at (unobserved) regions between the design points.
3. Extrapolation: predict what the response would be at (unobserved) regions beyond the design points.

For the discrete-factor case, the methods developed below to arrive at such a function still apply, and so the resulting model may be used for reproduction. However, the interpolation and extrapolation aspects do not apply.

In modeling, we seek a function f in the k factors $X_1, X_2, ..., X_k$ such that the predicted values

$$\hat{Y} = f(X_1, X_2, ..., X_k)$$

are "close" to the observed raw data values Y. To this end, two tasks exist:

1. Determine a good functional form f;
2. Determine good estimates for the coefficients in that function f.

For example, if we had two factors X_1 and X_2, our goal would be to

1. determine some function $f(X_1,X_2)$; and
2. estimate the parameters in f

such that the resulting model would yield predicted values $\hat{Y}$ that are as close as possible to the observed response values Y. If the form f has been wisely chosen, a good model will result and that model will have the characteristic that the differences ("residuals" = $Y - \hat{Y}$) will be uniformly near zero. On the other hand, a poor model (from a poor choice of the form f) will have the characteristic that some or all of the residuals will be "large".

For a given model, a statistic that summarizes the quality of the fit via the typical size of the n residuals is the residual standard deviation:

$$s_{res} = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n-p}}$$

with p denoting the number of terms in the model (including the constant term) and r denoting the ith residual. We are also assuming that the mean of the residuals is zero, which will be the case for models with a constant term that are fit using least squares.

If we have a good-fitting model, s_{res} will be small. If we have a poor-fitting model, s_{res} will be large.

For a given data set, each proposed model has its own quality of fit, and hence its own residual standard deviation. Clearly, the residual standard deviation is more of a model-descriptor than a data-descriptor. Whereas "nature" creates the data, the analyst creates the models. Theoretically, for the same data set, it is possible for the analyst to propose an indefinitely large number of models.

In practice, however, an analyst usually forwards only a small, finite number of plausible models for consideration. Each model will have its own residual standard deviation. The cumulative residual standard deviation plot is simply a graphical representation of this collection of residual standard deviations for various models. The plot is beneficial in that

1. good models are distinguished from bad models;
2. simple good models are distinguished from complicated good models.

In summary, then, the cumulative residual standard deviation plot is a graphical tool to help assess

1. which models are poor (least desirable); and
2. which models are good but complex (more desirable); and
3. which models are good and simple (most desirable).

Output

The outputs from the cumulative residual standard deviation plot are

1. Primary: A good-fitting prediction equation consisting of an additive constant plus the most important main effects and interactions.

2. Secondary: The residual standard deviation for this good-fitting model.

Definition A cumulative residual sd plot is formed by

1. Vertical Axis: Ordered (largest to smallest) residual standard deviations of a sequence of progressively more complicated fitted models.

2. Horizontal Axis: Factor/interaction identification of the last term included into the linear model:

   1 indicates factor X_1;
   2 indicates factor X_2;
   ...
   12 indicates the 2-factor X_1*X_2 interaction
   123 indicates the 3-factor $X_1*X_2*X_3$ interaction
   etc.

3. Far right margin: Factor/interaction identification (built-in redundancy):

   1 indicates factor X_1;
   2 indicates factor X_2;
   ...
   12 indicates the 2-factor X_1*X_2 interaction
   123 indicates the 3-factor $X_1*X_2*X_3$ interaction
   etc.

   If the design is a fractional factorial, the confounding structure is provided for main effects and 2-factor interactions.

The cumulative residual standard deviations plot is thus a Pareto-style, largest to smallest, graphical summary of residual standard deviations for a selected series of progressively more complicated linear models.

The plot shows, from left to right, a model with only a constant and the model then augmented by including, one at a time, remaining factors and interactions. Each factor and interaction is incorporated into the model in an additive (rather than in a multiplicative or logarithmic or power, etc. fashion). At any stage, the ordering of the next term to be added to the model is such that it will result in the maximal decrease in the resulting residual standard deviation.

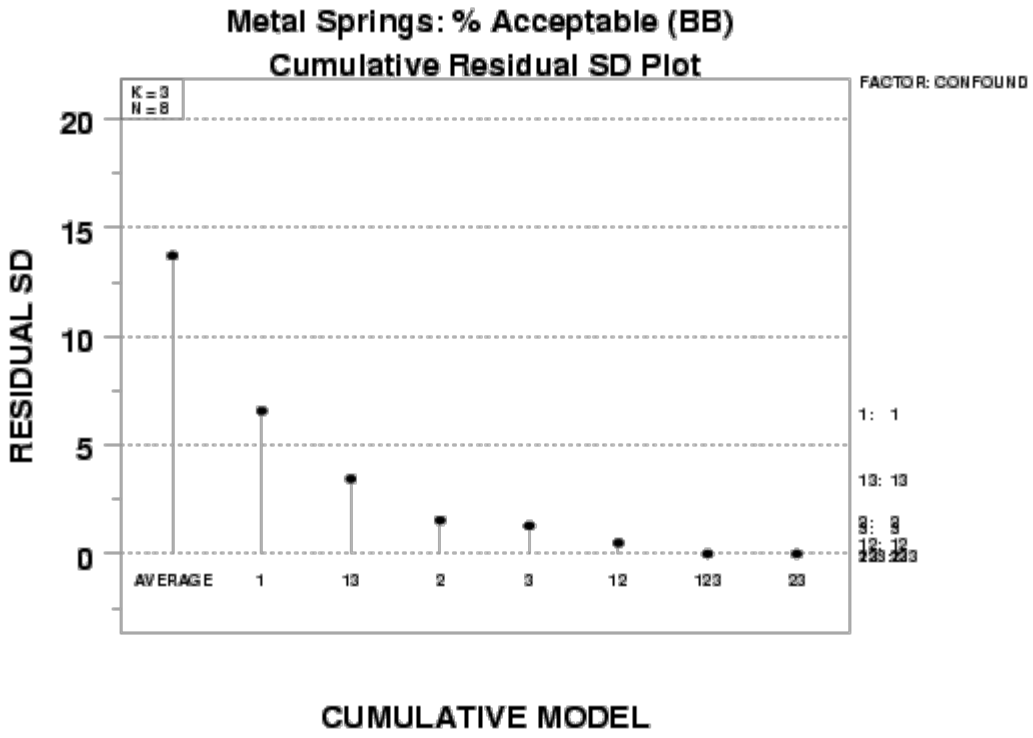
Motivation This section addresses the following questions:

1. What is a model?
2. How do we select a goodness-of-fit metric for a model?
3. How do we construct a good model?
4. How do we know when to stop adding terms?
5. What is the final form for the model?
6. What are the advantages of the linear model?
7. How do we use the model to generate predicted values?
8. How do we use the model beyond the data domain?
9. What is the best confirmation point for interpolation?
10. How do we use the model for interpolation?

11. How do we use the model for extrapolation?

Plot for defective springs data

Applying the cumulative residual standard deviation plot to the defective springs data set yields the following plot.



How to interpret

As discussed in detail under question 4 in the Motivation section, the cumulative residual standard deviation "curve" will characteristically decrease left to right as we add more terms to the model. The incremental improvement (decrease) tends to be large at the beginning when important factors are being added, but then the decrease tends to be marginal at the end as unimportant factors are being added.

Including all terms would yield a perfect fit (residual standard deviation = 0) but would also result in an unwieldy model. Including only the first term (the average) would yield a simple model (only one term!) but typically will fit poorly. Although a formal quantitative stopping rule can be developed based on statistical theory, a less-rigorous (but good) alternative stopping rule that is graphical, easy to use, and highly effective in practice is as follows:

> Keep adding terms to the model until the curve's "elbow" is encountered. The "elbow point" is that value in which there is a consistent, noticeably shallower slope (decrease) in the curve. Include all terms up to (and including) the elbow point (after all, each of these included terms decreased the residual standard deviation by a large amount). Exclude any terms **after** the elbow point since all such successive terms decreased the residual standard deviation so slowly that the terms were "not worth the complication of keeping".

From the residual standard deviation plot for the defective springs data, we note the following:

1. The residual standard deviation (rsd) for the "baseline" model

$$\hat{Y} = \bar{Y} = 71.25$$

   is s_{res} = 13.7.

2. As we add the next term, X_1, the rsd drops nearly 7 units (from 13.7 to 6.6).

3. If we add the term X_1*X_3, the rsd drops another 3 units (from 6.6 to 3.4).

4. If we add the term X_2, the rsd drops another 2 units (from 3.4 to 1.5).

5. When the term X_3 is added, the reduction in the rsd (from about 1.5 to 1.3) is negligible.

6. Thereafter to the end, the total reduction in the rsd is from only 1.3 to 0.

In step 5, note that when we have effects of equal magnitude (the X_3 effect is equal to the X_1*X_2 interaction effect), we prefer including a main effect before an interaction effect and a lower-order interaction effect before a higher-order interaction effect.

In this case, the "kink" in the residual standard deviation curve is at the X_2 term. Prior to that, all added terms (including X_2) reduced the rsd by a large amount (7, then 3, then 2). After the addition of X_2, the reduction in the rsd was small (all less than 1): 0.2, then 0.8, then 0.5, then 0.

The final recommended model in this case thus involves p = 4 terms:

1. the average
2. factor X_1
3. the X_1*X_3 interaction
4. factor X_2

The fitted model thus takes on the form

$$\hat{Y} = average + B_1*X_1 + B_{13}*X_1*X_3 + B_2*X_2$$

The least-squares estimates for the coefficients in this model are

average = 71.25
B_1 = 11.5
B_{13} = 5
B_2 = -2.5

The B_1 = 11.5, B_{13} = 5, and B_2 = -2.5 least-squares values are, of course, one half of the estimated effects E_1 = 23, E_{13} = 10, and E_2 = -5. Effects, calculated as $\hat{Y}(+1) - \hat{Y}(-1)$, were previously derived in step 7 of the recommended 10-step DOE analysis procedure.

The final fitted model is thus

$$\hat{Y} = 71.25 + 11.5*X_1 + 5*X_1*X_3 - 2.5*X_2$$

Applying this prediction equation to the 8 design points yields: predicted values $\hat{Y}$ that are close to the data Y, and residuals ($Res = Y - \hat{Y}$) that are close to zero:

X_1	X_2	X_3	Y	$\hat{Y}$	Res
-	-	-	67	67.25	-0.25
+	-	-	79	80.25	-1.25
-	+	-	61	62.25	-1.25
+	+	-	75	75.25	-0.25
-	-	+	59	57.25	+1.75
+	-	+	90	90.25	-0.25
-	+	+	52	52.25	-0.25
+	+	+	87	85.25	+1.75

Computing the residual standard deviation:

$$s_{res} = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n-p}}$$

with n = 8 data points, and p = 4 estimated coefficients (including the average) yields

s_{res} = 1.54 (or 1.5 if rounded to 1 decimal place)

The detailed s_{res} = 1.54 calculation brings us full circle, for 1.54 is the value given above the X_3 term on the cumulative residual standard deviation plot.

Conclusions for the defective springs data

The application of the Cumulative Residual Standard Deviation Plot to the defective springs data set results in the following conclusions:

1. Good-fitting Parsimonious (constant + 3 terms) Model:

   $$\hat{Y} = 71.25 + 11.5*X_1 + 5*X_1*X_3 - 2.5*X_2$$

2. Residual Standard Deviation for this Model (as a measure of the goodness-of-fit for the model):

   s_{res} = 1.54

5.5.9.9.1. Motivation: What is a Model?

Mathematical models: functional form and coefficients

A model is a mathematical function that relates the response Y to the factors X_1 to X_k. A model has a

1. functional form; and
2. coefficients.

An excellent and easy-to-use functional form that we find particularly useful is a linear combination of the main effects and the interactions (the selected model is a subset of the full model and almost always a proper subset). The coefficients in this linear model are easy to obtain via application of the least squares estimation criterion (regression). A given functional form with estimated coefficients is referred to as a "fitted model" or a "prediction equation".

Predicted values and residuals

For given settings of the factors X_1 to X_k, a fitted model will yield predicted values. For each (and every) setting of the X_i's, a "perfect-fit" model is one in which the predicted values are identical to the observed responses Y at these X_i's. In other words, a perfect-fit model would yield a vector of predicted values identical to the observed vector of response values. For these same X_i's, a "good-fitting" model is one that yields predicted values "acceptably near", but not necessarily identical to, the observed responses Y.

The residuals (= deviations = error) of a model are the vector of differences ($Y - \hat{Y}$) between the responses and the predicted values from the model. For a perfect-fit model, the vector of residuals would be all zeros. For a good-fitting model, the vector of residuals will be acceptably (from an engineering point of view) close to zero.

5.5.9.9.2. Motivation: How do we Construct a Goodness-of-fit Metric for a Model?

Motivation

This question deals with the issue of how to construct a metric, a statistic, that may be used to ascertain the quality of the fitted model. The statistic should be such that for one range of values, the implication is that the model is good, whereas for another range of values, the implication is that the model gives a poor fit.

Sum of absolute residuals

Since a model's adequacy is inversely related to the size of its residuals, one obvious statistic is the sum of the absolute residuals.

$$AR = \sum_{i=1}^{n} |r_i|$$

Clearly, for a fixed *n*,the smaller this sum is, the smaller are the residuals, which implies the closer the predicted values are to the raw data *Y*, and hence the better the fitted model. The primary disadvantage of this statistic is that it may grow larger simply as the sample size *n* grows larger.

Average absolute residual

A better metric that does not change (much) with increasing sample size is the average absolute residual:

$$AAR = \frac{\sum_{i=1}^{n} |r_i|}{n}$$

with *n* denoting the number of response values. Again, small values for this statistic imply better-fitting models.

Square root of the average squared residual

An alternative, but similar, metric that has better statistical properties is the square root of the average squared residual.

$$\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}}$$

As with the previous statistic, the smaller this statistic, the better the model.

Residual standard deviation

Our final metric, which is used directly in inferential statistics, is the residual standard deviation

$$s_{res} = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n-p}}$$

with p denoting the number of fitted coefficients in the model. This statistic is the standard deviation of the residuals from a given model. The smaller is this residual standard deviation, the better fitting is the model. We shall use the residual standard deviation as our metric of choice for evaluating and comparing various proposed models.

5.5.9.9.3. Motivation: How do we Construct a Good Model?

Models for 2^k and 2^{k-p} designs

Given that we have a statistic to measure the quality of a model, any model, we move to the question of how to construct reasonable models for fitting data from 2^k and 2^{k-p} designs.

Initial simple model

The simplest such proposed model is

$$Y = c + \epsilon$$

that is, the response Y = a constant + random error. This trivial model says that all of the factors (and interactions) are in fact worthless for prediction and so the best-fit model is one that consists of a simple horizontal straight line through the body of the data. The least squares estimate for this constant c in the above model is the sample mean $\bar{Y}$. The prediction equation for this model is thus

$$\hat{Y} = \bar{Y}$$

The predicted values $\hat{Y}$ for this fitted trivial model are thus given by a vector consisting of the same value (namely $\bar{Y}$) throughout. The residual vector for this model will thus simplify to simple deviations from the mean:

$$Y - \bar{Y}$$

Since the number of fitted coefficients in this model is 1 (namely the constant c), the residual standard deviation is the following:

$$s_{res} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$$

which is of course the familiar, commonly employed sample standard deviation. If the residual standard deviation for this trivial model were "small enough", then we could terminate the model-building process right there with no further inclusion of terms. In practice, however, this trivial model does **not** yield a residual standard deviation that is small

$\bar{Y}$

enough (because the common value will not be close enough to some of the raw responses *Y*) and so the model must be augmented--but how?

Next-step model

The logical next-step proposed model will consist of the above additive constant plus some term that will improve the predicted values the most. This will equivalently reduce the residuals the most and thus reduce the residual standard deviation the most.

Using the most important effects

As it turns out, it is a mathematical fact that the factor or interaction that has the largest estimated effect

$$\hat{E} = \bar{Y}(+) - \bar{Y}(-)$$

will necessarily, after being included in the model, yield the "biggest bang for the buck" in terms of improving the predicted values toward the response values *Y*. Hence at this point the model-building process and the effect estimation process merge.

In the previous steps in our analysis, we developed a ranked list of factors and interactions. We thus have a ready-made ordering of the terms that could be added, one at a time, to the model. This ranked list of effects is precisely what we need to cumulatively build more complicated, but better fitting, models.

Step through the ranked list of factors

Our procedure will thus be to step through, one by one, the ranked list of effects, cumulatively augmenting our current model by the next term in the list, and then compute (for all *n* design points) the predicted values, residuals, and residual standard deviation. We continue this one-term-at-a-time augmentation until the predicted values are acceptably close to the observed responses *Y* (and hence the residuals and residual standard deviation become acceptably close to zero).

Starting with the simple average, each cumulative model in this iteration process will have its own associated residual standard deviation. In practice, the iteration continues until the residual standard deviations become sufficiently small.

Cumulative residual standard deviation plot

The cumulative residual standard deviation plot is a graphical summary of the above model-building process. On the horizontal axis is a series of terms (starting with the average, and continuing on with various main effects and interactions). After the average, the ordering of terms on the horizontal axis is identical to the ordering of terms based on the half-normal probability plot ranking based on effect magnitude.

On the vertical axis is the corresponding residual standard deviation that results when the cumulative model has its coefficients fitted via least squares, and then has its predicted values, residuals, and residual standard deviations computed.

The first residual standard deviation (on the far left of the cumulative residual standard deviation plot) is that which results from the model consisting of

1. the average.

The second residual standard deviation plotted is from the model consisting of

1. the average, plus
2. the term with the largest |effect|.

The third residual standard deviation plotted is from the model consisting of

1. the average, plus
2. the term with the largest |effect|, plus
3. the term with the second largest |effect|.

and so forth.

5.5.9.9.4. Motivation: How do we Know When to Stop Adding Terms?

Cumulative residual standard deviation plot typically has a hockey stick appearance

Proceeding left to right, as we add more terms to the model, the cumulative residual standard deviation "curve" will typically decrease. At the beginning (on the left), as we add large-effect terms, the decrease from one residual standard deviation to the next residual standard deviation will be large. The incremental improvement (decrease) then tends to drop off slightly. At some point the incremental improvement will typically slacken off considerably. Appearance-wise, it is thus very typical for such a curve to have a "hockey stick" appearance:

1. starting with a series of large decrements between successive residual standard deviations; then
2. hitting an elbow; then
3. having a series of gradual decrements thereafter.

Stopping rule

The cumulative residual standard deviation plot provides a visual answer to the question:

> What is a good model?

by answering the related question:

> When do we stop adding terms to the cumulative model?

Graphically, the most common stopping rule for adding terms is to cease immediately upon encountering the "elbow". We include all terms up to and including the elbow point since each of these terms decreased the residual standard deviation by a large amount. However, we exclude any terms afterward since these terms do not decrease the residual standard deviation fast enough to warrant inclusion in the model.

5.5.9.9.5. Motivation: What is the Form of the Model?

Models for various values of k

From the above discussion, we thus note and recommend a form of the model that consists of an additive constant plus a linear combination of main effects and interactions. What then is the specific form for the linear combination?

The following are the full models for various values of k. The selected final model will be a subset of the full model.

- For the $k = 1$ factor case:

$$Y = f(X_1) + \varepsilon = c + B_1 \cdot X_1 + \varepsilon$$

- For the $k = 2$ factor case:

$$Y = f(X_1, X_2) + \varepsilon = c + B_1 \cdot X_1 + B_2 \cdot X_2 + B_{12} \cdot X_1 \cdot X_2 + \varepsilon$$

- For the $k = 3$ factor case:

$$Y = f(X_1, X_2, X_3) + \varepsilon = c + B_1 \cdot X_1 + B_2 \cdot X_2 + B_3 \cdot X_3 + B_{12} \cdot X_1 \cdot X_2 + B_{13} \cdot X_1 \cdot X_3 + B_{23} \cdot X_2 \cdot X_3 + B_{123} \cdot X_1 \cdot X_2 \cdot X_2 + \varepsilon$$

- and for the general k case:

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon = c + (\text{linear combination of all main effects and all interactions}) + \varepsilon$$

Note that the model equations shown above include coefficients that represent the change in Y for a one-unit change in X_i. To obtain an effect estimate, which represents a

two-unit change in X_i if the levels of X_i are +1 and -1, simply multiply the coefficient by two.

Ordered linear combination

The listing above has the terms ordered with the main effects, then the 2-factor interactions, then the 3-factor interactions, etc. In practice, it is recommended that the terms be ordered by importance (whether they be main effects or interactions). Aside from providing a functional representation of the response, models should help reinforce what is driving the response, which such a re-ordering does. Thus for $k = 2$, if factor 2 is most important, the 2-factor interaction is next in importance, and factor 1 is least important, then it is recommended that the above ordering of

$$Y = f(X_1,X_2) + \varepsilon = c + B_1 \cdot X_1 + B_2 \cdot X_2 + B_{12} \cdot X_1 \cdot X_2 + \varepsilon$$

be rewritten as

$$Y = f(X_1,X_2) + \varepsilon = c + B_2 \cdot X_2 + B_{12} \cdot X_1 \cdot X_2 + B_1 \cdot X_1 + \varepsilon$$

5.5.9.9.6. Motivation: What are the Advantages of the Linear Combinatoric Model?

Advantages: perfect fit and comparable coefficients

The linear model consisting of main effects and all interactions has two advantages:

1. Perfect Fit: If we choose to include in the model all of the main effects and all interactions (of all orders), then the resulting least squares fitted model will have the property that the predicted values will be **identical** to the raw response values *Y*. We will illustrate this in the next section.

2. Comparable Coefficients: Since the model fit has been carried out in the coded factor (-1, +1) units rather than the units of the original factor (temperature, time, pressure, catalyst concentration, etc.), the factor coefficients immediately become comparable to one another, which serves as an immediate mechanism for the scale-free ranking of the relative importance of the factors.

Example

To illustrate in detail the above latter point, suppose the (-1, +1) factor X_1 is really a coding of temperature T with the original temperature ranging from 300 to 350 degrees and the (-1, +1) factor X_2 is really a coding of time t with the original time ranging from 20 to 30 minutes. Given that, a linear model in the original temperature T and time t would yield coefficients whose magnitude depends on the magnitude of T (300 to 350) and t (20 to 30), and whose value would change if we decided to change the units of T (e.g., from Fahrenheit degrees to Celsius degrees) and t (e.g., from minutes to seconds). All of this is avoided by carrying out the fit not in the original units for T (300,350) and t (20, 30), but in the coded units of X_1 (-1, +1) and X_2 (-1, +1).

The resulting coefficients are unit-invariant, and thus the coefficient magnitudes reflect the true contribution of the factors and interactions without regard to the unit of measurement.

Coding

Such coding leads to no loss of generality since the coded

does not lead to loss of generality

factor may be expressed as a simple linear relation of the original factor (X_1 to T, X_2 to t). The unit-invariant coded coefficients may be easily transformed to unit-sensitive original coefficients if so desired.

5.5.9.9.7. Motivation: How do we use the Model to Generate Predicted Values?

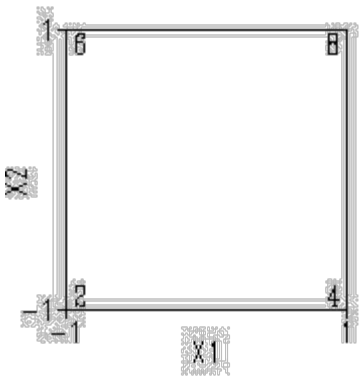
Design matrix with response for two factors

To illustrate the details as to how a model may be used for prediction, let us consider a simple case and generalize from it. Consider the simple Yates-order 2^2 full factorial design in X_1 and X_2, augmented with a response vector Y:

X_1	X_2	Y
-	-	2
+	-	4
-	+	6
+	+	8

Geometric representation

This can be represented geometrically



Determining the prediction equation

For this case, we might consider the model

$$Y = c + B_1*X_1 + B_2*X_2 + B_{12}*X_1*X_2 + \varepsilon$$

From the above diagram, we may deduce that the estimated factor effects are:

c = the average response = $\bar{Y}$
(2 + 4 + 6 + 8) / 4 = 5

=

E_1 = average change in Y as X_1 goes from -1 to +1

= ((4-2) + (8-6)) / 2 = (2 + 2) / 2 = 2

Note: the (4-2) is the change in Y (due to X_1) on the lower axis; the (8-6) is the change in Y (due to X_1) on the upper axis.

E_2 = average change in Y as X_2 goes from -1 to +1

= ((6-2) + (8-4)) / 2 = (4 + 4) / 2 = 4

E_{12} = interaction = (the less obvious) average change in Y as X_1*X_2 goes from -1 to +1

= ((2-4) + (8-6)) / 2 = (-2 + 2) / 2 = 0

For factors coded using +1 and -1, the least-squares estimate of a coefficient is one half of the effect estimate ($B_i = E_i / 2$), so the fitted model (that is, the prediction equation) is

$$\hat{Y} = 5 + 1*X_1 + 2*X_2 + 0*X_1*X_2$$

or with the terms rearranged in descending order of importance

$$\hat{Y} = 5 + 2*X_2 + X_1$$

Table of fitted values

Substituting the values for the four design points into this equation yields the following fitted values

X_1	X_2	Y	$\hat{Y}$
-	-	2	2
+	-	4	4
-	+	6	6
+	+	8	8

Perfect fit

This is a perfect-fit model. Such perfect-fit models will result anytime (in this orthogonal 2-level design family) we include all main effects and all interactions. Remarkably, this is true not only for $k = 2$ factors, but for general k.

Residuals

For a given model (any model), the difference between the response value Y and the predicted value $\hat{Y}$ is referred to as the "residual":

residual = Y - $\hat{Y}$

The perfect-fit full-blown (all main factors and all interactions of all orders) models will have all residuals identically zero.

The perfect fit is a mathematical property that comes if we choose to use the linear model with all possible terms.

Price for perfect fit

What price is paid for this perfect fit? One price is that the variance of $\hat{Y}$ is increased unnecessarily. In addition, we have a non-parsimonious model. We must compute and carry the average and the coefficients of all main effects and all interactions. Including the average, there will in general be 2^k coefficients to fully describe the fitting of the $n = 2^k$ points. This is very much akin to the $Y = f(X)$ polynomial fitting of n distinct points. It is well known that this may be done "perfectly" by fitting a polynomial of degree n-1. It is comforting to know that such perfection is mathematically attainable, but in practice do we want to do this all the time or even anytime? The answer is generally "no" for two reasons:

1. Noise: It is very common that the response data Y has noise (= error) in it. Do we want to go out of our way to fit such noise? Or do we want our model to filter out the noise and just fit the "signal"? For the latter, fewer coefficients may be in order, in the same spirit that we may forego a perfect-fitting (but jagged) 11-th degree polynomial to 12 data points, and opt out instead for an imperfect (but smoother) 3rd degree polynomial fit to the 12 points.

2. Parsimony: For full factorial designs, to fit the $n = 2^k$ points we would need to compute 2^k coefficients. We gain information by noting the magnitude and sign of such coefficients, but numerically we have n data values Y as input and n coefficients B as output, and so no numerical reduction has been achieved. We have simply used one set of n numbers (the data) to obtain another set of n numbers (the coefficients). Not all of these coefficients will be equally important. At times that importance becomes clouded by the sheer volume of the $n = 2^k$ coefficients. Parsimony suggests that our result should be simpler and more focused than our n starting points. Hence fewer retained coefficients are called for.

The net result is that in practice we almost always give up the perfect, but unwieldy, model for an imperfect, but parsimonious, model.

Imperfect fit

The above calculations illustrated the computation of predicted values for the full model. On the other hand, as

discussed above, it will generally be convenient for signal or parsimony purposes to deliberately omit some unimportant factors. When the analyst chooses such a model, we note that the methodology for computing predicted values $\hat{Y}$ is precisely the same. In such a case, however, the resulting predicted values will in general **not** be identical to the original response values *Y*; that is, we no longer obtain a perfect fit. Thus, linear models that omit some terms will have virtually all non-zero residuals.

5.5.9.9.8. Motivation: How do we Use the Model Beyond the Data Domain?

Interpolation and extrapolation

The previous section illustrated how to compute predicted values at the points included in the design. One of the virtues of modeling is that the resulting prediction equation is **not** restricted to the design data points. From the prediction equation, predicted values can be computed elsewhere and anywhere:

1. within the domain of the data (interpolation);
2. outside of the domain of the data (extrapolation).

In the hands of an expert scientist/engineer/analyst, the ability to predict elsewhere is extremely valuable. Based on the fitted model, we have the ability to compute predicted values for the response at a large number of internal and external points. Thus the analyst can go beyond the handful of factor combinations at hand and can get a feel (typically via subsequent contour plotting) as to what the nature of the entire response surface is.

This added insight into the nature of the response is "free" and is an incredibly important benefit of the entire model-building exercise.

Predict with caution

Can we be fooled and misled by such a mathematical and computational exercise? After all, is not the only thing that is "real" the data, and everything else artificial? The answer is "yes", and so such interpolation/extrapolation is a double-edged sword that must be wielded with care. The best attitude, and especially for extrapolation, is that the derived conclusions must be viewed with extra caution.

By construction, the recommended fitted models should be good at the design points. If the full-blown model were used, the fit will be perfect. If the full-blown model is reduced just a bit, then the fit will still typically be quite good. By continuity, one would expect perfection/goodness at the design points would lead to goodness in the immediate vicinity of the design points. However, such local goodness does **not** guarantee that the derived model will be good at some distance from the design points.

Do confirmation runs

Modeling and prediction allow us to go beyond the data to gain additional insights, but they must be done with great caution. Interpolation is generally safer than extrapolation, but mis-prediction, error, and misinterpretation are liable to occur in either case.

The analyst should definitely perform the model-building process and enjoy the ability to predict elsewhere, but the analyst must always be prepared to validate the interpolated and extrapolated predictions by collection of additional real, confirmatory data. The general empirical model that we recommend knows "nothing" about the engineering, physics, or chemistry surrounding your particular measurement problem, and although the model is the best generic model available, it must nonetheless be confirmed by additional data. Such additional data can be obtained pre-experimentally or post-experimentally. If done pre-experimentally, a recommended procedure for checking the validity of the fitted model is to augment the usual 2^k or 2^{k-p} designs with additional points at the center of the design. This is discussed in the next section.

Applies only for continuous factors

Of course, all such discussion of interpolation and extrapolation makes sense only in the context of continuous ordinal factors such as temperature, time, pressure, size, etc. Interpolation and extrapolation make no sense for discrete non-ordinal factors such as supplier, operators, design types, etc.

5.5.9.9.9. Motivation: What is the Best Confirmation Point for Interpolation?

Augment via center point

For the usual continuous factor case, the best (most efficient and highest leverage) additional model-validation point that may be added to a 2^k or 2^{k-p} design is at the center point. This center point augmentation "costs" the experimentalist only one additional run.

Example

For example, for the $k = 2$ factor (Temperature (300 to 350), and time (20 to 30)) experiment discussed in the previous sections, the usual 4-run 2^2 full factorial design may be replaced by the following 5-run 2^2 full factorial design with a center point.

X_1	X_2	Y
-	-	2
+	-	4
-	+	6
+	+	8
0	0	

Predicted value for the center point

Since "-" stands for -1 and "+" stands for +1, it is natural to code the center point as (0,0). Using the recommended model

$$\hat{Y} = 5 + 2*X_2 + X_1$$

we can substitute 0 for X_1 and X_2 to generate the predicted value of 5 for the confirmatory run.

Importance of the confirmatory run

The importance of the confirmatory run cannot be overstated. If the confirmatory run at the center point yields a data value of, say, $Y = 5.1$, since the predicted value at the center is 5 and we know the model is perfect at the corner points, that would give the analyst a greater confidence that the quality of the fitted model may extend over the entire interior (interpolation) domain. On the other hand, if the

confirmatory run yielded a center point data value quite different (e.g., $Y = 7.5$) from the center point predicted value of 5, then that would prompt the analyst to **not** trust the fitted model even for interpolation purposes. Hence when our factors are continuous, a single confirmatory run at the center point helps immensely in assessing the range of trust for our model.

Replicated center points

In practice, this center point value frequently has two, or even three or more, replications. This not only provides a reference point for assessing the interpolative power of the model at the center, but it also allows us to compute model-free estimates of the natural error in the data. This in turn allows us a more rigorous method for computing the uncertainty for individual coefficients in the model and for rigorously carrying out a lack-of-fit test for assessing general model adequacy.

5.5.9.9.10. Motivation: How do we Use the Model for Interpolation?

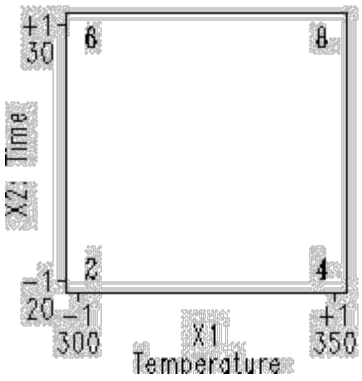
Design table in original data units

As for the mechanics of interpolation itself, consider a continuation of the prior $k = 2$ factor experiment. Suppose temperature T ranges from 300 to 350 and time t ranges from 20 to 30, and the analyst can afford $n = 4$ runs. A 2^2 full factorial design is run. Forming the coded temperature as X_1 and the coded time as X_2, we have the usual:

Temperature	Time	X_1	X_2	Y
300	20	-	-	2
350	20	+	-	4
300	30	-	+	6
350	30	+	+	8

Graphical representation

Graphically the design and data are as follows:



Typical interpolation question

As before, from the data, the prediction equation is

$$\hat{Y} = 5 + 2*X_2 + X_1$$

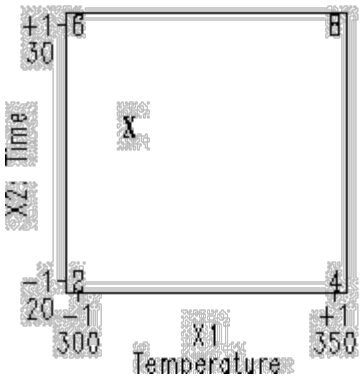
We now pose the following typical interpolation question:

> From the model, what is the predicted response at, say, temperature = 310 and time = 26?

In short:

$\hat{Y}(T = 310, t = 26) = ?$

To solve this problem, we first view the $k = 2$ design and data graphically, and note (via an "X") as to where the desired ($T = 310$, $t = 26$) interpolation point is:



Predicting the response for the interpolated point

The important next step is to convert the raw (in units of the original factors T and t) interpolation point into a coded (in units of X_1 and X_2) interpolation point. From the graph or otherwise, we note that a linear translation between T and X_1, and between t and X_2 yields

$T = 300 \Rightarrow X_1 = -1$
$T = 350 \Rightarrow X_1 = +1$

thus

$X_1 = 0$ is at $T = 325$

```
 |-------------|-------------|
-1     ?       0             +1
300   310     325           350
```

which in turn implies that

$T = 310 \Rightarrow X_1 = -0.6$

Similarly,

$t = 20 \Rightarrow X_2 = -1$
$t = 30 \Rightarrow X_2 = +1$

therefore,

$X_2 = 0$ is at $t = 25$

```
 |-------------|-------------|
-1             0   ?         +1
20             25 26         30
```

thus

$t = 26 \Rightarrow X_2 = +0.2$

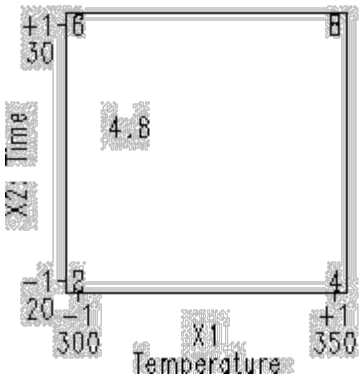
Substituting $X_1 = -0.6$ and $X_2 = +0.2$ into the prediction equation

$$\hat{Y} = 5 + 2*X_2 + X_1$$

yields a predicted value of 4.8.

Graphical representation of response value for interpolated data point

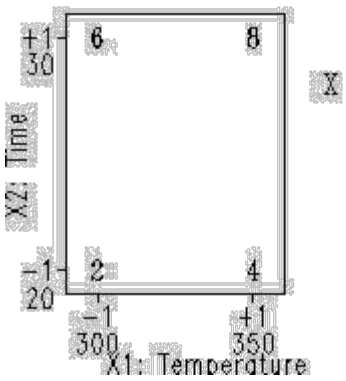
Thus



5.5.9.9.11. Motivation: How do we Use the Model for Extrapolation?

Graphical representation of extrapolation

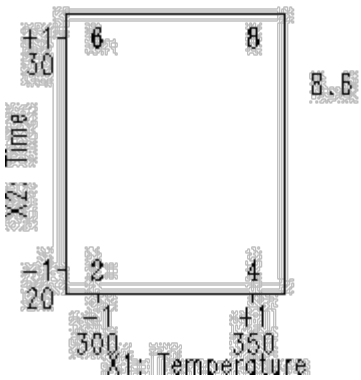
Extrapolation is performed similarly to interpolation. For example, the predicted value at temperature T = 375 and time t = 28 is indicated by the "X":



and is computed by substituting the values X_1 = +2.0 (T=375) and X_2 = +0.8 (t=28) into the prediction equation

$$\hat{Y} = 5 + 2*X_2 + X_1$$

yielding a predicted value of 8.6. Thus we have



Pseudo-data

The predicted value from the modeling effort may be viewed as pseudo-data, data obtained without the experimental effort. Such "free" data can add tremendously

to the insight via the application of graphical techniques (in particular, the contour plots and can add significant insight and understanding as to the nature of the response surface relating *Y* to the *X*'s.

But, again, a final word of caution: the "pseudo data" that results from the modeling process is exactly that, pseudo-data. It is **not** real data, and so the model and the model's predicted values must be validated by additional confirmatory (real) data points. A more balanced approach is that:

> Models may be trusted as "real" [that is, to generate predicted values and contour curves], but must always be verified [that is, by the addition of confirmatory data points].

The rule of thumb is thus to take advantage of the available and recommended model-building mechanics for these 2-level designs, but do treat the resulting derived model with an equal dose of both optimism and caution.

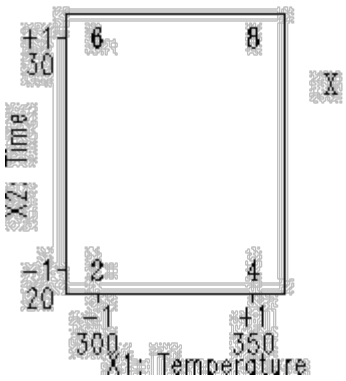
Summary

In summary, the motivation for model building is that it gives us insight into the nature of the response surface along with the ability to do interpolation and extrapolation; further, the motivation for the use of the cumulative residual standard deviation plot is that it serves as an easy-to-interpret tool for determining a good and parsimonious model.

5.5.9.9.12. Motivation: How do we Use the Model for Extrapolation?

Graphical representation of extrapolation

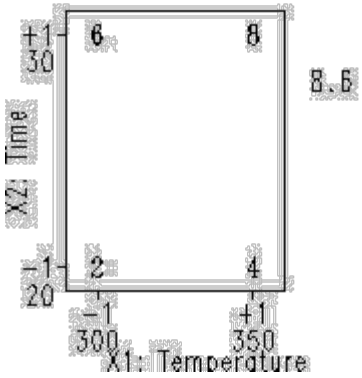
Extrapolation is performed similarly to interpolation. For example, the predicted value at temperature T = 375 and time t = 28 is indicated by the "X":



and is computed by substituting the values $X1$ = +2.0 (T=375) and $X2$ = +0.8 (t=28) into the prediction equation

$$\hat{Y} = 5 + 0.5 * (4 * X_2 + 2 * X_1)$$

yielding a predicted value of 8.6. Thus we have



Pseudo-data

The predicted value from the modeling effort may be viewed as pseudo-data, data obtained without the experimental effort. Such "free" data can add tremendously

to the insight via the application of graphical techniques (in particular, the contour plots and can add significant insight and understanding as to the nature of the response surface relating *Y* to the *X*'s.

But, again, a final word of caution: the "pseudo data" that results from the modeling process is exactly that, pseudo-data. It is **not** real data, and so the model and the model's predicted values must be validated by additional confirmatory (real) data points. A more balanced approach is that:

> Models may be trusted as "real" [that is, to generate predicted values and contour curves], but must always be verified [that is, by the addition of confirmatory data points].

The rule of thumb is thus to take advantage of the available and recommended model-building mechanics for these 2-level designs, but do treat the resulting derived model with an equal dose of both optimism and caution.

Summary

In summary, the motivation for model building is that it gives us insight into the nature of the response surface along with the ability to do interpolation and extrapolation; further, the motivation for the use of the cumulative residual standard deviation plot is that it serves as an easy-to-interpret tool for determining a good and parsimonious model.

5.5.9.10. DOE contour plot

Purpose The DOE contour plot answers the question:

> Where else could we have run the experiment to optimize the response?

Prior steps in this analysis have suggested the best setting for each of the *k* factors. These best settings may have been derived from

1. Data: which of the *n* design points yielded the best response, and what were the settings of that design point, or from
2. Averages: what setting of each factor yielded the best response "on the average".

This 10th (and last) step in the analysis sequence goes beyond the limitations of the *n* data points already chosen in the design and replaces the data-limited question

> "From among the *n* data points, what was the best setting?"

to a region-related question:

> "In general, what should the settings have been to optimize the response?"

Output The outputs from the DOE contour plot are

1. Primary: Best setting (X_{10}, X_{20}, ..., X_{k0}) for each of the *k* factors. This derived setting should yield an optimal response.
2. Secondary: Insight into the nature of the response surface and the importance/unimportance of interactions.

Definition A DOE contour plot is formed by

- Vertical Axis: The second most important factor in the experiment.
- Horizontal Axis: The most important factor in the experiment.

More specifically, the DOE contour plot is constructed and utilized via the following 7 steps:

1. Axes
2. Contour Curves
3. Optimal Response Value
4. Best Corner

5. Steepest Ascent/Descent
6. Optimal Curve
7. Optimal Setting

with

1. Axes: Choose the two most important factors in the experiment as the two axes on the plot.

2. Contour Curves: Based on the fitted model and the best data settings for all of the remaining factors, draw contour curves involving the two dominant factors. This yields a graphical representation of the response surface. The details for constructing linear contour curves are given in a later section.

3. Optimal Value: Identify the theoretical value of the response that constitutes "best." In particular, what value would we like to have seen for the response?

4. Best "Corner": The contour plot will have four "corners" for the two most important factors X_i and X_j: (X_i, X_j) = (-, -), (-, +), (+, -), and (+, +). From the data, identify which of these four corners yields the highest average response $\bar{Y}$.

5. Steepest Ascent/Descent: From this optimum corner point, and based on the nature of the contour lines near that corner, step out in the direction of steepest ascent (if maximizing) or steepest descent (if minimizing).

6. Optimal Curve: Identify the curve on the contour plot that corresponds to the ideal optimal value.

7. Optimal Setting: Determine where the steepest ascent/descent line intersects the optimum contour curve. This point represents our "best guess" as to where we could have run our experiment so as to obtain the desired optimal response.

Motivation

In addition to increasing insight, most experiments have a goal of optimizing the response. That is, of determining a setting $(X_{10}, X_{20}, ..., X_{k0})$ for which the response is optimized.

The tool of choice to address this goal is the DOE contour plot. For a pair of factors X_i and X_j, the DOE contour plot is a 2-dimensional representation of the 3-dimensional $Y = f(X_i, X_j)$ response surface. The position and spacing of the isocurves on the DOE contour plot are an easily interpreted reflection of the nature of the surface.

In terms of the construction of the DOE contour plot, there are three aspects of note:

1. Pairs of Factors: A DOE contour plot necessarily has two axes (only); hence only two out of the k factors can be represented on this plot. All other factors must be set at a fixed value (their optimum settings as determined by the ordered data plot, the DOE mean plot, and the interaction effects matrix plot).

2. Most Important Factor Pair: Many DOE contour plots are possible. For an

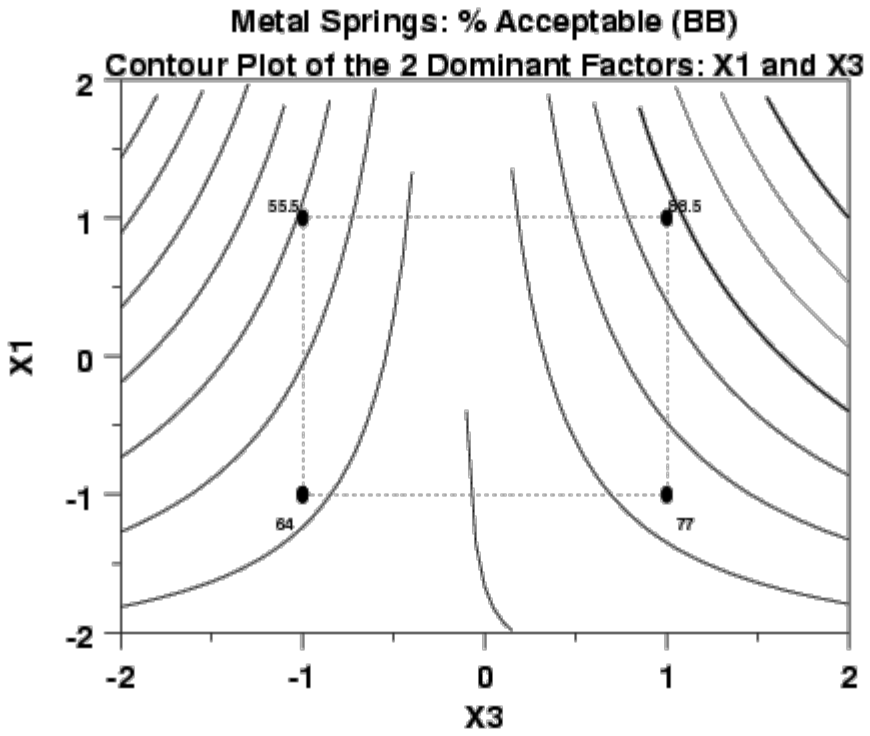
experiment with k factors, there are $\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$ possible contour plots. For example, for k = 4 factors there are 6 possible contour plots: X_1 and X_2, X_1 and X_3, X_1 and X_4, X_2 and X_3, X_2 and X_4, and X_3 and X_4. In practice, we usually generate only one contour plot involving the two most important factors.

3. Main Effects Only: The contour plot axes involve main effects only, not interactions. The rationale for this is that the "deliverable" for this step is k settings, a best setting for each of the k factors. These k factors are real and can be controlled, and so optimal settings can be used in production. Interactions are of a different nature as there is no "knob on the machine" by which an interaction may be set to -, or to +. Hence the candidates for the axes on contour plots are main effects only--no interactions.

In summary, the motivation for the DOE contour plot is that it is an easy-to-use graphic that provides insight as to the nature of the response surface, and provides a specific answer to the question "Where (else) should we have collected the data so to have optimized the response?".

Plot for defective springs data

Applying the DOE contour plot for the defective springs data set yields the following plot.



How to interpret

From the DOE contour plot for the defective springs data, we note the following regarding the 7 framework issues:

- Axes
- Contour curves

- Optimal response value
- Optimal response curve
- Best corner
- Steepest Ascent/Descent
- Optimal setting

Conclusions for the defective springs data

The application of the DOE contour plot to the defective springs data set results in the following conclusions:

1. Optimal settings for the "next" run:

   Coded : (X_1, X_2, X_3) = (+1.5, +1.0, +1.3)
   Uncoded: (OT, CC, QT) = (1637.5, 0.7, 127.5)

2. Nature of the response surface:

   The X_1*X_3 interaction is important, hence the effect of factor X_1 will change depending on the setting of factor X_3.

5.5.9.10.1. How to Interpret: Axes

What factors go on the two axes?

For this first item, we choose the two most important factors in the experiment as the plot axes.

These are determined from the ranked list of important factors as discussed in the previous steps. In particular, the |effects| plot includes a ranked factor table. For the defective springs data, that ranked list consists of

Factor/Interaction	Effect Estimate
X_1	23
X_1*X_3	10
X_2	-5
X_3	1.5
X_1*X_2	1.5
$X_1*X_2*X_3$	0.5
X_2*X_3	0

Possible choices

In general, the two axes of the contour plot could consist of

- X_1 and X_2,
- X_1 and X_3, or
- X_2 and X_3.

In this case, since X_1 is the top item in the ranked list, with an estimated effect of 23, X_1 is the most important factor and so will occupy the horizontal axis of the contour plot. The admissible list thus reduces to

- X_1 and X_2, or
- X_1 and X_3.

To decide between these two pairs, we look to the second item in the ranked list. This is the interaction term X_1*X_3, with an estimated effect of 10. Since interactions are **not** allowed as contour plot axes, X_1*X_3 must be set aside. On

the other hand, the components of this interaction (X_1 and X_3) are not to be set aside. Since X_1 has already been identified as one axis in the contour plot, this suggests that the other component (X_3) be used as the second axis. We do so. Note that X_3 itself does **not** need to be important (in fact, it is noted that X_3 is ranked fourth in the listed table with a value of 1.5).

In summary then, for this example the contour plot axes are:

Horizontal Axis: X_1
Vertical Axis: X_3

Four cases for recommended choice of axes

Other cases can be more complicated. In general, the recommended rule for selecting the two plot axes is that they be drawn from the first two items in the ranked list of factors. The following four cases cover most situations in practice:

- Case 1:
  1. Item 1 is a main effect (e.g., X_3)
  2. Item 2 is another main effect (e.g., X_5)

  Recommended choice:

  1. Horizontal axis: item 1 (e.g., X_3);
  2. Vertical axis: item 2 (e.g., X_5).

- Case 2:
  1. Item 1 is a main effect (e.g., X_3)
  2. Item 2 is a (common-element) interaction (e.g., X_3*X_4)

  Recommended choice:

  1. Horizontal axis: item 1 (e.g., X_3);
  2. Vertical axis: the remaining component in item 2 (e.g., X_4).

- Case 3:
  1. Item 1 is a main effect (e.g., X_3)
  2. Item 2 is a (non-common-element) interaction (e.g., X_2*X_4)

  Recommended choice:

  1. Horizontal axis: item 1 (e.g., X_3);
  2. Vertical axis: either component in item 2 (e.g., X_2, or X_4), but preferably the one with the

largest individual effect (thus scan the rest of the ranked factors and if the X_2 |effect| > X_4 |effect|, choose X_2; otherwise choose X_4).

- Case 4:
  1. Item 1 is a (2-factor) interaction (e.g., X_2*X_4)
  2. Item 2 is anything

  Recommended choice:

  1. Horizontal axis: component 1 from the item 1 interaction (e.g., X_2);
  2. Horizontal axis: component 2 from the item 1 interaction (e.g., X_4).

5. Process Improvement
5.5. Advanced topics
5.5.9. An EDA approach to experimental design
5.5.9.10. DOE contour plot

5.5.9.10.2. How to Interpret: Contour Curves

Non-linear appearance of contour curves implies strong interaction

Based on the fitted model (cumulative residual standard deviation plot) and the best data settings for all of the remaining factors, we draw contour curves involving the two dominant factors. This yields a graphical representation of the response surface.

Before delving into the details as to how the contour lines were generated, let us first note as to what insight can be gained regarding the general nature of the response surface. For the defective springs data, the dominant characteristic of the contour plot is the non-linear (fan-shaped, in this case) appearance. Such non-linearity implies a strong X_1*X_3 interaction effect. If the X_1*X_3 interaction were small, the contour plot would consist of a series of near-parallel lines. Such is decidedly not the case here.

Constructing the contour curves

As for the details of the construction of the contour plot, we draw on the model-fitting results that were achieved in the cumulative residual standard deviation plot. In that step, we derived the following good-fitting prediction equation:

$$\hat{Y} = 71.25 + 11.5*X_1 + 5*X_1*X_3 - 2.5*X_2$$

The contour plot has axes of X_1 and X_3. X_2 is not included and so a fixed value of X_2 must be assigned. The response variable is the percentage of acceptable springs, so we are attempting to maximize the response. From the ordered data plot, the main effects plot, and the interaction effects matrix plot of the general analysis sequence, we saw that the best setting for factor X_2 was "-". The best observed response data value (Y = 90) was achieved with the run (X_1, X_2, X_3) = (+, -, +), which has X_2 = "-". Also, the average response for X_2 = "-" was 73 while the average response for X_2 = "+" was 68. We thus set X_2 = -1 in the prediction equation to obtain

$$\hat{Y} = 71.25 + 11.5*X_1 + 5*X_1*X_3 - 2.5*(-1)$$
$$\hat{Y} = 73.75 + 11.5*X_1 + 5*X_1*X_3$$

This equation involves only X_1 and X_3 and is immediately usable for the X_1 and X_3 contour plot. The raw response values in the data ranged from 52 to 90. The response implies that the theoretical worst is $Y = 0$ and the theoretical best is $Y = 100$.

To generate the contour curve for, say, $Y = 70$, we solve

$$70 = 73.75 + 11.5*X_1 + 5*X_1*X_3$$

by rearranging the equation in X_3 (the vertical axis) as a function of X_1 (the horizontal axis). By substituting various values of X_1 into the rearranged equation, the above equation generates the desired response curve for $Y = 70$. We do so similarly for contour curves for any desired response value Y.

Values for X1

For these $X_3 = g(X_1)$ equations, what values should be used for X_1? Since X_1 is coded in the range -1 to +1, we recommend expanding the horizontal axis to -2 to +2 to allow extrapolation. In practice, for the DOE contour plot generated previously, we chose to generate X_1 values from -2, at increments of 0.05, up to +2. For most data sets, this gives a smooth enough curve for proper interpretation.

Values for Y

What values should be used for Y? Since the total theoretical range for the response Y (= percent acceptable springs) is 0 % to 100 %, we chose to generate contour curves starting with 0, at increments of 5, and ending with 100. We thus generated 21 contour curves. Many of these curves did not appear since they were beyond the -2 to +2 plot range for the X_1 and X_3 factors.

Summary

In summary, the contour plot curves are generated by making use of the (rearranged) previously derived prediction equation. For the defective springs data, the appearance of the contour plot implied a strong X_1*X_3 interaction.

5.5.9.10.3. How to Interpret: Optimal Response Value

Need to define "best"

We need to identify the theoretical value of the response that would constitute "best". What value would we like to have seen for the response?

For example, if the response variable in a chemical experiment is percent reacted, then the ideal theoretical optimum would be

100 %. If the response variable in a manufacturing experiment is amount of waste, then the ideal theoretical optimum would be zero. If the response variable in a flow experiment is the fuel flow rate in an engine, then the ideal theoretical optimum (as dictated by engine specifications) may be a specific value (e.g., 175 cc/sec). In any event, for the experiment at hand, select a number that represents the ideal response value.

Optimal value for this example

For the defective springs data, the response (percentage of acceptable springs) ranged from $Y = 52$ to 90. The theoretically worst value would be 0 (= no springs are acceptable), and the theoretically best value would be 100 (100 % of the springs are acceptable). Since we are trying to maximize the response, the selected optimal value is 100.

5.5.9.10.4. How to Interpret: Best Corner

Four corners representing 2 levels for 2 factors

The contour plot will have four "corners" (two factors times two settings per factor) for the two most important factors X_i and X_j: (X_i, X_j) = (-, -), (-, +), (+, -), or (+, +). Which of these four corners yields the highest average response $\bar{Y}$? That is, what is the "best corner"?

Use the raw data

This is done by using the raw data, extracting out the two "axes factors", computing the average response at each of the four corners, then choosing the corner with the best average.

For the defective springs data, the raw data were

X_1	X_2	X_3	Y
-	-	-	67
+	-	-	79
-	+	-	61
+	+	-	75
-	-	+	59
+	-	+	90
-	+	+	52
+	+	+	87

The two plot axes are X_1 and X_3 and so the relevant raw data collapses to

X_1	X_3	Y
-	-	67
+	-	79
-	-	61
+	-	75
-	+	59
+	+	90
-	+	52
+	+	87

Averages which yields averages

X_1	X_3	Y
-	-	(67 + 61)/2 = 64
+	-	(79 + 75)/2 = 77
-	+	(59 + 52)/2 = 55.5
+	+	(90 + 87)/2 = 88.5

These four average values for the corners are annotated on the plot. The best (highest) of these values is 88.5. This comes from the (+, +) upper right corner. We conclude that for the defective springs data the best corner is (+, +).

5.5.9.10.5. How to Interpret: Steepest Ascent/Descent

Start at optimum corner point

From the optimum corner point, based on the nature of the contour surface at that corner, step out in the direction of steepest ascent (if maximizing) or steepest descent (if minimizing).

Defective springs example

Since our goal for the defective springs problem is to maximize the response, we seek the path of steepest ascent. Our starting point is the best corner (the upper right corner (+, +)), which has an average response value of 88.5. The contour lines for this plot have increments of 5 units. As we move from left to right across the contour plot, the contour lines go from low to high response values. In the plot, we have drawn the maximum contour level, 105, as a thick line. For easier identification, we have also drawn the contour level of 90 as thick line. This contour level of 90 is immediately to the right of the best corner

Conclusions on steepest ascent for defective springs example

The nature of the contour curves in the vicinity of (+, +) suggests a path of steepest ascent

1. in the "northeast" direction
2. about 30 degrees above the horizontal.

5.5.9.10.6. How to Interpret: Optimal Curve

Corresponds to ideal optimum value	The optimal curve is the curve on the contour plot that corresponds to the ideal optimum value.
Defective springs example	For the defective springs data, we search for the Y = 100 contour curve. As determined in the steepest ascent/descent section, the Y = 90 curve is immediately outside the (+, +) point. The next curve to the right is the Y = 95 curve, and the next curve beyond that is the Y = 100 curve. This is the optimal response curve.

5.5.9.10.7. How to Interpret: Optimal Setting

Optimal setting

The "near-point" optimality setting is the intersection of the steepest-ascent line with the optimal setting curve.

Theoretically, any (X_1, X_3) setting along the optimal curve would generate the desired response of $Y = 100$. In practice, however, this is true only if our estimated contour surface is identical to "nature's" response surface. In reality, the plotted contour curves are truth estimates based on the available (and "noisy") $n = 8$ data values. We are confident of the contour curves in the vicinity of the data points (the four corner points on the chart), but as we move away from the corner points, our confidence in the contour curves decreases. Thus the point on the $Y = 100$ optimal response curve that is "most likely" to be valid is the one that is closest to a corner point. Our objective then is to locate that "near-point".

Defective springs example

In terms of the defective springs contour plot, we draw a line from the best corner, (+, +), outward and perpendicular to the $Y = 90$, $Y = 95$, and $Y = 100$ contour curves. The $Y = 100$ intersection yields the "nearest point" on the optimal response curve.

Having done so, it is of interest to note the coordinates of that optimal setting. In this case, from the graph, that setting is (in coded units) approximately at

$$(X_1 = 1.5, X_3 = 1.3)$$

Table of coded and uncoded factors

With the determination of this setting, we have thus, in theory, formally completed our original task. In practice, however, more needs to be done. We need to know "What is this optimal setting, not just in the coded units, but also in the original (uncoded) units"? That is, what does (X_1=1.5, X_3=1.3) correspond to in the units of the original data?

To deduce his, we need to refer back to the original (uncoded) factors in this problem. They were:

Coded Factor	Uncoded Factor
X_1	OT: Oven Temperature
X_2	CC: Carbon Concentration
X_3	QT: Quench Temperature

Uncoded and coded factor settings

These factors had settings-- what were the settings of the coded and uncoded factors? From the original description of the problem, the uncoded factor settings were:

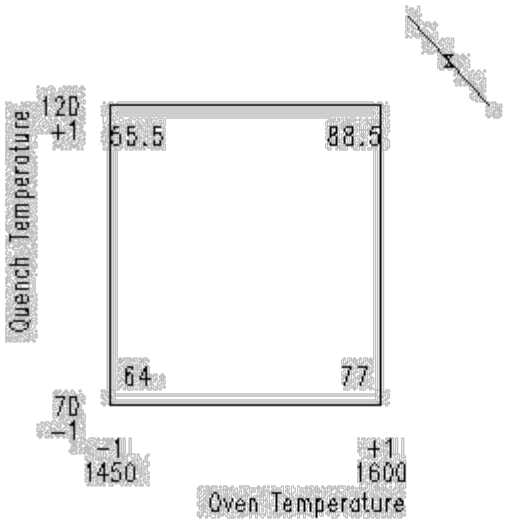
1. Oven Temperature (1450 and 1600 degrees)
2. Carbon Concentration (0.5 % and 0.7 %)
3. Quench Temperature (70 and 120 degrees)

with the usual settings for the corresponding coded factors:

1. X_1 (-1, +1)
2. X_2 (-1, +1)
3. X_3 (-1, +1)

Diagram

To determine the corresponding setting for (X_1=1.5, X_3=1.3), we thus refer to the following diagram, which mimics a scatter plot of response averages--oven temperature (OT) on the horizontal axis and quench temperature (QT) on the vertical axis:



The "X" on the chart represents the "near point" setting on the optimal curve.

Optimal setting for X_1 (oven

To determine what "X" is in uncoded units, we note (from the graph) that a linear transformation between OT and X_1 as defined by

temperature)

OT = 1450 => X_1 = -1
OT = 1600 => X_1 = +1

yields

X_1 = 0 being at OT = (1450 + 1600) / 2 = 1525

thus

```
            |-------------|-------------|
X1:        -1             0             +1
OT:       1450           1525           1600
```

and so X_1 = +2, say, would be at oven temperature OT = 1675:

```
            |-------------|-------------|---------
----|
X1:        -1             0             +1
+2
OT:       1450           1525           1600
1675
```

and hence the optimal X_1 setting of 1.5 must be at

OT = 1600 + 0.5*(1675-1600) = 1637.5

Optimal setting for X_3 (quench temperature)

Similarly, from the graph we note that a linear transformation between quench temperature QT and coded factor X_3 as specified by

QT = 70 => X_3 = -1
QT = 120 => X_3 = +1

yields

X_3 = 0 being at QT = (70 + 120) / 2 = 95

as in

```
          |-------------|-------------|
X3:      -1             0             +1
QT:      70            95            120
```

and so X_3 = +2, say, would be quench temperature = 145:

```
          |-------------|-------------|------------
-|
X3:      -1             0             +1
+2
QT:      70            95            120
145
```

Hence, the optimal X_3 setting of 1.3 must be at

QT = 120 + 0.3*(145-120)
QT = 127.5

Summary of optimal settings

In summary, the optimal setting is

coded : (X_1 = +1.5, X_3 = +1.3)
uncoded: (OT = 1637.5 degrees, QT = 127.5 degrees)

and finally, including the best setting of the fixed X_2 factor (carbon concentration CC) of X_2 = -1 (CC = 0.5 %), we thus have the final, complete recommended optimal settings for all three factors:

coded : (X_1 = +1.5, X_2 = -1.0, X_3 = +1.3)
uncoded: (OT = 1637.5, CC = 0.7 %, QT = 127.5)

If we were to run another experiment, this is the point (based on the data) that we would set oven temperature, carbon concentration, and quench temperature with the hope/goal of achieving 100 % acceptable springs.

Options for next step

In practice, we could either

1. collect a single data point (if money and time are an issue) at this recommended setting and see how close to 100 % we achieve, or

2. collect two, or preferably three, (if money and time are less of an issue) replicates at the center point (recommended setting).

3. if money and time are not an issue, run a 2^2 full factorial design with center point. The design is centered on the optimal setting (X_1 = +1, 5, X_3 = +1.3) with one overlapping new corner point at (X_1 = +1, X_3 = +1) and with new corner points at (X_1, X_3) = (+1, +1), (+2, +1), (+1, +1.6), (+2, +1.6). Of these four new corner points, the point (+1, +1) has the advantage that it overlaps with a corner point of the original design.

5. Process Improvement

5.6. Case Studies

Contents The purpose of this section is to illustrate the analysis of designed experiments with data collected from experiments run at the National Institute of Standards and Technology and SEMATECH.

1. Eddy current probe sensitivity study
2. Sonoluminescent light intensity study

5.6.1. Eddy Current Probe Sensitivity Case Study

Analysis of a 2^3 Full Factorial Design

This case study demonstrates the analysis of a 2^3 full factorial design.

The analysis for this case study is based on the EDA approach discussed in an earlier section.

Contents

The case study is divided into the following sections:

1. Background and data
2. Initial plots/main effects
3. Interaction effects
4. Main and interaction effects: block plots
5. Estimate main and interaction effects
6. Modeling and prediction equations
7. Intermediate conclusions
8. Important factors and parsimonious prediction
9. Validate the fitted model
10. Using the model
11. Conclusions and next step
12. Work this example yourself

5.6.1.1. Background and Data

Background

The data for this case study is a subset of a study performed by Capobianco, Splett, and Iyer. Capobianco was a member of the NIST Electromagnetics Division and Splett and Iyer were members of the NIST Statistical Engineering Division at the time of this study.

The goal of this project is to develop a nondestructive portable device for detecting cracks and fractures in metals. A primary application would be the detection of defects in airplane wings. The internal mechanism of the detector would be for sensing crack-induced changes in the detector's electromagnetic field, which would in turn result in changes in the impedance level of the detector. This change of impedance is termed "sensitivity" and it is a sub-goal of this experiment to maximize such sensitivity as the detector is moved from an unflawed region to a flawed region on the metal.

Statistical Goals

The case study illustrates the analysis of a 2^3 full factorial experimental design. The specific statistical goals of the experiment are:

1. Determine the important factors that affect sensitivity.
2. Determine the settings that maximize sensitivity.
3. Determine a predicition equation that functionally relates sensitivity to various factors.

Software

The analyses used in this case study can be generated using both Dataplot code and R code.

Data Used in the Analysis

There were three detector wiring component factors under consideration:

1. $X1$ = Number of wire turns
2. $X2$ = Wire winding distance
3. $X3$ = Wire gauge

Since the maximum number of runs that could be afforded timewise and costwise in this experiment was $n = 10$, a 2^3 full factoral experiment (involving $n = 8$ runs) was chosen.

With an eye to the usual monotonicity assumption for two-level factorial designs, the selected settings for the three factors were as follows:

1. $X1$ = Number of wire turns : -1 = 90, +1 = 180
2. $X2$ = Wire winding distance: -1 = 0.38, +1 = 1.14
3. $X3$ = Wire gauge : -1 = 40, +1 = 48

The experiment was run with the eight settings executed in random order. The following data resulted.

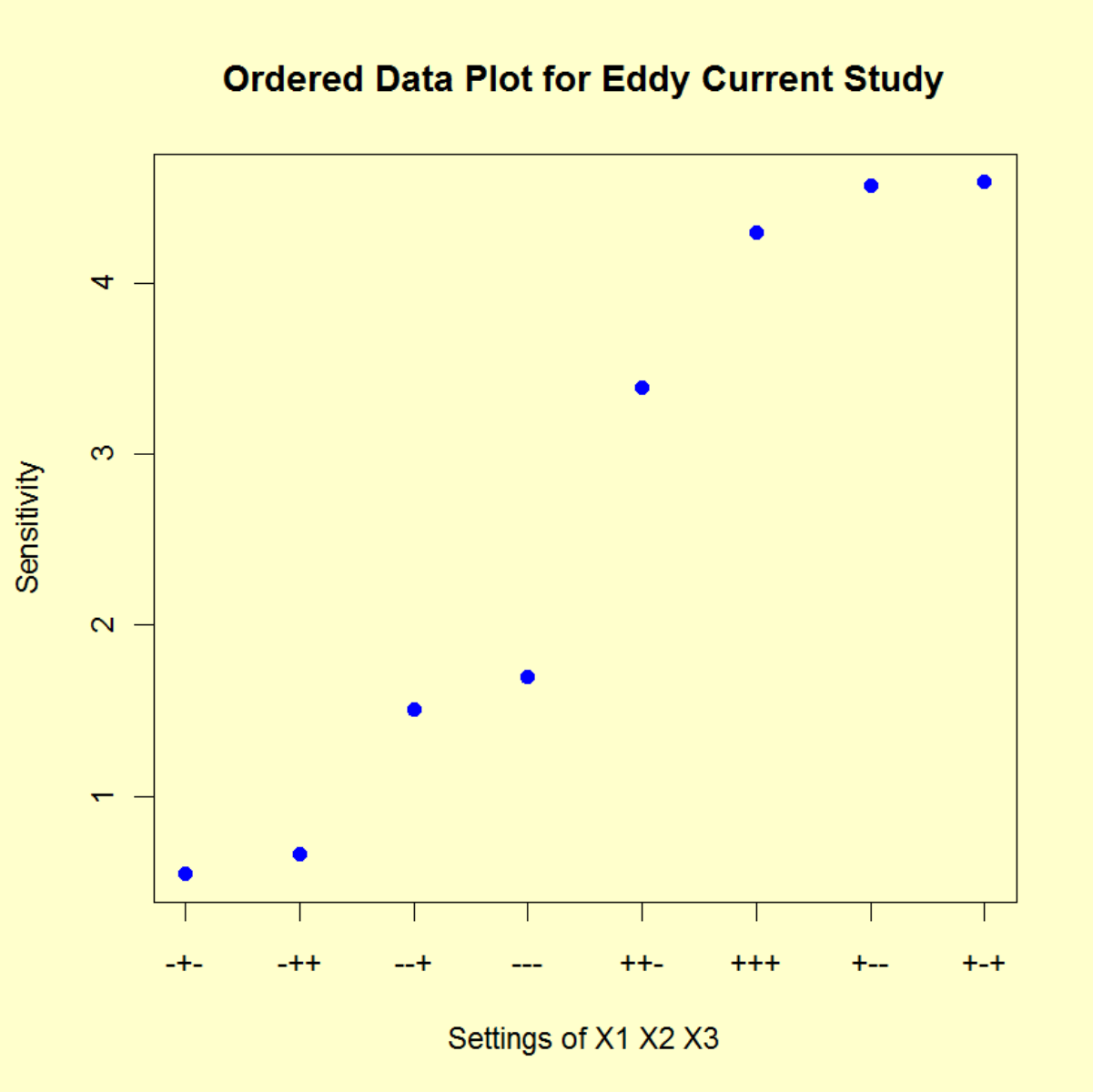
Y Probe Impedance	X1 Number of Turns	X2 Winding Distance	X3 Wire Gauge	Run Sequence
1.70	-1	-1	-1	2
4.57	+1	-1	-1	8
0.55	-1	+1	-1	3
3.39	+1	+1	-1	6
1.51	-1	-1	+1	7
4.59	+1	-1	+1	1
0.67	-1	+1	+1	4
4.29	+1	+1	+1	5

Note that the independent variables are coded as +1 and -1. These represent the low and high settings for the levels of each variable. Factorial designs often have two levels for each factor (independent variable) with the levels being coded as -1 and +1. This is a scaling of the data that can simplify the analysis. If desired, these scaled values can be converted back to the original units of the data for presentation.

5.6.1.2. Initial Plots/Main Effects

Plot the Data: Ordered Data Plot

The first step in the analysis is to generate an ordered data plot.



Conclusions from the Ordered Data Plot

We can make the following conclusions based on the ordered data plot.

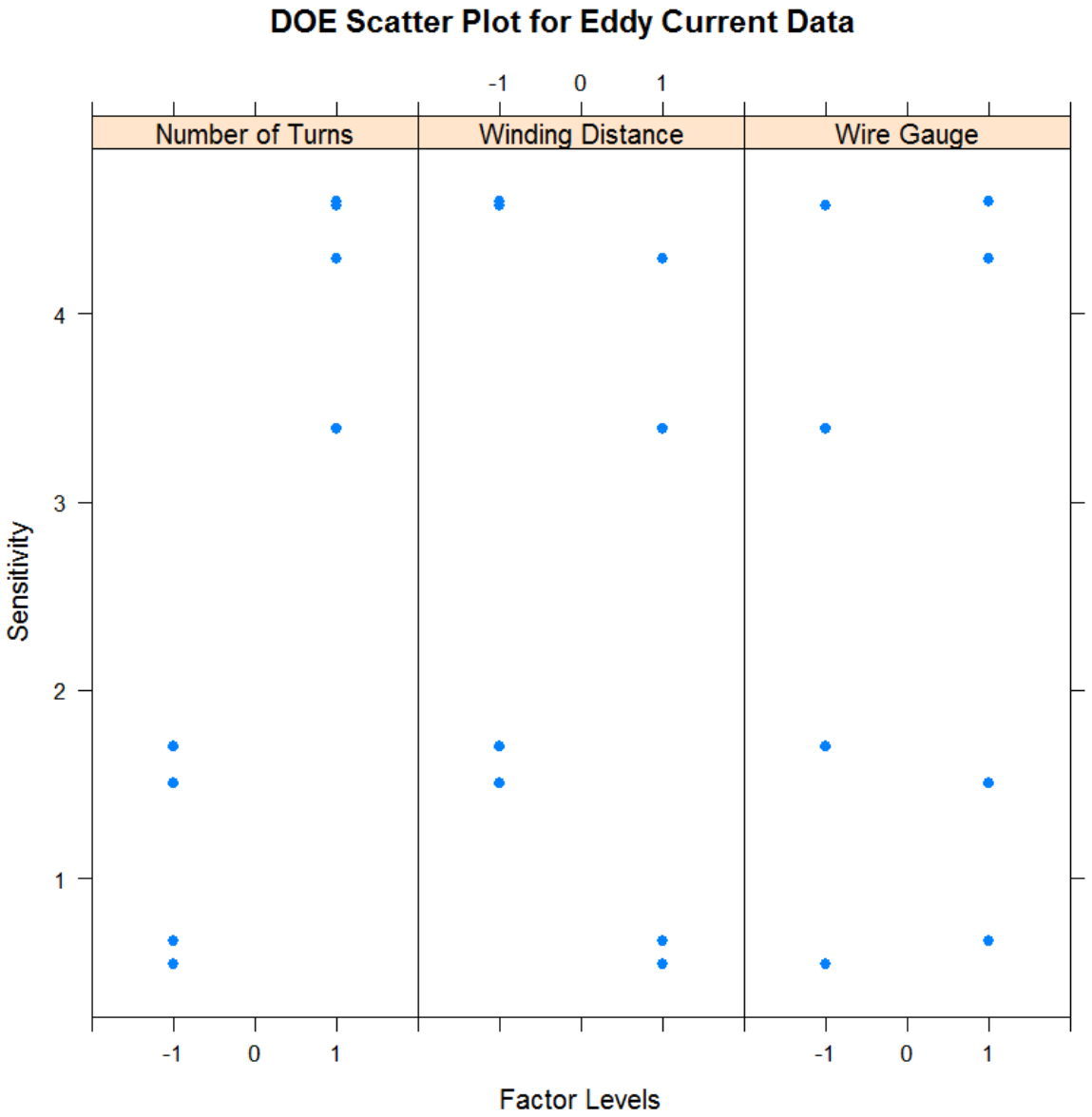
1. Important Factors: The four highest response values have $X1$ = + while the four lowest response values have $X1$ = -. This implies $X1$ is the most important factor. When $X1$ = -, the - values of $X2$ are higher than the + values of $X2$. Similarly, when

$X1$ = +, the - values of $X2$ are higher than the + values of $X2$. This implies $X2$ is important, but less so than $X1$. There is no clear pattern for $X3$.

2. Best Settings: In this experiment, we are using the device as a detector, and so high sensitivities are desirable. Given this, our first pass at best settings yields ($X1$ = +1, $X2$ = -1, $X3$ = either).

Plot the Data: DOE Scatter Plot

The next step in the analysis is to generate a DOE scatter plot.



Conclusions from the DOE Scatter Plot

We can make the following conclusions based on the DOE scatter plot.

1. Important Factors: $X1$ (Number of Turns) is clearly important. When $X1$ = -1, all four senstivities are low, and when $X1$ = +1, all four sensitivities are high. $X2$ (Winding Distance) is less important. The four sensitivities for $X2$ = -1 are slightly higher, as a group, than the four sensitivities for $X2$ = +1. $X3$ (Wire Gauge) does not appear to be important at all. The sensitivity is about the same (on the average) regardless of the settings for $X3$.

2. Best Settings: In this experiment, we are using the device as a detector, so high

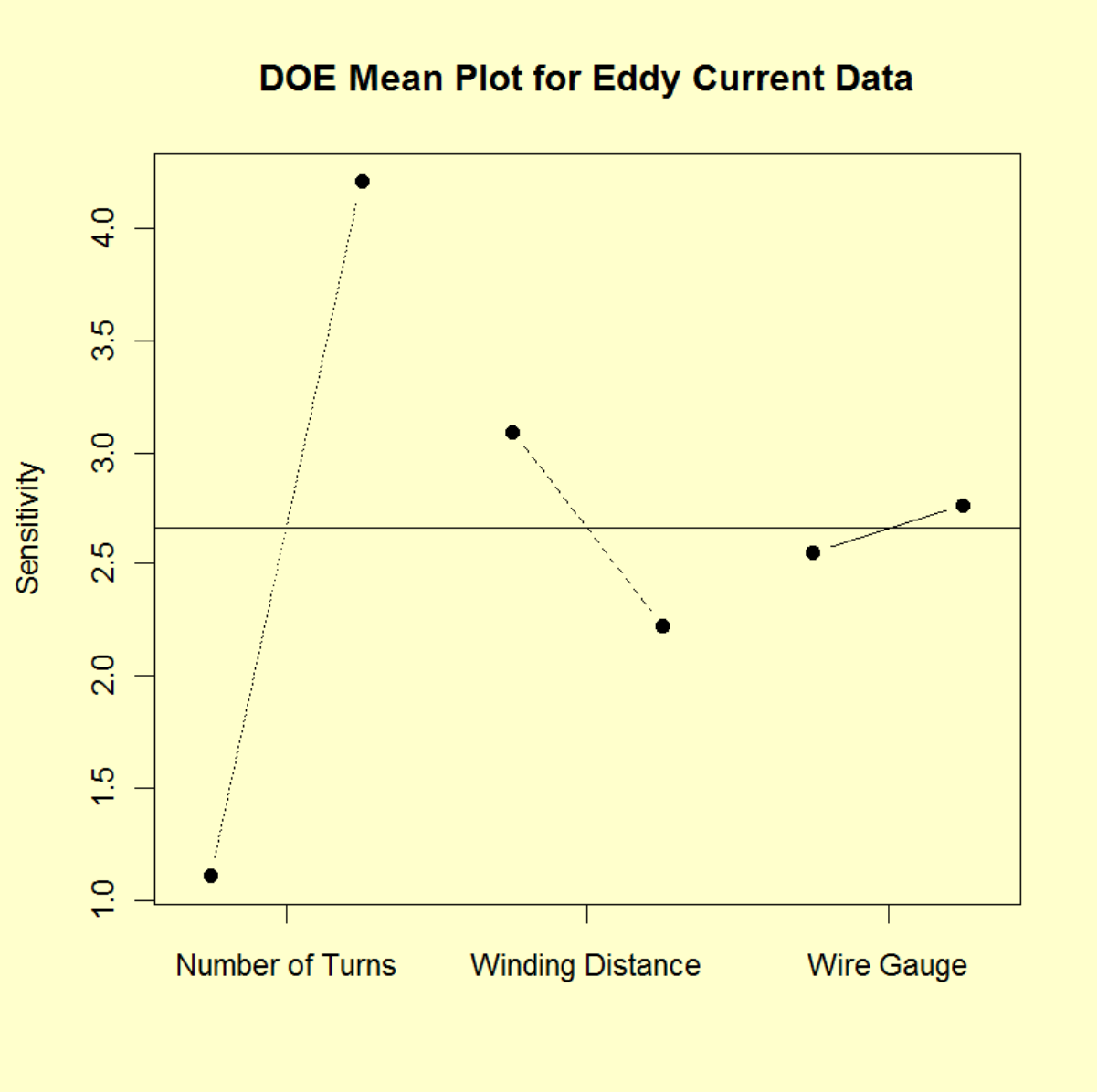
sensitivities are desirable. Given this, our first pass at best settings yields ($X1$ = +1, $X2$ = -1, $X3$ = either).

3. There does not appear to be any significant outliers.

Check for Main Effects: DOE Mean Plot

One of the primary questions is: what are the most important factors? The ordered data plot and the DOE scatter plot provide useful summary plots of the data. Both of these plots indicated that $X1$ is clearly important, $X2$ is somewhat important, and $X3$ is probably not important.

The DOE mean plot shows the main effects. This provides probably the easiest to interpret indication of the important factors.



Conclusions from the DOE Mean Plot

The DOE mean plot (or main effects plot) reaffirms the ordering of the DOE scatter plot, but additional information is gleaned because the eyeball distance between the mean values gives an approximation to the least-squares estimate of the factor effects.

We can make the following conclusions from the DOE mean plot.

1. Important Factors:
   $X1$ (effect = large: about 3 ohms)
   $X2$ (effect = moderate: about -1 ohm)
   $X3$ (effect = small: about 1/4 ohm)

2. Best Settings: As before, choose the factor settings that (on the average) maximize the sensitivity:

   $(X1,X2,X3) = (+,-,+)$

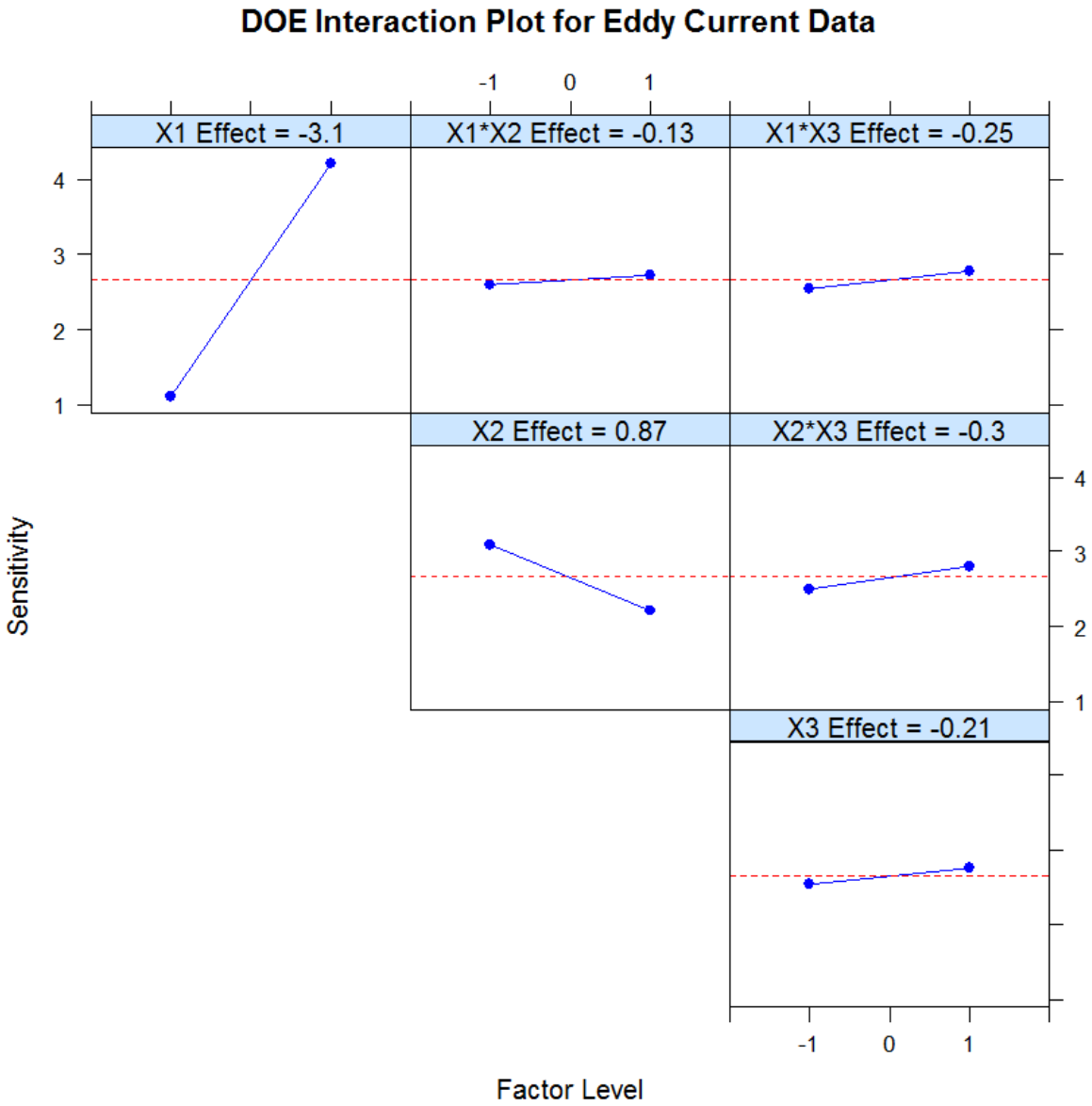
Comparison of Plots

All of these plots are used primarily to detect the most important factors. Because it plots a summary statistic rather than the raw data, the DOE mean plot shows the main effects most clearly. However, it is still recommended to generate either the ordered data plot or the DOE scatter plot (or both). Since these plot the raw data, they can sometimes reveal features of the data that might be masked by the DOE mean plot.

5.6.1.3. Interaction Effects

Check for Interaction Effects: DOE Interaction Plot

In addition to the main effects, it is also important to check for interaction effects, especially two-factor interaction effects. The DOE interaction effects plot is an effective tool for this. The effects on the plot represent the change in sensitivity from low to high levels of the factors.



Conclusions from the

We can make the following conclusions from the DOE interaction effects plot.

DOE Interaction Effects Plot

1. Important Factors: Looking for the plots that have the steepest lines (that is, largest effects), we note that:
    - $X1$ (number of turns) is the most important effect: estimated effect = -3.1025;
    - $X2$ (winding distance) is next most important: estimated effect = -0.8675;
    - $X3$ (wire gauge) is relatively unimportant;
    - All three two-factor interactions are relatively unimporant.
2. Best Settings: As with the main effects plot, the best settings to maximize the sensitivity are

    $(X1,X2,X3) = (+1,-1,+1)$

    but with the $X3$ setting of +1 mattering little.

5.6.1.4. Main and Interaction Effects: Block Plots

Block Plots

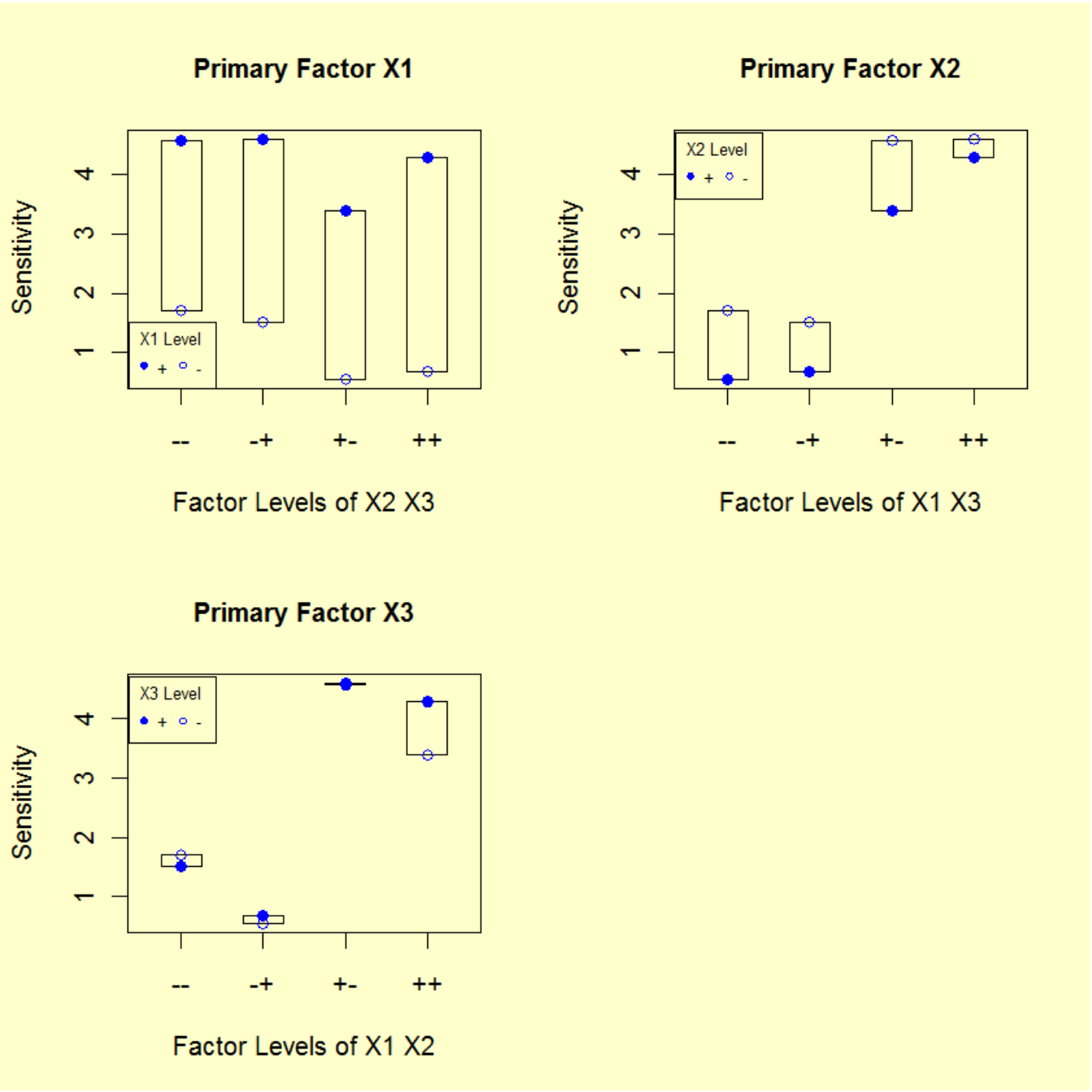
Block plots are a useful adjunct to the DOE mean plot and the DOE interaction effects plot to confirm the importance of factors, to establish the robustness of main effect conclusions, and to determine the existence of interactions. Specifically,

1. The first plot below answers the question: Is $X1$ important? If $X1$ is important, is this importance robust over all four settings of $X2$ and $X3$?

2. The second plot below answers the question: Is $X2$ important? If $X2$ is important, is this importance robust over all four settings of $X1$ and $X3$?

3. The third plot below answers the question: Is $X3$ important? If $X3$ is important, is this importance robust over all four settings of $X1$ and $X2$?

For block plots, it is the height of the bars that is important, not the relative positioning of each bar. Hence we focus on the size and internals of the blocks, not "where" the blocks are one relative to another.



Conclusions from the Block Plots

Recall that the block plot will access factor importance by the degree of consistency (robustness) of the factor effect over a variety of conditions. In this light, we can make the following conclusions from the block plots.

1. Relative Importance of Factors: All of the bar heights in plot 1 (turns) are greater than the bar heights in plots 2 and 3. Hence, $X1$ is more important than $X2$ and $X3$.

2. Statistical Significance: In plot 1, looking at the levels within each bar, we note that the response for level 2 is higher than level 1 in each of the four bars. By chance, this happens with probability $1/(2^4) = 1/16 = 6.25$ %. Hence, $X1$ is near-statistically significant at the 5 % level. Similarly, for plot 2, level 1 is greater than level 2 for all four bars. Hence, $X2$ is near-statistically significant. For $X3$, there is no consistent ordering within all four bars, and hence $X3$ is not statistically significant. Rigorously speaking then, $X1$ and $X2$ are not statistically significant (since 6.25 % is not < 5 %); on the other hand such near-significance is suggestive to the analyst that such factors may in fact be important, and hence warrant further attention.

   Note that the usual method for determining statistical significance is to perform an analysis of variance (ANOVA). ANOVA is based on normality assumptions. If these normality assumptions are valid, then ANOVA methods are the most powerful method for determining statistical signficance. The advantage of the block-plot method is that it is based on less rigorous assumptions than ANOVA. At an exploratory stage, it is useful to know that our conclusions regarding important factors are valid under a wide range of assumptions.

3. Interactions: For $X1$, the four bars do not change height in any systematic way and hence there is no evidence of $X1$ interacting with either $X2$ or $X3$. Similarly, there is no evidence of interactions for $X2$.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

5.6.1.5. Estimate Main and Interaction Effects

Effects Estimation

Although the effect estimates were given on the DOE interaction plot on a previous page, we also display them in tabular form.

The full model for the 2^3 factorial design is

$$Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{123} X_1 X_2 X_3 + \epsilon$$

Data from factorial designs with two levels can be analyzed using least-squares regression. The regresson coefficients represent the change per one unit of the factor variable, the effects shown on the interaction plot represent changes between high and low factor levels so they are twice as large as the regression coefficients.

Effect Estimates

The parameter estimates from a least-squares regression analysis for the full model are shown below.

Effect	Estimate
Mean	2.65875
X1	1.55125
X2	-0.43375
X3	0.10625
X1*X2	0.06375
X1*X3	0.12375
X2*X3	0.14875
X1*X2*X3	0.07125

Because we fit the full model to the data, there are no degrees of freedom for error and no significance tests are available.

If we sort the effects from largest to smallest (excluding the mean), the four most important factors are: X1 (number of turns), X2 (winding distance), X2*X3 (winding distance by wire gauge interaction), and X1*X3 (number of turns by wire gauge interaction).

5.6.1.6. Modeling and Prediction Equations

Parameter Estimates Don't Change as Additional Terms Added

In most cases of least-squares fitting, the model coefficient estimates for previously added terms change depending on what was successively added. For example, the estimate for the $X1$ coefficient might change depending on whether or not an $X2$ term was included in the model. This is **not** the case when the design is orthogonal, as is this 2^3 full factorial design. In such a case, the estimates for the previously included terms do not change as additional terms are added. This means the list of effect estimates in section 5.6.1.5 serves as the least-squares coefficient estimates for progressively more complicated models.

Default Model: Grand Mean

If none of the factors are important, the prediction equation defaults to the mean of all the response values (the overall or grand mean). That is,

$$\hat{Y} = 2.65875$$

For our example, the default model has a grand mean of 2.65875 with a residual standard deviation (a measure of goodness of fit) of 1.74106 ohms.

Possible Prediction Equations

We add effects to the default model in decreasing order of absolute magnitude and compute the residual standard deviation after adding each effect. The prediction equations and their residual standard deviations are shown below.

```
Residual
Model Terms
Std. Dev.
-------------------------------------------------
--      ---------
Mean + X1
0.57272
Mean + X1 + X2
0.30429
Mean + X1 + X2 + X2*X3
0.26737
Mean + X1 + X2 + X2*X3 + X1*X3
0.23341
Mean + X1 + X2 + X2*X3 + X1*X3 + X3
0.19121
Mean + X1 + X2 + X2*X3 + X1*X3 + X3 + X1*X2*X3
0.18031
Mean + X1 + X2 + X2*X3 + X1*X3 + X3 + X1*X2*X3 +
X1*X2             NA
```

Note that the full model is a perfect fit to the data.

5.6.1.7. Intermediate Conclusions

Important Factors

Taking stock from all of the graphical and quantitative analyses of the previous sections, we conclude that $X1$ (number of turns) is the most important engineering factor affecting sensitivity, followed by $X2$ (wire distance) as next in importance, followed then by some less important interactions and $X3$ (wire gauge).

Best Settings

Also, from the various analyses, we conclude that the best design settings (on the average) for a high-sensitivity detector are

$$(X1,X2,X3) = (+,-,+)$$

that is

number of turns = 180,
winding distance = 0.38, and
wire gauge = 48.

Can We Extract More From the Data?

Thus, in a very real sense, the analysis is complete. We have achieved the two most important stated goals of the experiment:

1. gaining insight into the most important factors, and
2. ascertaining the optimal production settings.

On the other hand, more information can be squeezed from the data, and that is what this section and the remaining sections address.

1. First of all, we focus on the problem of taking the ranked list of factors and objectively ascertaining which factors are "important" versus "unimportant".
2. In a parallel fashion, we use the subset of important factors derived above to form a "final" prediction equation that is good (that is, having a sufficiently small residual standard deviation) while being parsimonious (having a small number of terms), compared to the full model, which is perfect (having a residual standard

deviation = 0, that is, the predicted values = the raw data), but is unduly complicated (consisting of a constant + 7 terms).

5.6.1.8. Important Factors and Parsimonious Prediction

Identify Important Factors

The two problems discussed in the previous section (important factors and a parsimonious model) will be handled in parallel since determination of one yields the other. In regard to the "important factors", our immediate goal is to take the full subset of seven main effects and interactions and extract a subset that we will declare as "important", with the complementary subset being "unimportant". Seven criteria are discussed in detail in section 1.3.5.18.2 in Chapter 1. The relevant criteria will be applied here. These criteria are not all equally important, nor will they yield identical subsets, in which case a consensus subset or a weighted consensus subset must be extracted.

Criteria for Including Terms in the Model

The criteria that we can use in determining whether to keep a factor in the model can be summarized as follows.

1. Effects: Engineering Significance
2. Effects: 90 % Numerical Significance
3. Effects: Statistical Significance
4. Effects: Normal Probability Plot
5. Averages: Youden Plot

The first four criteria focus on effect estimates with three numerical criteria and one graphical criterion. The fifth criterion focuses on averages. We discuss each of these criteria in detail in the following sections.

The last section summarizes the conclusions based on all of the criteria.

Effects: Engineering Significance

The minimum engineering significant difference is defined as

$$|\hat{\beta}_i| > \Delta$$

where $|\hat{\beta}_i|$ is the absolute value of the parameter estimate (i.e., the effect) and Δ is the minimum engineering significant difference. That is, declare a factor as "important" if the effect is greater than some a priori declared engineering difference. We use a rough rule-of-thumb of keeping only those factors whose effect is greater than 10 % of the current production average. In this case, let's say that the average detector has a sensitivity of 1.25 ohms. This suggests that we would declare all factors whose effect is greater than 10 % of 1.25 ohms = 0.125 ohms to be significant from an engineering point of view.

Based on this minimum engineering-significant-difference criterion, we conclude to keep two terms: *X*1 (1.55125) and *X*2 (-0.43375).

Effects: 90 % Numerical

The 90 % numerical significance criterion is defined as

$$|\hat{\beta}_i| > (max|\hat{\beta}_i|)/10$$

Significance

That is, declare a factor as important if it exceeds 10 % of the largest effect. For the current case study, the largest effect is from *X*1 (1.55125 ohms), and so 10 % of that is 0.155 ohms. This suggests keeping all factors whose effects exceed 0.155 ohms.

Based on the 90 % numerical criterion, we would keep two terms: *X*1 (1.55125) and *X*2 (-0.43375). The *X*2**X*3 term, (0.14875), is just under the cutoff.

Effects: Statistical Significance

Statistical significance is defined as

$$|\hat{\beta}_i| > 2 \text{ s.e.}(\hat{\beta}_i)$$

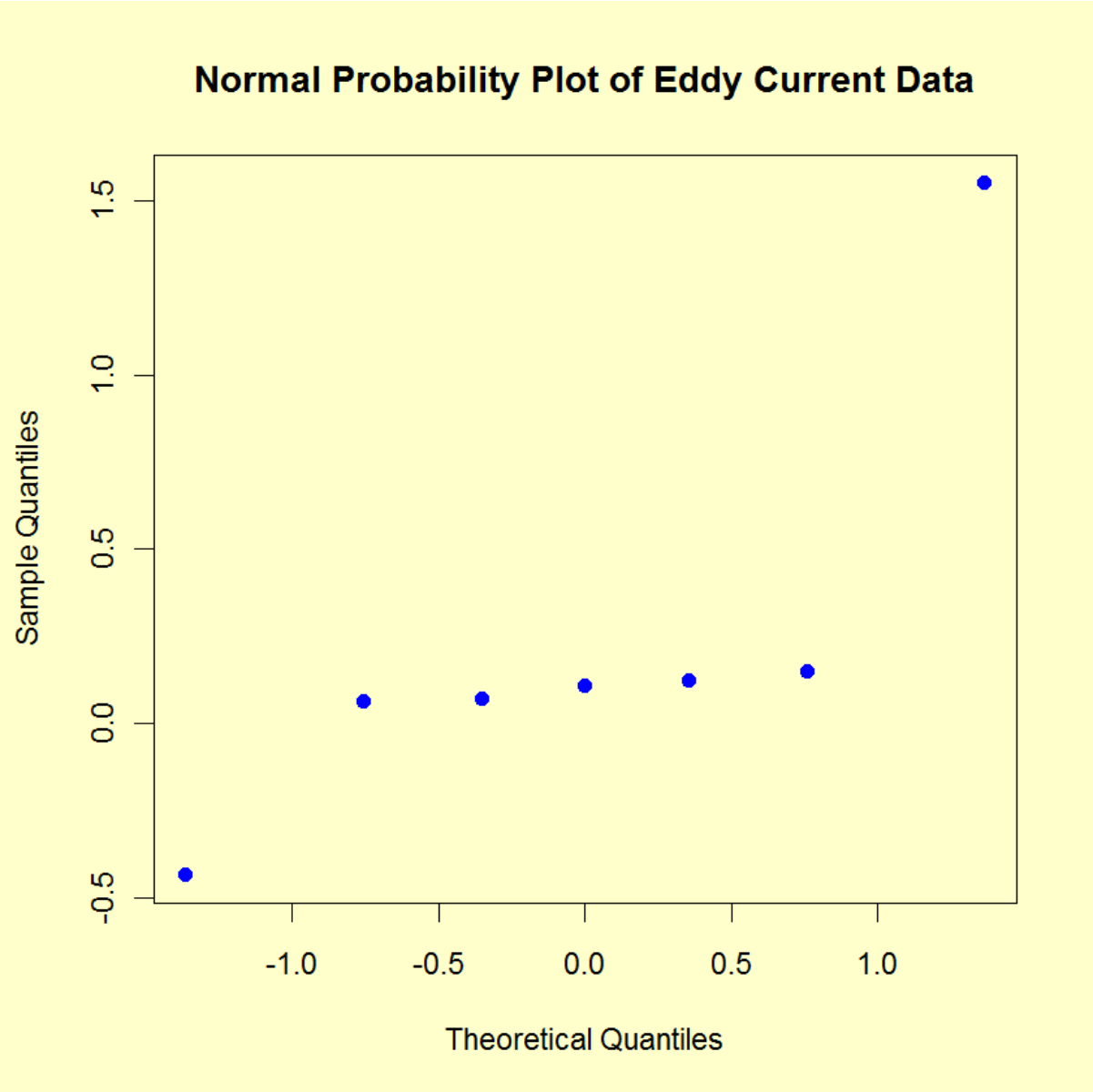
That is, declare a factor as "important" if its effect is more than 2 standard deviations away from 0 (0, by definition, meaning "no effect"). The difficulty with this is that in order to invoke this rule we need the σ is the standard deviation of an observation.

For the eddy current case study, ignoring three-factor and higher interactions leads to an estimate of σ based on omitting only a single term: the *X*1**X*2**X*3 interaction.

Thus for our example, if one assumes that the three-factor interaction is nil and hence represents a single drawing from a population centered at zero, an estimate of the standard deviation of an effect is simply the estimate of the interaction effect (0.07125). Two such effect standard deviations is 0.1425. This rule becomes to keep all $|\hat{\beta}_i|$ > 0.1425. This results in keeping three terms: *X*1 (1.55125), *X*2 (-0.43375), and *X*1**X*2 (0.14875).

Effects: Probability Plot

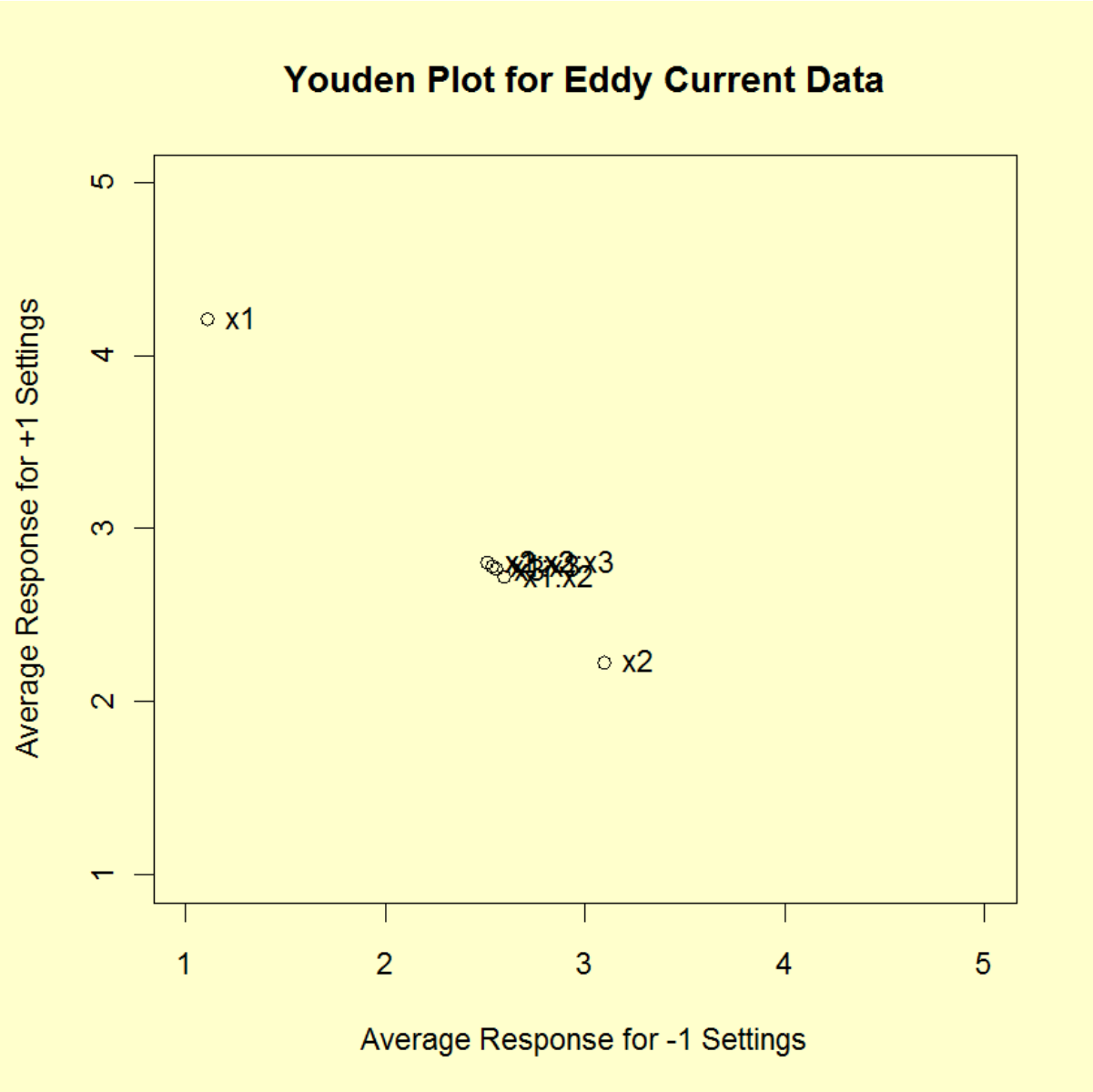
The normal probability plot can be used to identify important factors. The following graph shows the normal probability plot of the effects.



The normal probablity plot clearly shows two factors displaced off the line, and we see that those two factors are $X1$ and $X2$. Thus, we would keep $X1$ (1.55125) and $X2$ (-0.43375).

Effects: Youden Plot

A DOE Youden plot can be used in the following way. A factor is "important" if it is displaced away from the central-tendency bunch in a Youden plot of high and low averages.



For our example, the Youden plot clearly shows a cluster of points near the grand average (2.65875) with two displaced points above (*X*1) and below (*X*2). Based on the Youden plot, we keep two factors: *X*1 (1.55125) and *X*2 (-0.43375).

Conclusions In summary, the criterion for specifying "important" factors yielded the following:

1. Effects, Engineering Significant: *X*1 *X*2
2. Effects, Numerically Significant: *X*1 *X*2 (*X*2**X*3 is borderline)
3. Effects, Statistically Significant: *X*1 *X*2 *X*2**X*3
4. Effects, Normal Probability Plot: *X*1 *X*2
5. Averages, Youden Plot: *X*1 *X*2

All the criteria select *X*1 and *X*2. One also includes the *X*2**X*3 interaction term (and it is borderline for another criteria).

We thus declare the following consensus:

1. Important Factors: $X1$ and $X2$
2. Parsimonious Prediction Equation:

$$\hat{Y} = 2.65875 + 1.55125 \cdot X1 - 0.43375 \cdot X2$$

(with a residual standard deviation of 0.30429 ohms)

5.6.1.9. Validate the Fitted Model

Model Validation

In the Important Factors and Parsimonious Prediction section, we selected the following model

$$\hat{Y} = 2.65875 + 1.55125 \cdot X_1 - 0.43375 \cdot X_2$$

The residual standard deviation for this model is 0.30429.

The next step is to validate the model. The primary method of model validation is graphical residual analysis; that is, through an assortment of plots of the differences between the observed data Y and the predicted value $\hat{Y}$ from the model. For example, the design point (-1, -1, -1) has an observed data point (from the Background and data section) of $Y = 1.70$, while the predicted value from the above fitted model for this design point is

$$\hat{Y} = 2.65875 + 1.55125(-1) - 0.43375(-1) = 1.54125$$

which leads to the residual 0.15875.

Table of Residuals

If the model fits well, $\hat{Y}$ should be near Y for all eight design points. Hence the eight residuals should all be near zero. The eight predicted values and residuals for the model with these data are:

X1	X2	X3	Observed	Predicted	Residual
-1	-1	-1	1.70	1.54125	0.15875
+1	-1	-1	4.57	4.64375	-0.07375
-1	+1	-1	0.55	0.67375	-0.12375
+1	+1	-1	3.39	3.77625	-0.38625
-1	-1	+1	1.51	1.54125	-0.03125
+1	-1	+1	4.59	4.64375	-0.05375
-1	+1	+1	0.67	0.67375	-0.00375
+1	+1	+1	4.29	3.77625	0.51375

Residual Standard Deviation

What is the magnitude of the typical residual? There are several ways to compute this, but the statistically optimal measure is the residual standard deviation:

$$s_{res} = \sqrt{\frac{\sum_{i=1}^{N} r_i^2}{N - P}}$$

with r_i denoting the ith residual, $N = 8$ is the number of observations, and $P = 3$ is the number of fitted parameters. From the table of prediction equations, the residual standard deviation is 0.30429.

How Should Residuals Behave?

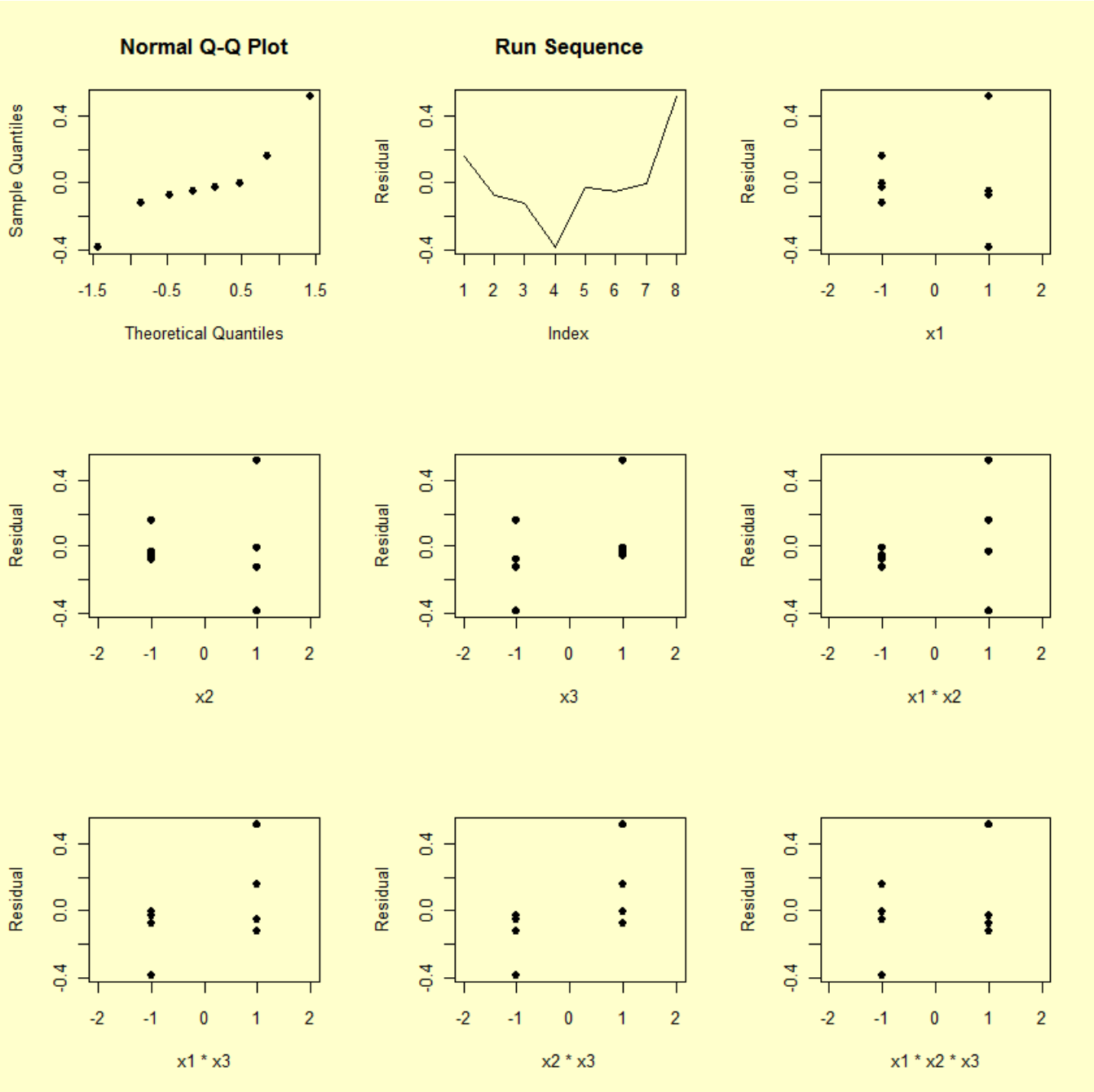
If the prediction equation is adequate, the residuals from that equation should behave like random drawings (typically from an approximately normal distribution), and should, since presumably random, have no structural relationship with any factor. This includes any and all potential terms ($X1$, $X2$, $X3$, $X1*X2$, $X1*X3$, $X2*X3$, $X1*X2*X3$).

Further, if the model is adequate and complete, the residuals should have no structural relationship with **any** other variables that may have been recorded. In particular, this includes the run sequence (time), which is really serving as a surrogate for any physical or environmental variable correlated with time. Ideally, all such residual scatter plots should appear structureless. Any scatter plot that exhibits structure suggests that the factor should have been formally included as part of the prediction equation.

Validating the prediction equation thus means that we do a final check as to whether any other variables may have been inadvertently left out of the prediction equation, including variables drifting with time.

The graphical residual analysis thus consists of scatter plots of the residuals versus all three factors and four interactions (all such plots should be structureless), a scatter plot of the residuals versus run sequence (which also should be structureless), and a normal probability plot of the residuals (which should be near linear). We present such plots below.

Residual Plots



The first plot is a normal probability plot of the residuals. The second plot is a run sequence plot of the residuals. The remaining plots show the residuals plotted against each of the factors and each of the interaction terms.

Conclusions We make the following conclusions based on the above plots.

1. Main Effects and Interactions: The *X*1 and *X*2 scatter plots are "flat" (as they must be since *X*1 and *X*2 were explicitly included in the model). The *X*3 plot shows some structure as does the *X*1**X*3, the *X*2**X*3, and the *X*1**X*2**X*3 plots. The *X*1**X*2 plot shows little structure. The net effect is that the relative ordering of these scatter plots is very much in agreement (again, as it must be) with the relative ordering of the "unimportant" factors. From the table of effects and the *X*2**X*3 residual plot, the third most influential term to be added to the model would be *X*2**X*3. In effect, these plots offer a higher-resolution confirmation of the ordering of effects. On the other hand, none of these other factors "passed" the criteria given in the previous section, and so these factors, suggestively influential as they might be, are still not influential enough to be added to the model.

2. Time Drift: The run sequence scatter plot is random. Hence there does not appear to be a drift either from time, or from any factor (e.g., temperature, humidity, pressure, etc.) possibly correlated with time.

3. Normality: The normal probability plot of the eight residuals has some trend, which suggests that additional terms might be added. On the other hand, the correlation coefficient of the 8 ordered residuals and the eight theoretical normal N(0,1) order statistic medians (which define the two axes of the plot) has the value 0.934, which is well within acceptable (5 %)

limits of the normal probability plot correlation coefficient test for normality. Thus, the plot is not so non-linear as to reject normality.

In summary, therefore, we accept the fitted model

$$\hat{Y} = 2.65875 + 1.55125 \cdot X_1 - 0.43375 \cdot X_2$$

as a parsimonious, but good, representation of the sensitivity phenomenon under study.

5.6.1.10. Using the Fitted Model

Model Provides Additional Insight

Although deriving the fitted model was not the primary purpose of the study, it does have two benefits in terms of additional insight:

1. Global prediction
2. Global determination of best settings

Global Prediction

How does one predict the response at points other than those used in the experiment? The prediction equation yields good results at the eight combinations of coded -1 and +1 values for the three factors:

1. $X1$ = Number of turns = 90 and 180
2. $X2$ = Winding distance = 0.38 and 1.14
3. $X3$ = Wire gauge = 40 and 48

What, however, would one expect the detector to yield at target settings of, say,

1. Number of turns = 150
2. Winding distance = 0.50
3. Wire gauge = 46

Based on the fitted equation, we first translate the target values into coded target values as follows:

coded target = -1 + 2*(target-low)/(high-low)

Hence the coded target values are

1. $X1$ = -1 + 2*(150-90)/(180-90) = 0.333333
2. $X2$ = -1 + 2*(0.50-0.38)/(1.14-0.38) = -0.684211
3. $X3$ = -1 + 2*(46-40)/(48-40) = 0.5000

Thus the raw data

(Number of turns, Winding distance, Wire gauge) = (150, 0.50, 46)

translates into the coded

($X1$, $X2$, $X3$) = (0.333333, -0.684211, 0.50000)

on the -1 to +1 scale.

Inserting these coded values into the fitted equation yields, as desired, a predicted value of

$$\hat{y} = 2.65875 + 1.55125(0.333333) - 0.43375(-0.684211) = 3.47261$$

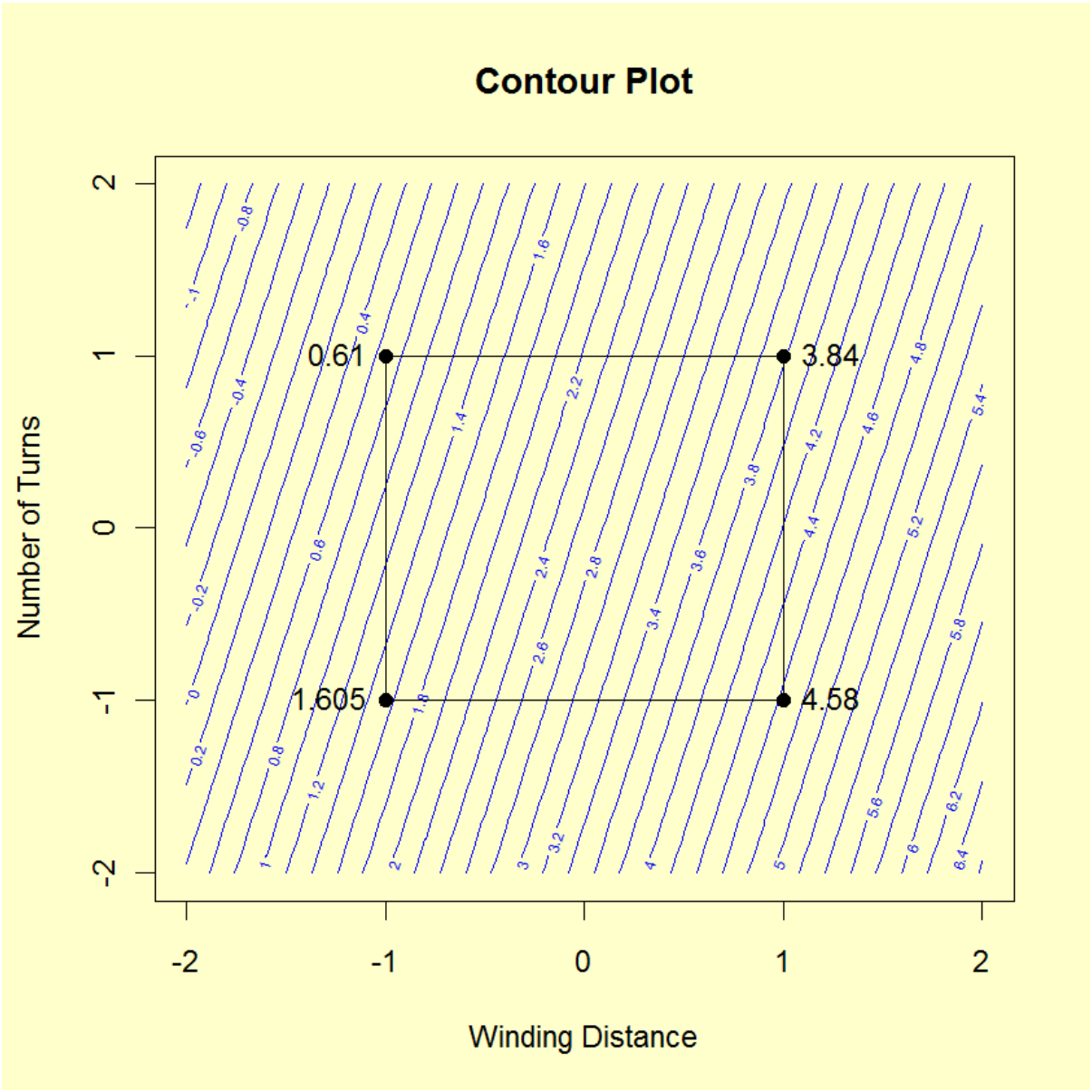
The above procedure can be carried out for any values of turns, distance, and gauge. This is subject to the usual cautions that equations that are good near the data point vertices may not necessarily be good everywhere in the factor space. Interpolation is a bit safer than extrapolation, but it is not guaranteed to provide good results, of course. One would feel more comfortable about interpolation (as in our example) if additional data had been collected at the center point and the center point data turned out to be in good agreement with predicted values at the center point based on the fitted model. In our case, we had no such data and so the sobering truth is that the user of the equation is assuming something in which the data set as given is not capable of suggesting one way or the other. Given that assumption, we have demonstrated how one may cautiously but insightfully generate predicted values that go well beyond our limited original data set of eight points.

Global Determination of Best Settings

In order to determine the best settings for the factors, we can use a DOE contour plot. The DOE contour plot is generated for the two most significant factors and shows the value of the response variable at the vertices (i.e, the -1 and +1 settings for the factor variables) and indicates the direction that maximizes (or minimizes) the response variable. If you have more than two significant factors, you can generate a series of DOE contour plots with each one using two of the important factors.

DOE Contour Plot

The following is the DOE contour plot of the number of turns and the winding distance.



The maximum value of the response variable (eddy current) corresponds to $X1$ (number of turns) equal to -1 and $X2$ (winding distance) equal to +1. The lower right corner of the contour plot corresponds to the direction that maximizes the response variable. This information can be used in planning the next phase of the experiment.

5.6.1.11. Conclusions and Next Step

Conclusions

The goals of this case study were:

1. Determine the most important factors.
2. Determine the best settings for the factors.
3. Determine a good prediction equation for the data.

The various plots and analysis showed that the number of turns ($X1$) and the winding distance ($X2$) were the most important factors and a good prediction equation for the data is:

$$\hat{Y} = 2.65875 + 1.55125 \cdot X1 - 0.43375 \cdot X2$$

The DOE contour plot gave us the best settings for the factors ($X1$ = -1 and $X2$ = 1).

Next Step

Full and fractional designs are typically used to identify the most important factors. In some applications, this is sufficient and no further experimentation is performed. In other applications, it is desired to maximize (or minimize) the response variable. This typically involves the use of response surface designs. The DOE contour plot can provide guidance on the settings to use for the factor variables in this next phase of the experiment.

This is a common sequence for designed experiments in engineering and scientific applications. Note the iterative nature of this approach. That is, you typically do not design one large experiment to answer all your questions. Rather, you run a series of smaller experiments. The initial experiment or experiments are used to identify the important factors. Once these factors are identified, follow-up experiments can be run to fine tune the optimal settings (in terms of maximizing/minimizing the response variable) for these most important factors.

For this particular case study, a response surface design was not used.

5.6.1.12. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot. It is required that you have already downloaded and installed Dataplot and configured your browser to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the Data Sheet window. Across the top of the main windows are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Get set up and started. 1. Read in the data.	1. You have read 4 columns of numbers into Dataplot: variables Y, X1, X2, and X3.
2. Plot the main effects. 1. Ordered data plot. 2. DOE scatter plot. 3. DOE mean plot.	1. Ordered data plot shows factor 1 clearly important, factor 2 somewhat important. 2. DOE scatter plot shows significant differences for factors 1 and 2.

	3. DOE mean plot shows significant differences in means for factors 1 and 2.
3. Plots for interaction effects 1. Generate a DOE interaction effects matrix plot.	1. The DOE interaction effects matrix plot does not show any major interaction effects.
4. Block plots for main and interaction effects 1. Generate block plots.	1. The block plots show that the factor 1 and factor 2 effects are consistent over all combinations of the other factors.
5. Estimate main and interaction effects 1. Perform a Yates fit to estimate the main effects and interaction effects.	1. The Yates analysis shows that the factor 1 and factor 2 main effects are significant, and the interaction for factors 2 and 3 is at the boundary of statistical significance.
6. Model selection 1. Generate half-normal probability plots of the effects. 2. Generate a Youden plot of the effects.	1. The probability plot indicates that the model should include main effects for factors 1 and 2. 2. The Youden plot indicates that the model should include main effects for factors 1 and 2.
7. Model validation 1. Compute residuals and predicted values from the partial model suggested by the Yates analysis.	1. Check the link for the values of the residual and predicted values.

2. Generate residual plots to validate the model.	2. The residual plots do not indicate any major problems with the model using main effects for factors 1 and 2.
8. DOE contour plot 1. Generate a DOE contour plot using factors 1 and 2.	1. The DOE contour plot shows X1 = -1 and X2 = +1 to be the best settings.

5.6.2. Sonoluminescent Light Intensity Case Study

Analysis of a 2^{7-3} Fractional Factorial Design

This case study demonstrates the analysis of a 2^{7-3} fractional factorial design. The purpose of the study is to optimize sonoluminescent light intensity.

The case study is based on the EDA approach to experimental design discussed in an earlier section.

Contents

The case study is divided into the following sections:

1. Background and data
2. Initial plots/main effects
3. Interaction effects
4. Main and interaction effects: block plots
5. Important Factors: Youden plot
6. Important Factors: |effects| plot
7. Important Factors: half-normal probability plot
8. Cumulative Residual SD plot
9. Next step: DOE contour plot
10. Summary of conclusions
11. Work this example yourself

5.6.2.1. Background and Data

Background and Motivation

Sonoluminescence is the process of turning sound energy into light. An ultrasonic horn is used to resonate a bubble of air in a medium, usually water. The bubble is ultrasonically compressed and then collapses to light-emitting plasma.

In the general physics community, sonoluminescence studies are being carried out to characterize it, to understand it, and to uncover its practical uses. An unanswered question in the community is whether sonoluminescence may be used for cold fusion.

NIST's motive for sonoluminescent investigations is to assess its suitability for the dissolution of physical samples, which is needed in the production of homogeneous Standard Reference Materials (SRMs). It is believed that maximal dissolution coincides with maximal energy and maximal light intensity. The ultimate motivation for striving for maximal dissolution is that this allows improved determination of alpha-and beta-emitting radionuclides in such samples.

The objectives of the NIST experiment were to determine the important factors that affect sonoluminescent light intensity and to ascertain optimal settings of such factors that will predictably achieve high intensities. An original list of 49 factors was reduced, based on physics reasons, to the following seven factors: molarity (amount of solute), solute type, pH, gas type in the water, water depth, horn depth, and flask clamping.

Time restrictions caused the experiment to be about one month, which in turn translated into an upper limit of roughly 20 runs. A 7-factor, 2-level fractional factorial design (Resolution IV) was constructed and run. The factor level settings are given below.

Eva Wilcox and Ken Inn of the NIST Physics Laboratory conducted this experiment during 1999. Jim Filliben of the NIST Statistical Engineering Division performed the analysis of the experimental data.

Software

The analyses used in this case study can be generated using both Dataplot code and R code.

Response Variable, Factor Variables, and Factor-Level Settings

This experiment utilizes the following response and factor variables.

1. Response Variable (Y) = The sonoluminescent light intensity.
2. Factor 1 ($X1$) = Molarity (amount of Solute). The coding is -1 for 0.10 mol and +1 for 0.33 mol.
3. Factor 2 ($X2$) = Solute type. The coding is -1 for sugar and +1 for glycerol.
4. Factor 3 ($X3$) = pH. The coding is -1 for 3 and +1 for 11.
5. Factor 4 ($X4$) = Gas type in water. The coding is -1 for helium and +1 for air.
6. Factor 5 ($X5$) = Water depth. The coding is -1 for half and +1 for full.
7. Factor 6 ($X6$) = Horn depth. The coding is -1 for 5 mm and +1 for 10 mm.
8. Factor 7 ($X7$) = Flask clamping. The coding is -1 for unclamped and +1 for clamped.

This data set has 16 observations. It is a 2^{7-3} design with no center points.

Goal of the Experiment

This case study demonstrates the analysis of a 2^{7-3} fractional factorial experimental design. The goals of this case study are:

1. Determine the important factors that affect the sonoluminescent light intensity. Specifically, we are trying to maximize this intensity.
2. Determine the best settings of the seven factors so as to maximize the sonoluminescent light intensity.

Data Used in the Analysis

The following are the data used for this analysis. This data set is given in Yates order.

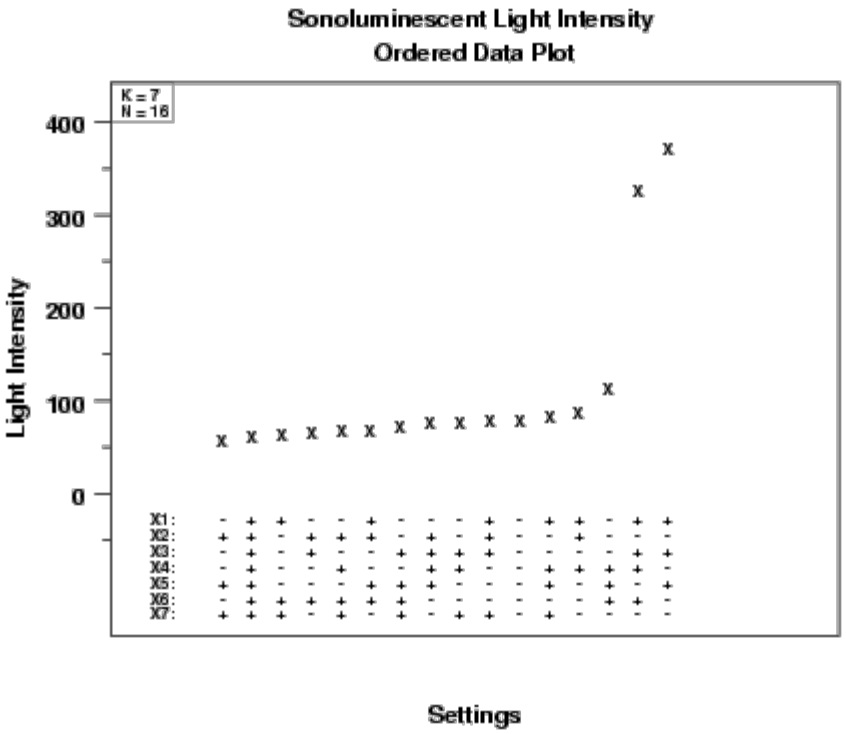
Y	X1	X2	X3	X4	X5	X6	X7
Light Intensity	Molarity	Solute type	pH	Gas Type	Water Depth	Horn Depth	Flask Clamping
80.6	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
66.1	1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0

```
 59.1        -1.0      1.0     -1.0     -1.0      1.0
-1.0      1.0
 68.9         1.0      1.0     -1.0     -1.0      1.0
1.0     -1.0
 75.1        -1.0     -1.0      1.0     -1.0      1.0
1.0      1.0
373.8         1.0     -1.0      1.0     -1.0      1.0
-1.0     -1.0
 66.8        -1.0      1.0      1.0     -1.0     -1.0
1.0     -1.0
 79.6         1.0      1.0      1.0     -1.0     -1.0
-1.0      1.0
114.3        -1.0     -1.0     -1.0      1.0      1.0
1.0     -1.0
 84.1         1.0     -1.0     -1.0      1.0      1.0
-1.0      1.0
 68.4        -1.0      1.0     -1.0      1.0     -1.0
1.0      1.0
 88.1         1.0      1.0     -1.0      1.0     -1.0
-1.0     -1.0
 78.1        -1.0     -1.0      1.0      1.0     -1.0
-1.0      1.0
327.2         1.0     -1.0      1.0      1.0     -1.0
1.0     -1.0
 77.6        -1.0      1.0      1.0      1.0      1.0
-1.0     -1.0
 61.9         1.0      1.0      1.0      1.0      1.0
1.0      1.0
```

5.6.2.2. Initial Plots/Main Effects

Plot the Data: Ordered Data Plot

The first step in the analysis is to generate an ordered data plot.



Conclusions from the Ordered Data Plot

We can make the following conclusions based on the ordered data plot.

1. Two points clearly stand out. The first 13 points lie in the 50 to 100 range, the next point is greater than 100, and the last two points are greater than 300.

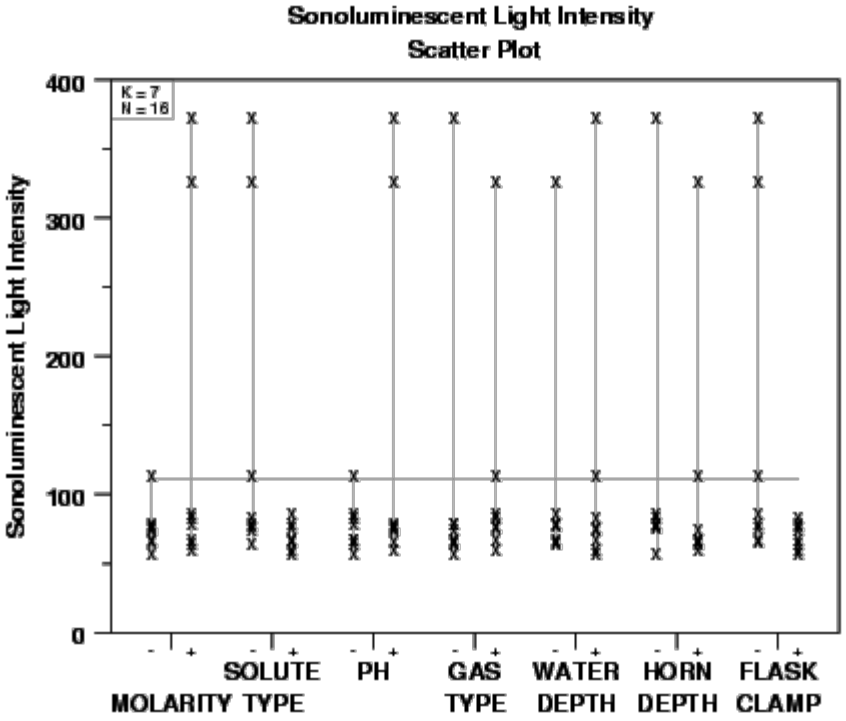
2. Important Factors: For these two highest points, factors *X*1, *X*2, *X*3, and *X*7 have the same value (namely, +, -, +, -, respectively) while *X*4, *X*5, and *X*6 have differing values. We conclude that *X*1, *X*2, *X*3, and *X*7 are potentially important factors, while *X*4, *X*5, and *X*6 are not.

3. Best Settings: Our first pass makes use of the settings at the observed maximum ($Y = 373.8$). The settings for this maximum are (+, -, +, -, +, -, -).

Plot the

The next step in the analysis is to generate a DOE scatter plot.

Data: DOE Scatter Plot



Conclusions from the DOE Scatter Plot

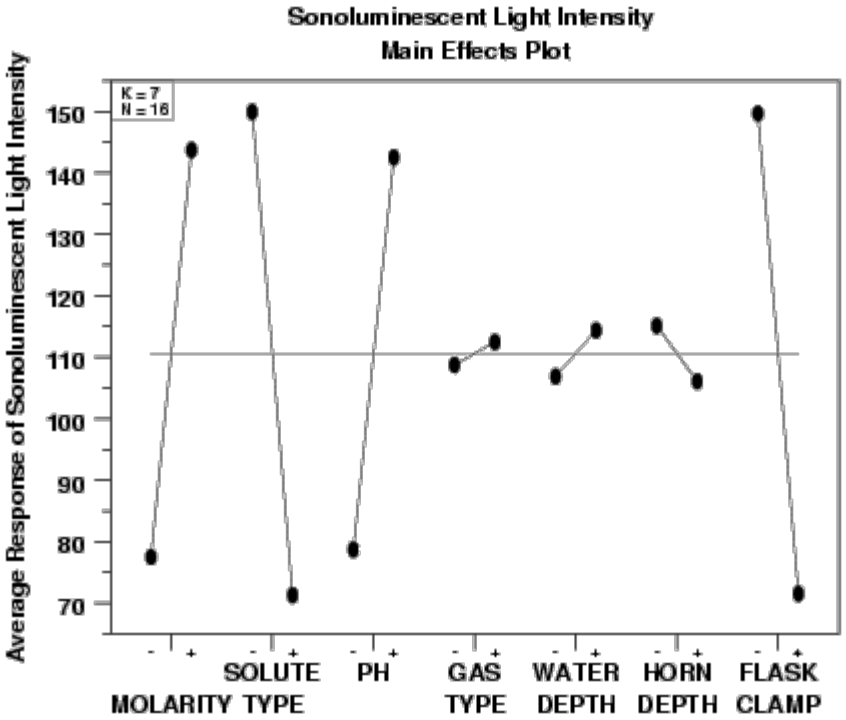
We can make the following conclusions based on the DOE scatter plot.

1. Important Factors: Again, two points dominate the plot. For $X1$, $X2$, $X3$, and $X7$, these two points emanate from the same setting, (+, -, +, -), while for $X4$, $X5$, and $X6$ they emanate from different settings. We conclude that $X1$, $X2$, $X3$, and $X7$ are potentially important, while $X4$, $X5$, and $X6$ are probably not important.

2. Best Settings: Our first pass at best settings yields ($X1$ = +, $X2$ = -, $X3$ = +, $X4$ = either, $X5$ = either, $X6$ = either, $X7$ = -).

Check for Main Effects: DOE Mean Plot

The DOE mean plot is generated to more clearly show the main effects:



Conclusions from the DOE Mean Plot

We can make the following conclusions from the DOE mean plot.

1. Important Factors:
   *X*2 (effect = large: about -80)
   *X*7 (effect = large: about -80)
   *X*1 (effect = large: about 70)
   *X*3 (effect = large: about 65)
   *X*6 (effect = small: about -10)
   *X*5 (effect = small: between 5 and 10)
   *X*4 (effect = small: less than 5)

2. Best Settings: Here we step through each factor, one by one, and choose the setting that yields the highest average for the sonoluminescent light intensity:

   $$(X1,X2,X3,X4,X5,X6,X7) = (+,-,+,+,+,-,-)$$

Comparison of Plots

All of the above three plots are used primarily to determine the most important factors. Because it plots a summary statistic rather than the raw data, the DOE mean plot shows the ordering of the main effects most clearly. However, it is still recommended to generate either the ordered data plot or the DOE scatter plot (or both). Since these plot the raw data, they can sometimes reveal features of the data that might be masked by the DOE mean plot.

In this case, the ordered data plot and the DOE scatter plot clearly show two dominant points. This feature would not be obvious if we had generated only the DOE mean plot.

Interpretation-wise, the most important factor *X*2 (solute) will, on the average,

change the light intensity by about 80 units regardless of the settings of the other factors. The other factors are interpreted similarly.

In terms of the best settings, note that the ordered data plot, based on the maximum response value, yielded

+, -, +, -, +, -, -

Note that a consensus best value, with "." indicating a setting for which the three plots disagree, would be

+, -, +, ., +, -, -

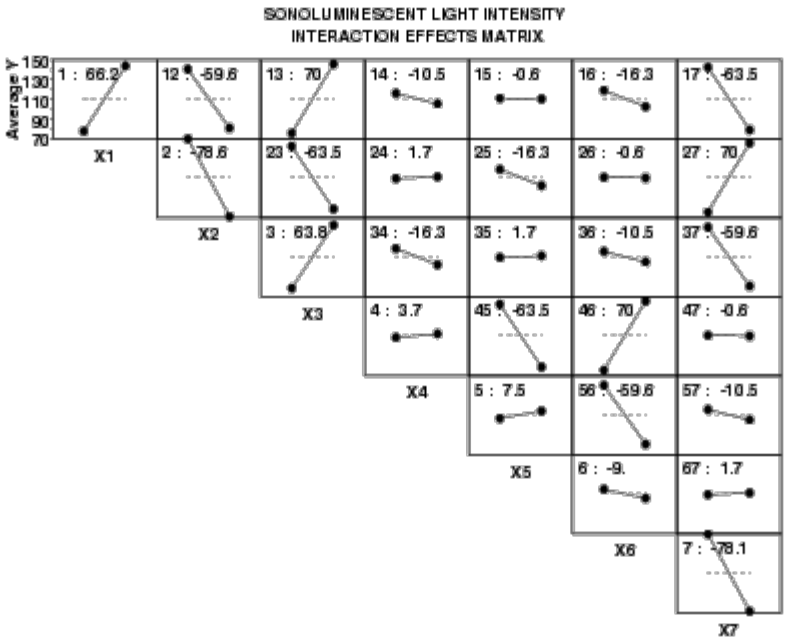
Note that the factor for which the settings disagree, *X*4, invariably defines itself as an "unimportant" factor.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

5.6.2.3. Interaction Effects

Check for Interaction Effects: DOE Interaction Plot

In addition to the main effects, it is also important to check for interaction effects, especially 2-factor interaction effects. The DOE interaction effects plot is an effective tool for this.



Conclusions from the DOE Interaction Effects Plot

We make the following conclusions from the DOE interaction effects plot.

1. Important Factors: Looking for the plots that have the steepest lines (that is, the largest effects), and noting that the legends on each subplot give the estimated effect, we have that
    - The diagonal plots are the main effects. The important factors are: *X*2, *X*7, *X*1, and *X*3. These four factors have |effect| > 60. The remaining three factors have |effect| < 10.
    - The off-diagonal plots are the 2-factor interaction effects. Of the 21 2-factor interactions, 9 are nominally important, but they fall into three groups of three:

- *X*1**X*3, *X*4**X*6, *X*2**X*7 (effect = 70)
- *X*2**X*3, *X*4**X*5, *X*1**X*7 (effect approximately 63.5)
- *X*1**X*2, *X*5**X*6, *X*3**X*7 (effect = -59.6)

All remaining 2-factor interactions are small having an |effect| < 20. A virtue of the interaction effects matrix plot is that the confounding structure of this Resolution IV design can be read off the plot. In this case, the fact that *X*1**X*3, *X*4**X*6, and *X*2**X*7 all have effect estimates identical to 70 is not a mathematical coincidence. It is a reflection of the fact that for this design, the three 2-factor interactions are confounded. This is also true for the other two sets of three (*X*2**X*3, *X*4**X*5, *X*1**X*7, and *X*1**X*2, *X*5**X*6, *X*3**X*7).

2. Best Settings: Reading down the diagonal plots, we select, as before, the best settings "on the average":

   (*X*1,*X*2,*X*3,*X*4,*X*5,*X*6,*X*7) = (+,-,+,+,+,-,-)

   For the more important factors (*X*1, *X*2, *X*3, *X*7), we note that the best settings (+, -, +, -) are consistent with the best settings for the 2-factor interactions (cross-products):

   *X*1: +, *X*2: - with *X*1**X*2: -
   *X*1: +, *X*3: + with *X*1**X*3: +
   *X*1: +, *X*7: - with *X*1**X*7: -
   *X*2: -, *X*3: + with *X*2**X*3: -
   *X*2: -, *X*7: - with *X*2**X*7: +
   *X*3: +, *X*7: - with *X*3**X*7: -

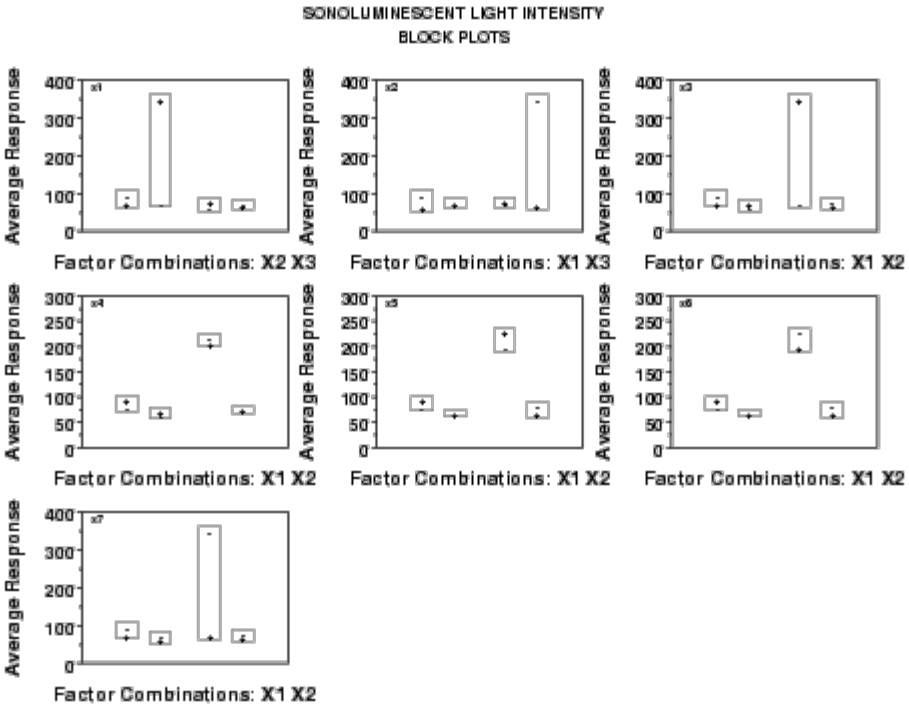
5.6.2.4. Main and Interaction Effects: Block Plots

Block Plots

Block plots are a useful adjunct to the DOE mean plot and the DOE interaction effects plot to confirm the importance of factors, to establish the robustness of main effect conclusions, and to determine the existence of interactions.

For block plots, it is the height of the bars that is important, not the relative positioning of each bar. Hence we focus on the size and internal signs of the blocks, not "where" the blocks are relative to each other.

We note in passing that for a fractional factorial design, we cannot display all combinations of the six remaining factors. We have arbitrarily chosen two robustness factors, which yields four blocks for comparison.



Conclusions from the Block Plots

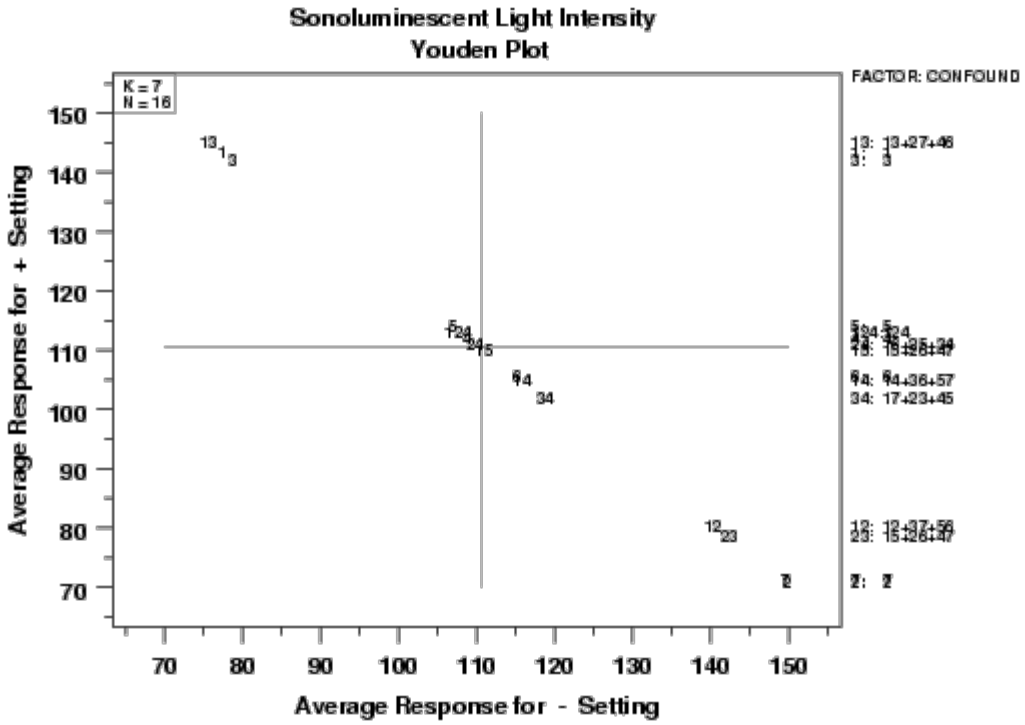
We can make the following conclusions from the block plots.

1. Relative Importance of Factors: Because of the expanded vertical axis, due to the two "outliers", the block plot is not particularly revealing. Block plots based on alternatively scaled data (e.g., LOG(Y)) would be more informative.

5.6.2.5. Important Factors: Youden Plot

Purpose The DOE Youden plot is used to distinguish between important and unimportant factors.

Sample Youden Plot



Conclusions from the Youden plot We can make the following conclusions from the Youden plot.

1. In the upper left corner are the interaction term $X1*X3$ and the main effects $X1$ and $X3$.

2. In the lower right corner are the main effects $X2$ and $X7$ and the interaction terms $X2*X3$ and $X1*X2$.

3. The remaining terms are clustered in the center, which indicates that such effects have averages that are similar (and hence the effects are near zero), and so such effects are relatively unimportant.

4. On the far right of the plot, the confounding structure is given (e.g., 13: 13+27+46), which suggests that the information on $X1*X3$ (on the plot) must

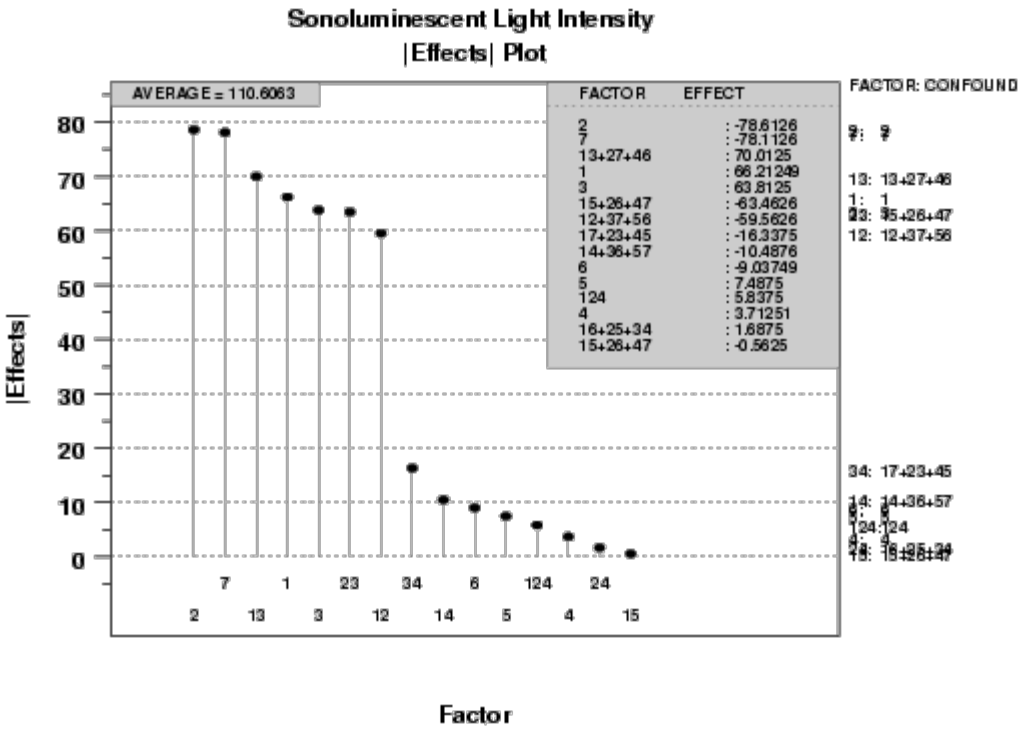
be tempered with the fact that $X1*X3$ is confounded with $X2*X7$ and $X4*X6$.

5.6.2.6. Important Factors: |Effects| Plot

Purpose

The |effects| plot displays the results of a Yates analysis in both a tabular and a graphical format. It is used to distinguish between important and unimportant effects.

Sample |Effects| Plot



Conclusions from the |effects| plot

We can make the following conclusions from the |effects| plot.

1. A ranked list of main effects and interaction terms is:
   *X*2
   *X*7
   *X*1**X*3 (confounded with *X*2**X*7 and *X*4**X*6)
   *X*1
   *X*3
   *X*2**X*3 (confounded with *X*4**X*5 and *X*1**X*7)
   *X*1**X*2 (confounded with *X*3**X*7 and *X*5**X*6)
   *X*3**X*4 (confounded with *X*1**X*6 and *X*2**X*5)
   *X*1**X*4 (confounded with *X*3**X*6 and *X*5**X*7)
   *X*6

*X*5
*X*1**X*2**X*4 (confounded with other 3-factor interactions)
*X*4
*X*2**X*4 (confounded with *X*3**X*5 and *X*6**X*7)
*X*1**X*5 (confounded with *X*2**X*6 and *X*4**X*7)

2. From the graph, there is a clear dividing line between the first seven effects (all |effect| > 50) and the last eight effects (all |effect| < 20). This suggests we retain the first seven terms as "important" and discard the remaining as "unimportant".

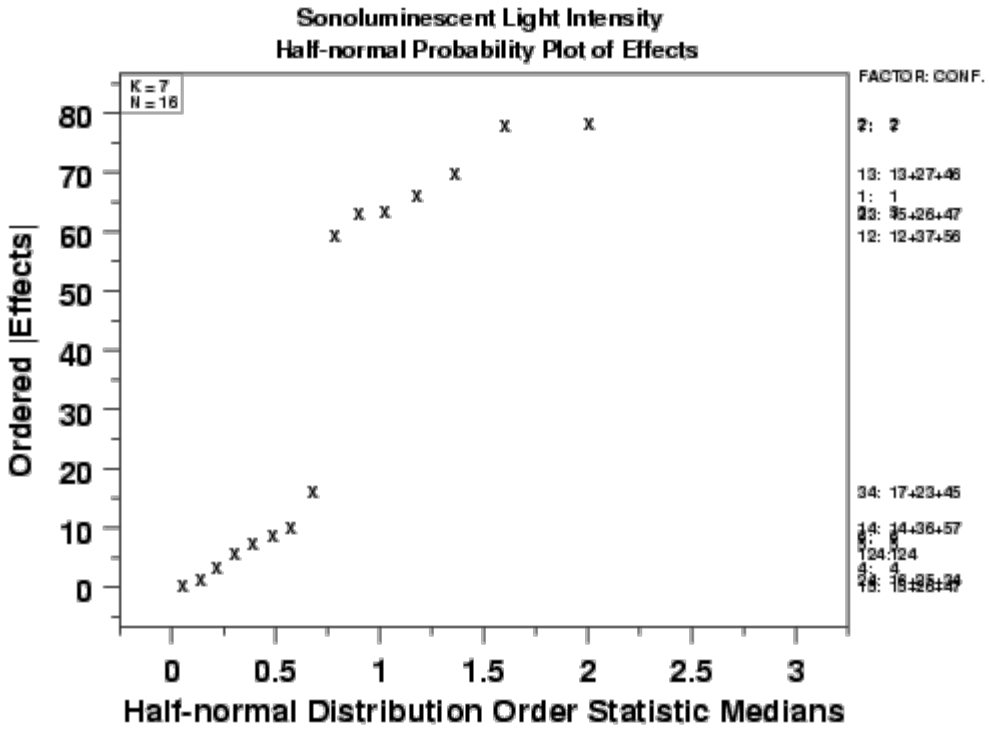
3. Again, the confounding structure on the right reminds us that, for example, the nominal effect size of 70.0125 for *X*1**X*3 (molarity*pH) can come from an *X*1**X*3 interaction, an *X*2**X*7 (solute*clamping) interaction, an *X*4**X*6 (gas*horn depth) interaction, or any mixture of the three interactions.

5. Process Improvement
5.6. Case Studies
5.6.2. Sonoluminescent Light Intensity Case Study

5.6.2.7. Important Factors: Half-Normal Probability Plot

Purpose The half-normal probability plot is used to distinguish between important and unimportant effects.

Sample Half-Normal Probability Plot



Conclusions from the Half-Normal Probability Plot We can make the following conclusions from the half-normal probability plot.

1. The points in the plot divide into two clear clusters:
   - An upper cluster (|effect| > 60).
   - A lower cluster (|effect| < 20).
2. The upper cluster contains the effects:

   *X*2, *X*7, *X*1**X*3 (and confounding), *X*1, *X*3, *X*2**X*3 (and confounding), *X*1**X*2 (and confounding)

   These effects should definitely be considered important.
3. The remaining effects lie on a line and form a lower cluster. These effects are

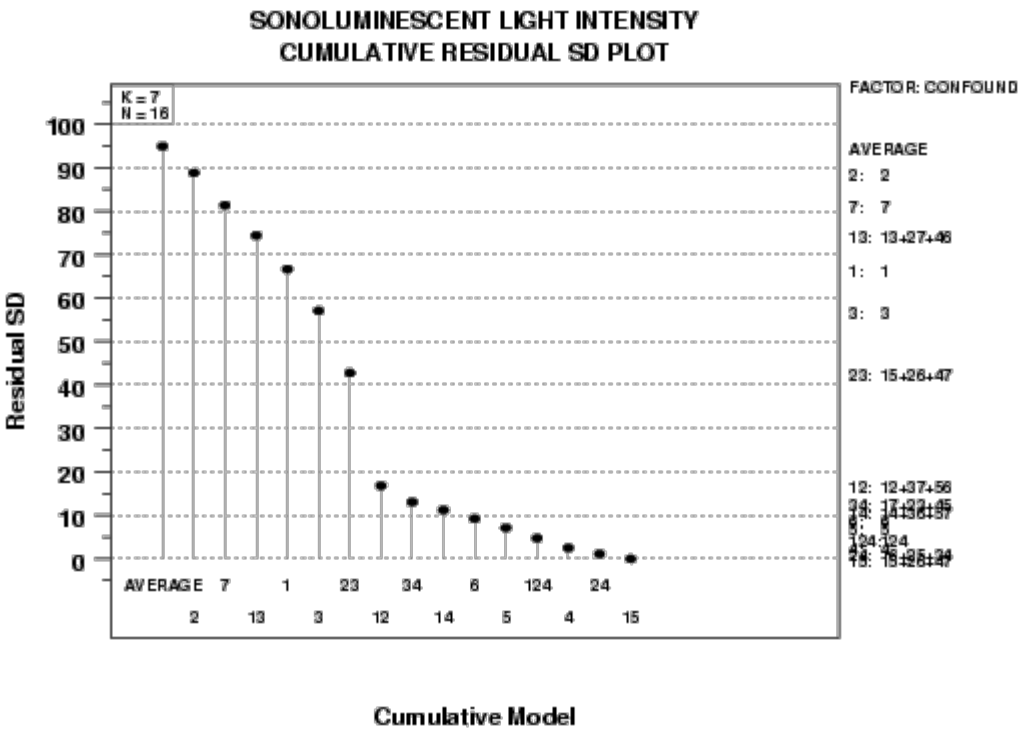
declared relatively unimportant.

4. The effect id's and the confounding structure are given on the far right (e.g., 13:13+27+46).

5.6.2.8. Cumulative Residual Standard Deviation Plot

Purpose The cumulative residual standard deviation plot is used to identify the best (parsimonious) model.

Sample Cumulative Residual Standard Deviation Plot



Conclusions from the Cumulative Residual SD Plot We can make the following conclusions from the cumulative residual standard deviation plot.

1. The baseline model consisting only of the average ($\hat{Y}$ = 110.6063) has a high residual standard deviation (95).

2. The cumulative residual standard deviation shows a significant and steady decrease as the following terms are added to the average: $X2$, $X7$, $X1*X3$, $X1$, $X3$, $X2*X3$, and $X1*X2$. Including these terms reduces the cumulative residual standard deviation from approximately 95 to approximately 17.

3. Exclude from the model any term after $X1*X2$ as the decrease in the residual standard deviation becomes relatively small.

4. From the |effects| plot, we see that the average is 110.6063, the estimated *X*2 effect is -78.6126, and so on. (The model coefficients are one half of the effect estimates.) We use this to from the following prediction equation:

$$\hat{Y} = 110.6063 - 39.3063X_2 - 39.0563X_7 + 35.00625X_1X_3 + 33.106245X_1 + 31.90625X_3 - 31.7313X_1X_5 - 29.781X_1X_2$$

Note that *X*1**X*3 is confounded with *X*2**X*7 and *X*4**X*6, *X*1**X*5 is confounded with *X*2*X6 and *X*4**X*7, and *X*1**X*2 is confounded with *X*3**X*7 and *X*5**X*6.

From the above graph, we see that the residual standard deviation for this model is approximately 17.

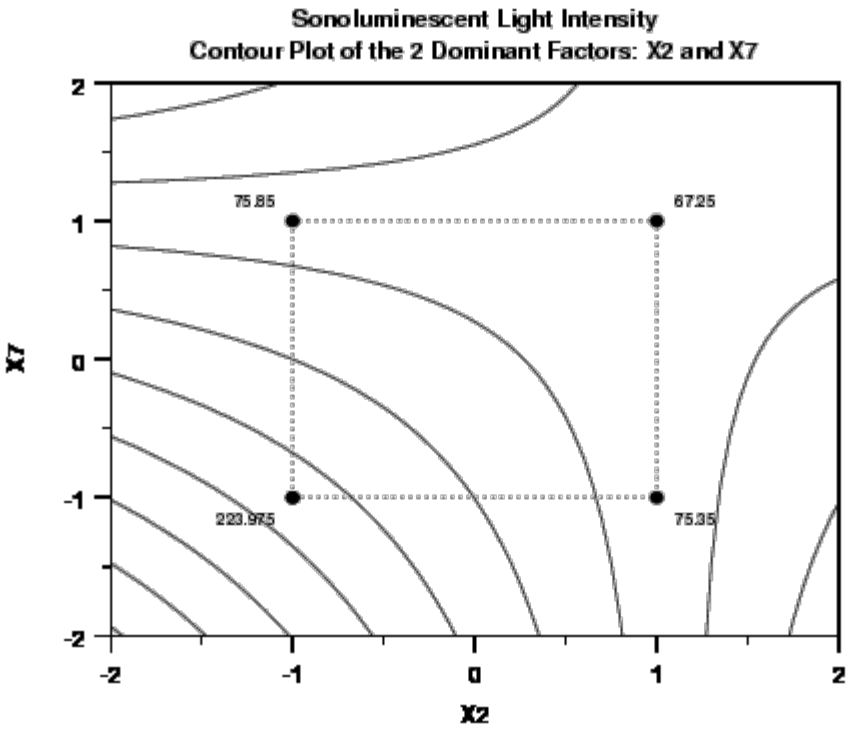
5.6.2.9. Next Step: DOE Contour Plot

Purpose

The DOE contour plot is used to determine the best factor settings for the two most important factors in the next iteration of the experiment.

From the previous plots, we identified *X*2 (solute) and *X*7 (horn depth) as the two most important factors.

Sample DOE Contour Plot



Conclusions from the DOE Contour Plot

We can make the following conclusions from the DOE contour plot.

1. The best (high light intensity) setting for *X*2 is "-" and the best setting for *X*7 is "-". This combination yields an average response of approximately 224. The next highest average response from any other combination of these factors is only 76.

2. The non-linear nature of the contour lines implies that the *X*2**X*7 interaction is important.

3. On the left side of the plot from top to bottom, the contour lines start at 0,

increment by 50 and stop at 400. On the bottom of the plot from right to left, the contour lines start at 0, increment by 50 and stop at 400.

To achieve a light intensity of, say 400, this suggests an extrapolated best setting of ($X2$, $X7$) = (-2,-2).

4. Such extrapolation only makes sense if $X2$ and $X7$ are continuous factors. Such is not the case here. In this example, $X2$ is solute (-1 = sugar and +1 = glycerol) and $X7$ is flask clamping (-1 is unclamped and +1 is clamped). Both factors are discrete, and so extrapolated settings are not possible.

5.6.2.10. Summary of Conclusions

Most Important Factors

The primary goal of this experiment was to identify the most important factors in maximizing the sonoluminescent light intensity.

Based on the preceding graphical analysis, we make the following conclusions.

- Four factors and three groups of 2-factor interactions are important. A rank-order listing of factors is:
  1. *X*2: Solute (effect = -78.6)
  2. *X*7: Clamping (effect = -78.1)
  3. *X*1**X*3 (Molarity*pH) or
     *X*2**X*7 (Solute*Clamping)
     (effect = 70.0)
  4. *X*1: Molarity (effect = 66.2)
  5. *X*3: pH (effect = 63.5)
  6. *X*2**X*3 (Solute*pH) or
     *X*4**X*5 (Gas*Water Depth)
     *X*1**X*7 (Molarity*Clamping)
     (effect = -63.5)
  7. *X*1**X*2 (Molarity*Solute) or
     *X*3**X*7 (Ph*Clamping)
     (effect = -59.6)
- Thus, of the seven factors and 21 2-factor interactions, it was found that four factors and at most seven 2-factor interactions seem important, with the remaining three factors and 14 interactions apparently being unimportant.

Best Settings

The best settings to maximize sonoluminescent light intensity are

- *X*1 (Molarity) + (0.33 mol)
- *X*2 (Solute) - (sugar)
- *X*3 (pH) + (11)
- *X*4 (Gas) . (either)
- *X*5 (Water Depth) + (full)
- *X*6 (Horn Depth) - (5 mm)
- *X*7 (Clamping) - (unclamped)

with the *X*1, *X*2, *X*3, and *X*7 settings especially important.

5.6.2.11. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the Data Sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Get set up and started. 1. Read in the data.	1. You have read 8 columns of numbers into Dataplot: variables Y, X1, X2, X3, X4, X5, X6, and X7.
2. Plot the main effects. 1. Ordered data plot. 2. DOE scatter plot. 3. DOE mean plot.	1. Ordered data plot shows 2 points that stand out. Potential important factors are X1, X2, X3, and X7. 2. DOE scatter plot identifies X1, X2, X3, and X7 as

	important factors. 3. DOE mean plot identifies X1, X2, X3, and X7 as important factors.
3. Plots for interaction effects 1. Generate a DOE interaction effects plot.	1. The DOE interaction effects plot shows several important interaction effects.
4. Block plots for main and interaction effects 1. Generate block plots.	1. The block plots are not particularly helpful in this case.
5. Youden plot to identify important factors 1. Generate a Youden plot.	1. The Youden plot identifies X1, X2, X3, and X7 as important factors. It also identifies a number of important interactions (X1*X3, X1*X2, X2*X3).
6. |Effects| plot to identify important factors 1. Generate |effects| plot.	1. The |effects| plot identifies X2, X7, X1*X3, X1, X3, X2*X3, and X1*X2 as important factors and interactions.
7. Half-normal probability plot to identify important factors 1. Generate half-normal probability plot.	1. The half-normal probability plot identifies X2, X7, X1*X3, X1, X3, X2*X3, and X1*X2 as important factors and interactions.
8. Cumulative residual standard deviation plot 1. Generate a cumulative residual standard deviation plot.	1. The cumulative residual standard

	deviation plot results in a model with 4 main effects and 3 2-factor interactions.
9. DOE contour plot 1. Generate a DOE contour plot using factors 2 and 7.	1. The DOE contour plot shows X2 = -1 and X7 = -1 to be the best settings.

5.7. A Glossary of DOE Terminology

Definitions for key DOE terms

This page gives definitions and information for many of the basic terms used in DOE.

- ***Alias***: When the estimate of an effect also includes the influence of one or more other effects (usually high order interactions) the effects are said to be *aliased* (see *confounding*). For example, if the estimate of effect D in a four factor experiment actually estimates (D + ABC), then the main effect D is aliased with the 3-way interaction ABC. **Note**: This causes no difficulty when the higher order interaction is either non-existent or insignificant.
- ***Analysis of Variance (ANOVA)***: A mathematical process for separating the variability of a group of observations into assignable causes and setting up various significance tests.
- ***Balanced Design***: An experimental design where all cells (i.e. treatment combinations) have the same number of observations.
- ***Blocking***: A schedule for conducting treatment combinations in an experimental study such that any effects on the experimental results due to a known change in raw materials, operators, machines, etc., become concentrated in the levels of the blocking variable. **Note**: the reason for blocking is to isolate a systematic effect and prevent it from obscuring the main effects. Blocking is achieved by restricting randomization.
- ***Center Points***: Points at the center value of all factor ranges.

Coding Factor Levels**:** Transforming the scale of measurement for a factor so that the high value becomes +1 and the low value becomes -1 (see *scaling*). After coding all factors in a 2-level full factorial experiment, the design matrix has all orthogonal columns.

Coding is a simple linear transformation of the original measurement scale. If the "high" value is X_h and the "low" value is X_L (in the original scale), then the scaling transformation takes any

original X value and converts it to (X - a)/b, where
$a = (X_h + X_L)/2$ and $b = (X_h - X_L)/2$.
To go back to the original measurement scale, just take the coded value and multiply it by "b" and add "a" or, X = b(coded value) + a.

As an example, if the factor is temperature and the high setting is 65^oC and the low setting is 55^oC, then a = (65 + 55)/2 = 60 and b = (65 - 55)/2 = 5. The center point (where the coded value is 0) has a temperature of 5(0) + 60 = 60^oC.

- ***Comparative Designs***: A design aimed at making conclusions about one a priori important factor, possibly in the presence of one or more other "nuisance" factors.
- ***Confounding***: A confounding design is one where some treatment effects (main or interactions) are estimated by the same linear combination of the experimental observations as some blocking effects. In this case, the treatment effect and the blocking effect are said to be *confounded*. Confounding is also used as a general term to indicate that the value of a main effect estimate comes from both the main effect itself and also contamination or bias from higher order interactions. Note: Confounding designs naturally arise when full factorial designs have to be run in blocks and the block size is smaller than the number of different treatment combinations. They also occur whenever a fractional factorial design is chosen instead of a full factorial design.
- ***Crossed Factors:*** See *factors below*.
- ***Design***: A set of experimental runs which allows you to fit a particular model and estimate your desired effects.
- ***Design Matrix***: A matrix description of an experiment that is useful for constructing and analyzing experiments.
- ***Effect***: How changing the settings of a factor changes the response. The effect of a single factor is also called a *main effect*. **Note:** For a factor A with two levels, scaled so that low = -1 and high = +1, the effect of A is estimated by subtracting the average response when A is -1 from the average response when A = +1 and dividing the result by 2 (division by 2 is needed because the -1 level is 2 scaled units away from the +1 level).
- ***Error***: Unexplained variation in a collection of observations. **Note**: DOE's typically require understanding of both random error and lack of fit

error.

- ***Experimental Unit***: The entity to which a specific treatment combination is applied. **Note**: an experimental unit can be a
    - PC board
    - silicon wafer
    - tray of components simultaneously treated
    - individual agricultural plants
    - plot of land
    - automotive transmissions
    - etc.

- ***Factors****: Process inputs* an investigator manipulates to cause a change in the output. Some factors cannot be controlled by the experimenter but may effect the responses. If their effect is significant, these *uncontrolled factors* should be measured and used in the data analysis. **Note**: The inputs can be discrete or continuous.
    - ***Crossed Factors***: Two factors are *crossed* if every level of one occurs with every level of the other in the experiment.
    - ***Nested Factors***: A factor "A" is nested within another factor "B" if the levels or values of "A" are different for every level or value of "B". **Note**: Nested factors or effects have a hierarchical relationship.

- ***Fixed Effect***: An effect associated with an input variable that has a limited number of levels or in which only a limited number of levels are of interest to the experimenter.
- ***Interactions***: Occurs when the effect of one factor on a response depends on the level of another factor(s).
- ***Lack of Fit Error***: Error that occurs when the analysis omits one or more important terms or factors from the process model. **Note**: Including replication in a DOE allows separation of experimental error into its components: lack of fit and random (pure) error.
- ***Model***: Mathematical relationship which relates changes in a given response to changes in one or more factors.
- ***Nested Factors***: See *factors* above.
- ***Orthogonality***: Two vectors of the same length are orthogonal if the sum of the products of their corresponding elements is 0. **Note**: An experimental design is orthogonal if the effects of any factor balance out (sum to zero) across the effects of the other factors.
- ***Random Effect***: An effect associated with input

variables chosen at random from a population having a large or infinite number of possible values.

- ***Random error***: Error that occurs due to natural variation in the process. Note: Random error is typically assumed to be normally distributed with zero mean and a constant variance. **Note**: Random error is also called experimental error.
- ***Randomization***: A schedule for allocating treatment material and for conducting treatment combinations in a DOE such that the conditions in one run neither depend on the conditions of the previous run nor predict the conditions in the subsequent runs. **Note**: The importance of randomization cannot be over stressed. Randomization is necessary for conclusions drawn from the experiment to be correct, unambiguous and defensible.
- ***Replication***: Performing the same treatment combination more than once. **Note**: Including replication allows an estimate of the random error independent of any lack of fit error.
- ***Resolution***: A term which describes the degree to which estimated main effects are aliased (or confounded) with estimated 2-level interactions, 3-level interactions, etc. In general, the resolution of a design is one more than the smallest order interaction that some main effect is confounded (aliased) with. If some main effects are confounded with some 2-level interactions, the resolution is 3. **Note**: Full factorial designs have no confounding and are said to have resolution "infinity". For most practical purposes, a resolution 5 design is excellent and a resolution 4 design may be adequate. Resolution 3 designs are useful as economical screening designs.
- ***Responses***: The output(s) of a process. Sometimes called dependent variable(s).
- ***Response Surface Designs***: A DOE that fully explores the process window and models the responses. **Note**: These designs are most effective when there are less than 5 factors. Quadratic models are used for response surface designs and at least three levels of every factor are needed in the design.
- ***Rotatability***: A design is *rotatable* if the variance of the predicted response at any point **x** depends only on the distance of **x** from the design center point. A design with this property can be rotated around its center point without changing the prediction variance at **x**. **Note**: Rotatability is a desirable property for response surface designs (i.e. quadratic model designs).
- ***Scaling Factor Levels***: Transforming factor

levels so that the high value becomes +1 and the low value becomes -1.

- ***Screening Designs***: A DOE that identifies which of many factors have a significant effect on the response. **Note**: Typically screening designs have more than 5 factors.
- ***Treatment***: A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.
- ***Treatment Combination***: The combination of the settings of several factors in a given experimental trial. Also known as a *run*.
- ***Variance Components***: Partitioning of the overall variation into assignable components.

5.8. References

Chapter specific references

Bisgaard, S. and Steinberg, D. M., (1997), "The Design and Analysis of 2^{k-p} Prototype Experiments," *Technometrics*, 39, 1, 52-62.

Box, G. E. P., and Draper, N. R., (1987), *Empirical Model Building and Response Surfaces*, John Wiley & Sons, New York, NY.

Box, G. E. P., and Hunter, J. S., (1954), "A Confidence Region for the Solution of a Set of Simultaneous Equations with an Application to Experimental Design," *Biometrika*, 41, 190-199

Box, G. E. P., and Wilson, K. B., (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society, Series B*, 13, 1-45.

Capobianco, T. E., Splett, J. D. and Iyer, H. K., "Eddy Current Probe Sensitivity as a Function of Coil Construction Parameters." *Research in Nondesructive Evaluation*, Vol. 2, pp. 169-186, December, 1990.

Cornell, J. A., (1990), *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, John Wiley & Sons, New York, NY.

Del Castillo, E., (1996), "Multiresponse Optimization Confidence Regions," *Journal of Quality Technology*, 28, 1, 61-70.

Derringer, G., and Suich, R., (1980), "Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology*, 12, 4, 214-219.

Draper, N.R., (1963), "Ridge Analysis of Response Surfaces," *Technometrics*, 5, 469-479.

Hoerl, A. E., (1959), "Optimum Solution of Many Variables Equations," *Chemical Engineering Progress*, 55, 67-78.

Hoerl, A. E., (1964), "Ridge Analysis," *Chemical Engineering Symposium Series*, 60, 67-77.

Khuri, A. I., and Cornell, J. A., (1987), *Response Surfaces*, Marcel Dekker, New York, NY.

Mee, R. W., and Peralta, M. (2000), "Semifolding 2^{k-p} Designs," *Technometrics, 42*, 2, p122.

Miller, A. (1997), "Strip-Plot Configuration of Fractional Factorials," *Technometrics,* 39, 2, p153.

Myers, R. H., and Montgomery, D. C., (1995), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments,* John Wiley & Sons, New York, NY.

Ryan, Thomas P., (2000), *Statistical Methods for Quality Improvement*, John Wiley & Sons, New York, NY.

Taguchi, G. and Konishi, S., (1987), *Orthogonal Arrays and Linear Graphs*, Dearborn, MI, ASI press.

Well Known General References

Box, G. E. P., Hunter, W. G., and Hunter, S. J. (1978), *Statistics for Experimenters*, John Wiley & Sons, Inc., New York, NY.

Diamond, W. J., (1989), *Practical Experimental Designs,* Second Ed., Van Nostrand Reinhold, NY.

John, P. W. M., (1971), *Statistical Design and Analysis of Experiments,* SIAM Classics in Applied Mathematics, Philadelphia, PA.

Milliken, G. A., and Johnson, D. E., (1984), *Analysis of Messy Data, Vol. 1,* Van Nostrand Reinhold, NY.

Montgomery, D. C., (2000), *Design and Analysis of Experiments*, Fifth Edition, John Wiley & Sons, New York, NY.

Case studies for different industries

Snee, R. D., Hare, L. B., and Trout, J. R.(1985), *Experiments in Industry. Design, Analysis and Interpretation of Results*, Milwaukee, WI, American Society for Quality.

Case studies in Process Improvement, including DOE, in the Semiconductor Industry

Czitrom, V., and Spagon, P. D., (1997), *Statistical Case Studies for Industrial process Improvement*, Philadelphia, PA, ASA-SIAM Series on Statistics and Applied Probability.

Software to design and analyze

In addition to the extensive design and analysis documentation and routines in Dataplot, there are many other good commercial DOE packages. This Chapter

experiments showed examples using "JMP" (by the SAS Institute, 100 SAS CampusDrive, Cary, North Carolina 27513-9905), as an illustration of a good commercial package.

6. Process or Product Monitoring and Control

This chapter presents techniques for monitoring and controlling processes and signaling when corrective actions are necessary.

1. **Introduction**

1. History
2. Process Control Techniques
3. Process Control
4. "Out of Control"
5. "In Control" but Unacceptable
6. Process Capability

2. **Test Product for Acceptability**

1. Acceptance Sampling
2. Kinds of Sampling Plans
3. Choosing a Single Sampling Plan
4. Double Sampling Plans
5. Multiple Sampling Plans
6. Sequential Sampling Plans
7. Skip Lot Sampling Plans

3. **Univariate and Multivariate Control Charts**

1. Control Charts
2. Variables Control Charts
3. Attributes Control Charts
4. Multivariate Control charts

4. **Time Series Models**

1. Definitions, Applications and Techniques
2. Moving Average or Smoothing Techniques
3. Exponential Smoothing
4. Univariate Time Series Models
5. Multivariate Time Series Models

5. **Tutorials**

1. What do we mean by "Normal" data?
2. What to do when data are non-normal
3. Elements of Matrix Algebra
4. Elements of Multivariate Analysis
5. Principal Components

6. **Case Study**

1. Lithography Process Data
2. Box-Jenkins Modeling Example

Detailed Table of Contents
References

6. Process or Product Monitoring and Control - Detailed Table of Contents [6.]

1. Introduction [6.1.]
    1. How did Statistical Quality Control Begin? [6.1.1.]
    2. What are Process Control Techniques? [6.1.2.]
    3. What is Process Control? [6.1.3.]
    4. What to do if the process is "Out of Control"? [6.1.4.]
    5. What to do if "In Control" but Unacceptable? [6.1.5.]
    6. What is Process Capability? [6.1.6.]

2. Test Product for Acceptability: Lot Acceptance Sampling [6.2.]
    1. What is Acceptance Sampling? [6.2.1.]
    2. What kinds of Lot Acceptance Sampling Plans (LASPs) are there? [6.2.2.]
    3. How do you Choose a Single Sampling Plan? [6.2.3.]
        1. Choosing a Sampling Plan: MIL Standard 105D [6.2.3.1.]
        2. Choosing a Sampling Plan with a given OC Curve [6.2.3.2.]
    4. What is Double Sampling? [6.2.4.]
    5. What is Multiple Sampling? [6.2.5.]
    6. What is a Sequential Sampling Plan? [6.2.6.]
    7. What is Skip Lot Sampling? [6.2.7.]

3. Univariate and Multivariate Control Charts [6.3.]
    1. What are Control Charts? [6.3.1.]
    2. What are Variables Control Charts? [6.3.2.]
        1. Shewhart X-bar and R and S Control Charts [6.3.2.1.]
        2. Individuals Control Charts [6.3.2.2.]
        3. Cusum Control Charts [6.3.2.3.]
            1. Cusum Average Run Length [6.3.2.3.1.]
        4. EWMA Control Charts [6.3.2.4.]
    3. What are Attributes Control Charts? [6.3.3.]
        1. Counts Control Charts [6.3.3.1.]
        2. Proportions Control Charts [6.3.3.2.]
    4. What are Multivariate Control Charts? [6.3.4.]
        1. Hotelling Control Charts [6.3.4.1.]
        2. Principal Components Control Charts [6.3.4.2.]
        3. Multivariate EWMA Charts [6.3.4.3.]

4. Introduction to Time Series Analysis [6.4.]
    1. Definitions, Applications and Techniques [6.4.1.]
    2. What are Moving Average or Smoothing Techniques? [6.4.2.]
        1. Single Moving Average [6.4.2.1.]
        2. Centered Moving Average [6.4.2.2.]

3. What is Exponential Smoothing? [6.4.3.]
    1. Single Exponential Smoothing [6.4.3.1.]
    2. Forecasting with Single Exponential Smoothing [6.4.3.2.]
    3. Double Exponential Smoothing [6.4.3.3.]
    4. Forecasting with Double Exponential Smoothing(LASP) [6.4.3.4.]
    5. Triple Exponential Smoothing [6.4.3.5.]
    6. Example of Triple Exponential Smoothing [6.4.3.6.]
    7. Exponential Smoothing Summary [6.4.3.7.]
4. Univariate Time Series Models [6.4.4.]
    1. Sample Data Sets [6.4.4.1.]
        1. Data Set of Monthly CO2 Concentrations [6.4.4.1.1.]
        2. Data Set of Southern Oscillations [6.4.4.1.2.]
    2. Stationarity [6.4.4.2.]
    3. Seasonality [6.4.4.3.]
        1. Seasonal Subseries Plot [6.4.4.3.1.]
    4. Common Approaches to Univariate Time Series [6.4.4.4.]
    5. Box-Jenkins Models [6.4.4.5.]
    6. Box-Jenkins Model Identification [6.4.4.6.]
        1. Model Identification for Southern Oscillations Data [6.4.4.6.1.]
        2. Model Identification for the CO_2 Concentrations Data [6.4.4.6.2.]
        3. Partial Autocorrelation Plot [6.4.4.6.3.]
    7. Box-Jenkins Model Estimation [6.4.4.7.]
    8. Box-Jenkins Model Diagnostics [6.4.4.8.]
        1. Box-Ljung Test [6.4.4.8.1.]
    9. Example of Univariate Box-Jenkins Analysis [6.4.4.9.]
    10. Box-Jenkins Analysis on Seasonal Data [6.4.4.10.]
5. Multivariate Time Series Models [6.4.5.]
    1. Example of Multivariate Time Series Analysis [6.4.5.1.]

5. Tutorials [6.5.]
    1. What do we mean by "Normal" data? [6.5.1.]
    2. What do we do when data are "Non-normal"? [6.5.2.]
    3. Elements of Matrix Algebra [6.5.3.]
        1. Numerical Examples [6.5.3.1.]
        2. Determinant and Eigenstructure [6.5.3.2.]
    4. Elements of Multivariate Analysis [6.5.4.]
        1. Mean Vector and Covariance Matrix [6.5.4.1.]
        2. The Multivariate Normal Distribution [6.5.4.2.]
        3. Hotelling's *T* squared [6.5.4.3.]
            1. T^2 Chart for Subgroup Averages -- Phase I [6.5.4.3.1.]
            2. T^2 Chart for Subgroup Averages -- Phase II [6.5.4.3.2.]
            3. Chart for Individual Observations -- Phase I [6.5.4.3.3.]
            4. Chart for Individual Observations -- Phase II [6.5.4.3.4.]
            5. Charts for Controlling Multivariate Variability [6.5.4.3.5.]
            6. Constructing Multivariate Charts [6.5.4.3.6.]
    5. Principal Components [6.5.5.]
        1. Properties of Principal Components [6.5.5.1.]
        2. Numerical Example [6.5.5.2.]

6. Case Studies in Process Monitoring [6.6.]
    1. Lithography Process [6.6.1.]
        1. Background and Data [6.6.1.1.]
        2. Graphical Representation of the Data [6.6.1.2.]

3. Subgroup Analysis [6.6.1.3.]
4. Shewhart Control Chart [6.6.1.4.]
5. Work This Example Yourself [6.6.1.5.]

2. Aerosol Particle Size [6.6.2.]
   1. Background and Data [6.6.2.1.]
   2. Model Identification [6.6.2.2.]
   3. Model Estimation [6.6.2.3.]
   4. Model Validation [6.6.2.4.]
   5. Work This Example Yourself [6.6.2.5.]

7. References [6.7.]

6. Process or Product Monitoring and Control

6.1. Introduction

Contents of Section

This section discusses the basic concepts of statistical process control, quality control and process capability.

1. How did Statistical Quality Control Begin?
2. What are Process Control Techniques?
3. What is Process Control?
4. What to do if the process is "Out of Control"?
5. What to do if "In Control" but Unacceptable?
6. What is Process Capability?

6.1.1. How did Statistical Quality Control Begin?

Historical perspective

Quality Control has been with us for a long time. How long? It is safe to say that when manufacturing began and competition accompanied manufacturing, consumers would compare and choose the most attractive product (barring a monopoly of course). If manufacturer A discovered that manufacturer B's profits soared, the former tried to improve his/her offerings, probably by improving the quality of the output, and/or lowering the price. Improvement of quality did not necessarily stop with the product - but also included the *process* used for making the product.

The process was held in high esteem, as manifested by the medieval guilds of the Middle Ages. These guilds mandated long periods of training for apprentices, and those who were aiming to become master craftsmen had to demonstrate evidence of their ability. Such procedures were, in general, aimed at the maintenance and improvement of the quality of the process.

In modern times we have professional societies, governmental regulatory bodies such as the Food and Drug Administration, factory inspection, etc., aimed at assuring the quality of products sold to consumers. *Quality Control* has thus had a long history.

Science of statistics is fairly recent

On the other hand, *statistical* quality control is comparatively new. The science of statistics itself goes back only two to three centuries. And its greatest developments have taken place during the 20th century. The earlier applications were made in astronomy and physics and in the biological and social sciences. It was not until the 1920s that statistical theory began to be applied effectively to quality control as a result of the development of sampling theory.

The concept of quality control in manufacturing was first

The first to apply the newly discovered statistical methods to the problem of quality control was Walter A. Shewhart of the Bell Telephone Laboratories. He issued a memorandum on May 16, 1924 that featured a sketch of a modern control chart.

advanced by Walter Shewhart

Shewhart kept improving and working on this scheme, and in 1931 he published a book on statistical quality control, *"Economic Control of Quality of Manufactured Product"*, published by Van Nostrand in New York. This book set the tone for subsequent applications of statistical methods to process control.

Contributions of Dodge and Romig to sampling inspection

Two other Bell Labs statisticians, H.F. Dodge and H.G. Romig spearheaded efforts in applying statistical theory to sampling inspection. The work of these three pioneers constitutes much of what nowadays comprises the theory of statistical quality and control. There is much more to say about the history of statistical quality control and the interested reader is invited to peruse one or more of the references. A very good summary of the historical background of SQC is found in chapter 1 of *"Quality Control and Industrial Statistics"*, by Acheson J. Duncan. See also Juran (1997).

6.1.2. What are Process Control Techniques?

Statistical Process Control (SPC)

Typical process control techniques

There are many ways to implement process control. Key monitoring and investigating tools include:

- Histograms
- Check Sheets
- Pareto Charts
- Cause and Effect Diagrams
- Defect Concentration Diagrams
- Scatter Diagrams
- Control Charts

All these are described in Montgomery (2000). This chapter will focus (Section 3) on control chart methods, specifically:

- Classical Shewhart Control charts,
- Cumulative Sum (CUSUM) charts
- Exponentially Weighted Moving Average (EWMA) charts
- Multivariate control charts

Underlying concepts

The underlying concept of statistical process control is based on a comparison of what is happening today with what happened previously. We take a snapshot of how the process typically performs or build a model of how we think the process will perform and calculate control limits for the expected measurements of the output of the process. Then we collect data from the process and compare the data to the control limits. The majority of measurements should fall within the control limits. Measurements that fall outside the control limits are examined to see if they belong to the same population as our initial snapshot or model. Stated differently, we use historical data to compute the initial control limits. Then the data are compared against these initial limits. Points that fall outside of the limits are investigated and, perhaps, some will later be discarded. If so, the limits would be recomputed and the process repeated. This is referred to as Phase I. Real-time process monitoring, using the limits from the end of Phase I, is Phase II.

Statistical Quality Control (SQC)

Tools of statistical quality control

Several techniques can be used to investigate the product for defects or defective pieces after all processing is complete. Typical tools of SQC (described in section 2) are:

- Lot Acceptance sampling plans
- Skip lot sampling plans
- Military (MIL) Standard sampling plans

Underlying concepts of statistical quality control

The purpose of statistical quality control is to ensure, in a cost efficient manner, that the product shipped to customers meets their specifications. Inspecting every product is costly and inefficient, but the consequences of shipping non conforming product can be significant in terms of customer dissatisfaction. Statistical Quality Control is the process of inspecting enough product from given lots to probabilistically ensure a specified quality level.

6.1.3. What is Process Control?

Two types of intervention are possible -- one is based on engineering judgment and the other is automated

Process Control is the active changing of the process based on the results of process monitoring. Once the process monitoring tools have detected an out-of-control situation, the person responsible for the process makes a change to bring the process back into control.

1. *Out-of-control Action Plans* (OCAPS) detail the action to be taken once an out-of-control situation is detected. A specific flowchart, that leads the process engineer through the corrective procedure, may be provided for each unique process.

2. *Advanced Process Control Loops* are automated changes to the process that are programmed to correct for the size of the out-of-control measurement.

6.1.4. What to do if the process is "Out of Control"?

Reactions to out-of-control conditions

If the process is out-of-control, the process engineer looks for an assignable cause by following the out-of-control action plan (OCAP) associated with the control chart. Out-of-control refers to rejecting the assumption that the current data are from the same population as the data used to create the initial control chart limits.

For classical Shewhart charts, a set of rules called the Western Electric Rules (WECO Rules) and a set of trend rules often are used to determine out-of-control.

6.1.5. What to do if "In Control" but Unacceptable?

In control means process is predictable	"In Control" only means that the process is predictable in a statistical sense. What do you do if the process is "in control" but the average level is too high or too low or the variability is unacceptable?
Process improvement techniques	Process improvement techniques such as • experiments • calibration • re-analysis of historical database can be initiated to put the process on target or reduce the variability.
Process must be stable	Note that the process must be stable before it can be centered at a target value or its overall variation can be reduced.

6.1.6. What is Process Capability?

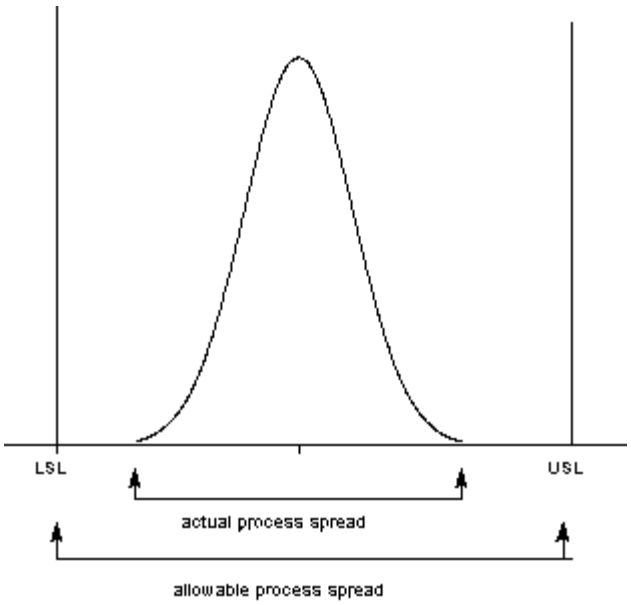
Process capability compares the output of an *in-control* process to the specification limits by using *capability indices*. The comparison is made by forming the ratio of the spread between the process specifications (the specification "width") to the spread of the process values, as measured by 6 process standard deviation units (the process "width").

Process Capability Indices

A process capability index uses both the process variability and the process specifications to determine whether the process is "capable"

We are often required to compare the output of a stable process with the process specifications and make a statement about how well the process meets specification. To do this we compare the natural variability of a stable process with the process specification limits.

A process where almost all the measurements fall inside the specification limits is a capable process. This can be represented pictorially by the plot below:



There are several statistics that can be used to measure the capability of a process: C_p, C_{pk}, C_{pm}.

Most capability indices estimates are valid only if the sample size used is 'large enough'. Large enough is generally thought to be about 50 independent data values.

The C_p, C_{pk}, *and* C_{pm} statistics assume that the population of data values is normally distributed. Assuming a two-sided specification, if μ and σ are the mean and

standard deviation, respectively, of the normal data and USL, LSL, and T are the upper and lower specification limits and the target value, respectively, then the population capability indices are defined as follows:

Definitions of various process capability indices

$$C_p = \frac{USL - LSL}{6\sigma}$$

$$C_{pk} = \min\left[\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right]$$

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}}$$

Sample estimates of capability indices

Sample estimators for these indices are given below. (Estimators are indicated with a "hat" over them).

$$\hat{C}_p = \frac{USL - LSL}{6s}$$

$$\hat{C}_{pk} = \min\left[\frac{USL - \bar{x}}{3s}, \frac{\bar{x} - LSL}{3s}\right]$$

$$\hat{C}_{pm} = \frac{USL - LSL}{6\sqrt{s^2 + (\bar{x} - T)^2}}$$

The estimator for C_{pk} can also be expressed as $C_{pk} = C_p(1-k)$, where k is a scaled distance between the midpoint of the specification range, m, and the process mean, μ.

Denote the midpoint of the specification range by m = (USL+LSL)/2. The distance between the process mean, μ, and the optimum, which is m, is μ - m, where $m \le \mu \le USL$. The scaled distance is

$$k = \frac{|m - \mu|}{(USL - LSL)/2}, \qquad 0 \le k \le 1$$

(the absolute sign takes care of the case when $LSL \le \mu \le m$). To determine the estimated value, $\hat{k}$, we estimate μ by $\bar{x}$. Note that $\bar{x} \le USL$.

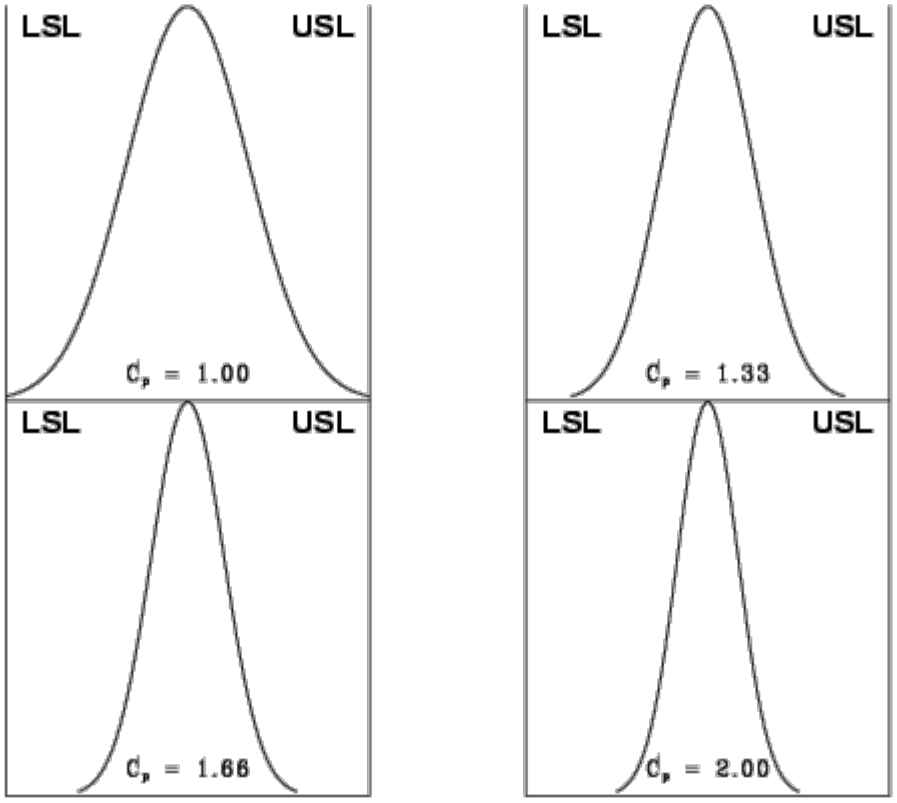
The estimator for the C_p index, adjusted by the k factor, is

$$\hat{C}_{pk} = \hat{C}_p(1 - \hat{k})$$

Since $0 \le k \le 1$, it follows that $\hat{C}_{pk} \le \hat{C}_p$.

Plot showing C_p for varying process widths

To get an idea of the value of the C_p statistic for varying process widths, consider the following plot



This can be expressed numerically by the table below:

Translating capability into "rejects"

USL - LSL	6σ	8σ	10σ	12σ
C_p	1.00	1.33	1.66	2.00
Rejects	.27%	64 ppm	.6 ppm	2 ppb
% of spec used	100	75	60	50

where ppm = parts per million and ppb = parts per billion. Note that the reject figures are based on the assumption that the distribution is centered at μ.

We have discussed the situation with two spec. limits, the USL and LSL. This is known as the *bilateral* or two-sided case. There are many cases where only the lower or upper specifications are used. Using one spec limit is called *unilateral* or one-sided. The corresponding capability indices are

One-sided specifications and the corresponding capability indices

$$C_{pu} = \frac{\text{allowable upper spread}}{\text{actual upper spread}} = \frac{USL - \mu}{3\sigma}$$

and

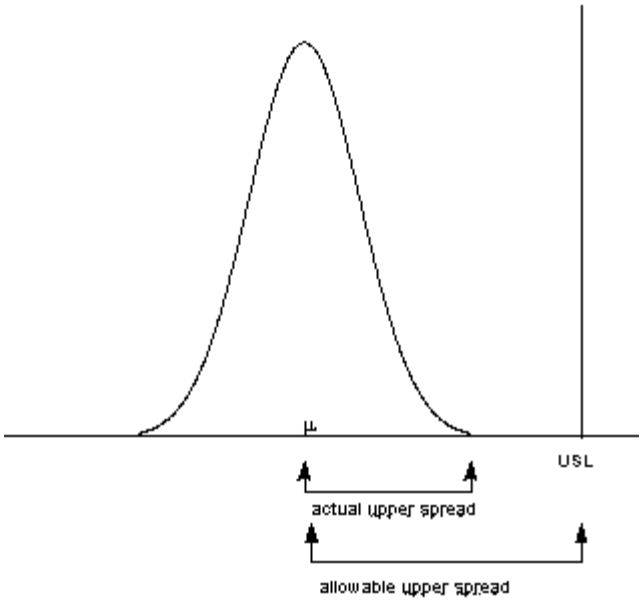
$$C_{pl} = \frac{\text{allowable lower spread}}{\text{actual lower spread}} = \frac{\mu - LSL}{3\sigma}$$

where μ and σ are the process mean and standard deviation, respectively.

Estimators of C_{pu} and C_{pl} are obtained by replacing μ and σ by $\bar{x}$ and s, respectively. The following relationship holds

$C_p = (C_{pu} + C_{pl})$ /2.

This can be represented pictorially by



Note that we also can write:

$C_{pk} = \min \{C_{pl}, C_{pu}\}$.

Confidence Limits For Capability Indices

Confidence intervals for indices

Assuming normally distributed process data, the distribution of the sample $\hat{C}_p$ follows from a Chi-square distribution and $\hat{C}_{pu}$ and $\hat{C}_{pl}$ have distributions related to the non-central *t* distribution. Fortunately, approximate confidence limits related to the normal distribution have been derived. Various approximations to the distribution of $\hat{C}_{pk}$ have been proposed, including those given by Bissell (1990), and we will use a normal approximation.

The resulting formulas for confidence limits are given below:

100(1-α)% Confidence Limits for C_p

$$Pr\{\hat{C}_p(L_1) \leq C_p \leq \hat{C}_p(L_2)\} = 1 - \alpha$$

where

$$L_1 = \sqrt{\frac{\chi^2_{\alpha/2,\nu}}{\nu}} \qquad L_2 = \sqrt{\frac{\chi^2_{1-\alpha/2,\nu}}{\nu}}$$

v = degrees of freedom.

Confidence Intervals for C_{pu} and C_{pl}

Approximate 100(1-α)% confidence limits for C_{pu} with sample size n are:

$$C_{pu}(lower) = \hat{C}_{pu} - z_{1-\beta}\sqrt{\frac{1}{9n} + \frac{\hat{C}_{pu}^2}{2(n-1)}}$$

$$C_{pu}(upper) = \hat{C}_{pu} + z_{1-\alpha}\sqrt{\frac{1}{9n} + \frac{\hat{C}_{pu}^2}{2(n-1)}}$$

with z denoting the percent point function of the standard normal distribution. If β is not known, set it to α.

Limits for C_{pl} are obtained by replacing $\hat{C}_{pu}$ by $\hat{C}_{pl}$.

Confidence Interval for C_{pk}

Zhang et al. (1990) derived the exact variance for the estimator of C_{pk} as well as an approximation for large n. The reference paper is Zhang, Stenback and Wardrop (1990), "Interval Estimation of the process capability index", *Communications in Statistics: Theory and Methods*, 19(21), 4455-4470.

The variance is obtained as follows:

Let

$$c = \sqrt{n}[\mu - (USL + LSL)/2]\sigma$$

$$d = (USL - LSL)/\sigma$$

$$\Phi(-c) = \int_{-inf}^{-c} \frac{1}{\sqrt{2\pi}} \exp -5z^2 dz$$

Then

$$Var(\hat{C}_{pk})$$

$$= (d^2/36)(n-1)(n-3)$$

$$-(d/9\sqrt{n})(n-1)(n-3)\{\sqrt{2\pi}\exp(-c^2/2) + c[1-2\Phi(-c)]\}$$

$$+[(1/9)(n-1)/(n(n-3))](1+c^2)$$

$$-[(n-1)/(72n)]\{\frac{\Gamma((n-2)/2)}{\Gamma((n-1)/2)}\}^2$$

$$*\{d\sqrt{n} - 2\sqrt{2\pi}\exp(-c^2/2) - 2c[1-2\Phi(-c)]\}^2$$

Their approximation is given by:

$$Var(\hat{C}_{pk}) = \frac{n-1}{n-3} - 0.5\{\frac{\Gamma((n-2)/2)}{\Gamma((n-1)/2)}\}^2$$

where

$$n \geq 25, 0.75 \leq C_{pk} \leq 4, \quad |c| \leq 100, \text{ and } \quad d \leq 24$$

The following approximation is commonly used in practice

$$C_{pk} = \hat{C}_{pk} \pm z_{1-\alpha/2}\sqrt{\frac{1}{9n} + \frac{\hat{C}_{pk}^2}{2(n-1)}}$$

It is important to note that the sample size should be at least 25 before these approximations are valid. In general, however, we need $n \geq 100$ for capability studies. Another point to observe is that variations are not negligible due to the randomness of capability indices.

Capability Index Example

An example

For a certain process the USL = 20 and the LSL = 8. The observed process average, $\overline{X}$ = 16, and the standard deviation, *s* = 2. From this we obtain

$$\hat{C}_p = \frac{USL - LSL}{6s} = \frac{20-8}{6(2)} = 1.0$$

This means that the process is capable as long as it is located at the midpoint, m = (USL + LSL)/2 = 14.

But it doesn't, since $\bar{x}$ = 16. The $\hat{k}$ factor is found by

$$\hat{k} = \frac{|m - \bar{x}|}{(USL - LSL)/2} = \frac{2}{6} = 0.3333$$

and

$$\hat{C}_{pk} = \hat{C}_p(1 - \hat{k}) = 0.6667$$

We would like to have $\hat{C}_{pk}$ at least 1.0, so this is not a good process. If possible, reduce the variability or/and center the process. We can compute the $\hat{C}_{pu}$ and $\hat{C}_{pl}$

$$\hat{C}_{pu} = \frac{USL - \bar{x}}{3s} = \frac{20-16}{3(2)} = 0.6667$$

$$\hat{C}_{pl} = \frac{\bar{x} - LSL}{3s} = \frac{16-8}{3(2)} = 1.3333$$

From this we see that the $\hat{C}_{pu}$, which is the smallest of the above indices, is 0.6667. Note that the formula $\hat{C}_{pk} = \hat{C}_p(1 - \hat{k})$ is the algebraic equivalent of the *min*{ $\hat{C}_{pu}$, $\hat{C}_{pl}$ } definition.

What happens if the process is not approximately normally distributed?

What you can

The indices that we considered thus far are based on normality of the process

do with non-normal data

distribution. This poses a problem when the process distribution is not normal. Without going into the specifics, we can list some remedies.

1. Transform the data so that they become approximately normal. A popular transformation is the Box-Cox transformation

2. Use or develop another set of indices, that apply to nonnormal distributions. One statistic is called C_{npk} (for non-parametric C_{pk}). Its estimator is calculated by

$$\hat{C}_{npk} = \min\left[\frac{USL - median}{p(.995) - median}, \frac{median - LSL}{median - p(.005)}\right]$$

   where $p(0.995)$ is the 99.5th percentile of the data and $p(.005)$ is the 0.5th percentile of the data.

   For additional information on nonnormal distributions, see Johnson and Kotz (1993).

There is, of course, much more that can be said about the case of nonnormal data. However, if a Box-Cox transformation can be successfully performed, one is encouraged to use it.

6. Process or Product Monitoring and Control

6.2. Test Product for Acceptability: Lot Acceptance Sampling

This section describes how to make decisions on a lot-by-lot basis whether to accept a lot as likely to meet requirements or reject the lot as likely to have too many defective units.

Contents of section 2

This section consists of the following topics.

1. What is Acceptance Sampling?
2. What kinds of Lot Acceptance Sampling Plans (LASPs) are there?
3. How do you Choose a Single Sampling Plan?
    1. Choosing a Sampling Plan: MIL Standard 105D
    2. Choosing a Sampling Plan with a given OC Curve
4. What is Double Sampling?
5. What is Multiple Sampling?
6. What is a Sequential Sampling Plan?
7. What is Skip Lot Sampling?

6.2.1. What is Acceptance Sampling?

Contributions of Dodge and Romig to acceptance sampling

Acceptance sampling is an important field of statistical quality control that was popularized by Dodge and Romig and originally applied by the U.S. military to the testing of bullets during World War II. If every bullet was tested in advance, no bullets would be left to ship. If, on the other hand, none were tested, malfunctions might occur in the field of battle, with potentially disastrous results.

Definintion of Lot Acceptance Sampling

Dodge reasoned that a sample should be picked at random from the lot, and on the basis of information that was yielded by the sample, a decision should be made regarding the disposition of the lot. In general, the decision is either to accept or reject the lot. This process is called *Lot Acceptance Sampling* or just *Acceptance Sampling*.

"Attributes" (i.e., defect counting) will be assumed

Acceptance sampling is "the middle of the road" approach between no inspection and 100% inspection. There are two major classifications of acceptance plans: by *attributes* ("go, no-go") and by *variables*. The attribute case is the most common for acceptance sampling, and will be assumed for the rest of this section.

Important point

A point to remember is that the main purpose of acceptance sampling is to decide whether or not the lot is likely to be acceptable, not to estimate the quality of the lot.

Scenarios leading to acceptance sampling

Acceptance sampling is employed when one or several of the following hold:

- Testing is destructive
- The cost of 100% inspection is very high
- 100% inspection takes too long

Acceptance Quality Control and Acceptance Sampling

It was pointed out by Harold Dodge in 1969 that Acceptance Quality Control is not the same as Acceptance Sampling. The latter depends on specific sampling plans, which when implemented indicate the conditions for acceptance or rejection of the immediate lot that is being inspected. The former may be implemented in the form of an Acceptance Control Chart. The control limits for the Acceptance Control Chart are computed using the

specification limits and the standard deviation of what is being monitored (see Ryan, 2000 for details).

An observation by Harold Dodge

In 1942, Dodge stated:

"....basically the "acceptance quality control" system that was developed encompasses the concept of protecting the consumer from getting unacceptable defective product, and encouraging the producer in the use of process quality control by: varying the quantity and severity of acceptance inspections in direct relation to the importance of the characteristics inspected, and in the inverse relation to the goodness of the quality level as indication by those inspections."

To reiterate the difference in these two approaches: acceptance sampling plans are one-shot deals, which essentially test short-run effects. Quality control is of the long-run variety, and is part of a well-designed system for lot acceptance.

An observation by Ed Schilling

Schilling (1989) said:

"An individual sampling plan has much the effect of a lone sniper, while the sampling plan scheme can provide a fusillade in the battle for quality improvement."

Control of product quality using acceptance control charts

According to the ISO standard on acceptance control charts (ISO 7966, 1993), an acceptance control chart combines consideration of control implications with elements of acceptance sampling. It is an appropriate tool for helping to make decisions with respect to process acceptance. The difference between acceptance sampling approaches and acceptance control charts is the emphasis on process acceptability rather than on product disposition decisions.

6.2.2. What kinds of Lot Acceptance Sampling Plans (LASPs) are there?

LASP is a sampling scheme and a set of rules

A lot acceptance sampling plan (LASP) is a sampling scheme and a set of rules for making decisions. The decision, based on counting the number of defectives in a sample, can be to accept the lot, reject the lot, or even, for multiple or sequential sampling schemes, to take another sample and then repeat the decision process.

Types of acceptance plans to choose from

LASPs fall into the following categories:

- **Single sampling plans:**. One sample of items is selected at random from a lot and the disposition of the lot is determined from the resulting information. These plans are usually denoted as (n,c) plans for a sample size n, where the lot is rejected if there are more than c defectives. *These are the most common (and easiest) plans to use although not the most efficient in terms of average number of samples needed.*

- **Double sampling plans:** After the first sample is tested, there are three possibilities:
  1. Accept the lot
  2. Reject the lot
  3. No decision

  If the outcome is (3), and a second sample is taken, the procedure is to combine the results of both samples and make a final decision based on that information.

- **Multiple sampling plans:** This is an extension of the double sampling plans where more than two samples are needed to reach a conclusion. The advantage of multiple sampling is smaller sample sizes.

- **Sequential sampling plans:** . This is the ultimate extension of multiple sampling where items are selected from a lot one at a time and after inspection of each item a decision is made to accept or reject the lot or select another unit.

- **Skip lot sampling plans:**. Skip lot sampling means that only a fraction of the submitted lots are inspected.

Definitions of basic Acceptance Sampling terms

Deriving a plan, within one of the categories listed above, is discussed in the pages that follow. All derivations depend on the properties you want the plan to have. These are described using the following terms:

- ***Acceptable Quality Level (AQL)***: The AQL is a percent defective that is the base line requirement for the quality of the producer's product. The producer would like to design a sampling plan such that there is a *high probability of accepting* a lot that has a defect level less than or equal to the AQL.

- ***Lot Tolerance Percent Defective (LTPD)***: The LTPD is a designated high defect level that would be unacceptable to the consumer. The consumer would like the sampling plan to have a *low probability of accepting* a lot with a defect level as high as the LTPD.

- ***Type I Error (Producer's Risk)***: This is the probability, for a given (n,c) sampling plan, of rejecting a lot that has a defect level equal to the AQL. The producer suffers when this occurs, because a lot with acceptable quality was rejected. The symbol α is commonly used for the Type I error and typical values for α range from 0.2 to 0.01.

- ***Type II Error (Consumer's Risk):*** This is the probability, for a given (n,c) sampling plan, of accepting a lot with a defect level equal to the LTPD. The consumer suffers when this occurs, because a lot with unacceptable quality was accepted. The symbol β is commonly used for the Type II error and typical values range from 0.2 to 0.01.

- ***Operating Characteristic (OC) Curve:*** This curve plots the probability of accepting the lot (Y-axis) versus the lot fraction or percent defectives (X-axis). *The OC curve is the primary tool for displaying and investigating the properties of a LASP*.

- ***Average Outgoing Quality (AOQ):*** A common procedure, when sampling and testing is non-destructive, is to 100% inspect rejected lots and replace all defectives with good units. In this case, all rejected lots are made perfect and the only defects left are those in lots that were accepted. *AOQ*'s refer to the long term defect level for this combined LASP and 100% inspection of rejected lots process. If all lots come in with a defect level of exactly p, and the OC curve for the chosen (n,c) LASP indicates a probability p_a of accepting such a lot, over the long run the *AOQ* can easily be shown to be:

$$AOQ = \frac{p_a p(N-n)}{N}$$

where N is the lot size.

- ***Average Outgoing Quality Level (AOQL):*** A plot of the *AOQ* (Y-axis) versus the incoming lot p (X-axis) will start at 0 for $p = 0$, and return to 0 for $p = 1$ (where every lot is 100% inspected and rectified). In between, it will rise to a maximum. This maximum, which is the worst possible long term *AOQ*, is called the *AOQL*.

- ***Average Total Inspection (ATI):*** When rejected lots are 100% inspected, it is easy to calculate the *ATI* if lots come consistently with a defect level of p. For a LASP (n,c) with a probability p_a of accepting a lot with defect level p, we have

  $$ATI = n + (1 - p_a)(N - n)$$

  where N is the lot size.

- ***Average Sample Number (ASN):*** For a single sampling LASP (n,c) we know each and every lot has a sample of size n taken and inspected or tested. For double, multiple and sequential LASP's, the amount of sampling varies depending on the number of defects observed. For any given double, multiple or sequential plan, a long term *ASN* can be calculated assuming all lots come in with a defect level of p. A plot of the *ASN*, versus the incoming defect level p, describes the sampling efficiency of a given LASP scheme.

The final choice is a tradeoff decision

Making a final choice between single or multiple sampling plans that have acceptable properties is a matter of deciding whether the average sampling savings gained by the various multiple sampling plans justifies the additional complexity of these plans and the uncertainty of not knowing how much sampling and inspection will be done on a day-by-day basis.

6.2.3. How do you Choose a Single Sampling Plan?

Two methods for choosing a single sample acceptance plan

A single sampling plan, as previously defined, is specified by the pair of numbers (n,c). The sample size is n, and the lot is rejected if there are more than c defectives in the sample; otherwise the lot is accepted.

There are two widely used ways of picking (n,c):

1. Use tables (such as MIL STD 105D) that focus on either the *AQL* or the *LTPD* desired.

2. Specify 2 desired points on the OC curve and solve for the (n,c) that uniquely determines an OC curve going through these points.

The next two pages describe these methods in detail.

6. Process or Product Monitoring and Control
6.2. Test Product for Acceptability: Lot Acceptance Sampling
6.2.3. How do you Choose a Single Sampling Plan?

6.2.3.1. Choosing a Sampling Plan: MIL Standard 105D

The AQL or Acceptable Quality Level is the baseline requirement

Sampling plans are typically set up with reference to an acceptable quality level, or *AQL* . The *AQL* is the base line requirement for the quality of the producer's product. The producer would like to design a sampling plan such that the OC curve yields a high probability of acceptance at the *AQL*. On the other side of the OC curve, the consumer wishes to be protected from accepting poor quality from the producer. So the consumer establishes a criterion, the *lot tolerance percent defective* or *LTPD* . Here the idea is to only accept poor quality product with a very low probability. Mil. Std. plans have been used for over 50 years to achieve these goals.

The U.S. Department of Defense Military Standard 105E

Military Standard 105E sampling plan

Standard military sampling procedures for inspection by attributes were developed during World War II. Army Ordnance tables and procedures were generated in the early 1940's and these grew into the Army Service Forces tables. At the end of the war, the Navy also worked on a set of tables. In the meanwhile, the Statistical Research Group at Columbia University performed research and outputted many outstanding results on attribute sampling plans.

These three streams combined in 1950 into a standard called Mil. Std. 105A. It has since been modified from time to time and issued as 105B, 195C and 105D. Mil. Std. 105D was issued by the U.S. government in 1963. It was adopted in 1971 by the American National Standards Institute as ANSI Standard Z1.4 and in 1974 it was adopted (with minor changes) by the International Organization for Standardization as ISO Std. 2859. The latest revision is Mil. Std 105E and was issued in 1989.

These three similar standards are continuously being updated and revised, but the basic tables remain the same. Thus the discussion that follows of the germane aspects of Mil. Std. 105E also applies to the other two standards.

Description of Mil. Std. 105D

Military Standard 105D sampling plan

This document is essentially a set of individual plans, organized in a system of sampling schemes. A sampling scheme consists of a combination of a normal sampling plan, a tightened sampling plan, and a reduced sampling plan plus rules for switching from one to the other.

AQL is foundation of standard

The foundation of the Standard is the acceptable quality level or *AQL*. In the following scenario, a certain military agency, called the Consumer from here on, wants to purchase a particular product from a supplier, called the Producer from here on.

In applying the Mil. Std. 105D it is expected that there is perfect agreement between Producer and Consumer regarding what the *AQL* is for a given product characteristic. It is understood by both parties that the Producer will be submitting for inspection a number of lots whose quality level is typically as good as specified by the Consumer. Continued quality is assured by the acceptance or rejection of lots following a particular sampling plan and also by providing for a shift to another, tighter sampling plan, when there is evidence that the Producer's product does not meet the agreed-upon *AQL*.

Standard offers 3 types of sampling plans

Mil. Std. 105E offers three types of sampling plans: single, double and multiple plans. The choice is, in general, up to the inspectors.

Because of the three possible selections, the standard does not give a sample size, but rather a sample code letter. This, together with the decision of the type of plan yields the specific sampling plan to be used.

Inspection level

In addition to an initial decision on an *AQL* it is also necessary to decide on an "inspection level". This determines the relationship between the lot size and the sample size. The standard offers three general and four special levels.

Steps in the standard

The steps in the use of the standard can be summarized as follows:

1. Decide on the *AQL*.
2. Decide on the inspection level.
3. Determine the lot size.
4. Enter the table to find sample size code letter.
5. Decide on type of sampling to be used.
6. Enter proper table to find the plan to be used.
7. Begin with normal inspection, follow the switching rules and the rule for stopping the inspection (if needed).

Additional

There is much more that can be said about Mil. Std. 105E,

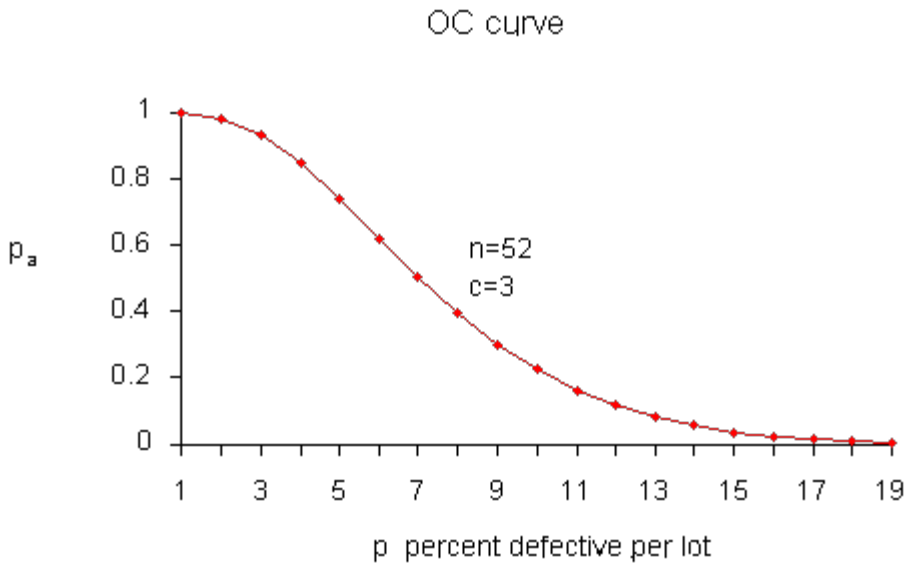
information (and 105D). The interested reader is referred to references such as (Montgomery (2000), Schilling, tables 11-2 to 11-17, and Duncan, pages 214 - 248).

There is also (currently) a web site developed by Galit Shmueli that will develop sampling plans interactively with the user, according to Military Standard 105E (ANSI/ASQC Z1.4, ISO 2859) Tables.

6.2.3.2. Choosing a Sampling Plan with a given OC Curve

Sample OC curve

We start by looking at a typical OC curve. The OC curve for a (52 ,3) sampling plan is shown below.



Number of defectives is approximately binomial

It is instructive to show how the points on this curve are obtained, once we have a sampling plan (*n*,*c*) - later we will demonstrate how a sampling plan (*n*,*c*) is obtained.

We assume that the lot size *N* is very large, as compared to the sample size *n*, so that removing the sample doesn't significantly change the remainder of the lot, no matter how many defects are in the sample. Then the distribution of the number of defectives, *d*, in a random sample of *n* items is approximately binomial with parameters *n* and *p*, where *p* is the fraction of defectives per lot.

The probability of observing exactly *d* defectives is given by

The binomial distribution

$$P(d) = f(d) = \frac{n!}{d!(n-d)!} p^d (1-p)^{n-d}$$

The probability of acceptance is the probability that *d*, the number of defectives, is less than or equal to *c*, the accept number. This means that

$$P_a = P\{d \le c\} = \sum_{d=0}^{c} \frac{n!}{d!(n-d)!} p^d (1-p)^{n-d}$$

Sample table for Pa, Pd using the binomial distribution

Using this formula with n = 52 and c=3 and p = .01, .02, ...,.12 we find

P_a	P_d
.998	.01
.980	.02
.930	.03
.845	.04
.739	.05
.620	.06
.502	.07
.394	.08
.300	.09
.223	.10
.162	.11
.115	.12

Solving for (*n*,*c*)

Equations for calculating a sampling plan with a given OC curve

In order to design a sampling plan with a specified OC curve one needs two designated points. Let us design a sampling plan such that the probability of acceptance is 1-α for lots with fraction defective p_1 and the probability of acceptance is β for lots with fraction defective p_2. Typical choices for these points are: p_1 is the *AQL*, p_2 is the *LTPD* and α, β are the Producer's Risk (Type I error) and Consumer's Risk (Type II error), respectively.

If we are willing to assume that binomial sampling is valid, then the sample size n, and the acceptance number c are the solution to

$$1-\alpha = \sum_{d=0}^{c} \frac{n!}{d!(n-d)!} p_1^d (1-p_1)^{n-d}$$

$$\beta = \sum_{d=0}^{c} \frac{n!}{d!(n-d)!} p_2^d (1-p_2)^{n-d}$$

These two simultaneous equations are nonlinear so there is no simple, direct solution. There are however a number of iterative techniques available that give approximate solutions so that composition of a computer program poses few problems.

Average Outgoing Quality (AOQ)

Calculating AOQ's

We can also calculate the *AOQ* for a (n,c) sampling plan, provided rejected lots are 100% inspected and defectives are replaced with good parts.

Assume all lots come in with exactly a p_0 proportion of defectives. After screening a rejected lot, the final fraction defectives will be zero for that lot. However, accepted lots have fraction defectivep_0. Therefore, the outgoing lots from the inspection stations are a mixture of lots with fractions defective p_0 and 0. Assuming the lot size is N, we have.

$$AOQ = \frac{p_a p(N-n)}{N}$$

For example, let N = 10000, n = 52, c = 3, and p, the quality of incoming lots, = 0.03. Now at p = 0.03, we glean from the OC curve table that p_a = 0.930 and

$$AOQ = (.930)(.03)(10000\text{-}52) / 10000 = 0.02775.$$

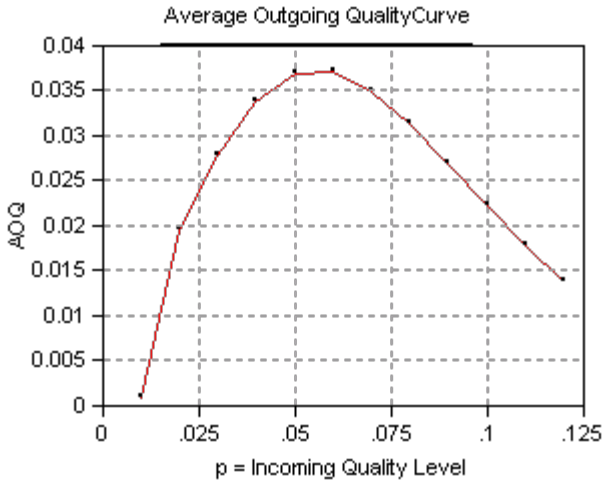
Sample table of AOQ versus p

Setting p = .01, .02, ..., .12, we can generate the following table

AOQ	p
.0010	.01
.0196	.02
.0278	.03
.0338	.04
.0369	.05
.0372	.06
.0351	.07
.0315	.08
.0270	.09
.0223	.10
.0178	.11
.0138	.12

Sample plot of AOQ versus p

A plot of the *AOQ* versus p is given below.



Interpretation of AOQ plot

From examining this curve we observe that when the incoming quality is very good (very small fraction of defectives coming in), then the outgoing quality is also very good (very small fraction of defectives going out). When the incoming lot quality is very bad, most of the lots are rejected and then inspected. The "duds" are eliminated or replaced by good ones, so that the quality of the outgoing lots, the *AOQ*, becomes very good. In between these extremes, the *AOQ* rises, reaches a maximum, and then drops.

The maximum ordinate on the *AOQ* curve represents the worst possible quality that results from the rectifying inspection program. It is called the **average outgoing quality limit,** (***AOQL***).

From the table we see that the *AOQL* = 0.0372 at p = .06 for the above example.

One final remark: if $N >> n$, then the $AOQ \sim p_a\, p$.

The Average Total Inspection (ATI)

Calculating the Average Total Inspection

What is the total amount of inspection when rejected lots are screened?

If all lots contain zero defectives, no lot will be rejected.

If all items are defective, all lots will be inspected, and the amount to be inspected is N.

Finally, if the lot quality is $0 < p < 1$, the average amount of inspection per lot will vary between the sample size n, and the lot size N.

Let the quality of the lot be p and the probability of lot acceptance be p_a, then the *ATI* per lot is

$$ATI = n + (1 - p_a)(N - n)$$

For example, let $N = 10000$, $n = 52$, $c = 3$, and $p = .03$ We know from the OC table that $p_a = 0.930$. Then $ATI = 52 + (1-.930)(10000 - 52) = 753$. (Note that while 0.930 was rounded to three decimal places, 753 was obtained using more decimal places.)

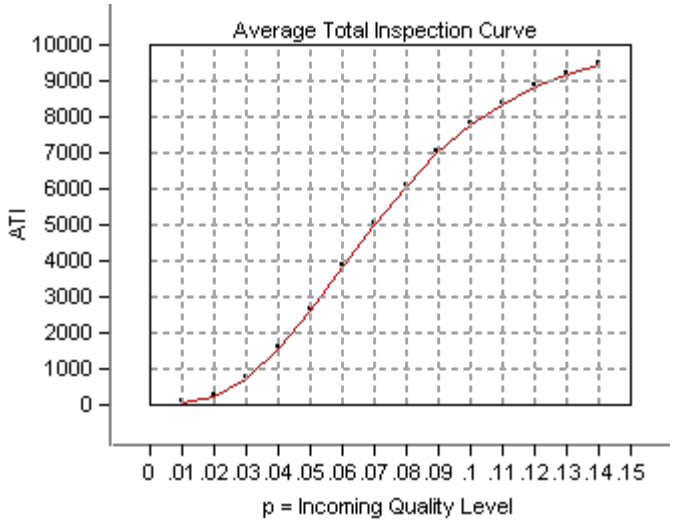
Sample table of ATI versus p

Setting p= .01, .02, ...14 generates the following table

ATI	P
70	.01
253	.02
753	.03
1584	.04
2655	.05
3836	.06
5007	.07
6083	.08
7012	.09
7779	.10
8388	.11
8854	.12
9201	.13
9453	.14

Plot of ATI versus p

A plot of *ATI* versus *p*, the Incoming Lot Quality (*ILQ*) is given below.



6.2.4. What is Double Sampling?

Double Sampling Plans

How double sampling plans work

Double and multiple sampling plans were invented to give a questionable lot another chance. For example, if in double sampling the results of the first sample are not conclusive with regard to accepting or rejecting, a second sample is taken. Application of double sampling requires that a first sample of size n_1 is taken at random from the (large) lot. The number of defectives is then counted and compared to the first sample's acceptance number a_1 and rejection number r_1. Denote the number of defectives in sample 1 by d_1 and in sample 2 by d_2, then:

If $d_1 \leq a_1$, the lot is accepted.
If $d_1 \geq r_1$, the lot is rejected.
If $a_1 < d_1 < r_1$, a second sample is taken.

If a second sample of size n_2 is taken, the number of defectives, d_2, is counted. The total number of defectives is $D_2 = d_1 + d_2$. Now this is compared to the acceptance number a_2 and the rejection number r_2 of sample 2. In double sampling, $r_2 = a_2 + 1$ to ensure a decision on the sample.

If $D_2 \leq a_2$, the lot is accepted.
If $D_2 \geq r_2$, the lot is rejected.

Design of a Double Sampling Plan

Design of a double sampling plan

The parameters required to construct the OC curve are similar to the single sample case. The two points of interest are (p_1, 1-α) and (p_2, β), where p_1 is the lot fraction defective for plan 1 and p_2 is the lot fraction defective for plan 2. As far as the respective sample sizes are concerned, the second sample size must be equal to, or an even multiple of, the first sample size.

There exist a variety of tables that assist the user in constructing double and multiple sampling plans. The index to these tables is the p_2/p_1 ratio, where $p_2 > p_1$. One set of tables, taken from the Army Chemical Corps Engineering Agency for α = .05 and β = .10, is

given below:

Tables for $n_1 = n_2$

$R = p_2/p_1$	accept numbers c_1	c_2	approximation of pn_1 P = .95	values for P = .10
11.90	0	1	0.21	2.50
7.54	1	2	0.52	3.92
6.79	0	2	0.43	2.96
5.39	1	3	0.76	4.11
4.65	2	4	1.16	5.39
4.25	1	4	1.04	4.42
3.88	2	5	1.43	5.55
3.63	3	6	1.87	6.78
3.38	2	6	1.72	5.82
3.21	3	7	2.15	6.91
3.09	4	8	2.62	8.10
2.85	4	9	2.90	8.26
2.60	5	11	3.68	9.56
2.44	5	12	4.00	9.77
2.32	5	13	4.35	10.08
2.22	5	14	4.70	10.45
2.12	5	16	5.39	11.41

Tables for $n_2 = 2n_1$

R = p_2/p_1	accept numbers c_1	c_2	approximation of pn_1 P = .95	values for P = .10
14.50	0	1	0.16	2.32
8.07	0	2	0.30	2.42
6.48	1	3	0.60	3.89
5.39	0	3	0.49	2.64
5.09	0	4	0.77	3.92
4.31	1	4	0.68	2.93
4.19	0	5	0.96	4.02
3.60	1	6	1.16	4.17
3.26	1	8	1.68	5.47
2.96	2	10	2.27	6.72
2.77	3	11	2.46	6.82
2.62	4	13	3.07	8.05
2.46	4	14	3.29	8.11
2.21	3	15	3.41	7.55

1.97	4	20	4.75	9.35
1.74	6	30	7.45	12.96

Example

Example of a double sampling plan

We wish to construct a double sampling plan according to

$p_1 = 0.01 \quad \alpha = 0.05 \quad p_2 = 0.05 \quad \beta = 0.10 \quad$ and $n_1 = n_2$

The plans in the corresponding table are indexed on the ratio

$R = p_2/p_1 = 5$

We find the row whose R is closet to 5. This is the 5th row (R = 4.65). This gives $c_1 = 2$ and $c_2 = 4$. The value of n_1 is determined from either of the two columns labeled pn_1.

The left holds α constant at 0.05 ($P = 0.95 = 1 - \alpha$) and the right holds β constant at 0.10. ($P = 0.10$). Then holding α constant we find $pn_1 = 1.16$ so $n_1 = 1.16/p_1 = 116$. And, holding β constant we find $pn_1 = 5.39$, so $n_1 = 5.39/p_2 = 108$. Thus the desired sampling plan is

$n_1 = 108 \quad c_1 = 2 \quad n_2 = 108 \quad c_2 = 4$

If we opt for $n_2 = 2n_1$, and follow the same procedure using the appropriate table, the plan is:

$n_1 = 77 \quad c_1 = 1 \quad n_2 = 154 \quad c_2 = 4$

The first plan needs less samples if the number of defectives in sample 1 is greater than 2, while the second plan needs less samples if the number of defectives in sample 1 is less than 2.

ASN Curve for a Double Sampling Plan

Construction of the ASN curve

Since when using a double sampling plan the sample size depends on whether or not a second sample is required, an important consideration for this kind of sampling is the Average Sample Number (*ASN*) curve. This curve plots the *ASN* versus p', the true fraction defective in an incoming lot.

We will illustrate how to calculate the *ASN* curve with an example. Consider a double-sampling plan $n_1 = 50$, $c_1 = 2$, $n_2 = 100$, $c_2 = 6$, where n_1 is the sample size for plan 1, with accept number c_1, and n_2, c_2, are the sample size and accept number, respectively, for plan 2.

Let $p' = .06$. Then the probability of acceptance on the first sample, which is the chance of getting two or less defectives, is .416 (using

binomial tables). The probability of rejection on the second sample, which is the chance of getting more than six defectives, is (1-.971) = .029. The probability of making a decision on the first sample is .445, equal to the sum of .416 and .029. With complete inspection of the second sample, the *average* size sample is equal to the size of the first sample times the probability that there will be only one sample plus the size of the combined samples times the probability that a second sample will be necessary. For the sampling plan under consideration, the *ASN* with complete inspection of the second sample for a p' of .06 is

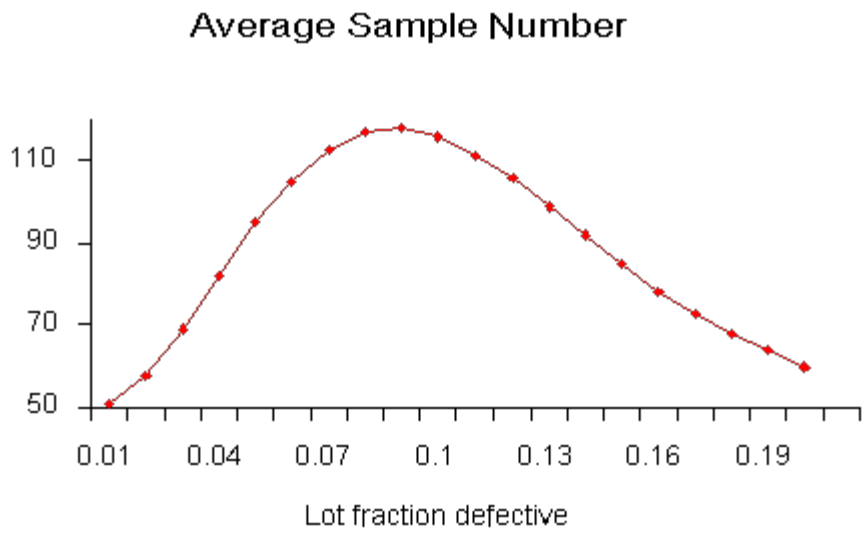
$$50(.445) + 150(.555) = 106$$

The general formula for an average sample number curve of a double-sampling plan with complete inspection of the second sample is

$$ASN = n_1P_1 + (n_1 + n_2)(1 - P_1) = n_1 + n_2(1 - P_1)$$

where P_1 is the probability of a decision on the first sample. The graph below shows a plot of the *ASN* versus p'.

The ASN curve for a double sampling plan



6.2.5. What is Multiple Sampling?

Multiple Sampling is an extension of the double sampling concept

Multiple sampling is an extension of double sampling. It involves inspection of 1 to k successive samples as required to reach an ultimate decision.

Mil-Std 105D suggests $k = 7$ is a good number. Multiple sampling plans are usually presented in tabular form:

Procedure for multiple sampling

The procedure commences with taking a random sample of size n_1 from a large lot of size N and counting the number of defectives, d_1.

if $d_1 \leq a_1$ the lot is accepted.
if $d_1 \geq r_1$ the lot is rejected.
if $a_1 < d_1 < r_1$, another sample is taken.

If subsequent samples are required, the first sample procedure is repeated sample by sample. For each sample, the total number of defectives found at any stage, say stage i, is

$$D_i = \sum_{j=1}^{i} d_j$$

This is compared with the acceptance number a_i and the rejection number r_i for that stage until a decision is made. Sometimes acceptance is not allowed at the early stages of multiple sampling; however, rejection can occur at any stage.

Efficiency measured by the ASN

Efficiency for a multiple sampling scheme is measured by the *average sample number (ASN)* required for a given Type I and Type II set of errors. The number of samples needed when following a multiple sampling scheme may vary from trial to trial, and the *ASN* represents the average of what might happen over many trials with a fixed incoming defect level.

6.2.6. What is a Sequential Sampling Plan?

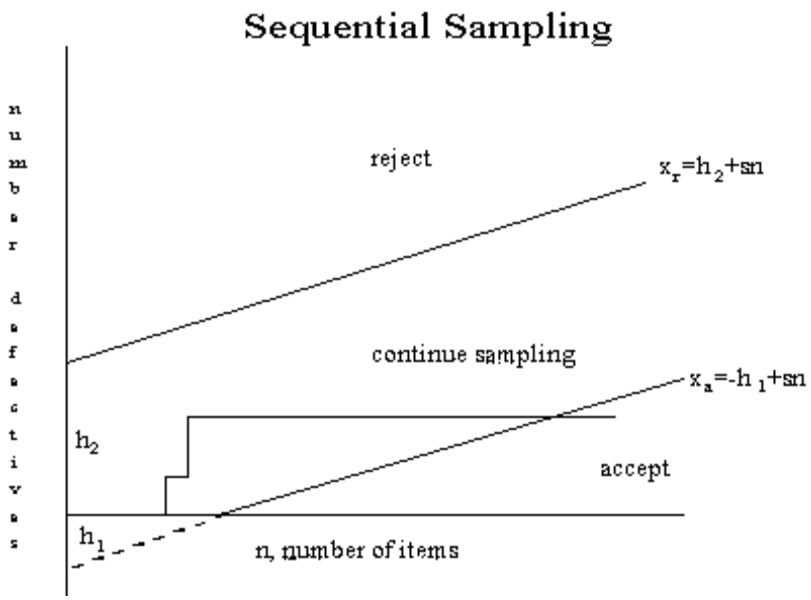
Sequential Sampling

Sequential sampling is different from single, double or multiple sampling. Here one takes a sequence of samples from a lot. How many total samples looked at is a function of the results of the sampling process.

Item-by-item and group sequential sampling

The sequence can be one sample at a time, and then the sampling process is usually called *item-by-item* sequential sampling. One can also select sample sizes greater than one, in which case the process is referred to as *group* sequential sampling. Item-by-item is more popular so we concentrate on it. The operation of such a plan is illustrated below:

Diagram of item-by-item sampling



Description of sequentail sampling graph

The cumulative observed number of defectives is plotted on the graph. For each point, the x-axis is the total number of items thus far selected, and the y-axis is the total number of observed defectives. If the plotted point falls within the parallel lines the process continues by drawing another sample. As soon as a point falls on or above the upper line, the lot is rejected. And when a point falls on or below the lower line, the lot is accepted. The process can theoretically last until the lot is 100% inspected. However, as a rule of thumb, sequential-sampling plans are truncated after the number inspected reaches three times the number that would have been inspected using a corresponding single sampling plan.

Equations for the limit lines

The equations for the two limit lines are functions of the parameters p_1, α, p_2, and β.

$$x_a = -h_1 + sn \quad \text{(acceptance line)}$$
$$x_r = h_2 + sn \quad \text{(rejection line)}$$

where

$$h_1 = (\log \frac{1-\alpha}{\beta})/k$$
$$h_2 = (\log \frac{1-\beta}{\alpha})/k$$
$$k = \log \frac{p_2(1-p_1)}{p_1(1-p_2)}$$
$$s = (\log[\frac{1-p_1}{1-p_2}])/k$$

Instead of using the graph to determine the fate of the lot, one can resort to generating tables (with the help of a computer program).

Example of a sequential sampling plan

As an example, let p_1 = .01, p_2 = .10, α = .05, β = .10. The resulting equations are

$$x_a = -0.939 + 0.04n$$
$$x_r = 1.205 + 0.04n$$

Both acceptance numbers and rejection numbers must be integers. The acceptance number is the next integer less than or equal to x_a and the rejection number is the next integer greater than or equal to x_r. Thus for n = 1, the acceptance number = -1, which is impossible, and the rejection number = 2, which is also impossible. For n = 24, the acceptance number is 0 and the rejection number = 3.

The results for n =1, 2, 3... 26 are tabulated below.

n inspect	*n* accept	*n* reject	*n* inspect	*n* accept	*n* reject
1	x	x	14	x	2
2	x	2	15	x	2
3	x	2	16	x	3
4	x	2	17	x	3
5	x	2	18	x	3
6	x	2	19	x	3
7	x	2	20	x	3
8	x	2	21	x	3
9	x	2	22	x	3

10	x	2	23	x	3
11	x	2	24	0	3
12	x	2	25	0	3
13	x	2	26	0	3

So, for n = 24 the acceptance number is 0 and the rejection number is 3. The "x" means that acceptance or rejection is not possible.

Other sequential plans are given below.

n inspect	n accept	n reject
49	1	3
58	1	4
74	2	4
83	2	5
100	3	5
109	3	6

The corresponding single sampling plan is (52,2) and double sampling plan is (21,0), (21,1).

Efficiency measured by ASN

Efficiency for a sequential sampling scheme is measured by the *average sample number (ASN)* required for a given Type I and Type II set of errors. The number of samples needed when following a sequential sampling scheme may vary from trial to trial, and the *ASN* represents the average of what might happen over many trials with a fixed incoming defect level. Good software for designing sequential sampling schemes will calculate the *ASN* curve as a function of the incoming defect level.

6.2.7. What is Skip Lot Sampling?

Skip Lot Sampling

Skip Lot sampling means that only a fraction of the submitted lots are inspected. This mode of sampling is of the cost-saving variety in terms of time and effort. However skip-lot sampling should only be used when it has been demonstrated that the quality of the submitted product is very good.

Implementation of skip-lot sampling plan

A skip-lot sampling plan is implemented as follows:

1. Design a single sampling plan by specifying the alpha and beta risks and the consumer/producer's risks. This plan is called "the reference sampling plan".
2. Start with normal lot-by-lot inspection, using the reference plan.
3. When a pre-specified number, *i*, of consecutive lots are accepted, switch to inspecting only a fraction *f* of the lots. The selection of the members of that fraction is done at random.
4. When a lot is rejected return to normal inspection.

The f and i parameters

The parameters *f* and *i* are essential to calculating the probability of acceptance for a skip-lot sampling plan. In this scheme, *i*, called the *clearance number*, is a positive integer and the sampling fraction *f* is such that $0 < f < 1$. Hence, when $f = 1$ there is no longer skip-lot sampling. The calculation of the acceptance probability for the skip-lot sampling plan is performed via the following formula

$$P_a(f,i) = \frac{fP + (1-f)P^i}{f + (1-f)P^i}$$

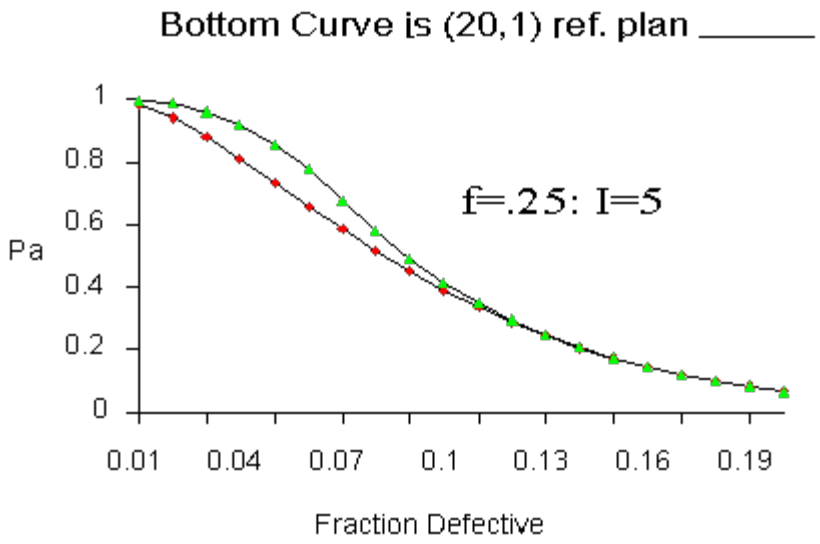
where *P* is the probability of accepting a lot with a given proportion of incoming defectives *p*, from the OC curve of the single sampling plan.

The following relationships hold:

for a given *i*, the smaller is *f*, the greater is P_a
for a given *f*, the smaller is *i*, the greater is P_a

Illustration of a skip lot sampling plan

An illustration of a a skip-lot sampling plan is given below.



ASN of skip-lot sampling plan

An important property of skip-lot sampling plans is the average sample number (*ASN*). The *ASN* of a skip-lot sampling plan is

$$ASN_{skip\text{-}lot} = (F)(ASN_{reference})$$

where F is defined by

$$F = \frac{f}{(1-f)P^i + f}$$

Therefore, since $0 < F < 1$, it follows that the *ASN* of skip-lot sampling is smaller than the *ASN* of the reference sampling plan.

In summary, skip-lot sampling is preferred when the quality of the submitted lots is excellent and the supplier can demonstrate a proven track record.

6.3. Univariate and Multivariate Control Charts

Contents of section 3

Control charts in this section are classified and described according to three general types: variables, attributes and multivariate.

1. What are Control Charts?
2. What are Variables Control Charts?
    1. Shewhart X bar and R and S Control Charts
    2. Individuals Control Charts
    3. Cusum Control Charts
        1. Cusum Average Run Length
    4. EWMA Control Charts
3. What are Attributes Control Charts?
    1. Counts Control Charts
    2. Proportions Control Charts
4. What are Multivariate Control Charts?
    1. Hotelling Control Charts
    2. Principal Components Control Charts
    3. Multivariate EWMA Charts

6.3.1. What are Control Charts?

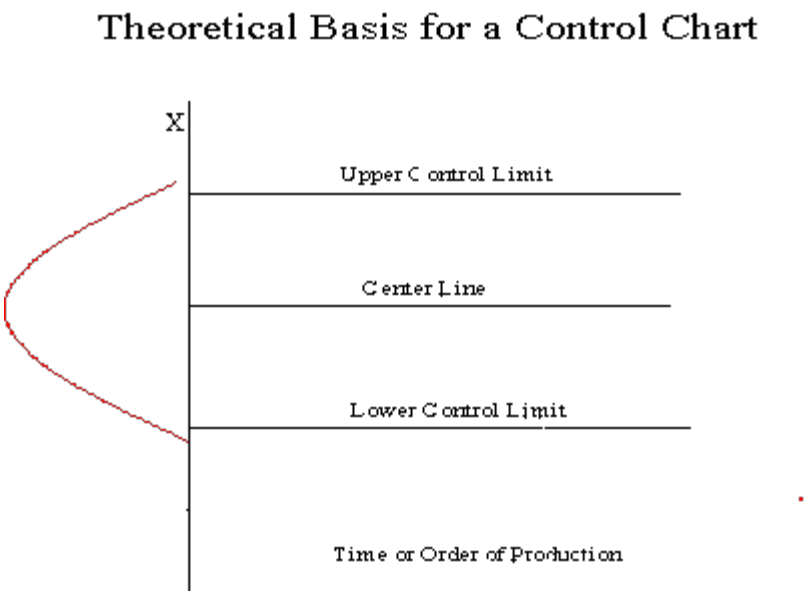
Comparison of univariate and multivariate control data

Control charts are used to routinely monitor quality. Depending on the number of process characteristics to be monitored, there are two basic types of control charts. The first, referred to as a univariate control chart, is a graphical display (chart) of one quality characteristic. The second, referred to as a multivariate control chart, is a graphical display of a statistic that summarizes or represents more than one quality characteristic.

Characteristics of control charts

If a single quality characteristic has been measured or computed from a sample, the control chart shows the value of the quality characteristic versus the sample number or versus time. In general, the chart contains a center line that represents the mean value for the in-control process. Two other horizontal lines, called the upper control limit (UCL) and the lower control limit (LCL), are also shown on the chart. These control limits are chosen so that almost all of the data points will fall within these limits as long as the process remains in-control. The figure below illustrates this.

Chart demonstrating basis of control chart



Why control charts "work"

The control limits as pictured in the graph might be .001 *probability* limits. If so, and if chance causes alone were present, the probability of a point falling above the upper limit would be one out of a thousand, and similarly, a point falling below the lower limit would be one out of a thousand.

We would be searching for an assignable cause if a point would fall outside these limits. Where we put these limits will determine the risk of undertaking such a search when in reality there is no assignable cause for variation.

Since two out of a thousand is a very small risk, the 0.001 limits may be said to give practical assurances that, if a point falls outside these limits, the variation was caused be an assignable cause. It must be noted that two out of one thousand is a purely arbitrary number. There is no reason why it could not have been set to one out a hundred or even larger. The decision would depend on the amount of risk the management of the quality control program is willing to take. In general (in the world of quality control) it is customary to use limits that approximate the 0.002 standard.

Letting X denote the value of a process characteristic, if the system of chance causes generates a variation in X that follows the normal distribution, the 0.001 probability limits will be very close to the 3σ limits. From normal tables we glean that the 3σ in one direction is 0.00135, or in both directions 0.0027. For normal distributions, therefore, the 3σ limits are the practical equivalent of 0.001 probability limits.

Plus or minus "3 sigma" limits are typical

In the U.S., whether X is normally distributed or not, it is an acceptable practice to base the control limits upon a multiple of the standard deviation. Usually this multiple is 3 and thus the limits are called 3-sigma limits. This term is used whether the standard deviation is the universe or population parameter, or some estimate thereof, or simply a "standard value" for control chart purposes. It should be inferred from the context what standard deviation is involved. (Note that in the U.K., statisticians generally prefer to adhere to probability limits.)

If the underlying distribution is skewed, say in the positive direction, the 3-sigma limit will fall short of the upper 0.001 limit, while the lower 3-sigma limit will fall below the 0.001 limit. This situation means that the risk of looking for assignable causes of positive variation when none exists will be greater than one out of a thousand. But the risk of searching for an assignable cause of negative variation, when none exists, will be reduced. The net result, however, will be an increase in the risk of a chance variation beyond the control limits. How much this risk will be increased will depend on the degree of skewness.

If variation in quality follows a Poisson distribution, for example, for which $np = .8$, the risk of exceeding the upper limit by chance would be raised by the use of 3-sigma limits from 0.001 to 0.009 and the lower limit reduces from 0.001 to 0. For a Poisson distribution the mean and variance both equal np. Hence the upper 3-sigma limit is $0.8 + 3\ sqrt(.8) = 3.48$ and the lower limit = 0 (here *sqrt* denotes "square root").

For *np* = .8 the probability of getting more than 3 successes = 0.009.

Strategies for dealing with out-of-control findings

If a data point falls outside the control limits, we assume that the process is probably out of control and that an investigation is warranted to find and eliminate the cause or causes.

Does this mean that when all points fall within the limits, the process is in control? Not necessarily. If the plot looks non-random, that is, if the points exhibit some form of systematic behavior, there is still something wrong. For example, if the first 25 of 30 points fall above the center line and the last 5 fall below the center line, we would wish to know why this is so. Statistical methods to detect sequences or nonrandom patterns can be applied to the interpretation of control charts. To be sure, "in control" implies that all points are between the control limits and they form a random pattern.

6.3.2. What are Variables Control Charts?

During the 1920's, Dr. Walter A. Shewhart proposed a general model for control charts as follows:

Shewhart Control Charts for variables

Let *w* be a sample statistic that measures some continuously varying quality characteristic of interest (e.g., thickness), and suppose that the mean of *w* is μ_w, with a standard deviation of σ_w. Then the center line, the UCL and the LCL are

$$\text{UCL} = \mu_w + k\sigma_w$$
$$\text{Center Line} = \mu_w$$
$$\text{LCL} = \mu_w - k\sigma_w$$

where *k* is the distance of the control limits from the center line, expressed in terms of standard deviation units. When *k* is set to 3, we speak of 3-sigma control charts.

Historically, k = 3 has become an accepted standard in industry.

The centerline is the process mean, which in general is unknown. We replace it with a *target* or the average of all the data. The quantity that we plot is the sample average, $\bar{X}$. The chart is called the $\bar{X}$ chart.

We also have to deal with the fact that σ is, in general, unknown. Here we replace σ_w with a given standard value, or we estimate it by a function of the *average standard deviation*. This is obtained by averaging the individual standard deviations that we calculated from each of *m* preliminary (or present) samples, each of size *n*. This function will be discussed shortly.

It is equally important to examine the standard deviations in ascertaining whether the process is in control. There is, unfortunately, a slight problem involved when we work with the usual estimator of σ. The following discussion will illustrate this.

Sample Variance

If σ^2 is the unknown variance of a probability distribution, then an unbiased estimator of σ^2 is the sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

However, *s*, the sample standard deviation is *not* an unbiased estimator of σ. If the underlying distribution is normal, then *s* actually estimates $c_4 \sigma$, where c_4 is a constant that depends on the sample size *n*. This constant is tabulated in most text books on statistical quality control and may be calculated using

C_4 factor

$$c_4 = \sqrt{\frac{2}{n-1}} \frac{\left(\frac{n}{2}-1\right)!}{\left(\frac{n-1}{2}-1\right)!}$$

To compute this we need a *non-integer factorial*, which is defined for *n*/2 as follows:

Fractional Factorials

$$\left(\frac{n}{2}\right)! = \left(\frac{n}{2}\right)\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{1}{2}\right)\sqrt{\pi}$$

For example, let *n* = 7. Then *n*/2 = 7/2 = 3.5 and

$$\left(\frac{7}{2}\right)! = (3.5)! = (3.5)(2.5)(1.5)(0.5)(1.77246) = 11.632$$

With this definition the reader should have no problem verifying that the c_4 factor for *n* = 10 is .9727.

Mean and standard deviation of the estimators

So the mean or expected value of the sample standard deviation is $c_4 \sigma$.

The standard deviation of the sample standard deviation is

$$\sigma_s = \sigma\sqrt{1-c_4^2}$$

What are the differences between control limits and specification limits ?

Control limits vs. specifications

Control Limits are used to determine if the process is in a state of statistical control (i.e., is producing consistent output).

Specification Limits are used to determine if the product will function in the intended fashion.

How many data points are needed to set up a control chart?

How many samples are needed?

Shewhart gave the following rule of thumb:

> *"It has also been observed that a person would seldom if ever be justified in concluding that a state of statistical control of a given repetitive operation or production process has been reached until he had obtained, under presumably the same essential conditions, a sequence of not less than twenty five samples of size four that are in control."*

It is important to note that control chart properties, such as false alarm probabilities, are generally given under the assumption that the parameters, such as μ and σ, are known. When the control limits are not computed from a large amount of data, the actual properties might be quite different from what is assumed (see, e.g., Quesenberry, 1993).

When do we recalculate control limits?

When do we recalculate control limits?

Since a control chart "compares" the current performance of the process characteristic to the past performance of this characteristic, changing the control limits frequently would negate any usefulness.

So, only change your control limits if you have a valid, compelling reason for doing so. Some examples of reasons:

- When you have at least 30 more data points to add to the chart and there have been no known changes to the process

  - you get a better estimate of the variability

- If a major process change occurs and affects the way your process runs.

- If a known, preventable act changes the way the tool or process would behave (power goes out, consumable is corrupted or bad quality, etc.)

What are the WECO rules for signaling "Out of Control"?

General rules for detecting out of control or non-random situaltions

WECO stands for Western Electric Company Rules

Any Point Above +3 Sigma
---------- +3 σ LIMIT
2 Out of the Last 3 Points Above +2 Sigma
---------- +2 σ LIMIT
4 Out of the Last 5 Points Above +1 Sigma
---------- +1 σ LIMIT
8 Consecutive Points on This Side of Control Line

========== CENTER LINE

8 Consecutive Points on This Side of Control Line

---------- -1 σ LIMIT

4 Out of the Last 5 Points Below - 1 Sigma

---------- -2 σ LIMIT

2 Out of the Last 3 Points Below -2 Sigma

---------- -3 σ LIMIT

Any Point Below -3 Sigma

Trend Rules:	6 in a row trending up or down. 14 in a row alternating up and down

WECO rules based on probabilities

The WECO rules are based on probability. We know that, for a normal distribution, the probability of encountering a point outside $\pm 3\sigma$ is 0.3%. This is a rare event. Therefore, if we observe a point outside the control limits, we conclude the process has shifted and is unstable. Similarly, we can identify other events that are equally rare and use them as flags for instability. The probability of observing two points out of three in a row between 2σ and 3σ and the probability of observing four points out of five in a row between 1σ and 2σ are also about 0.3%.

WECO rules increase false alarms

Note: While the WECO rules increase a Shewhart chart's sensitivity to trends or drifts in the mean, there is a severe downside to adding the WECO rules to an ordinary Shewhart control chart that the user should understand. When following the standard Shewhart "out of control" rule (i.e., signal if and only if you see a point beyond the plus or minus 3 sigma control limits) you will have "false alarms" every 371 points on the average (see the description of Average Run Length or ARL on the next page). Adding the WECO rules increases the frequency of false alarms to about once in every 91.75 points, on the average (see Champ and Woodall, 1987). The user has to decide whether this price is worth paying (some users add the WECO rules, but take them "less seriously" in terms of the effort put into troubleshooting activities when out of control signals occur).

With this background, the next page will describe how to construct Shewhart variables control charts.

6.3.2.1. Shewhart X-bar and R and S Control Charts

$\overline{X}$ and S Charts

$\overline{X}$ and S Shewhart Control Charts

We begin with $\overline{X}$ and *s* charts. We should use the *s* chart first to determine if the distribution for the process characteristic is stable.

Let us consider the case where we have to estimate σ by analyzing past data. Suppose we have *m* preliminary samples at our disposition, each of size *n*, and let s_i be the standard deviation of the *i*th sample. Then the average of the *m* standard deviations is

$$\bar{s} = \frac{1}{m}\sum_{i=1}^{m} s_i$$

Control Limits for $\overline{X}$ and S Control Charts

We make use of the factor c_4 described on the previous page.

The statistic $\bar{s}/c_4$ is an unbiased estimator of σ. Therefore, the parameters of the *S* chart would be

$$UCL = \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2}$$

$$\text{Center Line} = \bar{s}$$

$$LCL = \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2}$$

Similarly, the parameters of the $\overline{X}$ chart would be

$$UCL = \bar{\bar{x}} + 3\frac{\bar{s}}{c_4\sqrt{n}}$$

$$\text{Center Line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - 3\frac{\bar{s}}{c_4\sqrt{n}}$$

$\bar{\bar{x}}$, the "grand" mean is the average of all the observations.

It is often convenient to plot the $\overline{X}$ and *s* charts on one page.

$\overline{X}$ and R Control Charts

$\overline{X}$ and R control charts

If the sample size is relatively small (say equal to or less than 10), we can use the range instead of the standard deviation of a sample to construct control charts on $\overline{X}$ and *the range, R*. The range of a sample is simply the difference between the largest and smallest observation.

There is a statistical relationship (Patnaik, 1946) between the mean range for data from a normal distribution and σ, the standard deviation of that distribution. This relationship depends only on the sample size, *n*. The mean of *R* is $d_2 \sigma$, where the value of d_2 is also a function of *n*. *An estimator of* σ is therefore R / d_2.

Armed with this background we can now develop the $\overline{X}$ and *R* control chart.

Let R_1, R_2, ..., R_k, be the range of *k* samples. The average range is

$$\bar{R} = \frac{R_1 + R_2 + ... + R_k}{k}$$

Then an estimate of σ can be computed as

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

$\overline{X}$ control charts

So, if we use $\bar{\bar{x}}$ (or a given target) as an estimator of μ and $\bar{R} / d_2$ as an estimator of σ, then the parameters of the $\overline{X}$ chart are

$$UCL = \bar{\bar{x}} + \frac{3}{d_2\sqrt{n}}\bar{R}$$

$$\text{Center Line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - \frac{3}{d_2\sqrt{n}}\bar{R}$$

The simplest way to describe the limits is to define the factor $A_2 = 3/(d_2\sqrt{n})$ and the construction of the $\overline{X}$ becomes

$$UCL = \bar{\bar{x}} + A_2\bar{R}$$

$$\text{Center Line} = \bar{\bar{x}}$$

$$LCL = \bar{\bar{x}} - A_2\bar{R}$$

The factor A_2 depends only on *n*, and is tabled below.

The R chart

R control charts

This chart controls the process variability since the sample range is related to the process standard deviation. *The center line of the R chart is the average range.*

To compute the control limits we need an estimate of the true, but unknown standard deviation $W = R/\sigma$. This can be found from the distribution of $W = R/\sigma$ (assuming that the items that we measure follow a normal distribution). The standard deviation of W is d_3, and is a known function of the sample size, *n*. It is tabulated in many textbooks on statistical quality control.

Therefore since $R = W\sigma$, the standard deviation of R is $\sigma_R = d_3\sigma$. But since the true σ is unknown, we may estimate σ_R by

$$\hat{\sigma}_R = d_3\frac{\bar{R}}{d_2}$$

As a result, the parameters of the *R* chart with the customary 3-sigma control limits are

$$UCL = \bar{R} + 3\sigma_R = \bar{R} + 3d_3\frac{\bar{R}}{d_2}$$

$$\text{Center Line} = \bar{R}$$

$$LCL = \bar{R} - 3\sigma_R = \bar{R} - 3d_3\frac{\bar{R}}{d_2}$$

As was the case with the control chart parameters for the subgroup averages, defining another set of factors will ease the computations, namely:

$D_3 = 1 - 3\,d_3 / d_2$ and $D_4 = 1 + 3\,d_3 / d_2$. These yield

$$UCL = \bar{R}D_4$$

$$\text{Center Line} = \bar{R}$$

$$LCL = \bar{R}D_3$$

The factors D_3 and D_4 depend only on *n*, and are tabled below.

Factors for Calculating Limits for $\bar{X}$ and R Charts

n	A_2	D_3	D_4
2	1.880	0	3.267

3	1.023	0	2.575
4	0.729	0	2.282
5	0.577	0	2.115
6	0.483	0	2.004
7	0.419	0.076	1.924
8	0.373	0.136	1.864
9	0.337	0.184	1.816
10	0.308	0.223	1.777

In general, the range approach is quite satisfactory for sample sizes up to around 10. For larger sample sizes, using subgroup standard deviations is preferable. For small sample sizes, the relative efficiency of using the range approach as opposed to using standard deviations is shown in the following table.

Efficiency of R versus S/c_4

n	Relative Efficiency
2	1.000
3	0.992
4	0.975
5	0.955
6	0.930
10	0.850

A typical sample size is 4 or 5, so not much is lost by using the range for such sample sizes.

Time To Detection or Average Run Length (ARL)

Waiting time to signal "out of control"

Two important questions when dealing with control charts are:

1. How often will there be false alarms where we look for an assignable cause but nothing has changed?
2. How quickly will we detect certain kinds of systematic changes, such as mean shifts?

The ARL tells us, for a given situation, how long on the average we will plot successive control charts points before we detect a point beyond the control limits.

For an $\bar{X}$ chart, with no change in the process, we wait on the average $1/p$ points before a false alarm takes place, with p denoting the probability of an observation plotting outside the control limits. For a normal distribution, $p = .0027$ and the ARL is approximately 371.

A table comparing Shewhart $\bar{X}$ chart *ARL*'s to Cumulative Sum (CUSUM) *ARL*'s for various mean shifts is given later in

this section.

There is also (currently) a web site developed by Galit Shmueli that will do ARL calculations interactively with the user, for Shewhart charts with or without additional (Western Electric) rules added.

6.3.2.2. Individuals Control Charts

Samples are Individual Measurements

Moving range used to derive upper and lower limits

Control charts for individual measurements, e.g., the sample size = 1, use the *moving range* of two successive observations to measure the process variability.

The moving range is defined as

$$MR_i = |x_i - x_{i-1}|$$

which is the absolute value of the first difference (e.g., the difference between two consecutive data points) of the data. Analogous to the Shewhart control chart, one can plot both the data (which are the individuals) and the moving range.

Individuals control limits for an observation

For the control chart for individual measurements, the lines plotted are:

$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128}$$

$$\text{Center Line} = \bar{x}$$

$$LCL = \bar{x} - 3\frac{\overline{MR}}{1.128}$$

where $\bar{X}$ is the average of all the individuals and $\overline{MR}$ is the average of all the moving ranges of two observations. Keep in mind that either or both averages may be replaced by a standard or target, if available. (Note that 1.128 is the value of d_2 for $n = 2$).

Example of moving range

The following example illustrates the control chart for individual observations. A new process was studied in order to monitor flow rate. The first 10 batches resulted in

Batch Number	Flowrate x	Moving Range MR

1	49.6	
2	47.6	2.0
3	49.9	2.3
4	51.3	1.4
5	47.8	3.5
6	51.2	3.4
7	52.6	1.4
8	52.4	0.2
9	53.6	1.2
10	52.1	1.5
	$\overline{X}$ = 50.81	$\overline{MR}$ = 1.8778

Limits for the moving range chart

This yields the parameters below.

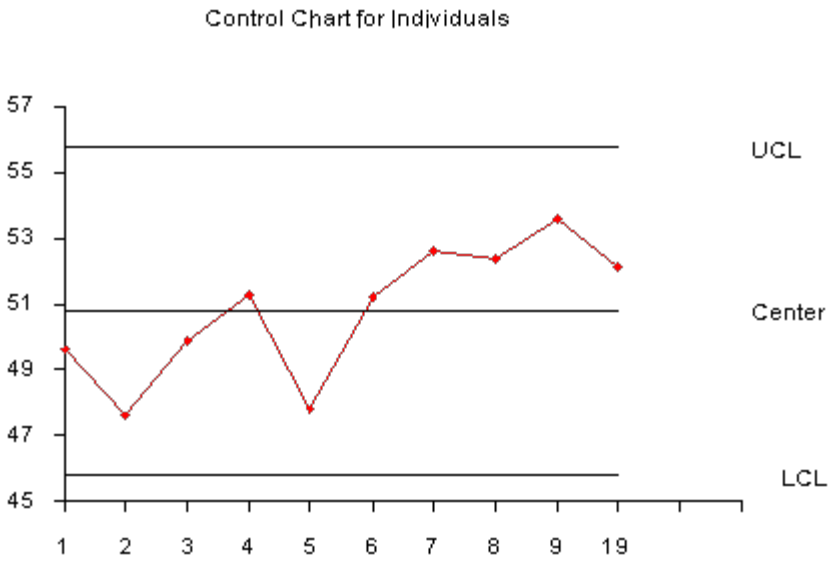
$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128} = 50.81 + 3\frac{1.8778}{1.128} = 55.8041$$

$$\text{Center Line} = \bar{x} = 50.81$$

$$LCL = \bar{x} - 3\frac{\overline{MR}}{1.128} = 50.81 - 3\frac{1.8778}{1.128} = 45.8159$$

Example of individuals chart

The control chart is given below



The process is in control, since none of the plotted points fall outside either the *UCL* or *LCL*.

Alternative for constructing individuals control

Note: Another way to construct the individuals chart is by using the standard deviation. Then we can obtain the chart from

$$\bar{x} \pm 3s/c_4$$

chart It is preferable to have the limits computed this way for the start of Phase 2.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

6.3.2.3. CUSUM Control Charts

CUSUM is an efficient alternative to Shewhart procedures

CUSUM charts, while not as intuitive and simple to operate as Shewhart charts, have been shown to be more efficient in detecting small shifts in the mean of a process. In particular, analyzing ARL's for CUSUM control charts shows that they are better than Shewhart control charts when it is desired to detect shifts in the mean that are 2 sigma or less.

CUSUM works as follows: Let us collect *m* samples, each of size *n*, and compute the mean of each sample. Then the cumulative sum (CUSUM) control chart is formed by plotting one of the following quantities:

Definition of cumulative sum

$$S_m = \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0) \quad \text{or} \quad S'_m = \frac{1}{\sigma_{\bar{x}}} \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0)$$

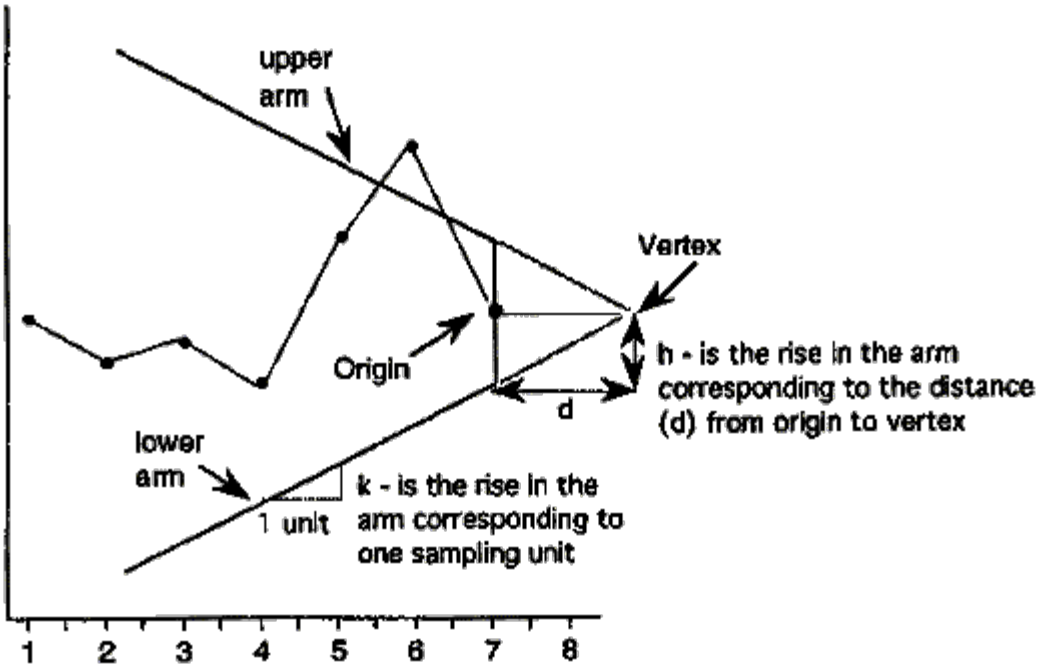
against the sample number *m*, where $\hat{\mu}_0$ is the estimate of the in-control mean and $\sigma_{\bar{x}}$ is the known (or estimated) standard deviation of the sample means. The choice of which of these two quantities is plotted is usually determined by the statistical software package. In either case, as long as the process remains in control centered at $\hat{\mu}_0$, the CUSUM plot will show variation in a random pattern centered about zero. If the process mean shifts upward, the charted CUSUM points will eventually drift upwards, and vice versa if the process mean decreases.

V-Mask used to determine if process is out of control

A visual procedure proposed by Barnard in 1959, known as *the V-Mask*, is sometimes used to determine whether a process is out of control. More often, the tabular form of the *V*-Mask is preferred. The tabular form is illustrated later in this section.

A *V*-Mask is an overlay shape in the form of a *V* on its side that is superimposed on the graph of the cumulative sums. The origin point of the *V*-Mask (see diagram below) is placed on top of the latest cumulative sum point and past points are examined to see if any fall above or below the sides of the *V*. As long as all the previous points lie between the sides of the *V*, the process is in control. Otherwise (even if one point lies outside) the process is suspected of being out of control.

Sample V-Mask demonstrating an out of control process



Interpretation of the V-Mask on the plot

In the diagram above, the *V*-Mask shows an out of control situation because of the point that lies above the upper arm. By sliding the *V*-Mask backwards so that the origin point covers other cumulative sum data points, we can determine the first point that signaled an out-of-control situation. This is useful for diagnosing what might have caused the process to go out of control.

From the diagram it is clear that the behavior of the *V*-Mask is determined by the distance *k* (which is the slope of the lower arm) and the rise distance *h*. These are the *design parameters* of the *V*-Mask. Note that we could also specify *d* and the vertex angle (or, as is more common in the literature, θ = 1/2 of the vertex angle) as the design parameters, and we would end up with the same *V*-Mask.

In practice, designing and manually constructing a *V*-Mask is a complicated procedure. A CUSUM spreadsheet style procedure shown below is more practical, unless you have statistical software that automates the *V*-Mask methodology. Before describing the spreadsheet approach, we will look briefly at an example of a *V*-Mask in graph form.

V-Mask Example

An example will be used to illustrate the construction and application of a *V*-Mask. The 20 data points

324.925, 324.675, 324.725, 324.350, 325.350, 325.225, 324.125, 324.525, 325.225, 324.600, 324.625, 325.150, 328.325, 327.250, 327.825, 328.500, 326.675, 327.775, 326.875, 328.350

are each the average of samples of size 4 taken from a

process that has an estimated mean of 325. Based on process data, the process standard deviation is 1.27 and therefore the sample means have a standard deviation of $1.27/(4^{1/2}) = 0.635$.

We can design a *V*-Mask using *h* and *k* or we can use an *alpha* and *beta* design approach. For the latter approach we must specify

- α: the probability of a false alarm, i.e., concluding that a shift in the process has occurred, while in fact it did not,
- β: the the probability of not detecting that a shift in the process mean has, in fact, occurred, and
- δ (delta): the amount of shift in the process mean that we wish to detect, expressed as a multiple of the standard deviation of the data points (which are the sample means).

Note: Technically, α and β are calculated in terms of one sequential trial where we monitor S_m until we have either an out-of-control signal or S_m returns to the starting point (and the monitoring begins, in effect, all over again).

The values of *h* and *k* are related to α, β, and δ based on the following equations (adapted from Montgomery, 2000).

$$k = \frac{\delta \sigma_x}{2}$$

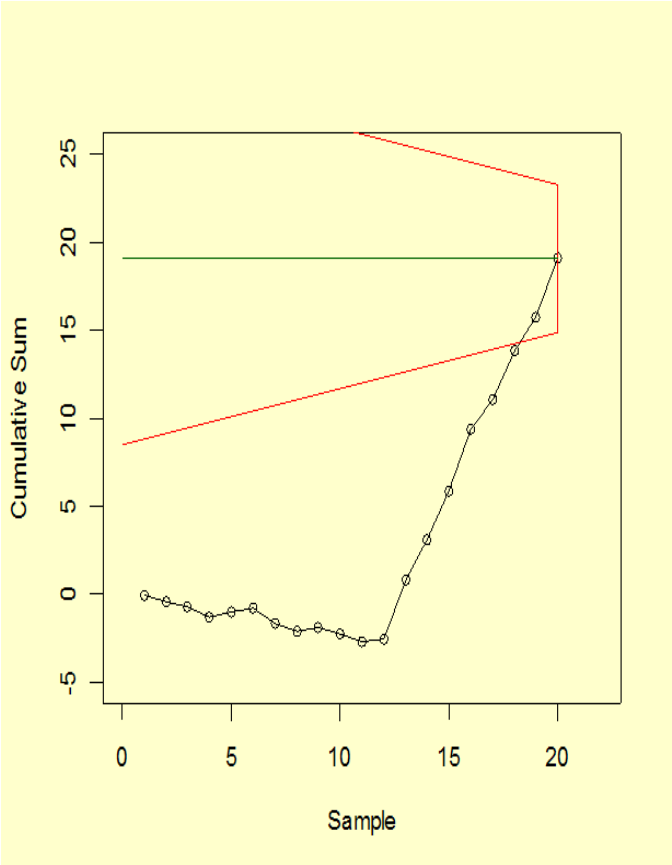
$$d = \frac{2}{\delta^2} \ln\left(\frac{1-\beta}{\alpha}\right)$$

$$h = dk$$

In our example we choose $\alpha = 0.0027$ (equivalent to the plus or minus 3 sigma criteria used in a standard Shewhart chart), and $\beta = 0.01$. Finally, we decide we want to quickly detect a shift as large as 1 sigma, which sets $\delta = 1$.

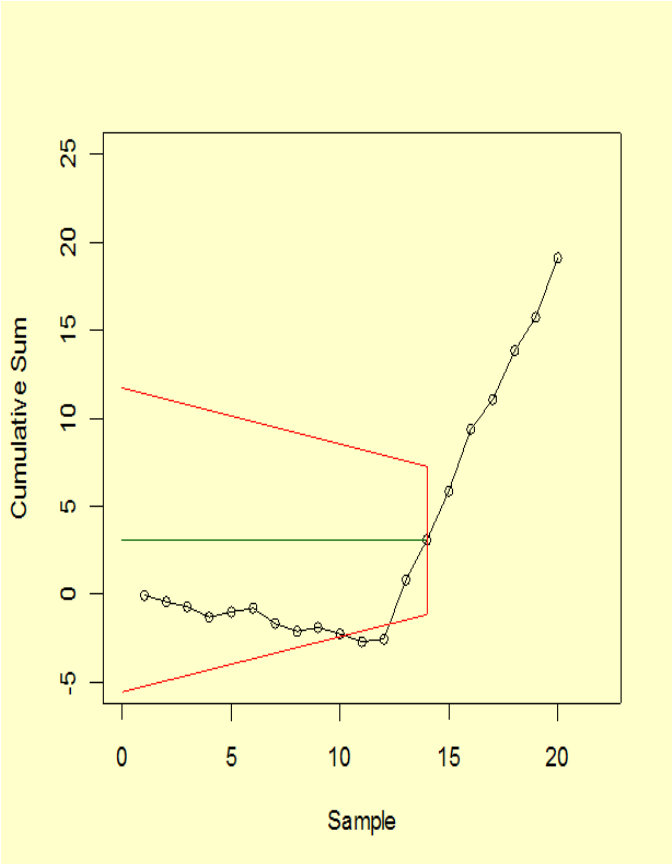
CUSUM Chart with V-Mask

When the *V*-Mask is placed over the last data point, the mask clearly indicates an out of control situation.



CUSUM chart after moving V-Mask to first out of control point

We next move the *V*-Mask and back to the first point that indicated the process was out of control. This is point number 14, as shown below.



Rule of thumb for choosing h and k

Note: A general rule of thumb (Montgomery) if one chooses to design with the h and k approach, instead of the α and β method illustrated above, is to choose k to be half the δ shift (0.5 in our example) and h to be around 4 or 5.

For more information on CUSUM chart design, see Woodall and Adams (1993).

Tabular or Spreadsheet Form of the *V*-Mask

A spreadsheet approach to CUSUM monitoring

Most users of CUSUM procedures prefer tabular charts over the *V*-Mask. The *V*-Mask is actually a carry-over of the pre-computer era. The tabular method can be quickly implemented by standard spreadsheet software.

To generate the tabular form we use the h and k parameters expressed in the original data units. It is also possible to use sigma units.

The following quantities are calculated:

$$S_{hi}(i) = \max(0, S_{hi}(i\text{-}1) + x_i - \hat{\mu}_0 - k)$$

$$S_{lo}(i) = \max(0, S_{lo}(i\text{-}1) + \hat{\mu}_0 - k - x_i))$$

where $S_{hi}(0)$ and $S_{lo}(0)$ are 0. When either S_{hi}(i) or S_{lo}(i) exceeds h, the process is out of control.

Example of spreadsheet calculations

We will construct a CUSUM tabular chart for the example described above. For this example, the parameter are h = 4.1959 and k = 0.3175. Using these design values, the tabular form of the example is

$\hat{\mu}_0$	h	k
325	4.1959	0.3175

Group	x	x-325	Increase in mean x-325-k	S_{hi}	Decrease in mean 325-k-x	S_{lo}	CUSUM
1	324.93	-0.07	-0.39	0.00	-0.24	0.00	-0.007
2	324.68	-0.32	-0.64	0.00	0.01	0.01	-0.40
3	324.73	-0.27	-0.59	0.00	-0.04	0.00	-0.67
4	324.35	-0.65	-0.97	0.00	0.33	0.33	-1.32
5	325.35	0.35	0.03	0.03	-0.67	0.00	-0.97

6	325.23	0.23	-0.09	0.00	-0.54	0.00	-0.75
7	324.13	-0.88	-1.19	0.00	0.56	0.56	-1.62
8	324.53	-0.48	-0.79	0.00	0.16	0.72	-2.10
9	325.23	0.23	-0.09	0.00	0.54	0.17	-1.87
10	324.60	-0.40	-0.72	0.00	0.08	0.25	-2.27
11	324.63	-0.38	-0.69	0.00	0.06	0.31	-2.65
12	325.15	0.15	-0.17	0.00	0.47	0.00	-2.50
13	328.33	3.32	3.01	3.01	-3.64	0.00	0.83
14	327.25	2.25	1.93	4.94*	-0.57	0.00	3.08
15	327.83	2.82	2.51	7.45*	-3.14	0.00	5.90
16	328.50	3.50	3.18	10.63*	-3.82	0.00	9.40
17	326.68	1.68	1.36	11.99*	-1.99	0.00	11.08
18	327.78	2.77	2.46	14.44*	-3.09	0.00	13.85
19	326.88	1.88	1.56	16.00*	-2.19	0.00	15.73
20	328.35	3.35	3.03	19.04*	-3.67	0.00	19.08

* = out of control signal

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

6.3.2.3.1. Cusum Average Run Length

The Average Run Length of Cumulative Sum Control Charts

The ARL of CUSUM

The operation of obtaining samples to use with a cumulative sum (CUSUM) control chart consists of taking samples of size n and plotting the cumulative sums

$$S_r = \sum_{i=1}^{r} (\bar{x}_i - k) \quad or \quad S_r = \sum_{i=1}^{r} (\bar{x}_i - k)/\sigma_{\bar{x}} \quad \text{(standardized)}$$

versus the sample number r, where $\bar{x}_i$ is the sample mean and k is a reference value.

In practice, k might be set equal to $(\hat{\mu}_0 + \mu_1)/2$, where $\hat{\mu}_0$ is the estimated in-control mean, which is sometimes known as the *acceptable quality level*, and μ_1 is referred to as the *rejectable quality level*.

If the distance between a plotted point and the lowest previous point is equal to or greater than h, one concludes that the process mean has shifted (increased).

h is decision limit

Hence, h is referred to as the *decision limit*. Thus the sample size n, reference value k, and decision limit h are the parameters required for operating a one-sided CUSUM chart. If one has to control both positive and negative deviations, as is usually the case, two one-sided charts are used, with respective values k_1, k_2, ($k_1 > k_2$) and respective decision limits h and $-h$.

Standardizing shift in mean and decision limit

The shift in the mean can be expressed as $\mu - k$. If we are dealing with normally distributed measurements, we can standardize this shift by

$$k_s = \frac{(\mu - k)}{\sigma / \sqrt{n}} \text{ or } (\mu - k)\sqrt{n} / \sigma$$

Similarly, the decision limit can be standardized by

$$h_s = \frac{h}{\sigma / \sqrt{n}} = h \sqrt{n} / \sigma$$

Determination of the ARL, given h and k

The average run length (ARL) at a given quality level is the average number of samples (subgroups) taken before an action signal is given. The standardized parameters k_s and h_s together with the sample size n are usually selected to yield approximate ARL's L_0 and L_1 at acceptable and rejectable quality levels μ_0 and μ_1 respectively. We would like to see a high ARL, L_0, when the process is on target, (i.e. in control), and a low ARL, L_1, when the process mean shifts to an unsatisfactory level.

In order to determine the parameters of a CUSUM chart, the acceptable and rejectable quality levels along with the desired respective ARL ' s are usually specified. The design parameters can then be obtained by a number of ways. Unfortunately, the calculations of the ARL for CUSUM charts are quite involved.

There are several nomographs available from different sources that can be utilized to find the ARL's when the standardized h and k are given. Some of the nomographs solve the unpleasant integral equations that form the basis of the exact solutions, using an approximation of Systems of Linear Algebraic Equations (SLAE). This Handbook used a computer program that furnished the required ARL's given the standardized h and k. An example is given below:

Example of finding ARL's given the standardized h and k

mean shift (k = .5)	$h\sqrt{n}/\sigma$ 4	5	Shewart $\bar{X}$
0	336	930	371.00
.25	74.2	140	281.14
.5	26.6	30.0	155.22
.75	13.3	17.0	81.22
1.00	8.38	10.4	44.0
1.50	4.75	5.75	14.97
2.00	3.34	4.01	6.30
2.50	2.62	3.11	3.24
3.00	2.19	2.57	2.00
4.00	1.71	2.01	1.19

Using the table

If k = .5, then the shift of the mean (in multiples of the standard deviation of the mean) is obtained by adding .5 to the first column. For example to detect a mean shift of 1 sigma at h = 4, the ARL = 8.38. (at first column entry of .5).

The last column of the table contains the ARL's for a Shewhart control chart at selected mean shifts. The ARL for Shewhart = 1/p, where p is the probability for a point to fall outside established

control limits. Thus, for 3-sigma control limits and assuming normality, the probability to exceed the upper control limit = .00135 and to fall below the lower control limit is also .00135 and their sum = .0027. (These numbers come from standard normal distribution tables or computer programs, setting $z = 3$). Then the ARL = 1/.0027 = 370.37. This says that when a process is in control one expects an out-of-control signal (false alarm) each 371 runs.

ARL if a 1 sigma shift has occurred

When the means shifts up by 1 sigma, then the distance between the upper control limit and the shifted mean is 2 sigma (instead of 3 σ). Entering normal distribution tables with $z = 2$ yields a probability of $p = .02275$ to exceed this value. The distance between the shifted mean and the lower limit is now 4 sigma and the probability of $\bar{X} < -4$ is only .000032 and can be ignored. The ARL is 1 / .02275 = 43.96 .

Shewhart is better for detecting large shifts, CUSUM is faster for small shifts

The conclusion can be drawn that the Shewhart chart is superior for detecting large shifts and the CUSUM scheme is faster for small shifts. The break-even point is a function of h, as the table shows.

6.3.2.4. EWMA Control Charts

EWMA statistic

The Exponentially Weighted Moving Average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time.

Comparison of Shewhart control chart and EWMA control chart techniques

For the Shewhart chart control technique, the decision regarding the state of control of the process at any time, *t*, depends solely on the most recent measurement from the process and, of course, the degree of 'trueness' of the estimates of the control limits from historical data. For the EWMA control technique, the decision depends on the EWMA statistic, which is an exponentially weighted average of all prior data, including the most recent measurement.

By the choice of weighting factor, λ, the EWMA control procedure can be made sensitive to a small or gradual drift in the process, whereas the Shewhart control procedure can only react when the last data point is outside a control limit.

Definition of EWMA

The statistic that is calculated is:

$$EWMA_t = \lambda Y_t + (1- \lambda) EWMA_{t-1} \quad \text{for } t = 1, 2, ..., n.$$

where

- $EWMA_0$ is the mean of historical data (target)
- Y_t is the observation at time *t*
- *n* is the number of observations to be monitored including $EWMA_0$
- $0 < \lambda \leq 1$ is a constant that determines the depth of memory of the EWMA.

The equation is due to Roberts (1959).

Choice of weighting factor

The parameter λ determines the rate at which 'older' data enter into the calculation of the EWMA statistic. A value of $\lambda = 1$ implies that only the most recent measurement influences the EWMA (degrades to Shewhart chart). Thus, a large value of $\lambda = 1$ gives more weight to recent data and less weight to older data; a small value of λ gives more weight to older data. The value of λ is usually set between 0.2 and 0.3

(Hunter) although this choice is somewhat arbitrary. Lucas and Saccucci (1990) give tables that help the user select λ.

Variance of EWMA statistic

The estimated variance of the EWMA statistic is approximately

$$s^2_{ewma} = (\lambda/(2-\lambda))\, s^2$$

when t is not small, where s is the standard deviation calculated from the historical data.

Definition of control limits for EWMA

The center line for the control chart is the target value or $EWMA_0$. The control limits are:

$$UCL = EWMA_0 + ks_{ewma}$$
$$LCL = EWMA_0 - ks_{ewma}$$

where the factor k is either set equal 3 or chosen using the Lucas and Saccucci (1990) tables. The data are assumed to be independent and these tables also assume a normal population.

As with all control procedures, the EWMA procedure depends on a database of measurements that are truly representative of the process. Once the mean value and standard deviation have been calculated from this database, the process can enter the monitoring stage, provided the process was in control when the data were collected. If not, then the usual Phase 1 work would have to be completed first.

Example of calculation of parameters for an EWMA control chart

To illustrate the construction of an EWMA control chart, consider a process with the following parameters calculated from historical data:

$$EWMA_0 = 50$$
$$s = 2.0539$$

with λ chosen to be 0.3 so that $\lambda / (2-\lambda) = .3 / 1.7 = 0.1765$ and the square root = 0.4201. The control limits are given by

$$UCL = 50 + 3\,(0.4201)(2.0539) = 52.5884$$
$$LCL = 50 - 3\,(0.4201)\,(2.0539) = 47.4115$$

Sample data

Consider the following data consisting of 20 points where 1 - 10 are on the top row from left to right and 11-20 are on the bottom row from left to right:

```
52.0 47.0 53.0 49.3 50.1 47.0
51.0 50.1 51.2 50.5 49.6 47.6
49.9 51.3 47.8 51.2 52.6 52.4
53.6 52.1
```

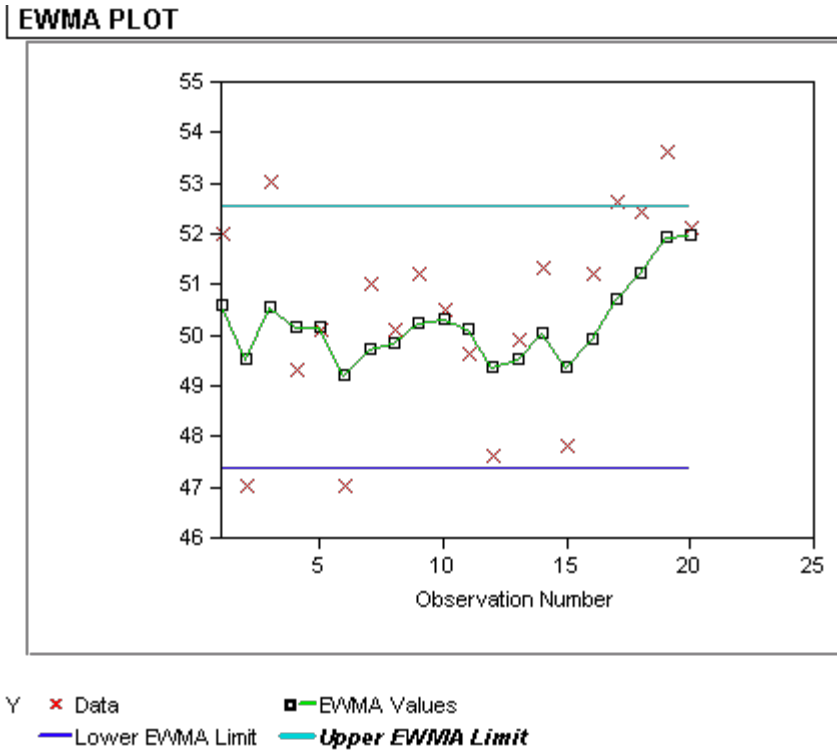
EWMA statistics for sample data

These data represent control measurements from the process which is to be monitored using the EWMA control chart technique. The corresponding EWMA statistics that are computed from this data set are:

```
50.00 50.60 49.52 50.56 50.18
50.16 49.12 49.75 49.85 50.26
50.33 50.11 49.36 49.52 50.05
49.34 49.92 50.73 51.23 51.94
```

Sample EWMA plot

The control chart is given below.



Interpretation of EWMA control chart

The red dots are the raw data; the jagged line is the EWMA statistic over time. The chart tells us that the process is in control because all $EWMA_t$ lie between the control limits.

However, there seems to be a trend upwards for the last 5 periods.

6.3.3. What are Attributes Control Charts?

Attributes data arise when classifying or counting observations

The Shewhart control chart plots quality characteristics that can be measured and expressed numerically. We measure weight, height, position, thickness, etc. If we cannot represent a particular quality characteristic numerically, or if it is impractical to do so, we then often resort to using a quality characteristic to sort or classify an item that is inspected into one of two "buckets".

An example of a common quality characteristic classification would be designating units as "conforming units" or "nonconforming units". Another quality characteristic criteria would be sorting units into "non defective" and "defective" categories. Quality characteristics of that type are called *attributes.*

Note that there is a difference between "nonconforming to an engineering specification" and "defective" -- a nonconforming unit may function just fine and be, in fact, not defective at all, while a part can be "in spec" and not fucntion as desired (i.e., be defective).

Examples of quality characteristics that are attributes are the number of failures in a production run, the proportion of malfunctioning wafers in a lot, the number of people eating in the cafeteria on a given day, etc.

Types of attribute control charts

Control charts dealing with the number of *defects* or *nonconformities* are called *c* charts (for count).

Control charts dealing with the *proportion* or *fraction* of defective product are called *p* charts (for proportion).

There is another chart which handles *defects per unit*, called the *u* chart (for unit). This applies when we wish to work with the average number of nonconformities per unit of product.

For additional references, see Woodall (1997) which reviews papers showing examples of attribute control charting, including examples from semiconductor manufacturing such as those examining the spatial depencence of defects.

6.3.3.1. Counts Control Charts

Defective items vs individual defects

The literature differentiates between *defect* and *defective*, which is the same as differentiating between *nonconformity* and *nonconforming units*. This may sound like splitting hairs, but in the interest of clarity let's try to unravel this man-made mystery.

Consider a wafer with a number of chips on it. The wafer is referred to as an "item of a product". The chip may be referred to as "a specific point". There exist certain specifications for the wafers. When a particular wafer (e.g., the item of the product) does not meet at least one of the specifications, it is classified as a nonconforming item. Furthermore, each chip, (e.g., the specific point) at which a specification is not met becomes a defect or nonconformity.

So, a nonconforming or defective item contains at least one defect or nonconformity. It should be pointed out that a wafer can contain several defects but still be classified as conforming. For example, the defects may be located at noncritical positions on the wafer. If, on the other hand, the number of the so-called "unimportant" defects becomes alarmingly large, an investigation of the production of these wafers is warranted.

Control charts involving counts can be either for the *total number* of nonconformities (defects) for the sample of inspected units, or for the *average number* of defects per inspection unit.

Poisson approximation for numbers or counts of defects

Let us consider an assembled product such as a microcomputer. The opportunity for the occurrence of any given defect may be quite large. However, the probability of occurrence of a defect in any one arbitrarily chosen spot is likely to be very small. In such a case, the incidence of defects might be modeled by a *Poisson* distribution. Actually, the Poisson distribution is an approximation of the *binomial* distribution and applies well in this capacity according to the following rule of thumb:

> The sample size n should be equal to or larger than 20 and the probability of a single success,

p, should be smaller than or equal to .05. If $n \geq 100$, the approximation is excellent if *np* is also ≤ 10.

Illustrate Poisson approximation to binomial

To illustrate the use of the Poisson distribution as an approximation of a binomial distribution, consider the following comparison: Let *p*, the probability of a single success in $n = 200$ trials, be .025.

Find the probability of exactly 3 successes. If we assume that *p* remains constant then the solution follows the binomial distribution rules, that is:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{200}{3} .025^3 .975^{197} = 0.1399995$$

By the Poisson approximation we have

$$c = (200)(.025)$$

and

$$p(x) = \frac{e^{-c} c^x}{x!} = \frac{e^{-5} 5^3}{3!} = 0.1403739$$

The inspection unit

Before the control chart parameters are defined there is one more definition: *the inspection unit.* We shall count the number of defects that occur in a so-called inspection unit. More often than not, an inspection unit is a single unit or item of product; for example, a wafer. However, sometimes the inspection unit could consist of five wafers, or ten wafers and so on. The size of the inspection units may depend on the recording facility, measuring equipment, operators, etc.

Suppose that defects occur in a given inspection unit according to the Poisson distribution, with parameter *c* (often denoted by *np* or the Greek letter λ). In other words

Control charts for counts, using the Poisson distribution

$$p(x) = \frac{e^{-c} c^x}{x!}$$

where *x* is the number of defects and $c > 0$ is the parameter of the Poisson distribution. It is known that both the mean and the variance of this distribution are equal to *c*. Then the *k*-sigma control chart is

$$UCL = c + k\sqrt{c}$$

$$CenterLine = c$$

$$LCL = c - k\sqrt{c}$$

If the LCL comes out negative, then there is no lower control limit. This control scheme assumes that a standard value for *c* is available. If this is not the case then *c* may be estimated as the average of the number of defects in a preliminary sample of inspection units, call it $\bar{c}$. Usually *k* is set to 3 by many practioners.

Control chart example using counts

An example may help to illustrate the construction of control limits for counts data. We are inspecting 25 successive wafers, each containing 100 chips. Here the wafer is the inspection unit. The observed number of defects are

Wafer Number	Number of Defects	Wafer Number	Number of Defects
1	16	14	16
2	14	15	15
3	28	16	13
4	16	17	14
5	12	18	16
6	20	19	11
7	10	20	20
8	12	21	11
9	10	22	19
10	17	23	16
11	19	24	31
12	17	25	13
13	14		

From this table we have

$$\bar{c} = \frac{\text{total number of defects}}{\text{total number of samples}} = \frac{400}{25} = 16$$

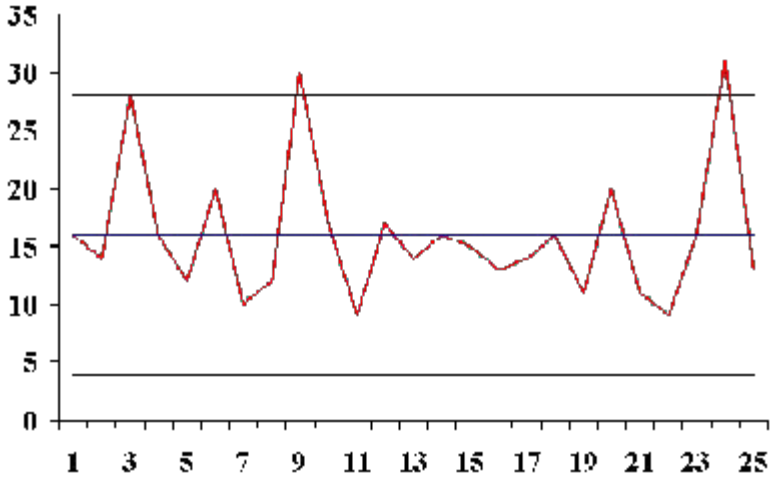
$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 16 + 2\sqrt{16} = 28$$

$$LCL = c - k\sqrt{c}$$

Sample counts control

Control Chart for Counts

chart



Transforming Poisson Data

Normal approximation to Poisson is adequate when the mean of the Poisson is at least 5

We have seen that the 3-sigma limits for a *c* chart, where *c* represents the number of nonconformities, are given by

$$\bar{c} \pm 3\sqrt{\bar{c}}$$

where it is assumed that the normal approximation to the Poisson distribution holds, hence the symmetry of the control limits. It is shown in the literature that the normal approximation to the Poisson is adequate when the mean of the Poisson is at least 5. When applied to the *c* chart this implies that the mean of the defects should be at least 5. This requirement will often be met in practice, but still, when the mean is smaller than 9 (solving the above equation) there will be no lower control limit.

Let the mean be 10. Then the lower control limit = 0.513. However, $P(c = 0) = .000045$, using the Poisson formula. This is only 1/30 of the assumed area of .00135. So one has to raise the lower limit so as to get as close as possible to .00135. From Poisson tables or computer software we find that $P(1) = .0005$ and $P(2) = .0027$, so the lower limit should actually be 2 or 3.

Transforming count data into approximately normal data

To avoid this type of problem, we may resort to a transformation that makes the transformed data match the normal distribution better. One such transformation described by Ryan (2000) is

$$Y = 2\sqrt{c}$$

which is, for a large sample, approximately normally distributed with mean = $2\sqrt{\lambda}$ and variace = 1, where λ is

the mean of the Poisson distribution.

Similar transformations have been proposed by Anscombe (1948) and Freeman and Tukey (1950). When applied to a *c* chart these are

$$y_1 = 2\sqrt{c + 3/8} \quad \text{and} \quad y_2 = \sqrt{c} + \sqrt{c+1}$$

The repspective control limits are

$$\bar{y} \pm 3, \quad \bar{y}_1 \pm 3, \quad \text{and } \bar{y}_2 \pm 3$$

While using transformations may result in meaningful control limits, one has to bear in mind that the user is now working with data on a different scale than the original measurements. There is another way to remedy the problem of symmetric limits applied to non symmetric cases, and that is to use *probability* limits. These can be obtained from tables given by Molina (1973). This allows the user to work with data on the original scale, but they require special tables to obtain the limits. Of course, software might be used instead.

Warning for highly skewed distributions

Note: In general, it is not a good idea to use 3-sigma limits for distributions that are highly skewed (see Ryan and Schwertman (1997) for more about the possibly extreme consequences of doing this).

6.3.3.2. Proportions Control Charts

p is the fraction defective in a lot or population

The proportion or fraction nonconforming (defective) in a population is defined as the ratio of the number of nonconforming items in the population to the total number of items in that population. The item under consideration may have one or more quality characteristics that are inspected simultaneously. If at least one of the characteristics does not conform to standard, the item is classified as nonconforming.

The fraction or proportion can be expressed as a decimal, or, when multiplied by 100, as a percent. The underlying statistical principles for a control chart for proportion nonconforming are based on the binomial distribution.

Let us suppose that the production process operates in a stable manner, such that the probability that a given unit will not conform to specifications is *p*. Furthermore, we assume that successive units produced are independent. Under these conditions, each unit that is produced is a realization of a Bernoulli random variable with parameter *p*. If a random sample of *n* units of product is selected and if *D* is the number of units that are nonconforming, the *D* follows a binomial distribution with parameters *n* and *p*

The binomial distribution model for number of defectives in a sample

$$p\{D = x\} = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \cdots, n$$

The mean of *D* is *np* and the variance is *np(1-p)*. The *sample* proportion nonconforming is the ratio of the number of nonconforming units in the sample, *D*, to the sample size *n*,

$$\hat{p} = \frac{D}{n}$$

The mean and variance of this estimator are

$$\mu = p$$

and

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

This background is sufficient to develop the control chart for proportion or fraction nonconforming. The chart is called the *p*-chart.

p control charts for lot proportion defective

If the true fraction conforming *p* is known (or a standard value is given), then the center line and control limits of the fraction nonconforming control chart is

$$UCL = p + 3\sqrt{\frac{p(1-p)}{n}}$$

$$CenterLine = p$$

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}}$$

When the process fraction (proportion) *p* is not known, it must be estimated from the available data. This is accomplished by selecting *m* preliminary samples, each of size *n*. If there are D_i defectives in sample *i*, the fraction nonconforming in sample *i* is

$$\hat{p}_i = \frac{D_i}{n} \quad i = 1, 2, ..., m$$

and the average of these individuals sample fractions is

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{mn} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}$$

The $\bar{p}$ is used instead of *p* in the control chart setup.

Example of a p-chart

A numerical example will now be given to illustrate the above mentioned principles. The location of chips on a wafer is measured on 30 wafers.

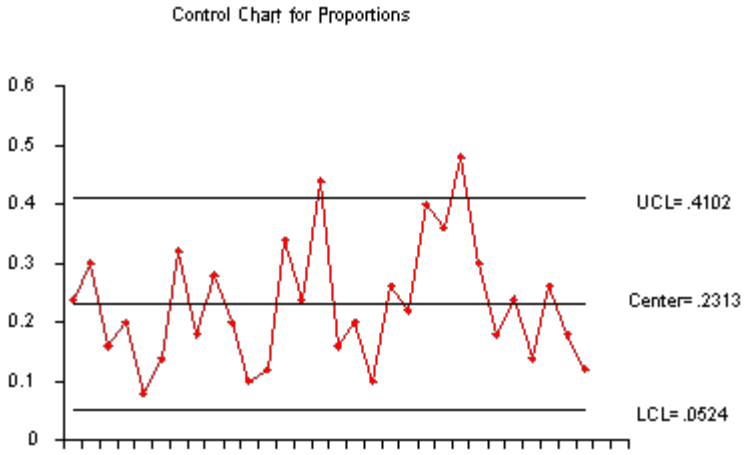
On each wafer 50 chips are measured and a defective is defined whenever a misregistration, in terms of horizontal and/or vertical distances from the center, is recorded. The results are

Sample Number	Fraction Defectives	Sample Number	Fraction Defectives	Sample Number	Fraction Defectives
1	.24	11	.10	21	.40
2	.30	12	.12	22	.36
3	.16	13	.34	23	.48
4	.20	14	.24	24	.30

5	.08	15	.44	25	.18
6	.14	16	.16	26	.24
7	.32	17	.20	27	.14
8	.18	18	.10	28	.26
9	.28	19	.26	29	.18
10	.20	20	.22	30	.12

Sample proportions control chart

The corresponding control chart is given below:



6.3.4. What are Multivariate Control Charts?

Multivariate control charts and Hotelling's T^2 statistic

It is a fact of life that most data are naturally multivariate. Hotelling in 1947 introduced a statistic which uniquely lends itself to plotting multivariate observations. This statistic, appropriately named Hotelling's T^2, is a scalar that combines information from the dispersion and mean of several variables. Due to the fact that computations are laborious and fairly complex and require some knowledge of matrix algebra, acceptance of multivariate control charts by industry was slow and hesitant.

Multivariate control charts now more accessible

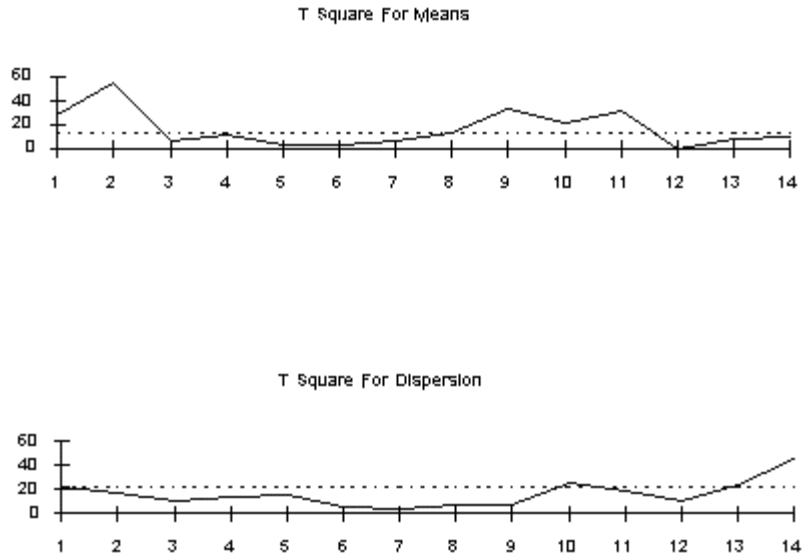
Nowadays, modern computers in general and the PC in particular have made complex calculations accessible and during the last decade, multivariate control charts were given more attention. In fact, the multivariate charts which display the Hotelling T^2 statistic became so popular that they sometimes are called Shewhart charts as well (e.g., Crosier, 1988), although Shewhart had nothing to do with them.

Hotelling charts for both means and dispersion

As in the univariate case, when data are grouped, the T^2 chart can be paired with a chart that displays a measure of variability within the subgroups for all the analyzed characteristics. The combined T^2 and T_d^2 (dispersion) charts are thus a multivariate counterpart of the univariate $\overline{X}$ and S (or $\overline{X}$ and R) charts.

Hotelling mean and dispersion control charts

An example of a Hotelling T^2 and T_d^2 pair of charts is given below:



Interpretation of sample Hotelling control charts

Each chart represents 14 consecutive measurements on the means of four variables. The T^2 chart for means indicates an out-of-control state for groups 1,2 and 9-11. The T^2_d chart for dispersions indicate that groups 10, 13 and 14 are also out of control. The interpretation is that the multivariate system is suspect. To find an assignable cause, one has to resort to the individual univariate control charts or some other univariate procedure that should accompany this multivariate chart.

Additional discussion

For more details and examples see the next page and also Tutorials, section 5, subsections 4.3, 4.3.1 and 4.3.2. An introduction to Elements of multivariate analysis is also given in the Tutorials.

6.3.4.1. Hotelling Control Charts

Definition of Hotelling's T^2 "distance" statistic

The Hotelling T^2 distance is a measure that accounts for the covariance structure of a multivariate normal distribution. It was proposed by Harold Hotelling in 1947 and is called Hotelling T^2. It may be thought of as the multivariate counterpart of the Student's-*t* statistic.

The T^2 distance is a constant multiplied by a quadratic form. This quadratic form is obtained by multiplying the following three quantities:

1. The vector of deviations between the observations and the mean **m,** which is expressed by **(X-m)'**,
2. The inverse of the covariance matrix, $\mathbf{S}^{-1}$,
3. The vector of deviations, **(X-m**).

It should be mentioned that for independent variables, the covariance matrix is a diagonal matrix and T^2 becomes proportional to the sum of squared standardized variables.

In general, the higher the T^2 value, the more distant is the observation from the mean. The formula for computing the T^2 is:

$$T^2 = c\left(\mathbf{X} - \mathbf{m}'\right)\mathbf{S}^{-1}\left(\mathbf{X} - \mathbf{m}\right)$$

The constant *c* is the sample size from which the covariance matrix was estimated.

T^2 readily graphable

The T^2 distances lend themselves readily to graphical displays and as a result the T^2-chart is the most popular among the multivariate control charts.

Estimation of the Mean and Covariance Matrix

Mean and Covariance matrices

Let $\mathbf{X}_1,...\mathbf{X}_n$ be *n p*-dimensional vectors of observations that are sampled independently from $N_p(\mathbf{m}, \boldsymbol{\Sigma})$ with $p < n-1$, with

Σ the covariance matrix of **X**. The observed mean vector $\bar{\mathbf{X}}$ and the sample dispersion matrix

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})'$$

are the unbiased estimators of **m** and Σ, respectively.

Additional discussion

See Tutorials (section 5), subsections 4.3, 4.3.1 and 4.3.2 for more details and examples. An introduction to Elements of multivariate analysis is also given in the Tutorials.

6.3.4.2. Principal Components Control Charts

Problems with T^2 charts

Although the T^2 chart is the most popular, easiest to use and interpret method for handling multivariate process data, and is beginning to be widely accepted by quality engineers and operators, it is not a panacea. First, unlike the univariate case, the scale of the values displayed on the chart is not related to the scales of any of the monitored variables. Secondly, when the T^2 statistic exceeds the upper control limit (UCL), the user does not know which particular variable(s) caused the out-of-control signal.

Run univariate charts along with the multivariate ones

With respect to scaling, we strongly advise to run individual univariate charts in tandem with the multivariate chart. This will also help in honing in on the culprit(s) that might have caused the signal. However, individual univariate charts cannot explain situations that are a result of some problems in the covariance or correlation between the variables. This is why a dispersion chart must also be used.

Another way to monitor multivariate data: Principal Components control charts

Another way to analyze the data is to use *principal components*. For each multivariate measurement (or observation), the principal components are linear combinations of the standardized p variables (to standardize subtract their respective targets and divide by their standard deviations). The principal components have two important advantages:

1. the new variables are uncorrelated (or almost)
2. very often, a few (sometimes 1 or 2) principal components may capture most of the variability in the data so that we do not have to use all of the p principal components for control.

Eigenvalues

Unfortunately, there is one big disadvantage: The identity of the original variables is lost! However, in some cases the specific linear combinations corresponding to the principal components with the largest *eigenvalues* may yield meaningful measurement units. What is being used in control charts are the *principal factors.*

A principal factor is the principal component divided by the square root of its eigenvalue.

Additional discussion

More details and examples are given in the Tutorials (section 5).

6. Process or Product Monitoring and Control
6.3. Univariate and Multivariate Control Charts
6.3.4. What are Multivariate Control Charts?

6.3.4.3. Multivariate EWMA Charts

Multivariate EWMA Control Chart

Univariate EWMA model

The model for a univariate EWMA chart is given by:

$$Z_i = \lambda X_i + (1 - \lambda) Z_{i-1} \quad i = 1, 2, ..., n$$

where Z_i is the ith EWMA, X_i is the the ith observation, Z_0 is the average from the historical data, and $0 < \lambda \leq 1$.

Multivariate EWMA model

In the multivariate case, one can extend this formula to

$$Z_i = \Lambda X_i + (1 - \Lambda) Z_{i-1}$$

where Z_i is the *i*th EWMA vector, X_i is the the *i*th observation vector i = 1, 2, ..., n, Z_0 is the vector of variable values from the historical data, Λ is the diag($\lambda_1, \lambda_2, ..., \lambda_p$) which is a diagonal matrix with $\lambda_1, \lambda_2, ..., \lambda_p$ on the main diagonal, and p is the number of variables; that is the number of elements in each vector.

Illustration of multivariate EWMA

The following illustration may clarify this. There are p variables and each variable contains n observations. The input data matrix looks like:

$$\begin{matrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{matrix}$$

The quantity to be plotted on the control chart is

$$T_i^2 = Z_i' \Sigma_{z_i}^{-1} Z_i$$

Simplification

It has been shown (Lowry et al., 1992) that the (k,l)th element of the covariance matrix of the ith EWMA, Σ_{zi}, is

$$\Sigma_{zi}(k,l) = \lambda_k \lambda_l \frac{[1-(1-\lambda_k)^i(1-\lambda_l)^i]}{[\lambda_k + \lambda_l - \lambda_k\lambda_l]}\sigma_{k,l}$$

where $\sigma_{k,l}$ is the (k,l)th element of Σ, the covariance matrix of the X's.

If $\lambda_1 = \lambda_2 = ... = \lambda_p = \lambda$, then the above expression simplifies to

$$\Sigma_{zi}(k,l) = \frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]\Sigma$$

where Σ is the covariance matrix of the input data.

Further simplification

There is a further simplification. When i becomes large, the covariance matrix may be expressed as:

$$\Sigma_{z_i} = \frac{\lambda}{2-\lambda}\Sigma$$

The question is "What is large?". When we examine the formula with the 2i in it, we observe that when 2i becomes sufficiently large such that $(1 - \lambda)^{2i}$ becomes almost zero, then we can use the simplified formula.

Table for selected values of λ and i

The following table gives the values of $(1-\lambda)^{2i}$ for selected values of λ and i.

					2*i*				
1 - λ	**4**	**6**	**8**	**10**	**12**	**20**	**30**	**40**	**50**
.9	.656	.531	.430	.349	.282	.122	.042	.015	.005
.8	.410	.262	.168	.107	.069	.012	.001	.000	.000
.7	.240	.118	.058	.028	.014	.001	.000	.000	.000
.6	.130	.047	.017	.006	.002	.000	.000	.000	.000
.5	.063	.016	.004	.001	.000	.000	.000	.000	.000
.4	.026	.004	.001	.000	.000	.000	.000	.000	.000
.3	.008	.001	.000	.000	.000	.000	.000	.000	.000
.2	.002	.000	.000	.000	.000	.000	.000	.000	.000
.1	.000	.000	.000	.000	.000	.000	.000	.000	.000

Simplified formuala not required

It should be pointed out that a well-meaning computer program does not have to adhere to the simplified formula, and potential inaccuracies for low values for λ and i can thus

be avoided.

MEWMA computer output for the Lowry data

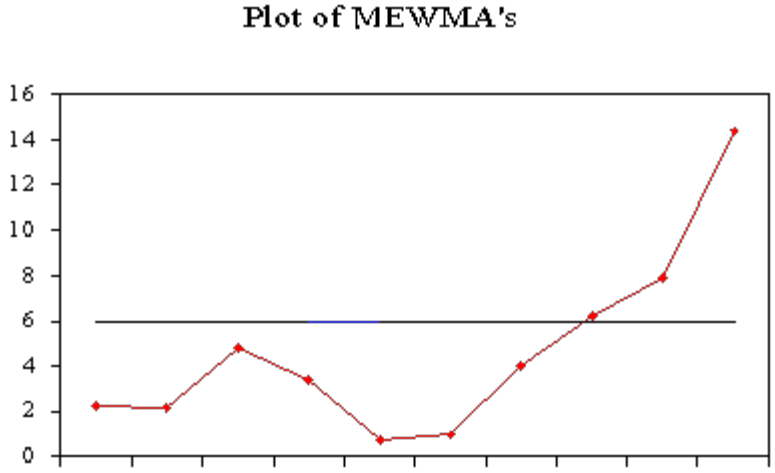
Here is an example of the application of an MEWMA control chart. To faciltate comparison with existing literature, we used data from Lowry et al. The data were simulated from a bivariate normal distribution with unit variances and a correlation coefficient of 0.5. The value for $\lambda = .10$ and the values for T_i^2 were obtained by the equation given above. The covariance of the MEWMA vectors was obtained by using the non-simplified equation. That means that for each MEWMA control statistic, the computer computed a covariance matrix, where i = 1, 2, ...10. The results of the computer routine are:

```
*****************************************************
*      Multi-Variate EWMA Control Chart
*
*****************************************************

DATA SERIES              MEWMA Vector           MEWMA
   1          2           1          2
STATISTIC
-1.190      0.590      -0.119      0.059         2.1886
 0.120      0.900      -0.095      0.143         2.0697
-1.690      0.400      -0.255      0.169         4.8365
 0.300      0.460      -0.199      0.198         3.4158
 0.890     -0.750      -0.090      0.103         0.7089
 0.820      0.980       0.001      0.191         0.9268
-0.300      2.280      -0.029      0.400         4.0018
 0.630      1.750       0.037      0.535         6.1657
 1.560      1.580       0.189      0.639         7.8554
 1.460      3.050       0.316      0.880        14.4158


VEC     XBAR       MSE       Lamda
 1       .260     1.200      0.100
 2     1.124      1.774      0.100
```

The UCL = 5.938 for $\alpha = .05$. Smaller choices of α are also used.

Sample MEWMA plot

The following is the plot of the above MEWMA.



6.4. Introduction to Time Series Analysis

Time series methods take into account possible internal structure in the data

Time series data often arise when monitoring industrial processes or tracking corporate business metrics. The essential difference between modeling data via time series methods or using the process monitoring methods discussed earlier in this chapter is the following:

> *Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.*

This section will give a brief overview of some of the more widely used techniques in the rich and rapidly growing field of time series modeling and analysis.

Contents for this section

Areas covered are:

1. Definitions, Applications and Techniques
2. What are Moving Average or Smoothing Techniques?
    1. Single Moving Average
    2. Centered Moving Average
3. What is Exponential Smoothing?
    1. Single Exponential Smoothing
    2. Forecasting with Single Exponential Smoothing
    3. Double Exponential Smoothing
    4. Forecasting with Double Exponential Smoothing
    5. Triple Exponential Smoothing
    6. Example of Triple Exponential Smoothing
    7. Exponential Smoothing Summary
4. Univariate Time Series Models
    1. Sample Data Sets
    2. Stationarity
    3. Seasonality
    4. Common Approaches
    5. Box-Jenkins Approach
    6. Box-Jenkins Model Identification
    7. Box-Jenkins Model Estimation
    8. Box-Jenkins Model Validation
    9. Example of Univariate Box-Jenkins Analysis

10. Box-Jenkins Model Analysis on Seasonal Data
5. Multivariate Time Series Models
    1. Example of Multivariate Time Series Analysis

6.4.1. Definitions, Applications and Techniques

Definition

Definition of Time Series: *An ordered sequence of values of a variable at equally spaced time intervals.*

Time series occur frequently when looking at industrial data

Applications: The usage of time series models is twofold:

- Obtain an understanding of the underlying forces and structure that produced the observed data
- Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control.

Time Series Analysis is used for many applications such as:

- Economic Forecasting
- Sales Forecasting
- Budgetary Analysis
- Stock Market Analysis
- Yield Projections
- Process and Quality Control
- Inventory Studies
- Workload Projections
- Utility Studies
- Census Analysis

and many, many more...

There are many methods used to model and forecast time series

Techniques: The fitting of time series models can be an ambitious undertaking. There are many methods of model fitting including the following:

- Box-Jenkins ARIMA models
- Box-Jenkins Multivariate Models
- Holt-Winters Exponential Smoothing (single, double, triple)

The user's application and preference will decide the selection of the appropriate technique. It is beyond the realm and intention of the authors of this handbook to cover all these methods. The overview presented here will start by looking at some basic smoothing techniques:

- Averaging Methods
- Exponential Smoothing Techniques.

Later in this section we will discuss the Box-Jenkins modeling methods and Multivariate Time Series.

6.4.2. What are Moving Average or Smoothing Techniques?

Smoothing data removes random variation and shows trends and cyclic components

Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of canceling the effect due to random variation. An often-used technique in industry is "smoothing". This technique, when properly applied, reveals more clearly the underlying trend, seasonal and cyclic components.

There are two distinct groups of smoothing methods

- Averaging Methods
- Exponential Smoothing Methods

Taking averages is the simplest way to smooth data

We will first investigate some averaging methods, such as the "simple" average of all past data.

A manager of a warehouse wants to know how much a typical supplier delivers in 1000 dollar units. He/she takes a sample of 12 suppliers, at random, obtaining the following results:

Supplier	Amount	Supplier	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

The computed mean or average of the data = 10. The manager decides to use this as the estimate for *expenditure of a typical supplier*.

Is this a good or bad estimate?

Mean squared error is a way to judge how

We shall compute the *"mean squared error"*:

- The "error" = true amount spent minus the estimated amount.
- The "error squared" is the error above, squared.

good a model is

- The "SSE" is the sum of the squared errors.
- The "MSE" is the mean of the squared errors.

MSE results for example

The results are:

Error and Squared Errors

The estimate = 10

Supplier	$	Error	Error Squared
1	9	-1	1
2	8	-2	4
3	9	-1	1
4	12	2	4
5	9	-1	1
6	12	2	4
7	11	1	1
8	7	-3	9
9	13	3	9
10	9	-1	1
11	11	1	1
12	10	0	0

The SSE = 36 and the MSE = 36/12 = 3.

Table of MSE results for example using different estimates

So how good was the estimator for the amount spent for each supplier? Let us compare the estimate (10) with the following estimates: 7, 9, and 12. That is, we estimate that each supplier will spend $7, or $9 or $12.

Performing the same calculations we arrive at:

Estimator	7	9	10	12
SSE	144	48	36	84
MSE	12	4	3	7

The estimator with the smallest MSE is the best. It can be shown mathematically that the estimator that minimizes the MSE for a set of random data is the mean.

Table showing squared error for the mean for sample data

Next we will examine the mean to see how well it predicts net income over time.

The next table gives the income before taxes of a PC manufacturer between 1985 and 1994.

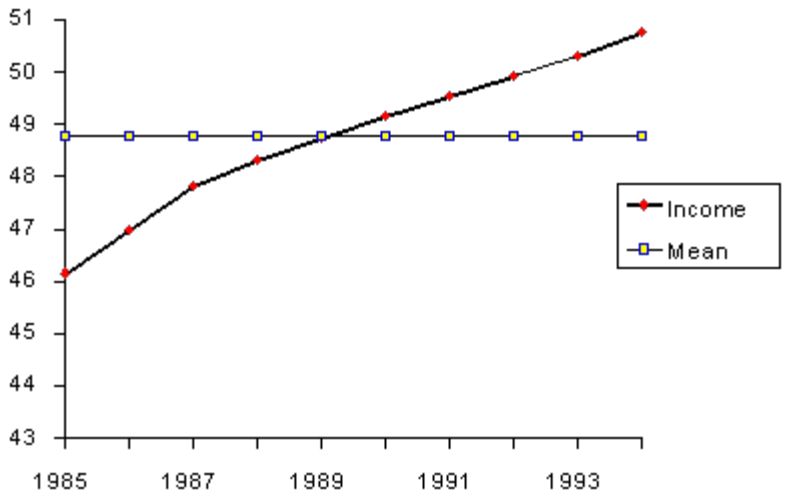
	Squared

Year	$ (millions)	Mean	Error	Error
1985	46.163	48.776	-2.613	6.828
1986	46.998	48.776	-1.778	3.161
1987	47.816	48.776	-0.960	0.922
1988	48.311	48.776	-0.465	0.216
1989	48.758	48.776	-0.018	0.000
1990	49.164	48.776	0.388	0.151
1991	49.548	48.776	0.772	0.596
1992	48.915	48.776	1.139	1.297
1993	50.315	48.776	1.539	2.369
1994	50.768	48.776	1.992	3.968

The MSE = 1.9508.

The mean is not a good estimator when there are trends

The question arises: *can we use the mean to forecast income if we suspect a trend?* A look at the graph below shows clearly that we should not do this.



Average weighs all past observations equally

In summary, we state that

1. The "simple" average or mean of all past observations is only a useful estimate for forecasting when there are no trends. If there are trends, use different estimates that take the trend into account.

2. The average "weighs" all past observations equally. For example, the average of the values 3, 4, 5 is 4. We know, of course, that an average is computed by adding all the values and dividing the sum by the number of values. Another way of computing the average is by adding each value divided by the number of values, or

   $$3/3 + 4/3 + 5/3 = 1 + 1.3333 + 1.6667 = 4.$$

   The multiplier 1/3 is called the *weight*. In general:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \left(\frac{1}{n}\right) x_1 + \left(\frac{1}{n}\right) x_2 + \ldots + \left(\frac{1}{n}\right) x_n$$

The $\left(\frac{1}{n}\right)$ are the weights and of course they sum to 1.

6.4.2.1. Single Moving Average

Taking a moving average is a smoothing process

An alternative way to summarize the past data is to compute the mean of successive smaller sets of numbers of past data as follows:

> Recall the set of numbers 9, 8, 9, 12, 9, 12, 11, 7, 13, 9, 11, 10 which were the dollar amount of 12 suppliers selected at random. Let us set *M*, the size of the "smaller set" equal to 3. Then the average of the first 3 numbers is: (9 + 8 + 9) / 3 = 8.667.

This is called "smoothing" (i.e., some form of averaging). This smoothing process is continued by advancing one period and calculating the next average of three numbers, dropping the first number.

Moving average example

The next table summarizes the process, which is referred to as *Moving Averaging*. The general expression for the moving average is

$$M_t = [X_t + X_{t-1} + ... + X_{t-N+1}] / N$$

Results of Moving Average

Supplier	$	MA	Error	Error squared
1	9			
2	8			
3	9	8.667	0.333	0.111
4	12	9.667	2.333	5.444
5	9	10.000	-1.000	1.000
6	12	11.000	1.000	1.000
7	11	10.667	0.333	0.111
8	7	10.000	-3.000	9.000
9	13	10.333	2.667	7.111
10	9	9.667	-0.667	0.444
11	11	11.000	0	0
12	10	10.000	0	0

The MSE = 2.018 as compared to 3 in the previous case.

6.4.2.2. Centered Moving Average

When computing a running moving average, placing the average in the middle time period makes sense

In the previous example we computed the average of the first 3 time periods and placed it next to period 3. We could have placed the average in the middle of the time interval of three periods, that is, next to period 2. This works well with odd time periods, but not so good for even time periods. So where would we place the first moving average when $M = 4$?

Technically, the Moving Average would fall at $t = 2.5, 3.5, ...$

To avoid this problem we smooth the MA's using $M = 2$. **Thus we smooth the smoothed values!**

If we average an even number of terms, we need to smooth the smoothed values

The following table shows the results using $M = 4$.

Period	Value	Interim Steps MA	Centered
1	9		
1.5			
2	8		
2.5		9.5	
3	9		9.5
3.5		9.5	
4	12		10.0
4.5		10.5	
5	9		10.750
5.5		11.0	
6	12		
6.5			
7	9		

Final table

This is the final table:

Period	Value	Centered MA
1	9	

2	8	
3	9	9.5
4	12	10.0
5	9	10.75
6	12	
7	11	

Double Moving Averages for a Linear Trend Process

Moving averages are still not able to handle significant trends when forecasting

Unfortunately, neither the mean of all data nor the moving average of the most recent *M* values, when used as forecasts for the next period, are able to cope with a significant trend.

There exists a variation on the MA procedure that often does a better job of handling trend. It is called *Double Moving Averages for a Linear Trend Process*. It calculates a second moving average from the original moving average, using the same value for *M*. As soon as both single and double moving averages are available, a computer routine uses these averages to compute a slope and intercept, and then forecasts one or more periods ahead.

6.4.3. What is Exponential Smoothing?

Exponential smoothing schemes weight past observations using exponentially decreasing weights

This is a very popular scheme to produce a smoothed Time Series. Whereas in Single Moving Averages the past observations are weighted equally, Exponential Smoothing assigns *exponentially decreasing weights* as the observation get older.

In other words, *recent observations are given relatively more weight in forecasting than the older observations*.

In the case of moving averages, the weights assigned to the observations are the same and are equal to $1/N$. In exponential smoothing, however, there are one or more *smoothing parameters* to be determined (or estimated) and these choices determine the weights assigned to the observations.

Single, double and triple Exponential Smoothing will be described in this section.

6.4.3.1. Single Exponential Smoothing

Exponential smoothing weights past observations with exponentially decreasing weights to forecast future values

This smoothing scheme begins by setting S_2 to y_1, where S_i stands for smoothed observation or EWMA, and y stands for the original observation. The subscripts refer to the time periods, 1, 2, ..., n. For the third period, $S_3 = \alpha y_2 + (1-\alpha) S_2$; and so on. There is no S_1; the smoothed series starts with the smoothed version of the second observation.

For any time period t, the smoothed value S_t is found by computing

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1} \qquad 0 < \alpha \leq 1 \qquad t \geq 3$$

This is the *basic equation of exponential smoothing* and the constant or parameter α is called the *smoothing constant*.

Note: There is an alternative approach to exponential smoothing that replaces y_{t-1} in the basic equation with y_t, the current observation. That formulation, due to Roberts (1959), is described in the section on EWMA control charts. The formulation here follows Hunter (1986).

Setting the first EWMA

The first forecast is very important

The initial EWMA plays an important role in computing all the subsequent EWMA's. Setting S_2 to y_1 is one method of initialization. Another way is to set it to the target of the process.

Still another possibility would be to average the first four or five observations.

It can also be shown that the smaller the value of α, the more important is the selection of the initial EWMA. The user would be wise to try a few methods, (assuming that the software has them available) before finalizing the settings.

Why is it called "Exponential"?

Expand basic

Let us expand the basic equation by first substituting for S

equation

in the basic equation to obtain

$$S_t = \alpha y_{t-1} + (1-\alpha)\,[\,\alpha y_{t-2} + (1-\alpha) S_{t-2}\,]$$
$$= \alpha y_{t-1} + \alpha(1-\alpha) y_{t-2} + (1-\alpha)^2 S_{t-2}$$

Summation formula for basic equation

By substituting for S_{t-2}, then for S_{t-3}, and so forth, until we reach S_2 (which is just y_1), it can be shown that the expanding equation can be written as:

$$S_t = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} S_2, \quad t \geq 2$$

Expanded equation for S_5

For example, the expanded equation for the smoothed value S_5 *is:*

$$S_5 = \alpha\left[(1-\alpha)^0 y_{5-1} + (1-\alpha)^1 y_{5-2} + (1-\alpha)^2 y_{5-3}\right] + (1-\alpha)^3 S_2$$

Illustrates exponential behavior

This illustrates the exponential behavior. The weights, $\alpha(1-\alpha)^t$ decrease geometrically, and their sum is unity as shown below, using a property of geometric series:

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = a\left[\frac{1-(1-\alpha)^t}{1-(1-\alpha)}\right] = 1-(1-\alpha)^t$$

From the last formula we can see that the summation term shows that the contribution to the smoothed value S_t becomes less at each consecutive time period.

Example for $\alpha = .3$

Let $\alpha = .3$. Observe that the weights $\alpha(1-\alpha)^t$ decrease exponentially (geometrically) with time.

		Value weight
last	y_1	.2100
	y_2	.1470
	y_3	.1029
	y_4	.0720

What is the "best" value for α?

How do you choose

The speed at which the older responses are dampened (smoothed) is a function of the value of α. When α is close to

the weight parameter?

1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below:

----------> towards past observations

α	$(1-\alpha)$	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
.9	.1	.01	.001	.0001
.5	.5	.25	.125	.0625
.1	.9	.81	.729	.6561

We choose the best value for α so the value which results in the smallest MSE.

Example

Let us illustrate this principle with an example. Consider the following data set consisting of 12 observations taken over time:

Time	y_t	***S*** **(α=.1)**	**Error**	**Error squared**
1	71			
2	70	71	-1.00	1.00
3	69	70.9	-1.90	3.61
4	68	70.71	-2.71	7.34
5	64	70.44	-6.44	41.47
6	65	69.80	-4.80	23.04
7	72	69.32	2.68	7.18
8	78	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	75	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

The sum of the squared errors (SSE) = 208.94. The mean of the squared errors (MSE) is the SSE /11 = 19.0.

Calculate for different values of α

The MSE was again calculated for α = .5 and turned out to be 16.29, so in this case we would prefer an α of .5. Can we do better? We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between .1 and .9. We determine the best initial choice for α and then search between α - Δ and α + Δ. We could repeat this perhaps one more time to find the best α to 3 decimal places.

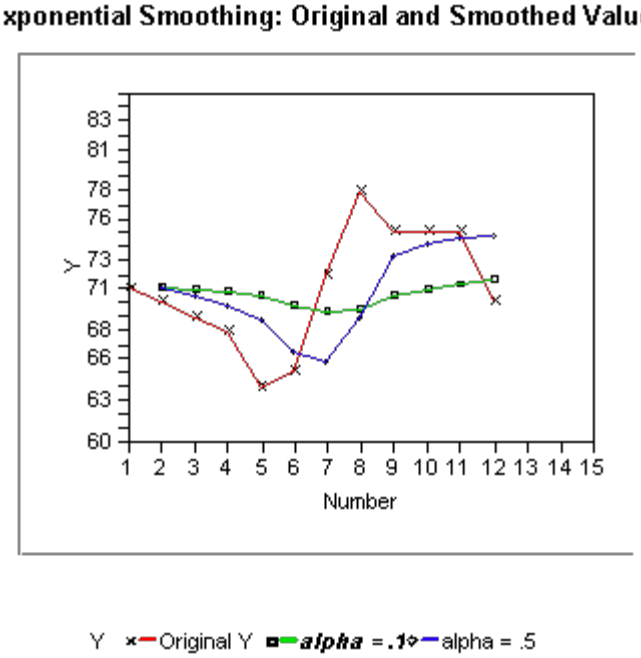
Nonlinear optimizers can be

But there are better search methods, such as the Marquardt procedure. This is a nonlinear optimizer that minimizes the sum of squares of residuals. In general, most well designed

used

statistical software programs should be able to find the value of α that minimizes the MSE.

Sample plot showing smoothed data for 2 values of α



6.4.3.2. Forecasting with Single Exponential Smoothing

Forecasting Formula

Forecasting the next point

The forecasting formula is the basic equation

$$S_{t+1} = \alpha y_t + (1-\alpha) S_t \qquad 0 < \alpha \le 1 \qquad t > 0$$

New forecast is previous forecast plus an error adjustment

This can be written as:

$$S_{t+1} = S_t + \alpha(\varepsilon_t)$$

where ϵ_t is the forecast error (actual - forecast) for period t.

In other words, the new forecast is the old one plus an adjustment for the error that occurred in the last forecast.

Bootstrapping of Forecasts

Bootstrapping forecasts

What happens if you wish to forecast from some origin, usually the last data point, and no actual observations are available? In this situation we have to modify the formula to become:

$$S_{t+1} = \alpha y_{origin} + (1-\alpha) S_t$$

where y_{origin} remains constant. This technique is known as *bootstrapping.*

Example of Bootstrapping

Example

The last data point in the previous example was 70 and its forecast (smoothed value S) was 71.7. Since we do have the data point **and** the forecast available, we can calculate the next forecast using the regular formula

$$S_{t+1} = \alpha y_{origin} + (1-\alpha)S_t$$

$$= .1(70) + .9(71.7) = 71.5 \qquad (\alpha = .1)$$

But for the next forecast we have no data point (observation). So now we compute:

$$S_{t+2} =. 1(70) + .9(71.5) = 71.35$$

Comparison between bootstrap and regular forecasting

Table comparing two methods

The following table displays the comparison between the two methods:

Period	Bootstrap forecast	Data	Single Smoothing Forecast
13	71.50	75	71.5
14	71.35	75	71.9
15	71.21	74	72.2
16	71.09	78	72.4
17	70.98	86	73.0

Single Exponential Smoothing with Trend

Single Smoothing (short for single exponential smoothing) is not very good when there is a trend. The single coefficient α is not enough.

Sample data set with trend

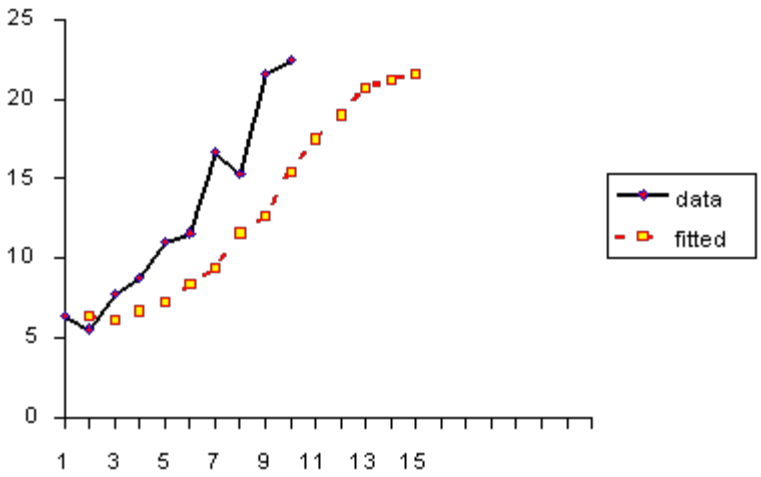
Let us demonstrate this with the following data set smoothed with an α of 0.3:

Data	Fit
6.4	
5.6	6.4
7.8	6.2
8.8	6.7
11.0	7.3
11.6	8.4
16.7	9.4
15.3	11.6
21.6	12.7
22.4	15.4

Plot demonstrating

The resulting graph looks like:

inadequacy of single exponential smoothing when there is trend



6.4.3.3. Double Exponential Smoothing

Double exponential smoothing uses two constants and is better at handling trends

As was previously observed, Single Smoothing does not excel in following the data when there is a trend. This situation can be improved by the introduction of a second equation with a second constant, γ, which must be chosen in conjunction with α.

Here are the two equations associated with Double Exponential Smoothing:

$$S_t = \alpha y_t + (1-\alpha)(S_{t-1} + b_{t-1}) \qquad 0 \le \alpha \le 1$$
$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)\, b_{t-1} \qquad 0 \le \gamma \le 1$$

Note that the current value of the series is used to calculate its smoothed value replacement in double exponential smoothing.

Initial Values

Several methods to choose the initial values

As in the case for single smoothing, there are a variety of schemes to set initial values for S_t and b_t in double smoothing.

S_1 is in general set to y_1. Here are three suggestions for b_1:

$$b_1 = y_2 - y_1$$

$$b_1 = [(y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3)]/3$$

$$b_1 = (y_n - y_1)/(n - 1)$$

Comments

Meaning of the smoothing equations

The first smoothing equation adjusts S_t directly for the trend of the previous period, b_{t-1}, by adding it to the last smoothed value, S_{t-1}. This helps to eliminate the lag and brings S_t to the appropriate base of the current value.

The second smoothing equation then updates the trend, which is expressed as the difference between the last two values. The equation is similar to the basic form of single smoothing, but here applied to the updating of the trend.

Non-linear optimization techniques can be used

The values for α and γ can be obtained via non-linear optimization techniques, such as the Marquardt Algorithm.

6.4.3.4. Forecasting with Double Exponential Smoothing(LASP)

Forecasting formula

The one-period-ahead forecast is given by:

$$F_{t+1} = S_t + b_t$$

The m-periods-ahead forecast is given by:

$$F_{t+m} = S_t + mb_t$$

Example

Example

Consider once more the data set:

6.4, 5.6, 7.8, 8.8, 11, 11.6, 16.7, 15.3, 21.6, 22.4.

Now we will fit a double smoothing model with α = .3623 and γ = 1.0. These are the estimates that result in the lowest possible MSE when comparing the orignal series to one step ahead at a time forecasts (since this version of double exponential smoothing uses the current series value to calculate a smoothed value, the smoothed series cannot be used to determine an α with minimum MSE). The chosen starting values are $S_1 = y_1 = 6.4$ and $b_1 = ((y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3))/3 = 0.8$.

For comparison's sake we also fit a single smoothing model with α = 0.977 (this results in the lowest MSE for single exponential smoothing).

The MSE for double smoothing is 3.7024.
The MSE for single smoothing is 8.8867.

Forecasting results for the example

The smoothed results for the example are:

Data	Double	Single
6.4	6.4	
5.6	6.6 (Forecast = 7.2)	6.4
7.8	7.2 (Forecast = 6.8)	5.6
8.8	8.1 (Forecast = 7.8)	7.8

11.0	9.8 (Forecast = 9.1)	8.8
11.6	11.5 (Forecast = 11.4)	10.9
16.7	14.5 (Forecast = 13.2)	11.6
15.3	16.7 (Forecast = 17.4)	16.6
21.6	19.9 (Forecast = 18.9)	15.3
22.4	22.8 (Forecast = 23.1)	21.5

Comparison of Forecasts

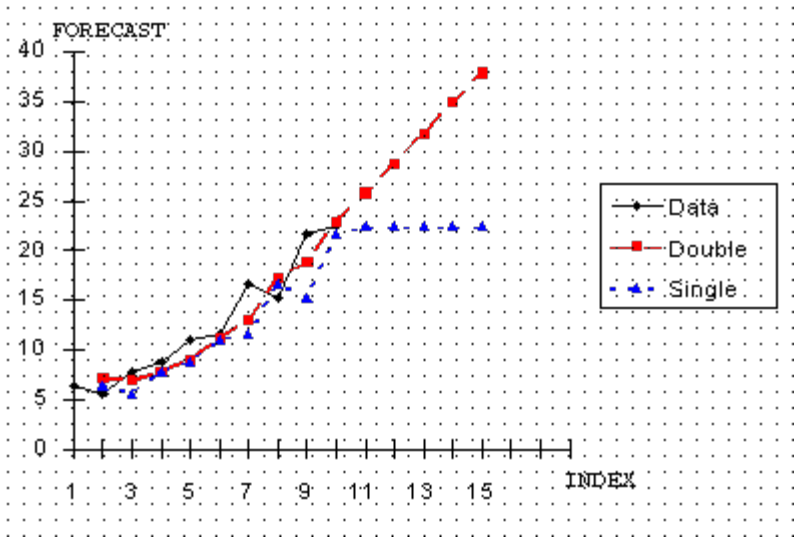
Table showing single and double exponential smoothing forecasts

To see how each method predicts the future, we computed the first five forecasts from the last observation as follows:

Period	Single	Double
11	22.4	25.8
12	22.4	28.7
13	22.4	31.7
14	22.4	34.6
15	22.4	37.6

Plot comparing single and double exponential smoothing forecasts

A plot of these results (using the forecasted double smoothing values) is very enlightening.

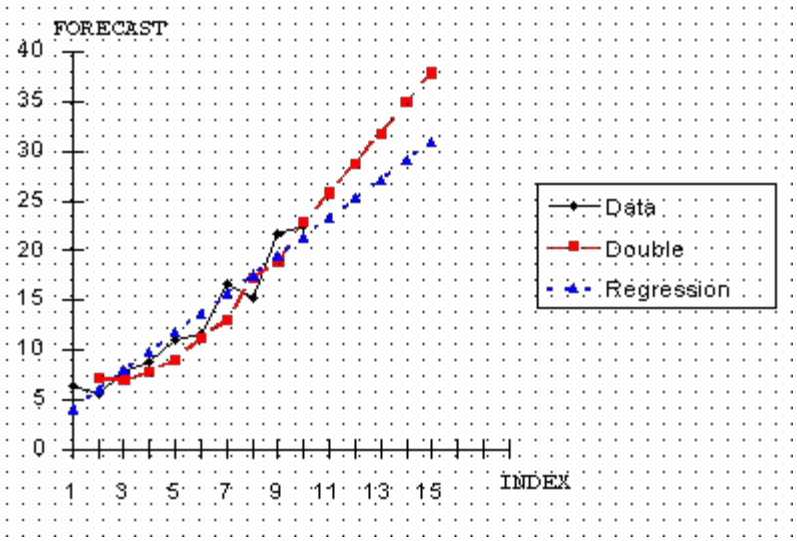


This graph indicates that double smoothing follows the data much closer than single smoothing. Furthermore, for forecasting single smoothing cannot do better than projecting a straight horizontal line, which is not very likely to occur in reality. So in this case double smoothing is preferred.

Plot comparing double exponential smoothing and

Finally, let us compare double smoothing with linear regression:

regression forecasts



This is an interesting picture. Both techniques follow the data in similar fashion, but the regression line is more conservative. That is, there is a slower increase with the regression line than with double smoothing.

Selection of technique depends on the forecaster

The selection of the technique depends on the forecaster. If it is desired to portray the growth process in a more aggressive manner, then one selects double smoothing. Otherwise, regression may be preferable. It should be noted that in linear regression "time" functions as the independent variable. Chapter 4 discusses the basics of linear regression, and the details of regression estimation.

6.4.3.5. Triple Exponential Smoothing

What happens if the data show trend **and** seasonality?

To handle seasonality, we have to add a third parameter

In this case double smoothing will not work. We now introduce a third equation to take care of seasonality (sometimes called periodicity). The resulting set of equations is called the "Holt-Winters" (HW) method after the names of the inventors.

The basic equations for their method are given by:

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \qquad \text{OVERALL SMOOTHING}$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \qquad \text{TREND SMOOTHING}$$

$$I_t = \beta \frac{y_t}{S_t} + (1-\beta) I_{t-L} \qquad \text{SEASONAL SMOOTHING}$$

$$F_{t+m} = (S_t + mb_t) I_{t-L+m} \qquad \text{FORECAST}$$

where

- y is the observation
- S is the smoothed observation
- b is the trend factor
- I is the seasonal index
- F is the forecast at m periods ahead
- t is an index denoting a time period

and α, β, and γ are constants that must be estimated in such a way that the MSE of the error is minimized. This is best left to a good software package.

Complete season needed

To initialize the HW method we need at least one complete season's data to determine initial estimates of the seasonal indices I_{t-L}.

L periods in a season

A complete season's data consists of L periods. And we need to estimate the trend factor from one period to the next. To accomplish this, it is advisable to use two complete seasons; that is, $2L$ periods.

Initial values for the trend factor

How to get initial estimates for trend and seasonality parameters

The general formula to estimate the initial trend is given by

$$b = \frac{1}{L}\left(\frac{y_{L+1} - y_1}{L} + \frac{y_{L+2} - y_2}{L} + \cdots + \frac{y_{L+L} - y_L}{L}\right)$$

Initial values for the Seasonal Indices

As we will see in the example, we work with data that consist of 6 years with 4 periods (that is, 4 quarters) per year. Then

Step 1: compute yearly averages

Step 1: Compute the averages of each of the 6 years

$$A_p = \frac{\sum_{i=1}^{4} y_i}{4} \qquad p = 1, 2, \cdots, 6$$

Step 2: divide by yearly averages

Step 2: Divide the observations by the appropriate yearly mean

1	**2**	**3**	**4**	**5**	**6**
y_1/A_1	y_5/A_2	y_9/A_3	y_{13}/A_4	y_{17}/A_5	y_{21}/A_6
y_2/A_1	y_6/A_2	y_{10}/A_3	y_{14}/A_4	y_{18}/A_5	y_{22}/A_6
y_3/A_1	y_7/A_2	y_{11}/A_3	y_{15}/A_4	y_{19}/A_5	y_{23}/A_6
y_4/A_1	y_8/A_2	y_{12}/A_3	y_{16}/A_4	y_{20}/A_5	y_{24}/A_6

Step 3: form seasonal indices

Step 3: Now the seasonal indices are formed by computing the average of each row. Thus the initial seasonal indices (symbolically) are:

$I_1 = (y_1/A_1 + y_5/A_2 + y_9/A_3 + y_{13}/A_4 + y_{17}/A_5 + y_{21}/A_6)/6$
$I_2 = (y_2/A_1 + y_6/A_2 + y_{10}/A_3 + y_{14}/A_4 + y_{18}/A_5 + y_{22}/A_6)/6$
$I_3 = (y_3/A_1 + y_7/A_2 + y_{11}/A_3 + y_{15}/A_4 + y_{19}/A_5 + y_{22}/A_6)/6$
$I_4 = (y_4/A_1 + y_8/A_2 + y_{12}/A_3 + y_{16}/A_4 + y_{20}/A_5 + y_{24}/A_6)/6$

We now know the algebra behind the computation of the initial estimates.

The next page contains an example of triple exponential smoothing.

The case of the Zero Coefficients

Zero coefficients for trend and seasonality parameters

Sometimes it happens that a computer program for triple exponential smoothing outputs a final coefficient for trend (γ) or for seasonality (β) of zero. Or worse, both are outputted as zero!

Does this indicate that there is no trend and/or no seasonality?

Of course not! It only means that the initial values for trend and/or seasonality were right on the money. No updating was necessary in order to arrive at the lowest possible MSE. We should inspect the updating formulas to verify this.

6.4.3.6. Example of Triple Exponential Smoothing

Example comparing single, double, triple exponential smoothing

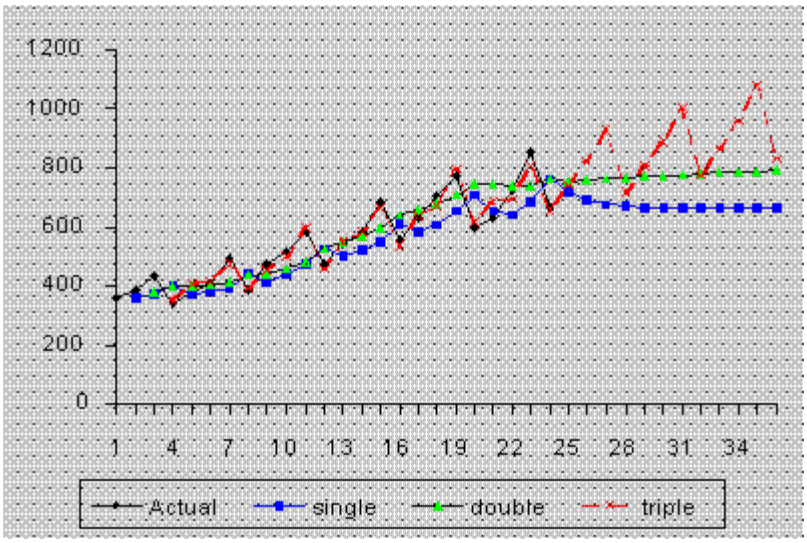
This example shows comparison of single, double and triple exponential smoothing for a data set.

The following data set represents 24 observations. These are six years of quarterly data (each year = 4 quarters).

Table showing the data for the example

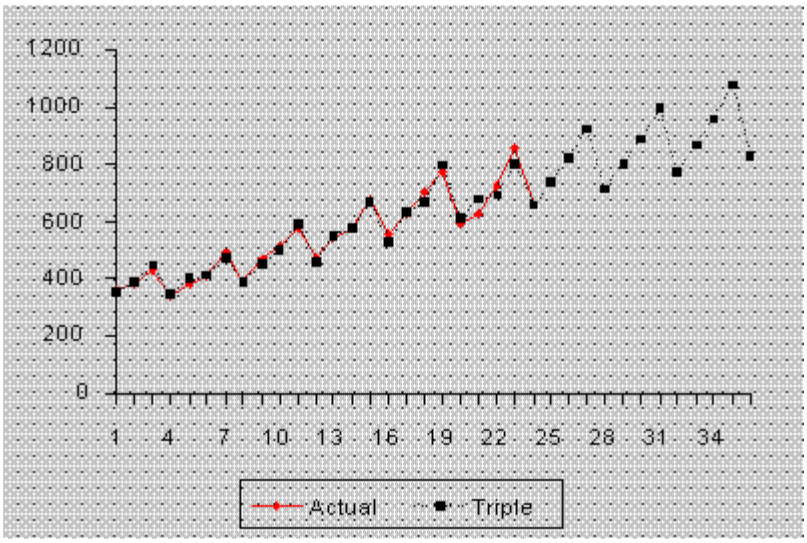
	Quarter	Period	Sales		Quarter	Period	Sales
90	1	1	362	**93**	1	13	544
	2	2	385		2	14	582
	3	3	432		3	15	681
	4	4	341		4	16	557
91	1	5	382	**94**	1	17	628
	2	6	409		2	18	707
	3	7	498		3	19	773
	4	8	387		4	20	592
92	1	9	473	**95**	1	21	627
	2	10	513		2	22	725
	3	11	582		3	23	854
	4	12	474		4	24	661

Plot of raw data with single, double, and triple exponential forecasts



Actual Time Series with forecasts

Plot of raw data with triple exponential forecasts



Comparison of MSE's

Comparison of MSE's

MSE	demand α	trend γ	seasonality β
6906	.4694		
5054	.1086	1.000	
936	1.000		1.000
520	.7556	0.000	.9837

The updating coefficients were chosen by a computer program such that the MSE for each of the methods was minimized.

Example of the computation of the Initial Trend

Computation of initial trend

The data set consists of quarterly sales data. The season is 1 year and since there are 4 quarters per year, $L = 4$. Using the formula we obtain:

$$
\begin{aligned}
b_1 &= \frac{1}{4}\left[\frac{x_5 - x_1}{4}\right] + \left[\frac{x_6 - x_2}{4}\right] + \left[\frac{x_7 - x_3}{4}\right] + \left[\frac{x_8 - x_4}{4}\right] \\
&= \frac{1}{4}\left[\frac{382 - 362}{4}\right] + \left[\frac{409 - 385}{4}\right] + \left[\frac{498 - 432}{4}\right] + \left[\frac{387 - 341}{4}\right] \\
&= \frac{5 + 6 + 16.5 + 11.5}{4} = 9.75
\end{aligned}
$$

Example of the computation of the Initial Seasonal Indices

Table of initial seasonal indices

	1	2	3	4	5	6
1	362	382	473	544	628	627
2	385	409	513	582	707	725
3	432	498	582	681	773	854
4	341	387	474	557	592	661
$\bar{X}$	**380**	**419**	**510.5**	**591**	**675**	**716.75**

In this example we used the full 6 years of data. Other schemes may use only 3, or some other number of years. There are also a number of ways to compute initial estimates.

6.4.3.7. Exponential Smoothing Summary

Summary

Exponential smoothing has proven to be a useful technique

Exponential smoothing has proven through the years to be very useful in many forecasting situations. It was first suggested by C.C. Holt in 1957 and was meant to be used for non-seasonal time series showing no trend. He later offered a procedure (1958) that does handle trends. Winters(1965) generalized the method to include seasonality, hence the name "Holt-Winters Method".

Holt-Winters has 3 updating equations

The Holt-Winters Method has 3 updating equations, each with a constant that ranges from 0 to 1. The equations are intended to give more weight to recent observations and less weights to observations further in the past.

These weights are geometrically decreasing by a constant ratio.

The HW procedure can be made fully automatic by user-friendly software.

6.4.4. Univariate Time Series Models

Univariate Time Series

The term "univariate time series" refers to a time series that consists of single (scalar) observations recorded sequentially over equal time increments. Some examples are monthly CO_2 concentrations and southern oscillations to predict el nino effects.

Although a univariate time series data set is usually given as a single column of numbers, time is in fact an implicit variable in the time series. If the data are equi-spaced, the time variable, or index, does not need to be explicitly given. The time variable may sometimes be explicitly used for plotting the series. However, it is not used in the time series model itself.

The analysis of time series where the data are not collected in equal time increments is beyond the scope of this handbook.

Contents

1. Sample Data Sets
2. Stationarity
3. Seasonality
4. Common Approaches
5. Box-Jenkins Approach
6. Box-Jenkins Model Identification
7. Box-Jenkins Model Estimation
8. Box-Jenkins Model Validation
9. SEMPLOT Sample Output for a Box-Jenkins Analysis
10. SEMPLOT Sample Output for a Box-Jenkins Analysis with Seasonality

6.4.4.1. Sample Data Sets

Sample Data Sets

The following two data sets are used as examples in the text for this section.

1. Monthly mean CO_2 concentrations.
2. Southern oscillations.

6.4.4.1.1. Data Set of Monthly CO2 Concentrations

Source and Background

This data set contains selected monthly mean CO2 concentrations at the Mauna Loa Observatory from 1974 to 1987. The CO2 concentrations were measured by the continuous infrared analyser of the Geophysical Monitoring for Climatic Change division of NOAA's Air Resources Laboratory. The selection has been for an approximation of 'background conditions'. See Thoning et al., "Atmospheric Carbon Dioxide at Mauna Loa Observatory: II Analysis of the NOAA/GMCC Data 1974-1985", *Journal of Geophysical Research* (submitted) for details.

This dataset was received from Jim Elkins of NOAA in 1988.

Data

Each line contains the CO2 concentration (mixing ratio in dry air, expressed in the WMO X85 mole fraction scale, maintained by the Scripps Institution of Oceanography). In addition, it contains the year, month, and a numeric value for the combined month and year. This combined date is useful for plotting purposes.

CO2	Year&Month	Year	Month
333.13	1974.38	1974	5
332.09	1974.46	1974	6
331.10	1974.54	1974	7
329.14	1974.63	1974	8
327.36	1974.71	1974	9
327.29	1974.79	1974	10
328.23	1974.88	1974	11
329.55	1974.96	1974	12
330.62	1975.04	1975	1
331.40	1975.13	1975	2
331.87	1975.21	1975	3
333.18	1975.29	1975	4
333.92	1975.38	1975	5
333.43	1975.46	1975	6
331.85	1975.54	1975	7
330.01	1975.63	1975	8
328.51	1975.71	1975	9
328.41	1975.79	1975	10
329.25	1975.88	1975	11
330.97	1975.96	1975	12
331.60	1976.04	1976	1
332.60	1976.13	1976	2
333.57	1976.21	1976	3
334.72	1976.29	1976	4
334.68	1976.38	1976	5

334.17	1976.46	1976	6
332.96	1976.54	1976	7
330.80	1976.63	1976	8
328.98	1976.71	1976	9
328.57	1976.79	1976	10
330.20	1976.88	1976	11
331.58	1976.96	1976	12
332.67	1977.04	1977	1
333.17	1977.13	1977	2
334.86	1977.21	1977	3
336.07	1977.29	1977	4
336.82	1977.38	1977	5
336.12	1977.46	1977	6
334.81	1977.54	1977	7
332.56	1977.63	1977	8
331.30	1977.71	1977	9
331.22	1977.79	1977	10
332.37	1977.88	1977	11
333.49	1977.96	1977	12
334.71	1978.04	1978	1
335.23	1978.13	1978	2
336.54	1978.21	1978	3
337.79	1978.29	1978	4
337.95	1978.38	1978	5
338.00	1978.46	1978	6
336.37	1978.54	1978	7
334.47	1978.63	1978	8
332.46	1978.71	1978	9
332.29	1978.79	1978	10
333.76	1978.88	1978	11
334.80	1978.96	1978	12
336.00	1979.04	1979	1
336.63	1979.13	1979	2
337.93	1979.21	1979	3
338.95	1979.29	1979	4
339.05	1979.38	1979	5
339.27	1979.46	1979	6
337.64	1979.54	1979	7
335.68	1979.63	1979	8
333.77	1979.71	1979	9
334.09	1979.79	1979	10
335.29	1979.88	1979	11
336.76	1979.96	1979	12
337.77	1980.04	1980	1
338.26	1980.13	1980	2
340.10	1980.21	1980	3
340.88	1980.29	1980	4
341.47	1980.38	1980	5
341.31	1980.46	1980	6
339.41	1980.54	1980	7
337.74	1980.63	1980	8
336.07	1980.71	1980	9
336.07	1980.79	1980	10
337.22	1980.88	1980	11
338.38	1980.96	1980	12
339.32	1981.04	1981	1
340.41	1981.13	1981	2
341.69	1981.21	1981	3
342.51	1981.29	1981	4
343.02	1981.38	1981	5
342.54	1981.46	1981	6
340.88	1981.54	1981	7
338.75	1981.63	1981	8
337.05	1981.71	1981	9
337.13	1981.79	1981	10
338.45	1981.88	1981	11
339.85	1981.96	1981	12
340.90	1982.04	1982	1
341.70	1982.13	1982	2
342.70	1982.21	1982	3
343.65	1982.29	1982	4
344.28	1982.38	1982	5
343.42	1982.46	1982	6
342.02	1982.54	1982	7
339.97	1982.63	1982	8

337.84	1982.71	1982	9
338.00	1982.79	1982	10
339.20	1982.88	1982	11
340.63	1982.96	1982	12
341.41	1983.04	1983	1
342.68	1983.13	1983	2
343.04	1983.21	1983	3
345.27	1983.29	1983	4
345.92	1983.38	1983	5
345.40	1983.46	1983	6
344.16	1983.54	1983	7
342.11	1983.63	1983	8
340.11	1983.71	1983	9
340.15	1983.79	1983	10
341.38	1983.88	1983	11
343.02	1983.96	1983	12
343.87	1984.04	1984	1
344.59	1984.13	1984	2
345.11	1984.21	1984	3
347.07	1984.29	1984	4
347.38	1984.38	1984	5
346.78	1984.46	1984	6
344.96	1984.54	1984	7
342.71	1984.63	1984	8
340.86	1984.71	1984	9
341.13	1984.79	1984	10
342.84	1984.88	1984	11
344.32	1984.96	1984	12
344.88	1985.04	1985	1
345.62	1985.13	1985	2
347.23	1985.21	1985	3
347.62	1985.29	1985	4
348.53	1985.38	1985	5
347.87	1985.46	1985	6
346.00	1985.54	1985	7
343.86	1985.63	1985	8
342.55	1985.71	1985	9
342.57	1985.79	1985	10
344.11	1985.88	1985	11
345.49	1985.96	1985	12
346.04	1986.04	1986	1
346.70	1986.13	1986	2
347.38	1986.21	1986	3
349.38	1986.29	1986	4
349.93	1986.38	1986	5
349.26	1986.46	1986	6
347.44	1986.54	1986	7
345.55	1986.63	1986	8
344.21	1986.71	1986	9
343.67	1986.79	1986	10
345.09	1986.88	1986	11
346.27	1986.96	1986	12
347.33	1987.04	1987	1
347.82	1987.13	1987	2
349.29	1987.21	1987	3
350.91	1987.29	1987	4
351.71	1987.38	1987	5
350.94	1987.46	1987	6
349.10	1987.54	1987	7
346.77	1987.63	1987	8
345.73	1987.71	1987	9

6. Process or Product Monitoring and Control
6.4. Introduction to Time Series Analysis
6.4.4. Univariate Time Series Models
6.4.4.1. Sample Data Sets

6.4.4.1.2. Data Set of Southern Oscillations

Source and Background

The southern oscillation is defined as the barametric pressure difference between Tahiti and the Darwin Islands at sea level. The southern oscillation is a predictor of el nino which in turn is thought to be a driver of world-wide weather. Specifically, repeated southern oscillation values less than -1 typically defines an el nino. Note: the decimal values in the second column of the data given below are obtained as (month number - 0.5)/12.

Data

Southern Oscillation	Year + fraction	Year	Month
-0.7	1955.04	1955	1
1.3	1955.13	1955	2
0.1	1955.21	1955	3
-0.9	1955.29	1955	4
0.8	1955.38	1955	5
1.6	1955.46	1955	6
1.7	1955.54	1955	7
1.4	1955.63	1955	8
1.4	1955.71	1955	9
1.5	1955.79	1955	10
1.4	1955.88	1955	11
0.9	1955.96	1955	12
1.2	1956.04	1956	1
1.1	1956.13	1956	2
0.9	1956.21	1956	3
1.1	1956.29	1956	4
1.4	1956.38	1956	5
1.2	1956.46	1956	6
1.1	1956.54	1956	7
1.0	1956.63	1956	8
0.0	1956.71	1956	9
1.9	1956.79	1956	10
0.1	1956.88	1956	11
0.9	1956.96	1956	12
0.4	1957.04	1957	1
-0.4	1957.13	1957	2
-0.4	1957.21	1957	3
0.0	1957.29	1957	4
-1.1	1957.38	1957	5
-0.4	1957.46	1957	6
0.1	1957.54	1957	7
-1.1	1957.63	1957	8
-1.0	1957.71	1957	9
-0.1	1957.79	1957	10
-1.2	1957.88	1957	11
-0.5	1957.96	1957	12
-1.9	1958.04	1958	1
-0.7	1958.13	1958	2
-0.3	1958.21	1958	3
0.1	1958.29	1958	4

-1.3	1958.38	1958	5
-0.3	1958.46	1958	6
0.3	1958.54	1958	7
0.7	1958.63	1958	8
-0.4	1958.71	1958	9
-0.4	1958.79	1958	10
-0.6	1958.88	1958	11
-0.8	1958.96	1958	12
-0.9	1959.04	1959	1
-1.5	1959.13	1959	2
0.8	1959.21	1959	3
0.2	1959.29	1959	4
0.2	1959.38	1959	5
-0.9	1959.46	1959	6
-0.5	1959.54	1959	7
-0.6	1959.63	1959	8
0.0	1959.71	1959	9
0.3	1959.79	1959	10
0.9	1959.88	1959	11
0.8	1959.96	1959	12
0.0	1960.04	1960	1
-0.2	1960.13	1960	2
0.5	1960.21	1960	3
0.9	1960.29	1960	4
0.2	1960.38	1960	5
-0.5	1960.46	1960	6
0.4	1960.54	1960	7
0.5	1960.63	1960	8
0.7	1960.71	1960	9
-0.1	1960.79	1960	10
0.6	1960.88	1960	11
0.7	1960.96	1960	12
-0.4	1961.04	1961	1
0.5	1961.13	1961	2
-2.6	1961.21	1961	3
1.1	1961.29	1961	4
0.2	1961.38	1961	5
-0.4	1961.46	1961	6
0.1	1961.54	1961	7
-0.3	1961.63	1961	8
0.0	1961.71	1961	9
-0.8	1961.79	1961	10
0.7	1961.88	1961	11
1.4	1961.96	1961	12
1.7	1962.04	1962	1
-0.5	1962.13	1962	2
-0.4	1962.21	1962	3
0.0	1962.29	1962	4
1.2	1962.38	1962	5
0.5	1962.46	1962	6
-0.1	1962.54	1962	7
0.3	1962.63	1962	8
0.5	1962.71	1962	9
0.9	1962.79	1962	10
0.2	1962.88	1962	11
0.0	1962.96	1962	12
0.8	1963.04	1963	1
0.3	1963.13	1963	2
0.6	1963.21	1963	3
0.9	1963.29	1963	4
0.0	1963.38	1963	5
-1.5	1963.46	1963	6
-0.3	1963.54	1963	7
-0.4	1963.63	1963	8
-0.7	1963.71	1963	9
-1.6	1963.79	1963	10
-1.0	1963.88	1963	11
-1.4	1963.96	1963	12
-0.5	1964.04	1964	1
-0.2	1964.13	1964	2
0.6	1964.21	1964	3
1.7	1964.29	1964	4
-0.2	1964.38	1964	5
0.7	1964.46	1964	6
0.5	1964.54	1964	7

1.4	1964.63	1964	8
1.3	1964.71	1964	9
1.3	1964.79	1964	10
0.0	1964.88	1964	11
-0.5	1964.96	1964	12
-0.5	1965.04	1965	1
0.0	1965.13	1965	2
0.2	1965.21	1965	3
-1.1	1965.29	1965	4
0.0	1965.38	1965	5
-1.5	1965.46	1965	6
-2.3	1965.54	1965	7
-1.3	1965.63	1965	8
-1.4	1965.71	1965	9
-1.2	1965.79	1965	10
-1.8	1965.88	1965	11
0.0	1965.96	1965	12
-1.4	1966.04	1966	1
-0.5	1966.13	1966	2
-1.6	1966.21	1966	3
-0.7	1966.29	1966	4
-0.6	1966.38	1966	5
0.0	1966.46	1966	6
-0.1	1966.54	1966	7
0.3	1966.63	1966	8
-0.3	1966.71	1966	9
-0.3	1966.79	1966	10
-0.1	1966.88	1966	11
-0.5	1966.96	1966	12
1.5	1967.04	1967	1
1.2	1967.13	1967	2
0.8	1967.21	1967	3
-0.2	1967.29	1967	4
-0.4	1967.38	1967	5
0.6	1967.46	1967	6
0.0	1967.54	1967	7
0.4	1967.63	1967	8
0.5	1967.71	1967	9
-0.2	1967.79	1967	10
-0.7	1967.88	1967	11
-0.7	1967.96	1967	12
0.5	1968.04	1968	1
0.8	1968.13	1968	2
-0.5	1968.21	1968	3
-0.3	1968.29	1968	4
1.2	1968.38	1968	5
1.4	1968.46	1968	6
0.6	1968.54	1968	7
-0.1	1968.63	1968	8
-0.3	1968.71	1968	9
-0.3	1968.79	1968	10
-0.4	1968.88	1968	11
0.0	1968.96	1968	12
-1.4	1969.04	1969	1
0.8	1969.13	1969	2
-0.1	1969.21	1969	3
-0.8	1969.29	1969	4
-0.8	1969.38	1969	5
-0.2	1969.46	1969	6
-0.7	1969.54	1969	7
-0.6	1969.63	1969	8
-1.0	1969.71	1969	9
-1.4	1969.79	1969	10
-0.1	1969.88	1969	11
0.3	1969.96	1969	12
-1.2	1970.04	1970	1
-1.2	1970.13	1970	2
0.0	1970.21	1970	3
-0.5	1970.29	1970	4
0.1	1970.38	1970	5
1.1	1970.46	1970	6
-0.6	1970.54	1970	7
0.3	1970.63	1970	8
1.2	1970.71	1970	9
0.8	1970.79	1970	10

1.8	1970.88	1970	11
1.8	1970.96	1970	12
0.2	1971.04	1971	1
1.4	1971.13	1971	2
2.0	1971.21	1971	3
2.6	1971.29	1971	4
0.9	1971.38	1971	5
0.2	1971.46	1971	6
0.1	1971.54	1971	7
1.4	1971.63	1971	8
1.5	1971.71	1971	9
1.8	1971.79	1971	10
0.5	1971.88	1971	11
0.1	1971.96	1971	12
0.3	1972.04	1972	1
0.6	1972.13	1972	2
0.1	1972.21	1972	3
-0.5	1972.29	1972	4
-2.1	1972.38	1972	5
-1.7	1972.46	1972	6
-1.9	1972.54	1972	7
-1.1	1972.63	1972	8
-1.5	1972.71	1972	9
-1.1	1972.79	1972	10
-0.4	1972.88	1972	11
-1.5	1972.96	1972	12
-0.4	1973.04	1973	1
-1.5	1973.13	1973	2
0.2	1973.21	1973	3
-0.4	1973.29	1973	4
0.3	1973.38	1973	5
1.2	1973.46	1973	6
0.5	1973.54	1973	7
1.2	1973.63	1973	8
1.3	1973.71	1973	9
0.6	1973.79	1973	10
2.9	1973.88	1973	11
1.7	1973.96	1973	12
2.2	1974.04	1974	1
1.5	1974.13	1974	2
2.1	1974.21	1974	3
1.3	1974.29	1974	4
1.3	1974.38	1974	5
0.1	1974.46	1974	6
1.2	1974.54	1974	7
0.5	1974.63	1974	8
1.1	1974.71	1974	9
0.8	1974.79	1974	10
-0.4	1974.88	1974	11
0.0	1974.96	1974	12
-0.6	1975.04	1975	1
0.4	1975.13	1975	2
1.1	1975.21	1975	3
1.5	1975.29	1975	4
0.5	1975.38	1975	5
1.7	1975.46	1975	6
2.1	1975.54	1975	7
2.0	1975.63	1975	8
2.2	1975.71	1975	9
1.7	1975.79	1975	10
1.3	1975.88	1975	11
2.0	1975.96	1975	12
1.2	1976.04	1976	1
1.2	1976.13	1976	2
1.3	1976.21	1976	3
0.2	1976.29	1976	4
0.6	1976.38	1976	5
-0.1	1976.46	1976	6
-1.2	1976.54	1976	7
-1.5	1976.63	1976	8
-1.2	1976.71	1976	9
0.2	1976.79	1976	10
0.7	1976.88	1976	11
-0.5	1976.96	1976	12

-0.5	1977.04	1977	1
0.8	1977.13	1977	2
-1.2	1977.21	1977	3
-1.3	1977.29	1977	4
-1.1	1977.38	1977	5
-2.3	1977.46	1977	6
-1.5	1977.54	1977	7
-1.4	1977.63	1977	8
-0.9	1977.71	1977	9
-1.4	1977.79	1977	10
-1.6	1977.88	1977	11
-1.3	1977.96	1977	12
-0.5	1978.04	1978	1
-2.6	1978.13	1978	2
-0.8	1978.21	1978	3
-0.9	1978.29	1978	4
1.3	1978.38	1978	5
0.4	1978.46	1978	6
0.4	1978.54	1978	7
0.1	1978.63	1978	8
0.0	1978.71	1978	9
-0.8	1978.79	1978	10
-0.1	1978.88	1978	11
-0.2	1978.96	1978	12
-0.5	1979.04	1979	1
0.6	1979.13	1979	2
-0.5	1979.21	1979	3
-0.7	1979.29	1979	4
0.5	1979.38	1979	5
0.6	1979.46	1979	6
1.3	1979.54	1979	7
-0.7	1979.63	1979	8
0.1	1979.71	1979	9
-0.4	1979.79	1979	10
-0.6	1979.88	1979	11
-0.9	1979.96	1979	12
0.3	1980.04	1980	1
0.0	1980.13	1980	2
-1.1	1980.21	1980	3
-1.7	1980.29	1980	4
-0.3	1980.38	1980	5
-0.7	1980.46	1980	6
-0.2	1980.54	1980	7
-0.1	1980.63	1980	8
-0.5	1980.71	1980	9
-0.3	1980.79	1980	10
-0.5	1980.88	1980	11
-0.2	1980.96	1980	12
0.3	1981.04	1981	1
-0.5	1981.13	1981	2
-2.0	1981.21	1981	3
-0.6	1981.29	1981	4
0.8	1981.38	1981	5
1.6	1981.46	1981	6
0.8	1981.54	1981	7
0.4	1981.63	1981	8
0.3	1981.71	1981	9
-0.7	1981.79	1981	10
0.1	1981.88	1981	11
0.4	1981.96	1981	12
1.0	1982.04	1982	1
0.0	1982.13	1982	2
0.0	1982.21	1982	3
-0.1	1982.29	1982	4
-0.6	1982.38	1982	5
-2.5	1982.46	1982	6
-2.0	1982.54	1982	7
-2.7	1982.63	1982	8
-1.9	1982.71	1982	9
-2.2	1982.79	1982	10
-3.2	1982.88	1982	11
-2.5	1982.96	1982	12
-3.4	1983.04	1983	1
-3.5	1983.13	1983	2
-3.2	1983.21	1983	3

-2.1	1983.29	1983	4
0.9	1983.38	1983	5
-0.5	1983.46	1983	6
-0.9	1983.54	1983	7
-0.4	1983.63	1983	8
0.9	1983.71	1983	9
0.3	1983.79	1983	10
-0.1	1983.88	1983	11
-0.1	1983.96	1983	12
0.0	1984.04	1984	1
0.4	1984.13	1984	2
-0.8	1984.21	1984	3
0.4	1984.29	1984	4
0.0	1984.38	1984	5
-1.2	1984.46	1984	6
0.0	1984.54	1984	7
0.1	1984.63	1984	8
0.1	1984.71	1984	9
-0.6	1984.79	1984	10
0.3	1984.88	1984	11
-0.3	1984.96	1984	12
-0.5	1985.04	1985	1
0.8	1985.13	1985	2
0.2	1985.21	1985	3
1.4	1985.29	1985	4
-0.2	1985.38	1985	5
-1.4	1985.46	1985	6
-0.3	1985.54	1985	7
0.7	1985.63	1985	8
0.0	1985.71	1985	9
-0.8	1985.79	1985	10
-0.4	1985.88	1985	11
0.1	1985.96	1985	12
0.8	1986.04	1986	1
-1.2	1986.13	1986	2
-0.1	1986.21	1986	3
0.1	1986.29	1986	4
-0.6	1986.38	1986	5
1.0	1986.46	1986	6
0.1	1986.54	1986	7
-0.9	1986.63	1986	8
-0.5	1986.71	1986	9
0.6	1986.79	1986	10
-1.6	1986.88	1986	11
-1.6	1986.96	1986	12
-0.7	1987.04	1987	1
-1.4	1987.13	1987	2
-2.0	1987.21	1987	3
-2.7	1987.29	1987	4
-2.0	1987.38	1987	5
-2.7	1987.46	1987	6
-1.8	1987.54	1987	7
-1.7	1987.63	1987	8
-1.1	1987.71	1987	9
-0.7	1987.79	1987	10
-0.1	1987.88	1987	11
-0.6	1987.96	1987	12
-0.3	1988.04	1988	1
-0.6	1988.13	1988	2
0.1	1988.21	1988	3
0.0	1988.29	1988	4
1.1	1988.38	1988	5
-0.3	1988.46	1988	6
1.1	1988.54	1988	7
1.4	1988.63	1988	8
1.9	1988.71	1988	9
1.5	1988.79	1988	10
1.9	1988.88	1988	11
1.1	1988.96	1988	12
1.5	1989.04	1989	1
1.1	1989.13	1989	2
0.6	1989.21	1989	3
1.6	1989.29	1989	4
1.2	1989.38	1989	5
0.5	1989.46	1989	6

0.8	1989.54	1989	7
-0.8	1989.63	1989	8
0.6	1989.71	1989	9
0.6	1989.79	1989	10
-0.4	1989.88	1989	11
-0.7	1989.96	1989	12
-0.2	1990.04	1990	1
-2.4	1990.13	1990	2
-1.2	1990.21	1990	3
0.0	1990.29	1990	4
1.1	1990.38	1990	5
0.0	1990.46	1990	6
0.5	1990.54	1990	7
-0.5	1990.63	1990	8
-0.8	1990.71	1990	9
0.1	1990.79	1990	10
-0.7	1990.88	1990	11
-0.4	1990.96	1990	12
0.6	1991.04	1991	1
-0.1	1991.13	1991	2
-1.4	1991.21	1991	3
-1.0	1991.29	1991	4
-1.5	1991.38	1991	5
-0.5	1991.46	1991	6
-0.2	1991.54	1991	7
-0.9	1991.63	1991	8
-1.8	1991.71	1991	9
-1.5	1991.79	1991	10
-0.8	1991.88	1991	11
-2.3	1991.96	1991	12
-3.4	1992.04	1992	1
-1.4	1992.13	1992	2
-3.0	1992.21	1992	3
-1.4	1992.29	1992	4
0.0	1992.38	1992	5
-1.2	1992.46	1992	6
-0.8	1992.54	1992	7
0.0	1992.63	1992	8
0.0	1992.71	1992	9
-1.9	1992.79	1992	10
-0.9	1992.88	1992	11
-1.1	1992.96	1992	12

6.4.4.2. Stationarity

Stationarity

A common assumption in many time series techniques is that the data are stationary.

A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

For practical purposes, stationarity can usually be determined from a run sequence plot.

Transformations to Achieve Stationarity

If the time series is not stationary, we can often transform it to stationarity with one of the following techniques.

1. We can difference the data. That is, given the series Z_t, we create the new series

$$Y_i = Z_i - Z_{i-1}$$

   The differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is usually sufficient.

2. If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove long term trend, a simple fit, such as a straight line, is typically used.

3. For non-constant variance, taking the logarithm or square root of the series may stabilize the variance. For negative data, you can add a suitable constant to make all the data positive before applying the transformation. This constant can then be subtracted from the model to obtain predicted (i.e., the fitted) values and forecasts for future points.

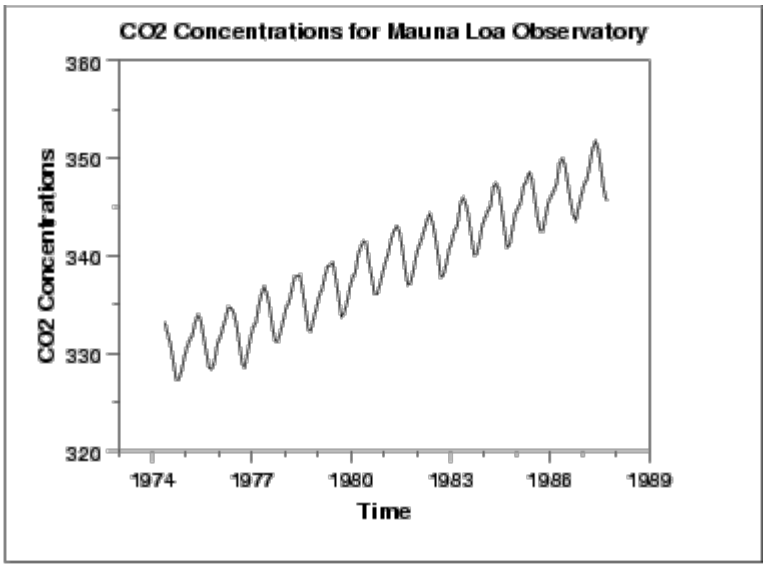
The above techniques are intended to generate series with

constant location and scale. Although seasonality also violates stationarity, this is usually explicitly incorporated into the time series model.

Example

The following plots are from a data set of monthly CO2 concentrations.

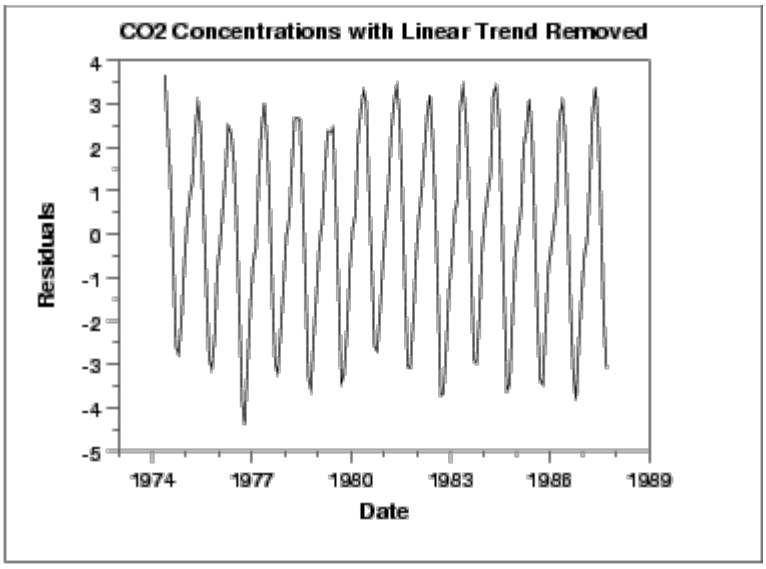
Run Sequence Plot



The initial run sequence plot of the data indicates a rising trend. A visual inspection of this plot indicates that a simple linear fit should be sufficient to remove this upward trend.

This plot also shows periodical behavior. This is discussed in the next section.

Linear Trend Removed



This plot contains the residuals from a linear fit to the original data. After removing the linear trend, the run sequence plot indicates that the data have a constant location and variance, although the pattern of the residuals shows that the data depart from the model in a systematic way.

6.4.4.3. Seasonality

Seasonality

Many time series display seasonality. By seasonality, we mean periodic fluctuations. For example, retail sales tend to peak for the Christmas season and then decline after the holidays. So time series of retail sales will typically show increasing sales from September through December and declining sales in January and February.

Seasonality is quite common in economic time series. It is less common in engineering and scientific data.

If seasonality is present, it must be incorporated into the time series model. In this section, we discuss techniques for detecting seasonality. We defer modeling of seasonality until later sections.

Detecting Seasonality

he following graphical techniques can be used to detect seasonality.

1. A run sequence plot will often show seasonality.
2. A seasonal subseries plot is a specialized technique for showing seasonality.
3. Multiple box plots can be used as an alternative to the seasonal subseries plot to detect seasonality.
4. The autocorrelation plot can help identify seasonality.

Examples of each of these plots will be shown below.

The run sequence plot is a recommended first step for analyzing any time series. Although seasonality can sometimes be indicated with this plot, seasonality is shown more clearly by the seasonal subseries plot or the box plot. The seasonal subseries plot does an excellent job of showing both the seasonal differences (between group patterns) and also the within-group patterns. The box plot shows the seasonal difference (between group patterns) quite well, but it does not show within group patterns. However, for large data sets, the box plot is usually easier to read than the seasonal subseries plot.

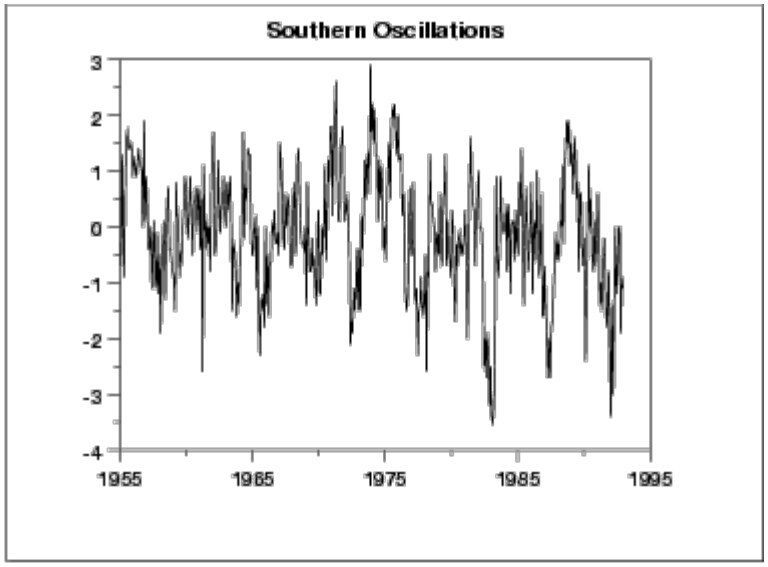
Both the seasonal subseries plot and the box plot assume that

the seasonal periods are known. In most cases, the analyst will in fact know this. For example, for monthly data, the period is 12 since there are 12 months in a year. However, if the period is not known, the autocorrelation plot can help. If there is significant seasonality, the autocorrelation plot should show spikes at lags equal to the period. For example, for monthly data, if there is a seasonality effect, we would expect to see significant peaks at lag 12, 24, 36, and so on (although the intensity may decrease the further out we go).

Example without Seasonality

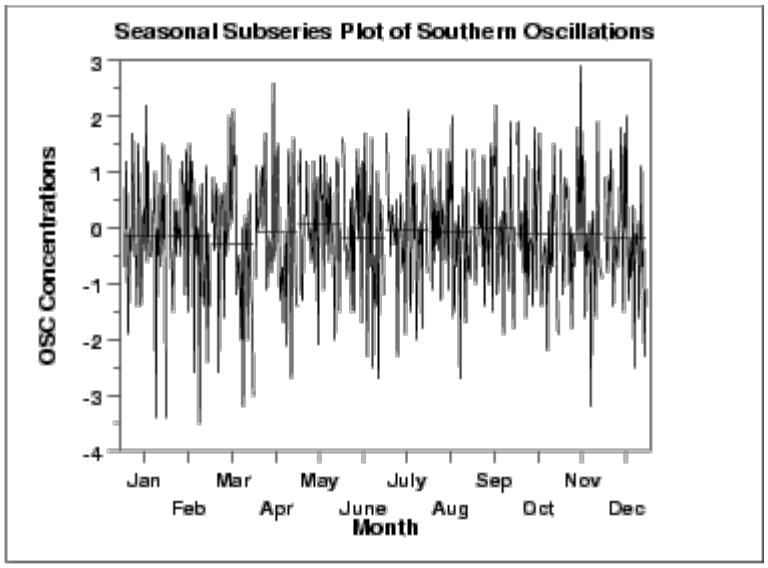
The following plots are from a data set of southern oscillations for predicting el nino.

Run Sequence Plot



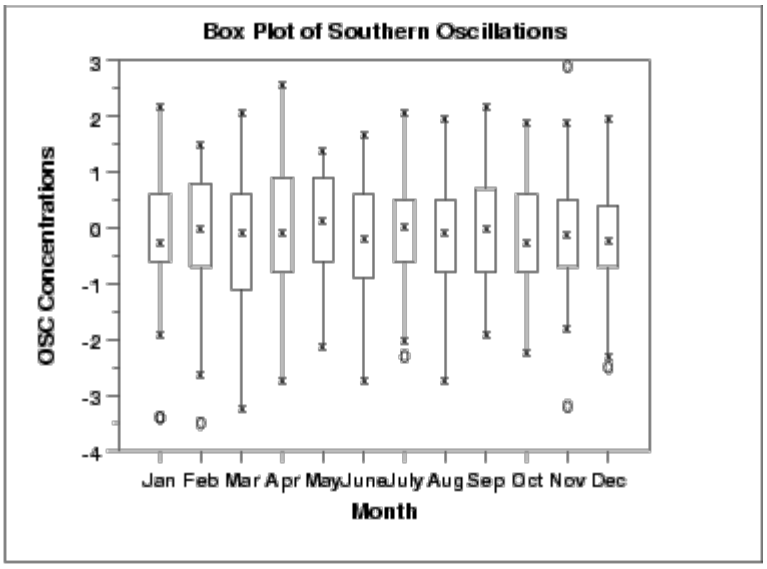
No obvious periodic patterns are apparent in the run sequence plot.

Seasonal Subseries Plot



The means for each month are relatively close and show no obvious pattern.

Box Plot



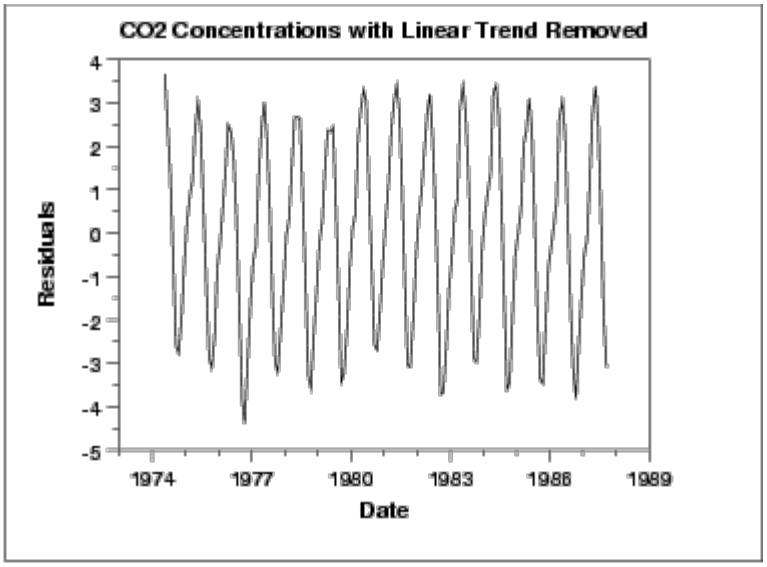
As with the seasonal subseries plot, no obvious seasonal pattern is apparent.

Due to the rather large number of observations, the box plot shows the difference between months better than the seasonal subseries plot.

Example with Seasonality

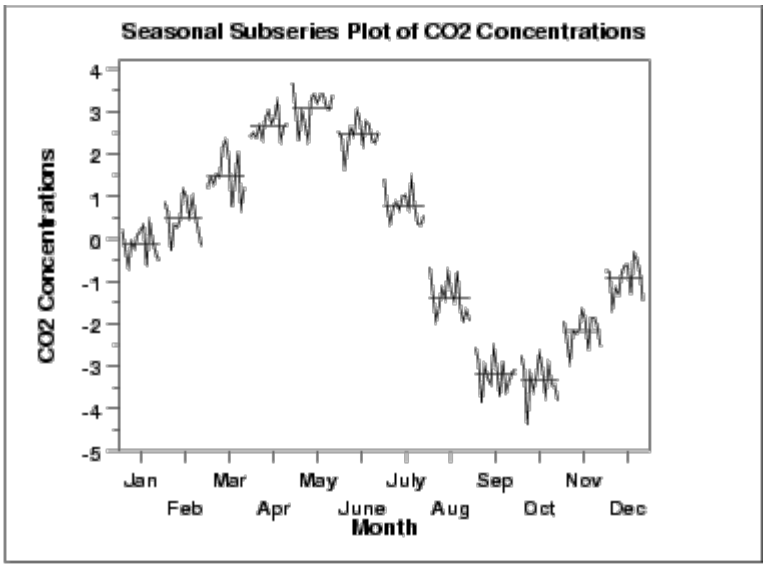
The following plots are from a data set of monthly CO2 concentrations. A linear trend has been removed from these data.

Run Sequence Plot



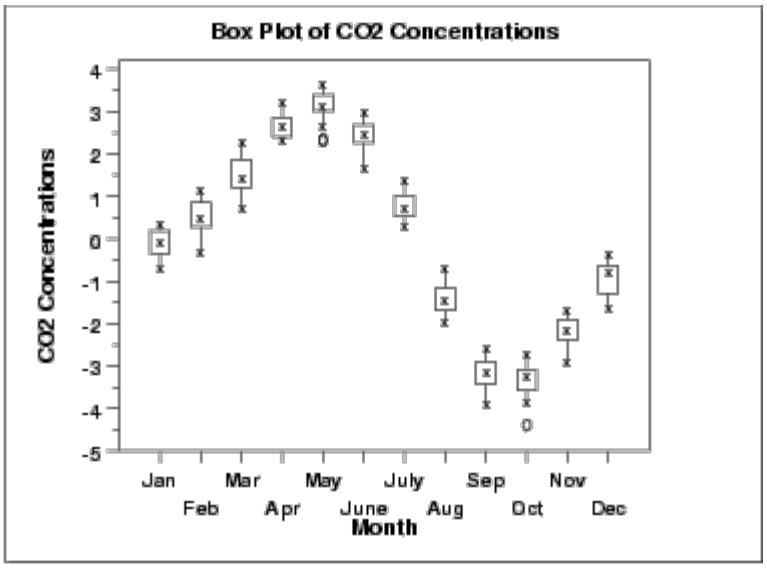
This plot shows periodic behavior. However, it is difficult to determine the nature of the seasonality from this plot.

Seasonal Subseries Plot



The seasonal subseries plot shows the seasonal pattern more clearly. In this case, the CO_2 concentrations are at a minimun in September and October. From there, steadily the concentrations increase until June and then begin declining until September.

Box Plot



As with the seasonal subseries plot, the seasonal pattern is quite evident in the box plot.

6.4.4.3.1. Seasonal Subseries Plot

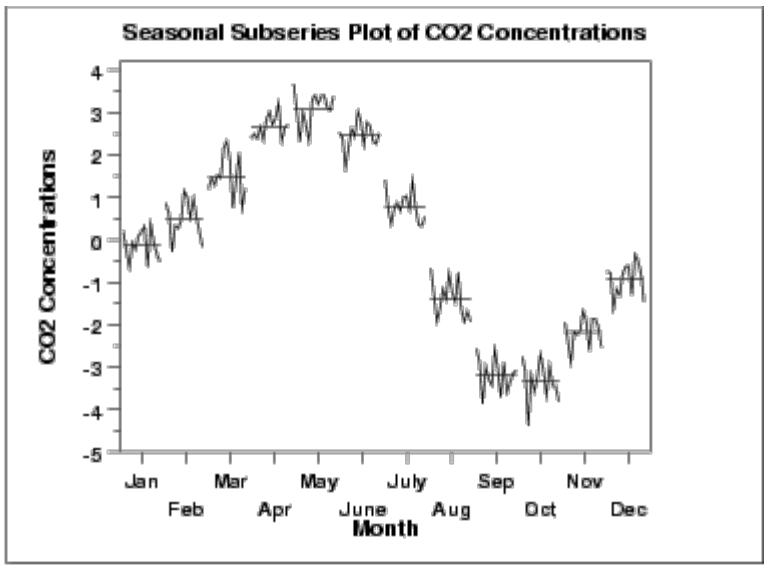
Purpose

Seasonal subseries plots (Cleveland 1993) are a tool for detecting seasonality in a time series.

This plot is only useful if the period of the seasonality is already known. In many cases, this will in fact be known. For example, monthly data typically has a period of 12.

If the period is not known, an autocorrelation plot or spectral plot can be used to determine it.

Sample Plot



This seasonal subseries plot containing monthly data of CO_2 concentrations reveals a strong seasonality pattern. The CO_2 concentrations peak in May, steadily decrease through September, and then begin rising again until the May peak.

This plot allows you to detect both between group and within group patterns.

If there is a large number of observations, then a box plot may be preferable.

Definition

Seasonal subseries plots are formed by

Vertical Response variable

axis:

Horizontal axis: Time ordered by season. For example, with monthly data, all the January values are plotted (in chronological order), then all the February values, and so on.

In addition, a reference line is drawn at the group means.

The user must specify the length of the seasonal pattern before generating this plot. In most cases, the analyst will know this from the context of the problem and data collection. Sometimes the series will need to be detrended before generating the plot, as was the case for the CO_2 data.

Questions

The seasonal subseries plot can provide answers to the following questions:

1. Do the data exhibit a seasonal pattern?
2. What is the nature of the seasonality?
3. Is there a within-group pattern (e.g., do January and July exhibit similar patterns)?
4. Are there any outliers once seasonality has been accounted for?

Importance

It is important to know when analyzing a time series if there is a significant seasonality effect. The seasonal subseries plot is an excellent tool for determining if there is a seasonal pattern.

Related Techniques

Box Plot
Run Sequence Plot
Autocorrelation Plot

Software

Seasonal subseries plots are available in a few general purpose statistical software programs. It may possible to write macros to generate this plot in most statistical software programs that do not provide it directly. Seasonal subseries plots can be generated using both Dataplot code and R code.

6.4.4.4. Common Approaches to Univariate Time Series

There are a number of approaches to modeling time series. We outline a few of the most common approaches below.

Trend, Seasonal, Residual Decompositions

One approach is to decompose the time series into a trend, seasonal, and residual component.

Triple exponential smoothing is an example of this approach. Another example, called seasonal *loess*, is based on locally weighted least squares and is discussed by Cleveland (1993). We do not discuss seasonal *loess* in this handbook.

Frequency Based Methods

Another approach, commonly used in scientific and engineering applications, is to analyze the series in the frequency domain. An example of this approach in modeling a sinusoidal type data set is shown in the beam deflection case study. The spectral plot is the primary tool for the frequency analysis of time series.

Detailed discussions of frequency-based methods are included in Bloomfield (1976), Jenkins and Watts (1968), and Chatfield (1996).

Autoregressive (AR) Models

A common approach for modeling univariate time series is the autoregressive (AR) model:

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t$$

where X_t is the time series, A_t is white noise, and

$$\delta = (1 - \sum_{i=1}^{p} \phi_i)\mu$$

with μ denoting the process mean.

An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. The value of p is called the order of the AR model.

AR models can be analyzed with one of various methods,

including standard linear least squares techniques. They also have a straightforward interpretation.

Moving Average (MA) Models

Another common approach for modeling univariate time series models is the moving average (MA) model:

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q}$$

where X_t is the time series, μ is the mean of the series, A_{t-i} are white noise, and $\theta_1, \ldots, \theta_q$ are the parameters of the model. The value of q is called the order of the MA model.

That is, a moving average model is conceptually a linear regression of the current value of the series against the white noise or random shocks of one or more prior values of the series. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propogated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.

Sometimes the ACF and PACF will suggest that a MA model would be a better model choice and sometimes both AR and MA terms should be used in the same model (see Section 6.4.4.5).

Note, however, that the error terms *after* the model is fit should be independent and follow the standard assumptions for a univariate process.

Box-Jenkins Approach

Box and Jenkins popularized an approach that combines the moving average and the autoregressive approaches in the book "*Time Series Analysis: Forecasting and Control*" (Box, Jenkins, and Reinsel, 1994).

Although both autoregressive and moving average approaches were already known (and were originally investigated by Yule), the contribution of Box and Jenkins was in developing a systematic methodology for identifying and estimating models that could incorporate both approaches. This makes Box-Jenkins models a powerful class of models. The next several sections will discuss these models in detail.

6.4.4.5. Box-Jenkins Models

Box-Jenkins Approach

The Box-Jenkins ARMA model is a combination of the AR and MA models (described on the previous page):

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \ldots - \theta_q A_{t-q}$$

where the terms in the equation have the same meaning as given for the AR and MA model.

Comments on Box-Jenkins Model

A couple of notes on this model.

1. The Box-Jenkins model assumes that the time series is stationary. Box and Jenkins recommend differencing non-stationary series one or more times to achieve stationarity. Doing so produces an ARIMA model, with the "I" standing for "Integrated".

2. Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero. Whether you need to do this or not is dependent on the software you use to estimate the model.

3. Box-Jenkins models can be extended to include seasonal autoregressive and seasonal moving average terms. Although this complicates the notation and mathematics of the model, the underlying concepts for seasonal autoregressive and seasonal moving average terms are similar to the non-seasonal autoregressive and moving average terms.

4. The most general Box-Jenkins model includes difference operators, autoregressive terms, moving average terms, seasonal difference operators, seasonal autoregressive terms, and seasonal moving average terms. As with modeling in general, however, only necessary terms should be included in the model. Those interested in the mathematical details can consult Box, Jenkins and Reisel (1994), Chatfield (1996), or Brockwell and Davis (2002).

Stages in Box-Jenkins Modeling

There are three primary stages in building a Box-Jenkins time series model.

1. Model Identification
2. Model Estimation
3. Model Validation

Remarks

The following remarks regarding Box-Jenkins models should be noted.

1. Box-Jenkins models are quite flexible due to the inclusion of both autoregressive and moving average terms.

2. Based on the Wold decomposition thereom (not discussed in the Handbook), a stationary process can be approximated by an ARMA model. In practice, finding that approximation may not be easy.

3. Chatfield (1996) recommends decomposition methods for series in which the trend and seasonal components are dominant.

4. Building good ARIMA models generally requires more experience than commonly used statistical methods such as regression.

Sufficiently Long Series Required

Typically, effective fitting of Box-Jenkins models requires at least a moderately long series. Chatfield (1996) recommends at least 50 observations. Many others would recommend at least 100 observations.

6.4.4.6. Box-Jenkins Model Identification

Stationarity and Seasonality	The first step in developing a Box-Jenkins model is to determine if the series is stationary and if there is any significant seasonality that needs to be modeled.
Detecting stationarity	Stationarity can be assessed from a run sequence plot. The run sequence plot should show constant location and scale. It can also be detected from an autocorrelation plot. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay.
Detecting seasonality	Seasonality (or periodicity) can usually be assessed from an autocorrelation plot, a seasonal subseries plot, or a spectral plot.
Differencing to achieve stationarity	Box and Jenkins recommend the differencing approach to achieve stationarity. However, fitting a curve and subtracting the fitted values from the original data can also be used in the context of Box-Jenkins models.
Seasonal differencing	At the model identification stage, our goal is to detect seasonality, if it exists, and to identify the order for the seasonal autoregressive and seasonal moving average terms. For many series, the period is known and a single seasonality term is sufficient. For example, for monthly data we would typically include either a seasonal AR 12 term or a seasonal MA 12 term. For Box-Jenkins models, we do not explicitly remove seasonality before fitting the model. Instead, we include the order of the seasonal terms in the model specification to the ARIMA estimation software. However, it may be helpful to apply a seasonal difference to the data and regenerate the autocorrelation and partial autocorrelation plots. This may help in the model idenfitication of the non-seasonal component of the model. In some cases, the seasonal differencing may remove most or all of the seasonality effect.
Identify p and q	Once stationarity and seasonality have been addressed, the next step is to identify the order (i.e., the ***p*** and ***q***) of the autoregressive and moving average terms.
Autocorrelation	The primary tools for doing this are the autocorrelation

and Partial Autocorrelation Plots

plot and the partial autocorrelation plot. The sample autocorrelation plot and the sample partial autocorrelation plot are compared to the theoretical behavior of these plots when the order is known.

Order of Autoregressive Process (p)

Specifically, for an AR(1) process, the sample autocorrelation function should have an exponentially decreasing appearance. However, higher-order AR processes are often a mixture of exponentially decreasing and damped sinusoidal components.

For higher-order autoregressive processes, the sample autocorrelation needs to be supplemented with a partial autocorrelation plot. The partial autocorrelation of an AR(***p***) process becomes zero at lag ***p*+1** and greater, so we examine the sample partial autocorrelation function to see if there is evidence of a departure from zero. This is usually determined by placing a 95% confidence interval on the sample partial autocorrelation plot (most software programs that generate sample autocorrelation plots will also plot this confidence interval). If the software program does not generate the confidence band, it is approximately $\pm 2/\sqrt{N}$, with N denoting the sample size.

Order of Moving Average Process (q)

The autocorrelation function of a MA(***q***) process becomes zero at lag ***q*+1** and greater, so we examine the sample autocorrelation function to see where it essentially becomes zero. We do this by placing the 95% confidence interval for the sample autocorrelation function on the sample autocorrelation plot. Most software that can generate the autocorrelation plot can also generate this confidence interval.

The sample partial autocorrelation function is generally not helpful for identifying the order of the moving average process.

Shape of Autocorrelation Function

The following table summarizes how we use the sample autocorrelation function for model identification.

SHAPE	INDICATED MODEL
Exponential, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to identify the order of the autoregressive model.
Alternating positive and negative, decaying to zero	Autoregressive model. Use the partial autocorrelation plot to help identify the order.

One or more spikes, rest are essentially zero	Moving average model, order identified by where plot becomes zero.
Decay, starting after a few lags	Mixed autoregressive and moving average model.
All zero or close to zero	Data is essentially random.
High values at fixed intervals	Include seasonal autoregressive term.
No decay to zero	Series is not stationary.

Mixed Models Difficult to Identify

In practice, the sample autocorrelation and partial autocorrelation functions are random variables and will not give the same picture as the theoretical functions. This makes the model identification more difficult. In particular, mixed models can be particularly difficult to identify.

Although experience is helpful, developing good models using these sample plots can involve much trial and error. For this reason, in recent years information-based criteria such as FPE (Final Prediction Error) and AIC (Aikake Information Criterion) and others have been preferred and used. These techniques can help automate the model identification process. These techniques require computer software to use. Fortunately, these techniques are available in many commerical statistical software programs that provide ARIMA modeling capabilities.

For additional information on these techniques, see Brockwell and Davis (1987, 2002).

Examples

We show a typical series of plots for performing the initial model identification for

1. the southern oscillations data and
2. the CO_2 monthly concentrations data.

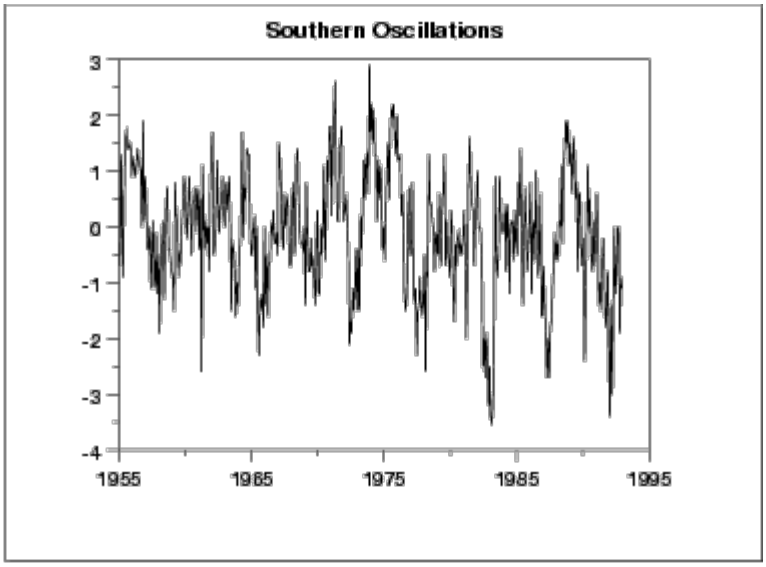
6.4.4.6.1. Model Identification for Southern Oscillations Data

Example for Southern Oscillations

We show typical series of plots for the initial model identification stages of Box-Jenkins modeling for two different examples.

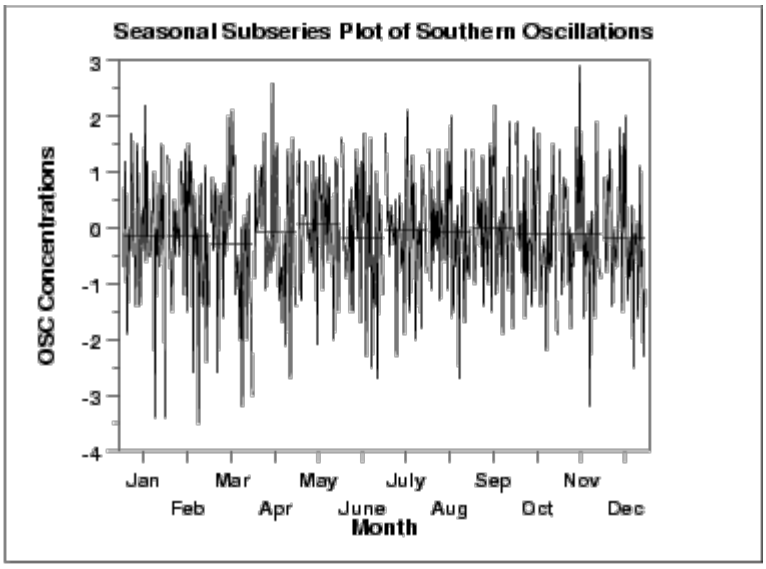
The first example is for the southern oscillations data set. We start with the run sequence plot and seasonal subseries plot to determine if we need to address stationarity and seasonality.

Run Sequence Plot



The run sequence plot indicates stationarity.

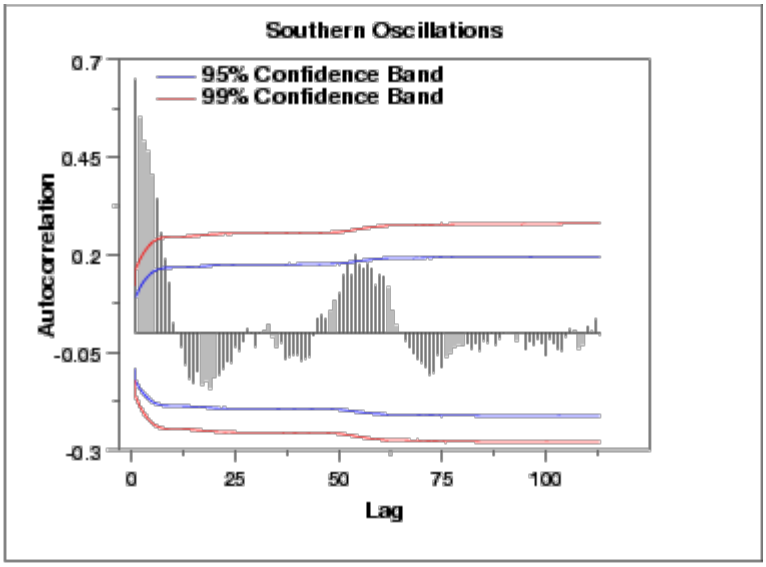
Seasonal Subseries Plot



The seasonal subseries plot indicates that there is no significant seasonality.

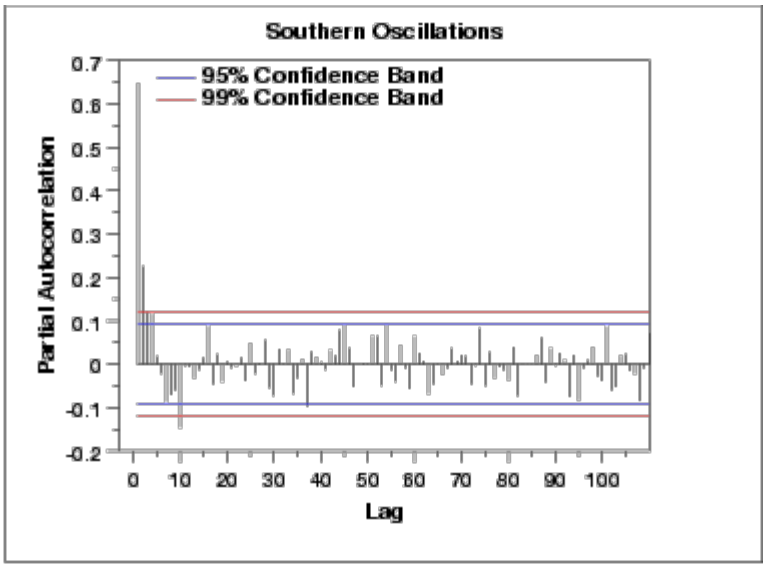
Since the above plots show that this series does not exhibit any significant non-stationarity or seasonality, we generate the autocorrelation and partial autocorrelation plots of the raw data.

Autocorrelation Plot



The autocorrelation plot shows a mixture of exponentially decaying and damped sinusoidal components. This indicates that an autoregressive model, with order greater than one, may be appropriate for these data. The partial autocorrelation plot should be examined to determine the order.

Partial Autocorrelation Plot



The partial autocorrelation plot suggests that an AR(2) model might be appropriate.

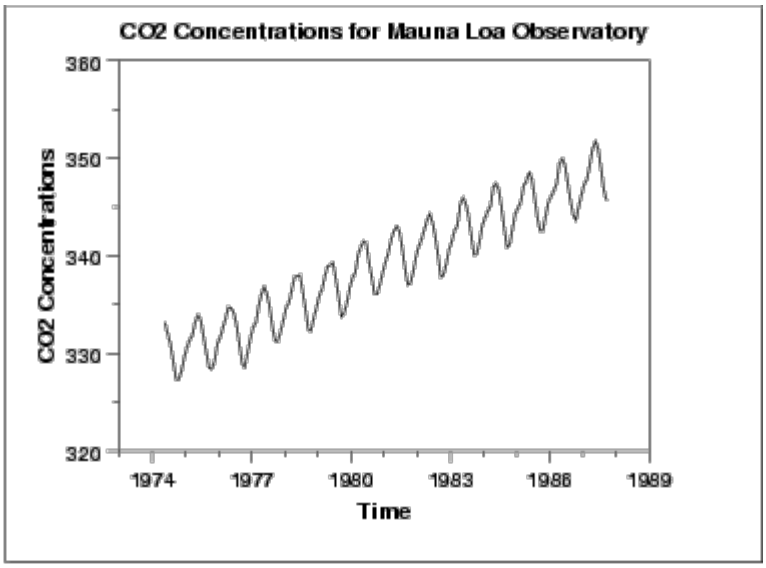
In summary, our intial attempt would be to fit an AR(2) model with no seasonal terms and no differencing or trend removal. Model validation should be performed before accepting this as a final model.

6.4.4.6.2. Model Identification for the CO_2 Concentrations Data

Example for Monthly CO_2 Concentrations

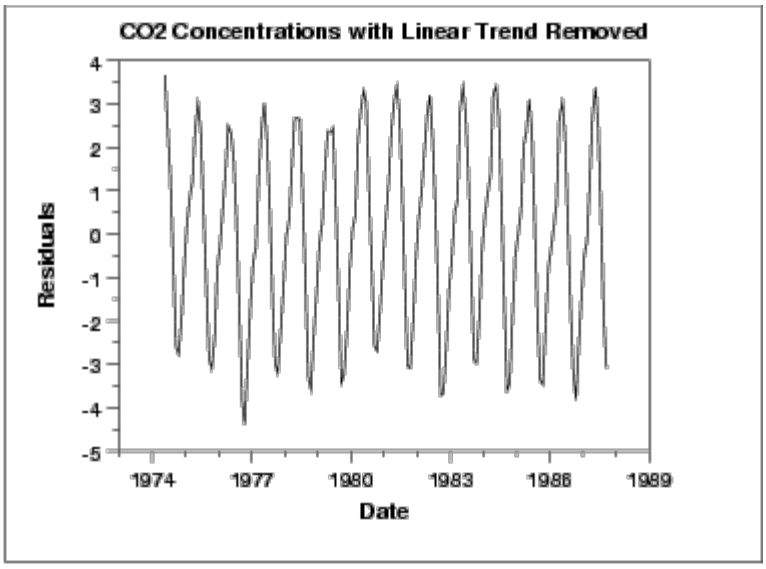
The second example is for the monthly CO_2 concentrations data set. As before, we start with the run sequence plot to check for stationarity.

Run Sequence Plot



The initial run sequence plot of the data indicates a rising trend. A visual inspection of this plot indicates that a simple linear fit should be sufficient to remove this upward trend.

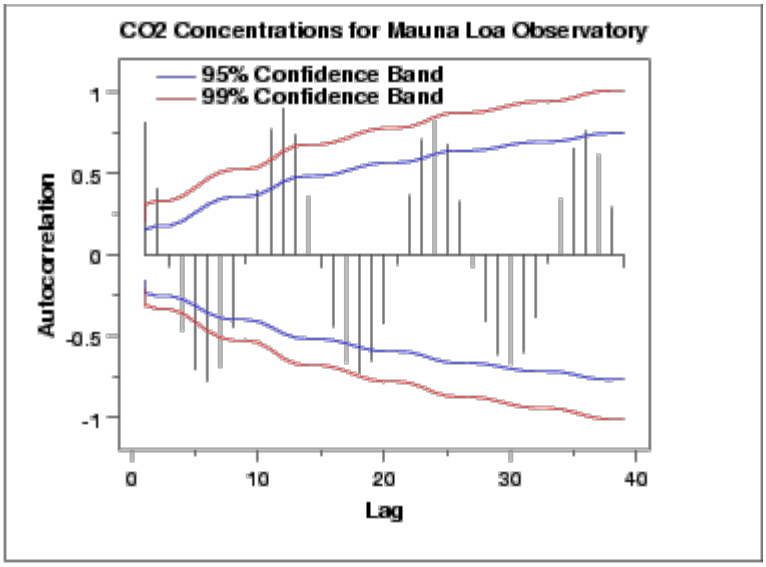
Linear Trend Removed



This plot contains the residuals from a linear fit to the original data. After removing the linear trend, the run sequence plot indicates that the data have a constant location and variance, which implies stationarity.

However, the plot does show seasonality. We generate an autocorrelation plot to help determine the period followed by a seasonal subseries plot.

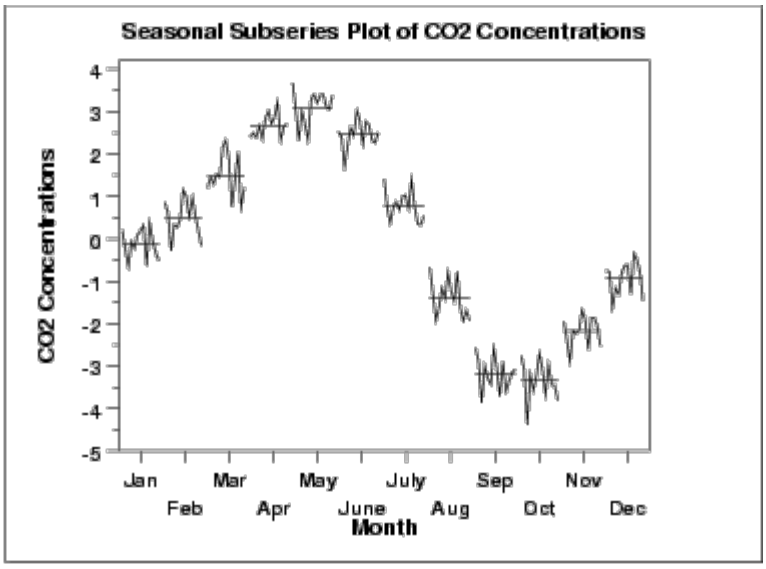
Autocorrelation Plot



The autocorrelation plot shows an alternating pattern of positive and negative spikes. It also shows a repeating pattern every 12 lags, which indicates a seasonality effect.

The two connected lines on the autocorrelation plot are 95% and 99% confidence intervals for statistical significance of the autocorrelations.

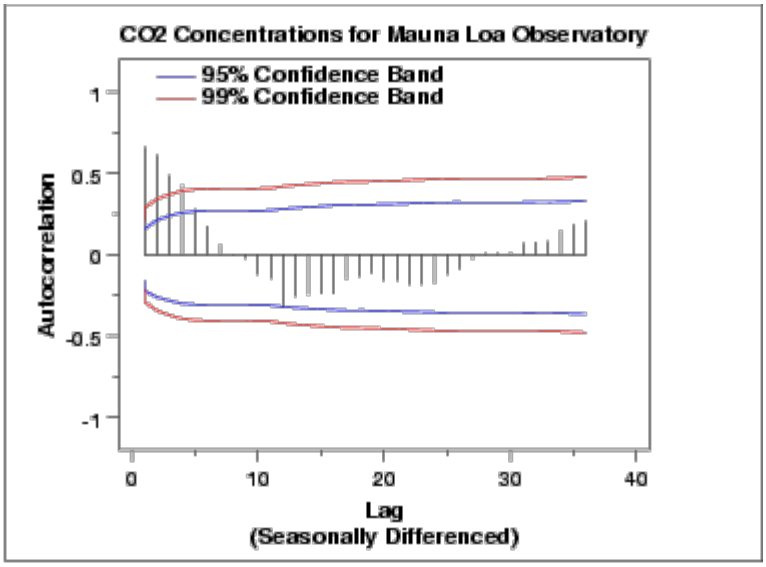
Seasonal Subseries Plot



A significant seasonal pattern is obvious in this plot, so we need to include seasonal terms in fitting a Box-Jenkins model. Since this is monthly data, we would typically include either a lag 12 seasonal autoregressive and/or moving average term.

To help identify the non-seasonal components, we will take a seasonal difference of 12 and generate the autocorrelation plot on the seasonally differenced data.

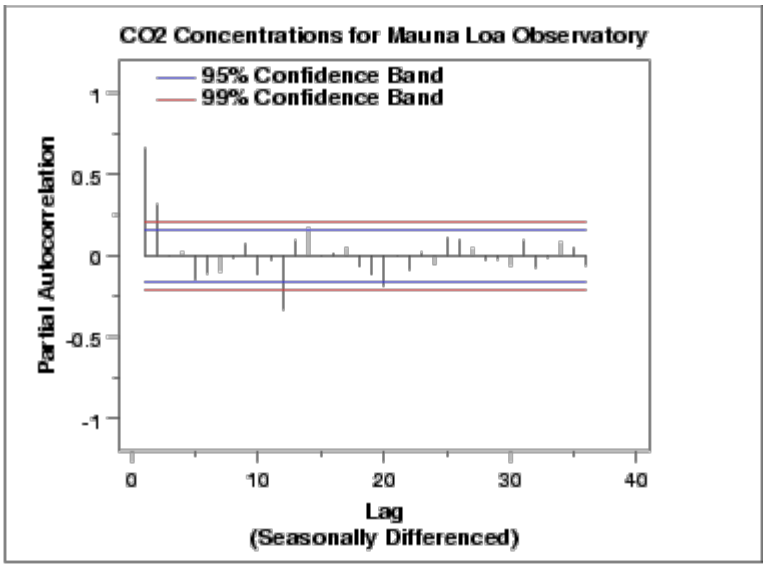
Autocorrelation Plot for Seasonally Differenced Data



This autocorrelation plot shows a mixture of exponential decay and a damped sinusoidal pattern. This indicates that an AR model, with order greater than one, may be appropriate. We generate a partial autocorrelation plot to help identify the order.

Partial Autocorrelation Plot of Seasonally Differenced

Data



The partial autocorrelation plot suggests that an AR(2) model might be appropriate since the partial autocorrelation becomes zero after the second lag. The lag 12 is also significant, indicating some remaining seasonality.

In summary, our intial attempt would be to fit an AR(2) model with a seasonal AR(12) term on the data with a linear trend line removed. We could try the model both with and without seasonal differencing applied. Model validation should be performed before accepting this as a final model.

6.4.4.6.3. Partial Autocorrelation Plot

Purpose: Model Identification for Box-Jenkins Models

Partial autocorrelation plots (Box and Jenkins, pp. 64-65, 1970) are a commonly used tool for model identification in Box-Jenkins models.

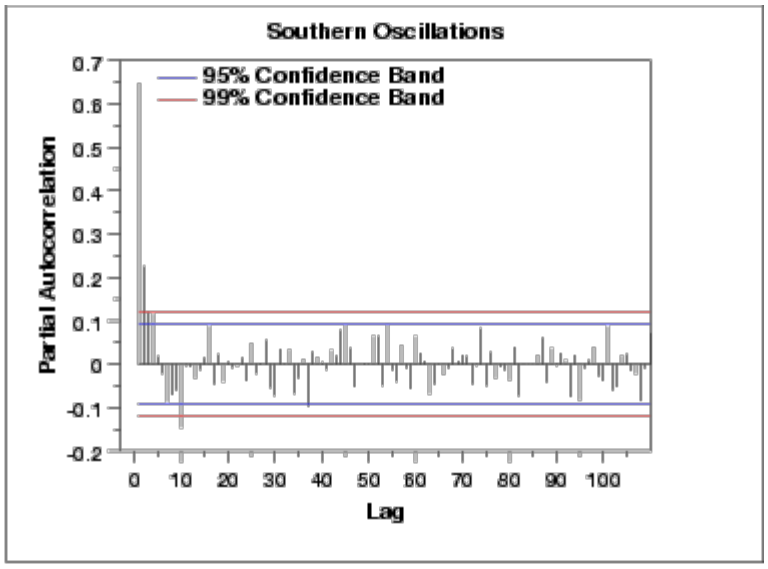
The partial autocorrelation at lag ***k*** is the autocorrelation between X_t and X_{t-k} that is not accounted for by lags 1 through ***k*-1**.

There are algorithms, not discussed here, for computing the partial autocorrelation based on the sample autocorrelations. See (Box, Jenkins, and Reinsel 1970) or (Brockwell, 1991) for the mathematical details.

Specifically, partial autocorrelations are useful in identifying the order of an autoregressive model. The partial autocorrelation of an AR(***p***) process is zero at lag ***p*+1** and greater. If the sample autocorrelation plot indicates that an AR model may be appropriate, then the sample partial autocorrelation plot is examined to help identify the order. We look for the point on the plot where the partial autocorrelations essentially become zero. Placing a 95% confidence interval for statistical significance is helpful for this purpose.

The approximate 95% confidence interval for the partial autocorrelations are at $\pm 2/\sqrt{N}$.

Sample Plot



This partial autocorrelation plot shows clear statistical significance for lags 1 and 2 (lag 0 is always 1). The next few lags are at the borderline of statistical significance. If the autocorrelation plot indicates that an AR model is appropriate, we could start our modeling with an AR(2) model. We might compare this with an AR(3) model.

Definition

Partial autocorrelation plots are formed by

Vertical axis:	Partial autocorrelation coefficient at lag *h*.
Horizontal axis:	Time lag *h* (*h* = 0, 1, 2, 3, ...).

In addition, 95% confidence interval bands are typically included on the plot.

Questions

The partial autocorrelation plot can help provide answers to the following questions:

1. Is an AR model appropriate for the data?
2. If an AR model is appropriate, what order should we use?

Related Techniques

Autocorrelation Plot
Run Sequence Plot
Spectral Plot

Case Study

The partial autocorrelation plot is demonstrated in the Negiz data case study.

Software

Partial autocorrelation plots are available in many general purpose statistical software programs.

6.4.4.7. Box-Jenkins Model Estimation

Use Software	Estimating the parameters for the Box-Jenkins models is a quite complicated non-linear estimation problem. For this reason, the parameter estimation should be left to a high quality software program that fits Box-Jenkins models. Fortunately, many commerical statistical software programs now fit Box-Jenkins models.
Approaches	The main approaches to fitting Box-Jenkins models are non-linear least squares and maximum likelihood estimation. Maximum likelihood estimation is generally the preferred technique. The likelihood equations for the full Box-Jenkins model are complicated and are not included here. See (Brockwell and Davis, 1991) for the mathematical details.
Model Estimation Example	The Negiz case study shows an example of the Box-Jenkins model-fitting.

6.4.4.8. Box-Jenkins Model Diagnostics

Assumptions for a Stable Univariate Process

Model diagnostics for Box-Jenkins models is similar to model validation for non-linear least squares fitting.

That is, the error term A_t is assumed to follow the assumptions for a stationary univariate process. The residuals should be white noise (or independent when their distributions are normal) drawings from a fixed distribution with a constant mean and variance. If the Box-Jenkins model is a good model for the data, the residuals should satisfy these assumptions.

If these assumptions are not satisfied, we need to fit a more appropriate model. That is, we go back to the model identification step and try to develop a better model. Hopefully the analysis of the residuals can provide some clues as to a more appropriate model.

4-Plot of Residuals

As discussed in the EDA chapter, one way to assess if the residuals from the Box-Jenkins model follow the assumptions is to generate a 4-plot of the residuals and an autocorrelation plot of the residuals. One could also look at the value of the Box-Ljung (1978) statistic.

An example of analyzing the residuals from a Box-Jenkins model is given in the Negiz data case study.

Topics

Biometrics

Computer Security

Math

Networking

Statistics

Software

Usability

Divisions

Applied and Computational Mathematics

Advanced Network Technologies

Computer Security

Information Access

Software and Systems

Statistical Engineering

Staff Directory

NIST Digital Library of Mathematical Functions Receives IT Award 1 2 3 4

News

New Quantum Computing Algorithm Could Simulate Giant Particle Accelerators

NIST Releases Final Version of Revised Bluetooth Security Guide

NIST Will Hold 'Big Data' Workshop June 13-14

>> see all news ...

Conferences and Events

ICT Supply Chain Risk Management (SCRM) Workshop - POSTPONED
July 11-12, 2012

Revised Draft FIPS 201–2 Workshop
July 25, 2012

Popular Links

Computer Security Resource Center

Biometrics Resource Center

ITL Brochure - January 2012

NIST Voting Activities

Federal Processing Information Standards (FIPS)

ANSI Accredited Standards Development

Current ITL Newsletter

ITL Bulletins

ITL SURF Program

Comments received in Response to Internet Policy Task Force Cybersecurity Green Paper

NIST Special Publication 500 Series

Comments received in Response to Models To Advance Voluntary Corporate Notification to Consumers Regarding the Illicit Use of Computer Equipment by Botnets and Related Malware Federal Register Notice

Programs/Projects

Complex Systems

Pervasive Information Technology

Virtual Measurements

Health Information Technology

Voting

Quantum

Cloud Computing

Contact

Charles H. Romine, Director
James St. Pierre, Deputy Director

General Information:
301-975-2900 Telephone
301-975-2378 Facsimile

100 Bureau Drive, M/S 8900
Gaithersburg, MD 20899-8900

Sign Up for ITL Newsletters & Bulletins:
Enter email address Go

The National Institute of Standards and Technology (NIST) is an agency of the U.S. Department of Commerce.

Privacy Policy / Security Notice / Accessibility Statement / Disclaimer / Freedom of Information Act (FOIA) / Environmental Policy Statement / No Fear Act Policy / NIST Information Quality Standards / Scientific Integrity Summary

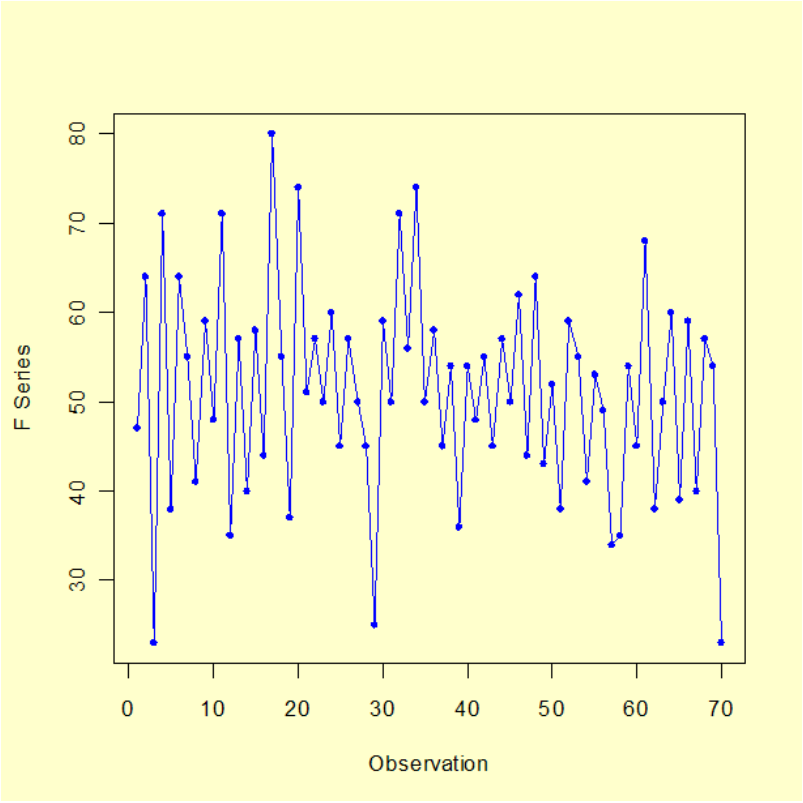
Sign Up for NIST E-mail alerts:
Enter email address Go

Date created: August 24, 2010 | Last updated: May 23, 2012 Contact: **Webmaster**

6.4.4.9. Example of Univariate Box-Jenkins Analysis

Series F

We analyze the series F data set in Box, Jenkins, and Reinsel, 1994. A plot of the 70 raw data points is shown below.



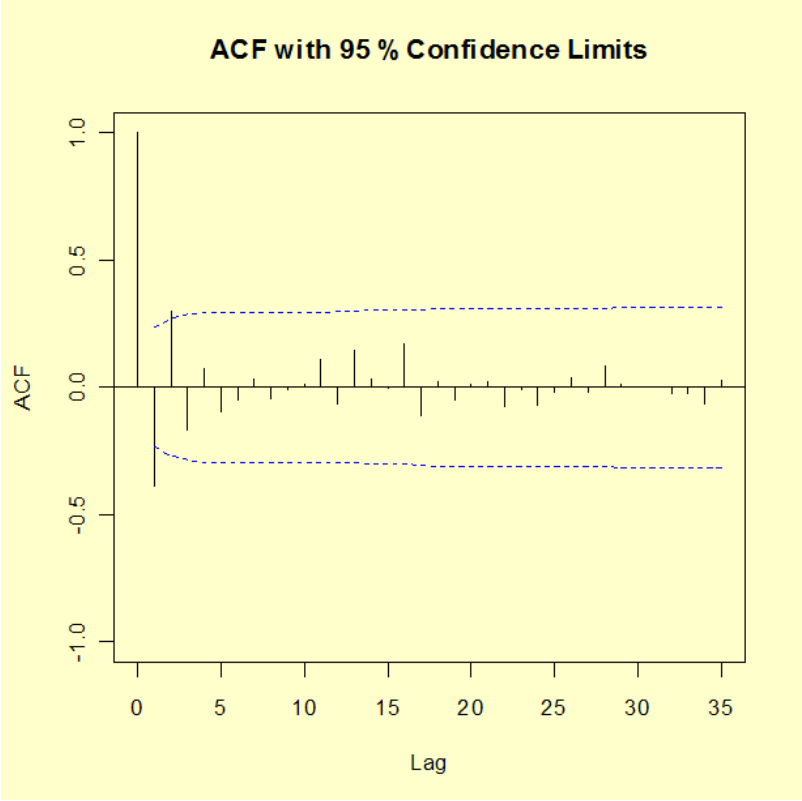
The data do not appear to have a seasonal component or a noticeable trend. (The stationarity of the series was verified by fitting a straight line to the data versus time period. The slope was not found to be significantly different from zero (p-value = 0.2).)

Model Identification

We compute the autocorrelation function (ACF) of the data for the first 35 lags to determine the type of model to fit to the data. We list the numeric results and plot the ACF (along with 95 % confidence limits) versus the lag number.

Lag	ACF
0	1.000000000
1	-0.389878319
2	0.304394082
3	-0.165554717
4	0.070719321
5	-0.097039288
6	-0.047057692

```
 7  0.035373112
 8 -0.043458199
 9 -0.004796162
10  0.014393137
11  0.109917200
12 -0.068778492
13  0.148034489
14  0.035768581
15 -0.006677806
16  0.173004275
17 -0.111342583
18  0.019970791
19 -0.047349722
20  0.016136806
21  0.022279561
22 -0.078710582
23 -0.009577413
24 -0.073114034
25 -0.019503289
26  0.041465024
27 -0.022134370
28  0.088887299
29  0.016247148
30  0.003946351
31  0.004584069
32 -0.024782198
33 -0.025905040
34 -0.062879966
35  0.026101117
```



The ACF values alternate in sign and decay quickly after lag 2, indicating that an AR(2) model is appropriate for the data.

Model Fitting

We fit an AR(2) model to the data.

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + A_t$$

The model fitting results are shown below.

Source	Estimate	Standard Error
φ_1	-0.3198	0.1202

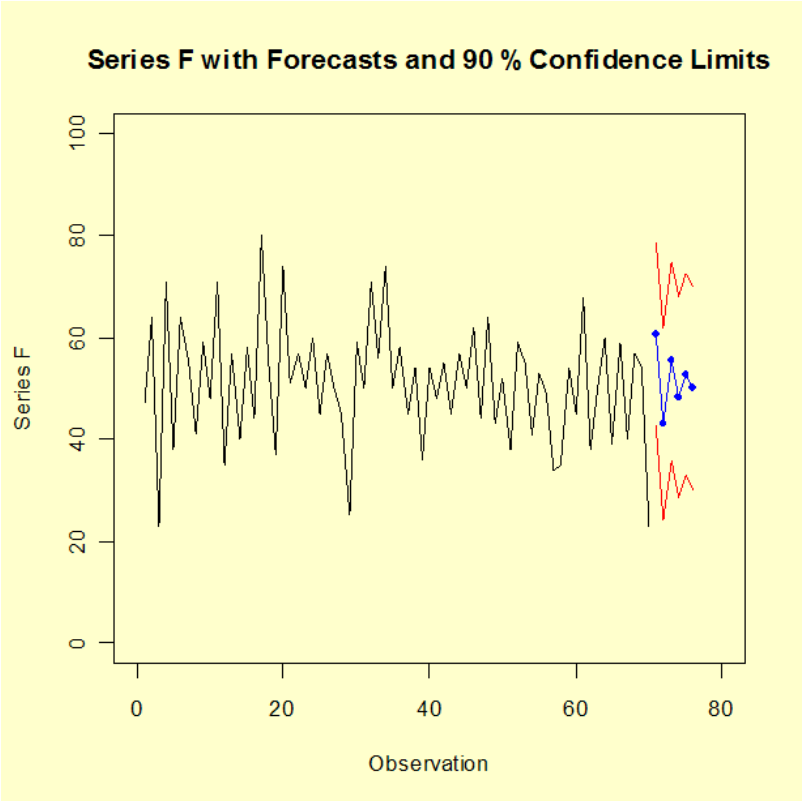
```
φ2          0.1797      0.1202

δ = 51.1286
Residual standard deviation = 10.9599

Test randomness of residuals:
Standardized Runs Statistic Z = 0.4887, p-value =
0.625
```

Forecasting

Using our AR(2) model, we forcast values six time periods into the future.

Period	Prediction	Standard Error
71	60.6405	10.9479
72	43.0317	11.4941
73	55.4274	11.9015
74	48.2987	12.0108
75	52.8061	12.0585
76	50.0835	12.0751

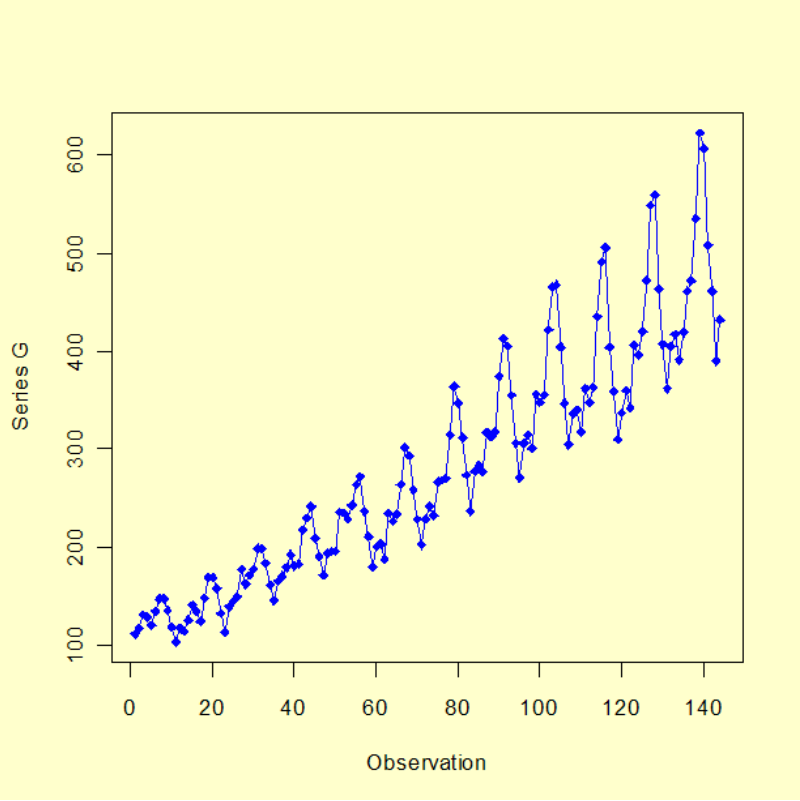
The "historical" data and forecasted values (with 90 % confidence limits) are shown in the graph below.



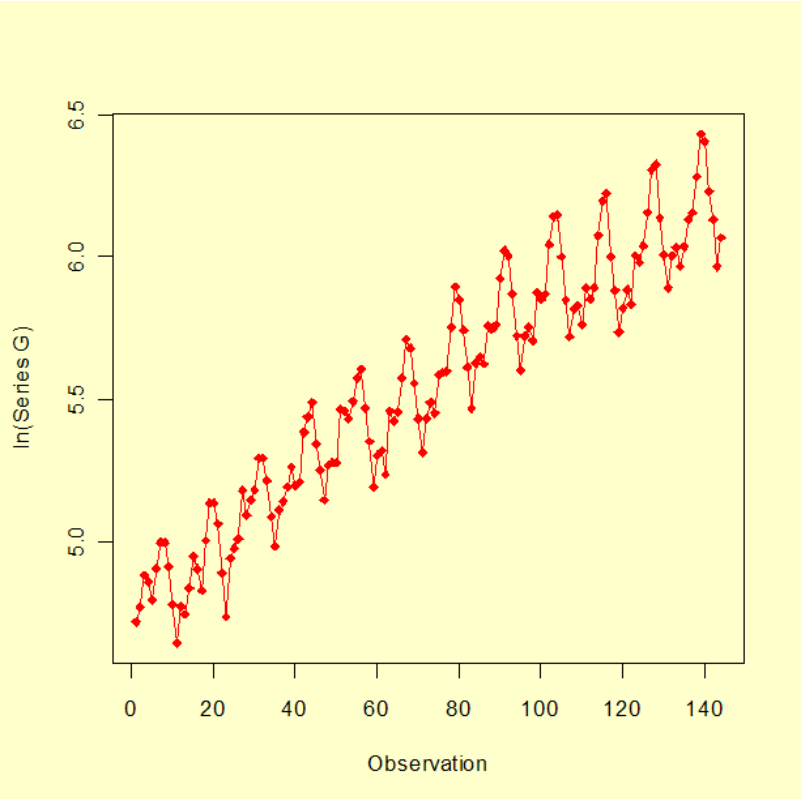
6.4.4.10. Box-Jenkins Analysis on Seasonal Data

Series G

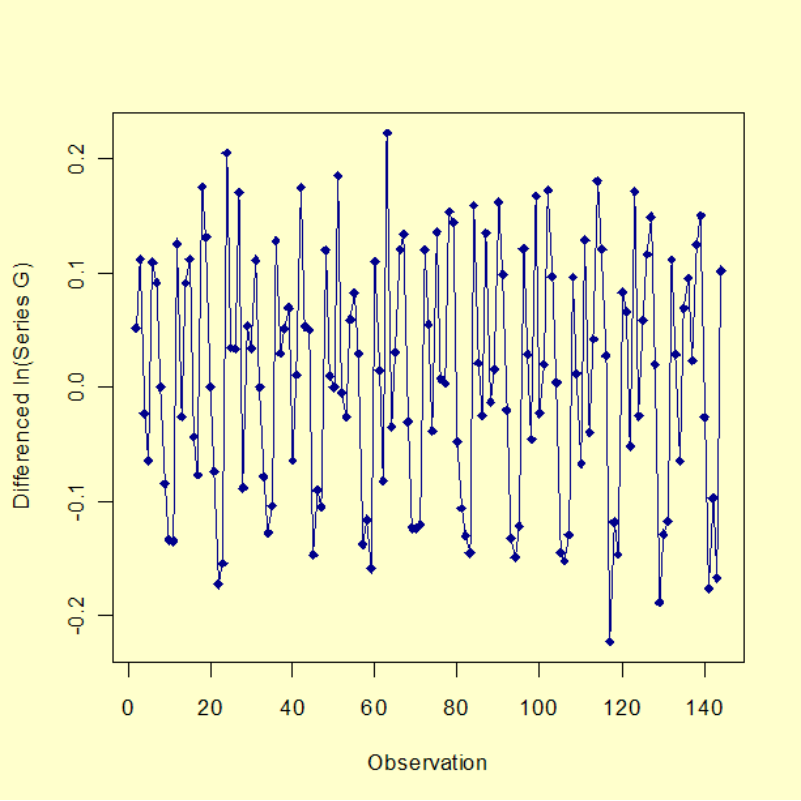
This example illustrates a Box-Jenkins time series analysis for seasonal data using the series G data set in Box, Jenkins, and Reinsel, 1994. A plot of the 144 observations is shown below.



Non-constant variance can be removed by performing a natural log transformation.

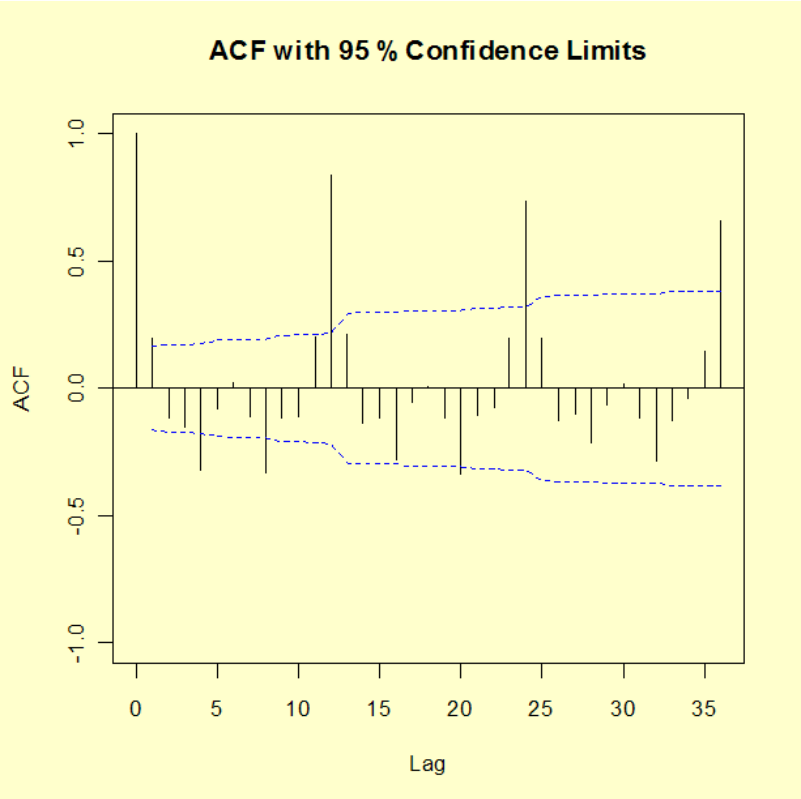


Next, we remove trend in the series by taking first differences. The resulting series is shown below.



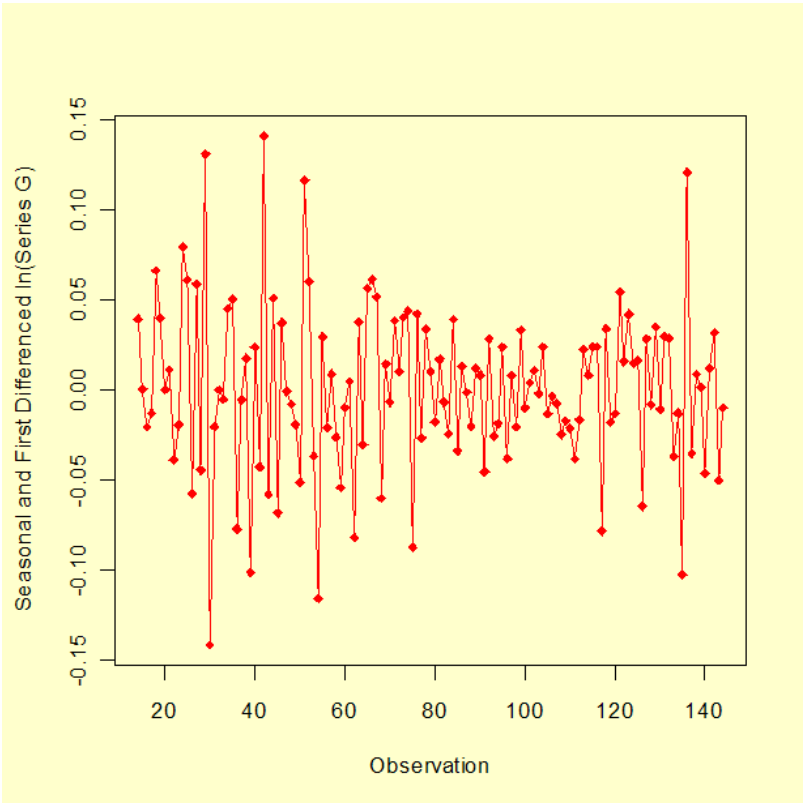
Analyzing Autocorrelation Plot for Seasonality

To identify an appropriate model, we plot the ACF of the time series.



If very large autocorrelations are observed at lags spaced *n* periods apart, for example at lags 12 and 24, then there is evidence of periodicity. That effect should be removed, since the objective of the identification stage is to reduce the autocorrelation throughout. So if simple differencing is not enough, try seasonal differencing at a selected period, such as 4, 6, or 12. In our example, the seasonal period is 12.

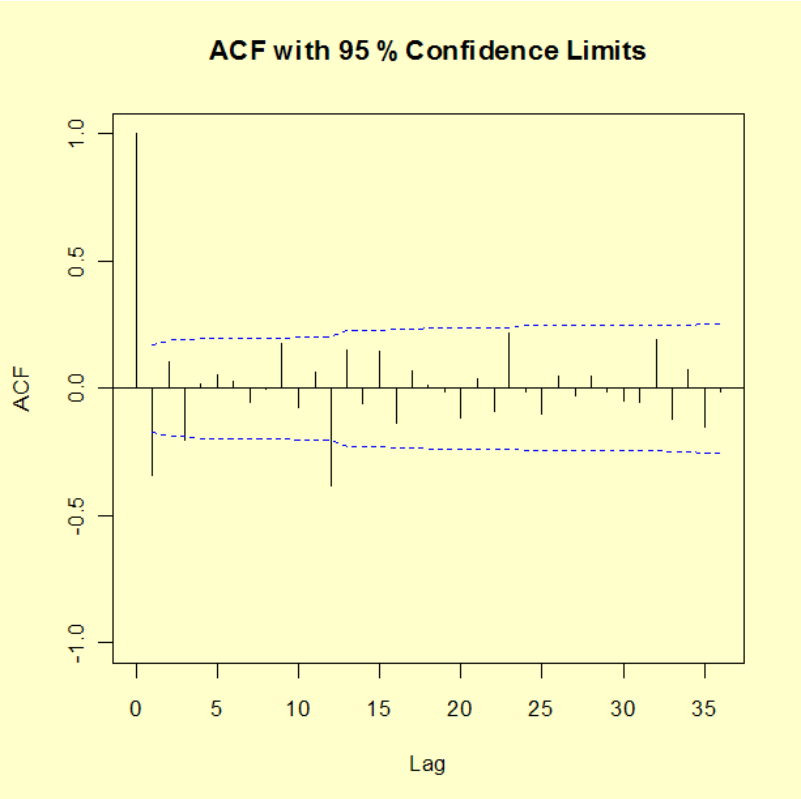
A plot of Series G after taking the natural log, first differencing, and seasonal differencing is shown below.



The number of seasonal terms is rarely more than one. If you know the shape of your forecast function, or you wish to assign a particular shape to the forecast function, you can select the appropriate number of terms for seasonal AR or seasonal MA models.

The book by Box and Jenkins, *Time Series Analysis Forecasting and Control* (the later edition is Box, Jenkins and Reinsel, 1994) has a discussion on these forecast functions on pages 326 - 328. Again, if you have only a faint notion, but you do know that there was a trend upwards before differencing, pick a seasonal MA term and see what comes out in the diagnostics.

An ACF plot of the seasonal and first differenced natural log of series G is shown below.



The plot has a few spikes, but most autocorrelations are near zero, indicating that a seasonal MA(1) model is appropriate.

Model Fitting

We fit an MA(1) model to the data.

$$X_t = \mu + A_t - \theta_1 A_{t-1}$$

The model fitting results are shown below.

Estimate	MA(1)	Seasonal MA(1)
Parameter	-0.4018	-0.5569
Standard Error	0.0896	0.0731

Residual standard deviation = 0.0367
Log likelihood = 244.7
AIC = -483.4

Test the randomness of the residuals up to 30 lags using the Box-Ljung test. Recall that the degrees of freedom for the critical region must be adjusted to account for two estimated parameters.

H_0: The residuals are random.
H_a: The residuals are not random.

Test statistic: $Q = 29.4935$
Significance level: $\alpha = 0.05$
Degrees of freedom: $h = 30 - 2 = 28$
Critical value: $X^2_{1-\alpha,h} = 41.3371$
Critical region: Reject H_0 if $Q > 41.3371$

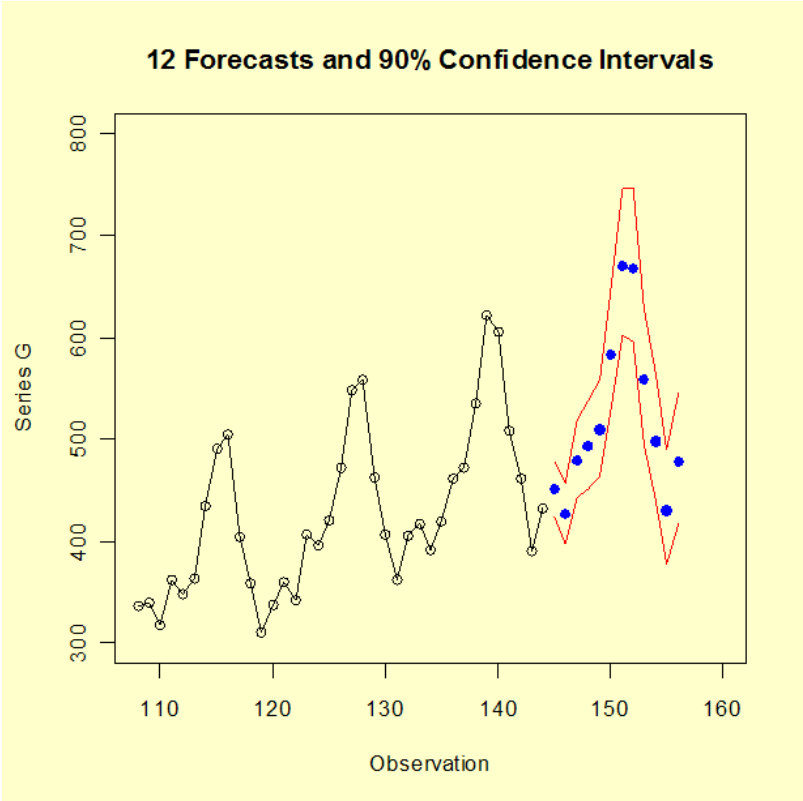
Since the null hypothesis of the Box-Ljung test is not rejected we conclude that the fitted model is adequate.

Forecasting

Using our seasonal MA(1) model, we forcast values 12

periods into the future and compute 90 % confidence limits.

Period	Lower Limit	Forecast	Upper Limit
145	424.0234	450.7261	478.4649
146	396.7861	426.0042	456.7577
147	442.5731	479.3298	518.4399
148	451.3902	492.7365	537.1454
149	463.3034	509.3982	559.3245
150	527.3754	583.7383	645.2544
151	601.9371	670.4625	745.7830
152	595.7602	667.5274	746.9323
153	495.7137	558.5657	628.5389
154	439.1900	497.5430	562.8899
155	377.7598	430.1618	489.1730
156	417.3149	477.5643	545.7760



6.4.5. Multivariate Time Series Models

If each time series observation is a vector of numbers, you can model them using a multivariate form of the Box-Jenkins model

The multivariate form of the Box-Jenkins univariate models is sometimes called the ARMAV model, for AutoRegressive Moving Average Vector or simply vector ARMA process.

The ARMAV model for a stationary multivariate time series, with a zero mean vector, represented by

$$x_t = (x_{1t}, x_{2t}, ..., x_{nt})^T \qquad -\infty < t < \infty$$

is of the form

$$\mathbf{x}_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

where

- $\boldsymbol{x_t}$ and $\boldsymbol{a_t}$ are n x 1 column vectors with $\boldsymbol{a_t}$ representing multivariate white noise
- $\phi_k = \{\phi_{k,jj}\}, \qquad k = 1, 2, \ldots, p$

  $\theta_k = \{\theta_{k,jj}\}, \qquad k = 1, 2, \ldots, q$

  are n x n matrices for autoregressive and moving average parameters
- E$[\boldsymbol{a_t}]$ = **0**
- $E(a_t a'_{t-k}) = 0 \qquad k \neq 0$

  $E(a_t a'_{t-k}) = \Sigma_a \qquad k = 0$

  where Σ_a is the dispersion or covariance matrix of a_t

As an example, for a bivariate series with n = 2, p = 2, and q = 1, the ARMAV(2,1) model is:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} \phi_{1.11} & \phi_{1.12} \\ \phi_{1.21} & \phi_{1.22} \end{pmatrix} \begin{pmatrix} x_{1t-1} \\ x_{2t-1} \end{pmatrix} + \begin{pmatrix} \phi_{2.11} & \phi_{2.12} \\ \phi_{2.21} & \phi_{2.22} \end{pmatrix} \begin{pmatrix} x_{1t-2} \\ x_{2t-2} \end{pmatrix} + \begin{pmatrix} a_{1t} \\ a_{2t} \end{pmatrix} - \begin{pmatrix} \phi_{1.11} & \phi_{1.12} \\ \phi_{1.21} & \phi_{1.22} \end{pmatrix} \begin{pmatrix} a_{1t-1} \\ a_{2t-1} \end{pmatrix}$$

with

$$a_t = \begin{pmatrix} a_{1t} \\ a_{2t} \end{pmatrix}$$

Estimation of parameters and covariance matrix difficult

The estimation of the matrix parameters and covariance matrix is complicated and very difficult without computer software. The estimation of the Moving Average matrices is especially an ordeal. If we opt to ignore the MA component(s) we are left with the ARV model given by:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + a_t$$

where

- $\boldsymbol{x_t}$ is a vector of observations, $\boldsymbol{x_{1t}}$, $\boldsymbol{x_{2t}}$, ... , $\boldsymbol{x_{nt}}$ at time *t*
- $\boldsymbol{a_t}$ is a vector of white noise, $\boldsymbol{a_{1t}}$, $\boldsymbol{a_{2t}}$, ... , $\boldsymbol{a_{nt}}$ at time *t*
- $\phi_k = \{\phi_{k.jj}\}, \quad k = 1, 2, \ldots, p$
  is a *n* x *n* matrix of autoregressive parameters
- E$[\boldsymbol{a_t}]$ = **0**
- $E(a_t a'_{t-k}) = 0 \quad k \neq 0$

  $E(a_t a'_{t-k}) = \Sigma_a \quad k = 0$

  where Σ_a is the dispersion or covariance matrix

A model with *p* autoregressive matrix parameters is an ARV(*p*) model or a vector AR model.

The parameter matrices may be estimated by multivariate least squares, but there are other methods such as maximium likelihood estimation.

Interesting properties of parameter matrices

There are a few interesting properties associated with the phi or AR parameter matrices. Consider the following example for a bivariate series with *n* =2, *p* = 2, and *q* = 0. The ARMAV(2,0) model is:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \phi_{1.11} & \phi_{1.12} \\ \phi_{1.21} & \phi_{1.22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \phi_{2.11} & \phi_{2.12} \\ \phi_{2.21} & \phi_{2.22} \end{pmatrix} \begin{pmatrix} x_{t-2} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} a_{1t} \\ a_{2t} \end{pmatrix}$$

Without loss of generality, assume that the *X* series is input and the *Y* series are output and that the mean vector = (0,0).

Therefore, tranform the observation by subtracting their respective averages.

Diagonal terms of Phi matrix

The diagonal terms of each Phi matrix are the scalar estimates for each series, in this case:

$\phi_{1.11}, \phi_{2.11}$ for the input series *X*,
$\phi_{1.22}, \phi_{.2.22}$ for the output series *Y*.

Transfer mechanism

The lower off-diagonal elements represent the influence of the input on the output.

This is called the "transfer" mechanism or transfer-function model as discussed by Box and Jenkins in Chapter 11. The ϕ terms here correspond to their δ terms.

The upper off-diagonal terms represent the influence of the output on the input.

Feedback

This is called "feedback". The presence of feedback can also be seen as a high value for a coefficient in the correlation matrix of the residuals. A "true" transfer model exists when there is no feedback.

This can be seen by expressing the matrix form into scalar form:

$$x_t = \phi_{1.11}x_{t-1} + \phi_{2.11}x_{t-2} + \phi_{1.12}y_{t-1} + \phi_{2.12}y_{t-2} + a_{1t}$$
$$y_t = \phi_{1.22}y_{t-1} + \phi_{2.22}y_{t-2} + \phi_{1.21}x_{t-1} + \phi_{2.21}x_{t-2} + a_{2t}$$

Delay

Finally, delay or "dead' time can be measured by studying the lower off-diagonal elements again.

If, for example, $\phi_{1.21}$ is non-significant, the delay is 1 time period.

6.4.5.1. Example of Multivariate Time Series Analysis

Bivariate Gas Furance Example

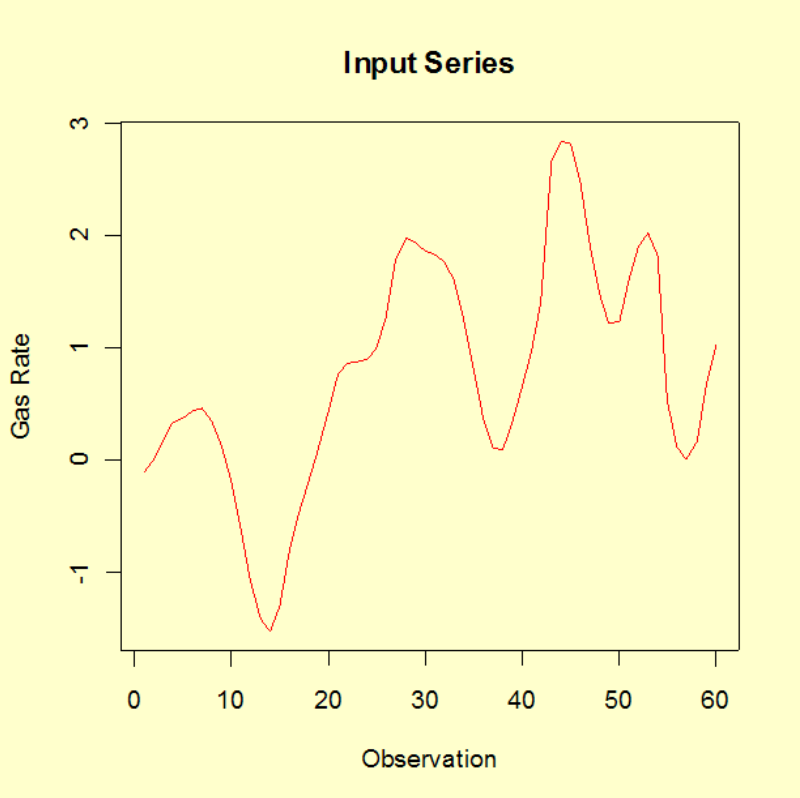
The gas furnace data from Box, Jenkins, and Reinsel, 1994 is used to illustrate the analysis of a bivariate time series. Inside the gas furnace, air and methane were combined in order to obtain a mixture of gases containing CO_2 (carbon dioxide). The input series is the methane gas feedrate described by

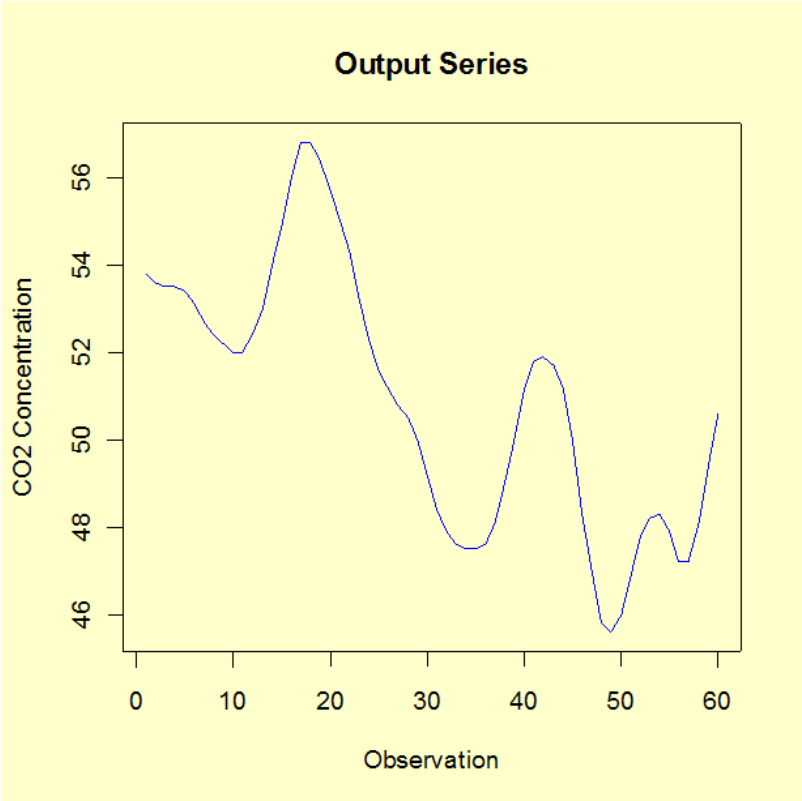
$$\text{Methane Gas Input Feed} = 0.60 - 0.04\,X(t)$$

the CO_2 concentration was the output series, $Y(t)$. In this experiment 296 successive pairs of observations (X_t, Y_t) were collected from continuous records at 9-second intervals. For the analysis described here, only the first 60 pairs were used. We fit an ARV(2) model as described in 6.4.5.

Plots of input and output series

The plots of the input and output series are displayed below.





Model Fitting

The scalar form of the ARV(2) model is the following.

$$x_t = \phi_{1.11}x_{t-1} + \phi_{2.11}x_{t-2} + \phi_{1.12}y_{t-1} + \phi_{2.12}y_{t-2} + a_{1t}$$

$$y_t = \phi_{1.22}y_{t-1} + \phi_{2.22}y_{t-2} + \phi_{1.21}x_{t-1} + \phi_{2.21}x_{t-2} + a_{2t}$$

The equation for x_t corresponds to gas rate while the equation for y_t corresponds to CO_2 concentration.

The parameter estimates for the equation associated with gas rate are the following.

	Estimate	Std. Err.	*t* value	Pr(>|*t*|)
a_{1t}	0.003063	0.035769	0.086	0.932
$\varphi_{1.11}$	1.683225	0.123128	13.671	< 2e-16
$\varphi_{2.11}$	-0.860205	0.165886	-5.186	3.44e-06
$\varphi_{1.12}$	-0.076224	0.096947	-0.786	0.435
$\varphi_{2.12}$	0.044774	0.082285	0.544	0.589

```
Residual standard error:  0.2654 based on 53 degrees of freedom
Multiple R-Squared:  0.9387
Adjusted R-squared:  0.9341
F-statistic:  203.1 based on 4 and 53 degrees of freedom
p-value:  < 2.2e-16
```

The parameter estimates for the equation associated with CO_2 concentration are the following.

	Estimate	Std. Err.	*t* value	Pr(>|*t*|)
a_{2t}	-0.03372	0.01615	-2.088	0.041641
$\varphi_{1.22}$	1.22630	0.04378	28.013	< 2e-16
$\varphi_{2.22}$	-0.40927	0.03716	-11.015	2.57e-15
φ	0.22898	0.05560	4.118	0.000134

```
 1.21
φ2.21     -0.80532    0.07491  -10.751   6.29e-15

Residual standard error:  0.1198 based on 53 degrees of freedom
Multiple R-Squared:  0.9985
Adjusted R-squared:  0.9984
F-statistic:  8978 based on 4 and 53 degrees of freedom
p-value:  < 2.2e-16
```

Box-Ljung tests performed for each series to test the randomness of the first 24 residuals were not significant. The *p*-values for the tests using CO_2 concentration residuals and gas rate residuals were 0.4 and 0.6, respectively.

Forecasting

The forecasting method is an extension of the model and follows the theory outlined in the previous section. The forecasted values of the next six observations (61-66) and the associated 90 % confidence limits are shown below for each series.

Observation	90% Lower Limit	Concentration Forecast	90% Upper Limit
61	51.0	51.2	51.4
62	51.0	51.3	51.6
63	50.6	51.0	51.4
64	49.8	50.5	51.1
65	48.7	50.0	51.3
66	47.6	49.7	51.8

Observation	90% Lower Limit	Rate Forecast	90% Upper Limit
61	0.795	1.231	1.668
62	0.439	1.295	2.150
63	0.032	1.242	2.452
64	-0.332	1.128	2.588
65	-0.605	1.005	2.614
66	-0.776	0.908	2.593

6.5. Tutorials

Tutorial contents

1. What do we mean by "Normal" data?
2. What do we do when data are "Non-normal"?
3. Elements of Matrix Algebra
    1. Numerical Examples
    2. Determinant and Eigenstructure
4. Elements of Multivariate Analysis
    1. Mean vector and Covariance Matrix
    2. The Multivariate Normal Distribution
    3. Hotelling's T^2
        1. Example of Hotelling's T^2 Test
        2. Example 1 (continued)
        3. Example 2 (multiple groups)
    4. Hotelling's T^2 Chart
5. Principal Components
    1. Properties of Principal Components
    2. Numerical Example

6.5.1. What do we mean by "Normal" data?

The Normal distribution model

"Normal" data are data that are drawn (come from) a population that has a normal distribution. This distribution is inarguably the most important and the most frequently used distribution in both the theory and application of statistics. If X is a normal random variable, then the probability distribution of X is

Normal probability distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

Parameters of normal distribution

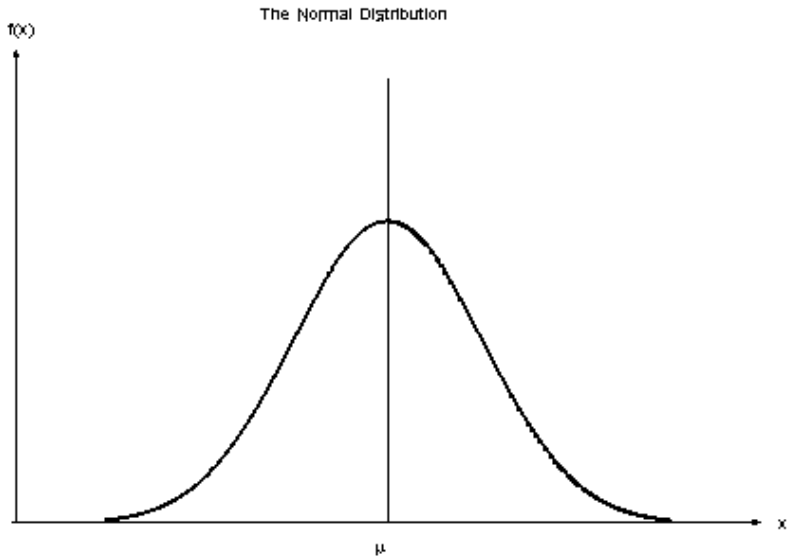
The parameters of the normal distribution are the mean μ and the standard deviation σ (or the variance σ^2). A special notation is employed to indicate that X is normally distributed with these parameters, namely

$$X \sim N(\mu, \sigma) \text{ or } X \sim N(\mu, \sigma^2).$$

Shape is symmetric and unimodal

The shape of the normal distribution is symmetric and unimodal. It is called the *bell-shaped or Gaussian* distribution after its inventor, Gauss (although De Moivre also deserves credit).

The visual appearance is given below.



Property of probability distributions is that area under curve equals one

A property of a special class of non-negative functions, called probability distributions, is that the area under the curve equals unity. One finds the area under any portion of the curve by integrating the distribution between the specified limits. The area under the bell-shaped curve of the normal distribution can be shown to be equal to 1, and therefore the normal distribution is a probability distribution.

Interpretation of σ

There is a simple interpretation of σ

68.27% of the population fall between μ +/- 1 σ
95.45% of the population fall between μ +/- 2 σ
99.73% of the population fall between μ +/- 3 σ

The cumulative normal distribution

The cumulative normal distribution is defined as the probability that the normal variate is less than or equal to some value *v*, or

$$P\{X \le v\} = F(v) = \int_{-\infty}^{v} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Unfortunately this integral cannot be evaluated in closed form and one has to resort to numerical methods. But even so, tables for all possible values of μ and σ would be required. A change of variables rescues the situation. We let

$$z = \frac{x - \mu}{\sigma}.$$

Now the evaluation can be made independently of μ and σ; that is,

$$P\{X \le v\} = P\left\{z \le \frac{v - \mu}{\sigma}\right\} = \Phi\left(\frac{v - \mu}{\sigma}\right)$$

where Φ(.) is the cumulative distribution function of the *standard normal distribution* (μ = 0, σ = 1).

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Tables for the cumulative standard normal distribution

Tables of the cumulative standard normal distribution are given in every statistics textbook and in the handbook. A rich variety of approximations can be found in the literature on numerical methods.

For example, if μ = 0 and σ = 1 then the area under the curve from μ - 1σ to μ + 1σ is the area from 0 - 1 to 0 + 1, which is 0.6827. Since most standard normal tables give area to the left of the lookup value, they will have for *z* = 1 an area of

.8413 and for z = -1 an area of .1587. By subtraction we obtain the area between -1 and +1 to be .8413 - .1587 = .6826.

6.5.2. What to do when data are non-normal

Often it is possible to transform non-normal data into approximately normal data

Non-normality is a way of life, since no characteristic (height, weight, etc.) will have *exactly* a normal distribution. One strategy to make non-normal data resemble normal data is by using a transformation. There is no dearth of transformations in statistics; the issue is which one to select for the situation at hand. Unfortunately, the choice of the "best" transformation is generally not obvious.

This was recognized in 1964 by G.E.P. Box and D.R. Cox. They wrote a paper in which a useful family of power transformations was suggested. These transformations are defined only for positive data values. This should not pose any problem because a constant can always be added if the set of observations contains one or more negative values.

The Box-Cox power transformations are given by

The Box-Cox Transformation

$$x(\lambda) = \frac{(x^\lambda - 1)}{\lambda} \qquad \lambda \neq 0$$
$$x(\lambda) = \ln(x) \qquad \lambda = 0$$

Given the vector of data observations $\mathbf{x} = x_1, x_2, \ldots x_n$, one way to select the power λ is to use the λ that maximizes the logarithm of the likelihood function

The logarithm of the likelihood function

$$f(x, \lambda) = -\frac{n}{2}\ln\left[\sum_{i=1}^{n}\frac{(x_i(\lambda) - \bar{x}(\lambda))^2}{n}\right] + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i)$$

where

$$\bar{x}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$$

is the arithmetic mean of the transformed data.

Confidence bound for λ

In addition, a confidence bound (based on the likelihood ratio statistic) can be constructed for λ as follows: A set of λ values that represent an approximate 100(1- α)% confidence bound for λ is formed from those λ that satisfy

$$f(x,\lambda) \geq f(x,\hat{\lambda}) - 0.5\chi^2_{1-\alpha,1}$$

where $\hat{\lambda}$ denotes the maximum likelihood estimator for λ and $X^2_{1-\alpha,1}$ is the 100(1-α) percentile of the chi-square distribution with 1 degree of freedom.

Example of the Box-Cox scheme

To illustrate the procedure, we used the data from Johnson and Wichern's textbook (Prentice Hall 1988), Example 4.14. The observations are microwave radiation measurements.

Sample data

.15 .09 .18 .10 .05 .12 .08
.05 .08 .10 .07 .02 .01 .10
.10 .10 .02 .10 .01 .40 .10
.05 .03 .05 .15 .10 .15 .09
.08 .18 .10 .20 .11 .30 .02
.20 .20 .30 .30 .40 .30 .05

Table of log-likelihood values for various values of λ

The values of the log-likelihood function obtained by varying λ from -2.0 to 2.0 are given below.

λ	*LLF*	λ	*LLF*	λ	*LLF*
-2.0	7.1146	-0.6	89.0587	0.7	103.0322
-1.9	14.1877	-0.5	92.7855	0.8	101.3254
-1.8	21.1356	-0.4	96.0974	0.9	99.3403
-1.7	27.9468	-0.3	98.9722	1.0	97.1030
-1.6	34.6082	-0.2	101.3923	1.1	94.6372
-1.5	41.1054	-0.1	103.3457	1.2	91.9643
-1.4	47.4229	0.0	104.8276	1.3	89.1034
-1.3	53.5432	0.1	105.8406	1.4	86.0714
1.2	59.4474	0.2	106.3947	1.5	82.8832
-1.1	65.1147	0.3	106.5069	1.6	79.5521
-0.9	75.6471	0.4	106.1994	1.7	76.0896
-0.8	80.4625	0.5	105.4985	1.8	72.5061
-0.7	84.9421	0.6	104.4330	1.9	68.8106

This table shows that λ = .3 maximizes the log-likelihood function (LLF). This becomes 0.28 if a second digit of accuracy is calculated.

The Box-Cox transform is also discussed in Chapter 1 under the Box Cox Linearity Plot and the Box Cox Normality Plot. The Box-Cox normality plot discussion provides a graphical method for choosing λ to transform a data set to normality. The criterion used to choose λ for the Box-Cox linearity plot is the value of λ that maximizes the correlation between the transformed x-values and the y-values when making a normal probability plot of the

(transformed) data.

6.5.3. Elements of Matrix Algebra

Elementary Matrix Algebra

Basic definitions and operations of matrix algebra - needed for multivariate analysis

Vectors and matrices are arrays of numbers. The algebra for symbolic operations on them is different from the algebra for operations on scalars, or single numbers. For example there is no division in matrix algebra, although there is an operation called "multiplying by an inverse". It is possible to express the exact equivalent of matrix algebra equations in terms of scalar algebra expressions, but the results look rather messy.

It can be said that the matrix algebra notation is shorthand for the corresponding scalar longhand.

Vectors

A *vector* is a column of numbers

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

The scalars a_i are the elements of vector **a**.

Transpose

The *transpose* of **a,** denoted by **a'**, is the row arrangement of the elements of **a**.

$$\mathbf{a}' = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}$$

Sum of two vectors

The **sum of two vectors** (say, *a* and *b*) is the vector of sums of corresponding elements.

$$\mathbf{a}+\mathbf{b} = \begin{bmatrix} a_1+b_1 \\ a_2+b_2 \\ \vdots \\ a_p+b_p \end{bmatrix}$$

The difference of two vectors is the vector of differences of corresponding elements.

Product of a'b

The **product a'b** is a scalar formed by

$$\mathbf{a'b} = \left[a_1b_1 + \quad a_2b_2 + \quad \cdots \quad + a_pb_p\right]$$

which may be written in shortcut notation as

$$c = \sum_{i=1}^{p} a_ib_i$$

where a_i and b_i are the ith elements of vector **a** and **b**, respectively.

Product of ab'

The **product ab'** is a *square matrix*

$$\mathbf{ab'} = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_p \\ a_2b_1 & a_2b_2 & \cdots & a_2b_p \\ \vdots & \vdots & & \vdots \\ a_pb_1 & a_pb_2 & \cdots & a_pb_p \end{bmatrix}$$

Product of scalar times a vector

The product of a scalar k, times a vector **a** is k times each element of **a**

$$k\mathbf{a} = \mathbf{a}k = \begin{bmatrix} ka_1 \\ ka_2 \\ \vdots \\ ka_p \end{bmatrix}$$

A matrix is a rectangular table of numbers

A matrix is a rectangular table of numbers, with p rows and n columns. It is also referred to as an array of n column vectors of length p. Thus

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{bmatrix}$$

is a p by n matrix. The typical element of **A** is a_{ij}, denoting the element of row i and column j.

Matrix addition and

Matrices are **added and subtracted** on an element-by-element basis. Thus

subtraction

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{p1}+b_{p1} & a_{p2}+b_{p2} & \cdots & a_{pn}+b_{pn} \end{bmatrix}$$

Matrix multiplication

Matrix **multiplication** involves the computation of the sum of the products of elements from a row of the first matrix (the premultiplier on the left) and a column of the second matrix (the postmultiplier on the right). This sum of products is computed for every combination of rows and columns. For example, if **A** is a 2 x 3 matrix and **B** is a 3 x 2 matrix, the product **AB** is

$$\mathbf{AB}=\begin{bmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31} & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32} \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31} & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32} \end{bmatrix}$$

Thus, the product is a 2 x 2 matrix. This came about as follows: The number of columns of **A** must be equal to the number of rows of **B.** In this case this is 3. If they are not equal, multiplication is impossible. If they are equal, then the number of rows of the product **AB** is equal to the number of rows of **A** and the number of columns is equal to the number of columns of **B**.

Example of 3x2 matrix multiplied by a 2x3

It follows that the result of the product **BA** is a 3 x 3 matrix

$$\mathbf{BA}=\begin{bmatrix} b_{11}a_{11}+b_{12}a_{21} & b_{11}a_{12}+b_{12}a_{22} & b_{11}a_{13}+b_{12}a_{23} \\ b_{21}a_{11}+b_{22}a_{21} & b_{21}a_{12}+b_{22}a_{22} & b_{21}a_{13}+b_{22}a_{23} \\ b_{31}a_{11}+b_{32}a_{21} & b_{31}a_{12}+b_{32}a_{22} & b_{31}a_{13}+b_{32}a_{23} \end{bmatrix}$$

General case for matrix multiplication

In general, if **A** is a *k x p* matrix and **B** is a *p* x *n* matrix, the product **AB** is a *k x n* matrix. If *k* = *n*, then the product **BA** can also be formed. We say that matrices conform for the operations of addition, subtraction or multiplication when their respective orders (numbers of row and columns) are such as to permit the operations. Matrices that do not conform for addition or subtraction cannot be added or subtracted. Matrices that do not conform for multiplication cannot be multiplied.

6.5.3.1. Numerical Examples

Numerical examples of matrix operations

Numerical examples of the matrix operations described on the previous page are given here to clarify these operations.

Sample matrices

If

$$\mathbf{A} = \begin{bmatrix} 5 & 6 \\ 3 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 3 & 2 \\ 1 & 5 \end{bmatrix}$$

then

Matrix addition, subtraction, and multipication

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 8 & 8 \\ 4 & 12 \end{bmatrix} \quad \text{and} \quad \mathbf{A} - \mathbf{B} = \begin{bmatrix} 2 & 4 \\ 2 & 2 \end{bmatrix}$$

and

$$\mathbf{AB} = \begin{bmatrix} 21 & 40 \\ 16 & 41 \end{bmatrix} \quad \text{and} \quad \mathbf{BA} = \begin{bmatrix} 21 & 32 \\ 20 & 41 \end{bmatrix}$$

Multiply matrix by a scalar

To multiply a a matrix by a given scalar, each element of the matrix is multiplied by that scalar

$$2\mathbf{A} = \begin{bmatrix} 10 & 12 \\ 6 & 14 \end{bmatrix} \quad \text{and} \quad .5\mathbf{B} = \begin{bmatrix} 1.5 & 1.0 \\ 0.5 & 2.5 \end{bmatrix}$$

Pre-multiplying matrix by transpose of a vector

Pre-multiplying a *p x n* matrix by the transpose of a *p*-element vector yields a *n*-element transpose

$$\mathbf{c}' = \mathbf{a}'\mathbf{B} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}' = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}$$

Post-multiplying matrix by vector

Post-multiplying a *p x n* matrix by an *n*-element vector yields an *n*-element vector

$$\mathbf{c} = \mathbf{BA} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

Quadratic form

It is not possible to pre-multiply a matrix by a column vector, nor to post-multiply a matrix by a row vector. The matrix product **a'Ba** yields a scalar and is called a quadratic form. Note that **B** must be a square matrix if **a'Ba** is to conform to multiplication. Here is an example of a quadratic form

$$\mathbf{a'Ba} = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 & 7 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 43$$

Inverting a matrix

The matrix analog of division involves an operation called *inverting a matrix. Only* square matrices can be inverted. Inversion is a tedious numerical procedure and it is best performed by computers. There are many ways to invert a matrix, but ultimately whichever method is selected by a program is immaterial. If you wish to try one method by hand, a very popular numerical method is the Gauss-Jordan method.

Identity matrix

To augment the notion of the inverse of a matrix, $\mathbf{A}^{-1}$ (**A** inverse) we notice the following relation

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\,\mathbf{A}^{-1} = \mathbf{I}$$

I is a matrix of form

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

I is called the identity matrix and is a special case of a *diagonal* matrix. Any matrix that has zeros in all of the off-diagonal positions is a diagonal matrix.

6.5.3.2. Determinant and Eigenstructure

A matrix determinant is difficult to define but a very useful number

Unfortunately, not every square matrix has an inverse (although most do). Associated with any square matrix is a single number that represents a unique function of the numbers in the matrix. This scalar function of a square matrix is called the *determinant*. *The determinant of a matrix* **A** *is denoted by* |**A**|. A formal definition for the deteterminant of a square matrix **A** = (a_{ij}) is somewhat beyond the scope of this Handbook. Consult any good linear algebra textbook if you are interested in the mathematical details.

Singular matrix

As is the case of inversion of a square matrix, calculation of the determinant is tedious and computer assistance is needed for practical calculations. If the determinant of the (square) matrix is exactly zero, the matrix is said to be *singular* and it has no inverse.

Determinant of variance-covariance matrix

Of great interest in statistics is the determinant of a square symmetric matrix **D** whose diagonal elements are sample variances and whose off-diagonal elements are sample covariances. Symmetry means that the matrix and its transpose are identical (i.e., **A = A'**). An example is

$$\mathbf{D} = \begin{bmatrix} s_1^2 & s_1 s_2 r_{12} & \cdots & s_1 s_p r_{1p} \\ s_2 s_1 r_{21} & s_2^2 & \cdots & s_2 s_p r_{2p} \\ \vdots & \vdots & & \vdots \\ s_p s_1 r_{p1} & s_p s_2 r_{p2} & \cdots & s_p^2 \end{bmatrix}$$

where s_1 and s_2 are sample standard deviations and r_{ij} is the sample correlation.

D is the *sample variance-covariance matrix* for observations of a multivariate vector of *p* elements. The determinant of **D**, in this case, is sometimes called the *generalized variance.*

Characteristic equation

In addition to a determinant and possibly an inverse, every square matrix has associated with it a *characteristic equation*. The characteristic equation of a matrix is formed

by subtracting some particular value, usually denoted by the greek letter λ (lambda), from each diagonal element of the matrix, such that the determinant of the resulting matrix is equal to zero. For example, the characteristic equation of a second order (2 x 2) matrix **A** may be written as

Definition of the characteristic equation for 2x2 matrix

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$$

Eigenvalues of a matrix

For a matrix of order *p*, there may be as many as *p* different values for λ that will satisfy the equation. These different values are called the eigenvalues of the matrix.

Eigenvectors of a matrix

Associated with each eigenvalue is a vector, **v**, called the *eigenvector*. The eigenvector satisfies the equation

$$\mathbf{Av} = \lambda\mathbf{v}$$

Eigenstructure of a matrix

If the complete set of eigenvalues is arranged in the diagonal positions of a diagonal matrix **V**, the following relationship holds

$$\mathbf{AV} = \mathbf{VL}$$

This equation specifies the complete *eigenstructure* of **A**. Eigenstructures and the associated theory figure heavily in multivariate procedures and the numerical evaluation of **L** and **V** is a central computing problem.

6.5.4. Elements of Multivariate Analysis

Multivariate analysis

Multivariate analysis is a branch of statistics concerned with the analysis of multiple measurements, made on one or several samples of individuals. For example, we may wish to measure length, width and weight of a product.

Multiple measurement, or observation, as row or column vector

A multiple measurement or observation may be expressed as

$$\mathbf{x} = [4 \quad 2 \quad 0.6]$$

referring to the physical properties of length, width and weight, respectively. It is customary to denote multivariate quantities with bold letters. The collection of measurements on **x** is called a vector. In this case it is a row vector. We could have written **x** as a column vector.

$$\mathbf{x} = \begin{bmatrix} 4 \\ 2 \\ 0.6 \end{bmatrix}$$

Matrix to represent more than one multiple measurement

If we take several such measurements, we record them in a rectangular array of numbers. For example, the **X** matrix below represents 5 observations, on each of three variables.

$$\mathbf{X} = \begin{bmatrix} 4.0 & 2.0 & .60 \\ 4.2 & 2.1 & .59 \\ 3.9 & 2.0 & .58 \\ 4.3 & 2.1 & .62 \\ 4.1 & 2.2 & .63 \end{bmatrix}$$

By convention, rows typically represent

In this case the number of rows, (n = 5), is the number of observations, and the number of columns, (p = 3), is the number of variables that are measured. The rectangular array is an assembly of n row vectors of length p. This array is called a matrix, or, more specifically, a n by p matrix. Its

observations and columns represent variables

name is **X**. The names of matrices are usually written in bold, uppercase letters, as in Section 6.5.3. We could just as well have written **X** as a *p* (variables) by *n* (measurements) matrix as follows:

$$\mathbf{X} = \begin{bmatrix} 4.0 & 4.2 & 3.9 & 4.3 & 4.1 \\ 2.0 & 2.1 & 2.0 & 2.1 & 2.2 \\ .60 & .59 & .58 & .62 & .63 \end{bmatrix}$$

Definition of Transpose

A matrix with rows and columns exchanged in this manner is called the transpose of the original matrix.

6.5.4.1. Mean Vector and Covariance Matrix

The first step in analyzing multivariate data is computing the mean vector and the variance-covariance matrix.

Sample data matrix

Consider the following matrix:

$$\mathbf{X} = \begin{bmatrix} 4.0 & 2.0 & .60 \\ 4.2 & 2.1 & .59 \\ 3.9 & 2.0 & .58 \\ 4.3 & 2.1 & .62 \\ 4.1 & 2.2 & .63 \end{bmatrix}$$

The set of 5 observations, measuring 3 variables, can be described by its *mean vector* and *variance-covariance matrix*. The three variables, from left to right are length, width, and height of a certain object, for example. Each row vector $\mathbf{X_i}$ is another observation of the three variables (or components).

Definition of mean vector and variance-covariance matrix

The mean vector consists of the means of each variable and the variance-covariance matrix consists of the variances of the variables along the main diagonal and the covariances between each pair of variables in the other matrix positions.

The formula for computing the covariance of the variables X and Y is

$$COV = \frac{\sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y})}{n - 1}$$

with $\bar{x}$ and $\bar{y}$ denoting the means of X and Y, respectively.

Mean vector and variance-covariance matrix for sample data matrix

The results are:

$$\bar{\mathbf{x}} = \begin{bmatrix} 4.10 & 2.08 & .604 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 0.025 & 0.0075 & 0.00175 \\ 0.0075 & 0.0070 & 0.00135 \\ 0.00175 & 0.00135 & 0.00043 \end{bmatrix}$$

where the mean vector contains the arithmetic averages of the three variables and the (unbiased) variance-covariance matrix **S** is calculated by

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'.$$

where $n = 5$ for this example.

Thus, 0.025 is the variance of the length variable, 0.0075 is the covariance between the length and the width variables, 0.00175 is the covariance between the length and the height variables, 0.007 is the variance of the width variable, 0.00135 is the covariance between the width and height variables and .00043 is the variance of the height variable.

Centroid, dispersion matix

The mean vector is often referred to as the *centroid* and the variance-covariance matrix as the *dispersion* or dispersion matrix. Also, the terms variance-covariance matrix and covariance matrix are used interchangeably.

6.5.4.2. The Multivariate Normal Distribution

Multivariate normal model

When multivariate data are analyzed, the multivariate normal model is the most commonly used model.

The multivariate normal distribution model extends the univariate normal distribution model to fit vector observations.

Definition of multivariate normal distribution

A p-dimensional vector of random variables

$$X = X_1, X_2, \ldots, X_p \qquad -\infty < X_i < \infty, i = 1, \ldots, p$$

is said to have a multivariate normal distribution if its density function f(X) is of the form

$$\begin{aligned} f(X) &= f(X_1, X_2, \ldots, X_p) \\ &= \left(\tfrac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\left\{-\tfrac{1}{2}(X-m)'\Sigma^{-1}(X-m)\right\} \end{aligned}$$

where $\mathbf{m} = (m_1, ..., m_p)$ is the vector of means and Σ is the variance-covariance matrix of the multivariate normal distribution. The shortcut notation for this density is

$$X = \mathrm{N}_p(m, \Sigma)$$

Univariate normal distribution

When $p = 1$, the one-dimensional vector $\mathbf{X} = X_1$ has the normal distribution with mean m and variance σ^2

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

Bivariate normal distribution

When $p = 2$, $\mathbf{X} = (X_1, X_2)$ has the bivariate normal distribution with a two-dimensional vector of means, $\boldsymbol{m} = (m_1, m_2)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

The correlation between the two random variables is given by

$$\rho = \frac{\sigma_{21}}{\sigma_1 \sigma_2}$$

6.5.4.3. Hotelling's *T* squared

Hotelling's T^2 distribution

A multivariate method that is the multivariate counterpart of Student's-*t* and which also forms the basis for certain multivariate control charts is based on Hotelling's T^2 distribution, which was introduced by Hotelling (1947).

Univariate t-test for mean

Recall, from Section 1.3.5.2,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a *t* distribution provided that *X* is normally distributed, and can be used as long as *X* doesn't differ greatly from a normal distribution. If we wanted to test the hypothesis that $\mu = \mu_0$, we would then have

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

so that

$$\begin{aligned} t^2 &= (\bar{x} - \mu_0)^2/(s^2/n) \\ &= n(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0) \end{aligned}$$

Generalize to p variables

When t^2 is generalized to *p* variables it becomes

$$T^2 = n(\bar{x} - \mu_0)S^{-1}(\bar{x} - \mu_0)$$

with

$$\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \qquad \mu_0 = \begin{bmatrix} \mu_1^0 \\ \mu_2^0 \\ \vdots \\ \mu_p^0 \end{bmatrix}$$

S^{-1} is the inverse of the sample variance-covariance matrix, *S*, and *n* is the sample size upon which each $\bar{x}_i$, $i = 1, 2, ..., p$, is based. (The diagonal elements of *S* are the variances and the off-diagonal elements are the covariances for the *p*

variables. This is discussed further in Section 6.5.4.3.1.)

Distribution of T^2

It is well known that when $\mu = \mu_0$

$$T^2 \sim \frac{p(n-1)}{n-p} F_{(p,n-p)}$$

with $F_{(p,n-p)}$ representing the *F* distribution with p degrees of freedom for the numerator and $n - p$ for the denominator. Thus, if μ were specified to be μ_0, this could be tested by taking a single p-variate sample of size n, then computing T^2 and comparing it with

$$\frac{p(n-1)}{n-p} F_{\alpha(p,n-p)}$$

for a suitably chosen α.

Result does not apply directly to multivariate Shewhart-type charts

Although this result applies to hypothesis testing, it does not apply directly to multivariate Shewhart-type charts (for which there is no μ_0), although the result might be used as an approximation when a large sample is used and data are in subgroups, with the upper control limit (UCL) of a chart based on the approximation.

Three-sigma limits from univariate control chart

When a univariate control chart is used for Phase I (analysis of historical data), and subsequently for Phase II (real-time process monitoring), the general form of the control limits is the same for each phase, although this need not be the case. Specifically, three-sigma limits are used in the univariate case, which skirts the relevant distribution theory for each Phase.

Selection of different control limit forms for each Phase

Three-sigma units are generally not used with multivariate charts, however, which makes the selection of different control limit forms for each Phase (based on the relevant distribution theory), a natural choice.

6.5.4.3.1. T^2 Chart for Subgroup Averages -- Phase I

Estimate μ with $\bar{\bar{x}}$

Since μ is generally unknown, it is necessary to estimate μ analogous to the way that μ is estimated when an $\bar{X}$ chart is used. Specifically, when there are rational subgroups, μ is estimated by $\bar{\bar{x}}$, with

$$\bar{\bar{x}} = \begin{bmatrix} \bar{\bar{x}}_1 \\ \bar{\bar{x}}_2 \\ \vdots \\ \bar{\bar{x}}_p \end{bmatrix}$$

Obtaining the $\bar{\bar{x}}_i$

Each $\bar{\bar{x}}_i$, i = 1, 2, ..., p, is obtained the same way as with an $\bar{X}$ chart, namely, by taking k subgroups of size n and computing

$$\bar{\bar{x}}_i = \left(\frac{1}{k}\right) \sum_{l=1}^{k} \bar{x}_{il}$$

Here $\bar{x}_{il}$ is used to denote the average for the lth subgroup of the ith variable. That is,

$$\bar{x}_{il} = \sum_{r=1}^{n} x_{ilr}$$

with x_{ilr} denoting the rth observation (out of n) for the ith variable in the lth subgroup.

Estimating the variances and covariances

The variances and covariances are similarly averaged over the subgroups. Specifically, the s_{ij} elements of the variance-covariance matrix S are obtained as

$$s_{ij} = \left(\frac{1}{k}\right) \sum_{l=1}^{k} s_{ijl}$$

with s_{ijl} for $i \neq j$ denoting the sample covariance between variables X_i and X_j for the lth subgroup, and s_{ij} for $i = j$

denotes the sample variance of X_i. The variances s_{il}^2 (= s_{iil}) for subgroup l and for variables i = 1, 2, ..., p are computed as

$$\frac{1}{n-1}\sum_{r=1}^{n}(x_{ilr}-\bar{x}_{il})^2 .$$

Similarly, the covariances s_{ijl} between variables X_i and X_j for subgroup l are computed as

$$\frac{1}{n-1}\sum_{r=1}^{n}(x_{ilr}-\bar{x}_{il})(x_{jlr}-\bar{x}_{jl}) .$$

Compare T^2 against control values

As with an $\overline{X}$ chart (or any other chart), the k subgroups would be tested for control by computing k values of T^2 and comparing each against the UCL. If any value falls above the UCL (there is no lower control limit), the corresponding subgroup would be investigated.

Formula for plotted T^2 values

Thus, one would plot

$$T_j^2 = n(\bar{x}^{(j)}-\bar{\bar{x}})' S_p^{-1}(\bar{x}^{(j)}-\bar{\bar{x}})$$

for the jth subgroup (j = 1, 2, ..., k), with $\bar{x}$ denoting a vector with p elements that contains the subgroup averages for each of the p characteristics for the jth subgroup. (S_p^{-1} is the inverse matrix of the "pooled" variance-covariance matrix, S_p, which is obtained by averaging the subgroup variance-covariance matrices over the k subgroups.)

Formula for the upper control limit

Each of the k values of T_j^2 given in the equation above would be compared with

$$UCL = \left(\frac{knp-kp-np+p}{kn-k-p+1}\right)F_{\alpha,(p,kn-k-p+1)}$$

Lower control limits

A lower control limit is generally not used in multivariate control chart applications, although some control chart methods do utilize a LCL. Although a small value for T_j^2 might seem desirable, a value that is very small would likely indicate a problem of some type as we would not expect every element of $\bar{x}^{(j)}$ to be virtually equal to every element in $\bar{\bar{x}}$.

Delete out-of-control points once cause discovered and

As with any Phase I control chart procedure, if there are any points that plot above the UCL and can be identified as corresponding to out-of-control conditions that have been corrected, the point(s) should be deleted and the UCL recomputed. The remaining points would then be compared with the new UCL and the process continued as long as

corrected necessary, remembering that points should be deleted only if their correspondence with out-of-control conditions can be identified and the cause(s) of the condition(s) were removed.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

6.5.4.3.2. T^2 Chart for Subgroup Averages -- Phase II

Phase II requires recomputing S_p and $\bar{\bar{x}}$ and different control limits

Determining the UCL that is to be subsequently applied to *future* subgroups entails recomputing, if necessary, S_p and $\bar{\bar{x}}$, *and* using a constant and an *F*-value that are different from the form given for the Phase I control limits. The form is different because different distribution theory is involved since future subgroups are assumed to be independent of the "current" set of subgroups that is used in calculating S_p and $\bar{\bar{x}}$. (The same thing happens with $\bar{X}$ charts; the problem is simply ignored through the use of 3-sigma limits, although a different approach should be used when there is a small number of subgroups -- and the necessary theory has been worked out.)

Illustration

To illustrate, assume that *a* subgroups had been discarded (with possibly *a* = 0) so that *k* - *a* subgroups are used in obtaining S_p and $\bar{\bar{x}}$. We shall let these two values be represented by $S_p^{\bullet}$ and $\bar{\bar{x}}^{\bullet}$ to distinguish them from the original values, S_p and $\bar{\bar{x}}$, before any subgroups are deleted. Future values to be plotted on the multivariate chart would then be obtained from

$$n(\bar{x}^{(future)} - \bar{\bar{x}}^{\bullet})'(S_p^{\bullet})^{-1}(\bar{x}^{(future)} - \bar{\bar{x}}^{\bullet})$$

with $\bar{x}^{(future)}$ denoting an arbitrary vector containing the averages for the *p* characteristics for a single subgroup obtained in the future. Each of these future values would be plotted on the multivariate chart and compared with

Phase II control limits

$$UCL = \left(\frac{p(k-a+1)(n-1)}{(k-a)n-k+a-p+1}\right) F_{\alpha,(p,(k-a)n-k+a-p+1)}$$

with *a* denoting the number of the original subgroups that are deleted before computing $S_p^{\bullet}$ and $\bar{\bar{x}}^{\bullet}$. Notice that the equation for the control limits for Phase II given here does *not* reduce to the equation for the control limits for Phase I when *a* = 0, nor should we expect it to since the Phase I UCL is used when testing for control of the entire *set* of subgroups that is used in computing S_p and $\bar{\bar{x}}$.

6.5.4.3.3. Chart for Individual Observations -- Phase I

Multivariate individual control charts

Control charts for multivariate individual observations can be constructed, just as charts can be constructed for univariate individual observations.

Constructing the control chart

Assume there are m historical multivariate observations to be tested for control, so that Q_j, j = 1, 2, ..., m are computed, with

$$Q_j = (x - \bar{x}_m)' S_m^{-1} (x - \bar{x}_m)$$

Control limits

Each value of Q_j is compared against control limits of

$$LCL = \left(\frac{(m-1)^2}{m} \right) B\left(1 - \frac{\alpha}{2}; \frac{p}{2}; \frac{m-p-1}{2}\right)$$

$$UCL = \left(\frac{(m-1)^2}{m} \right) B\left(\frac{\alpha}{2}; \frac{p}{2}; \frac{m-p-1}{2}\right)$$

with $B(\cdot)$ denoting the beta distribution with parameters $p/2$ and $(m\text{-}p\text{-}1)/2$. These limits are due to Tracy, Young and Mason (1992). Note that a LCL is stated, unlike the other multivariate control chart procedures given in this section. Although interest will generally be centered at the UCL, a value of Q below the LCL should also be investigated, as this could signal problems in data recording.

Delete points if special cause(s) are identified and corrected

As in the case when subgroups are used, if any points plot outside these control limits and special cause(s) that were subsequently removed can be identified, the point(s) would be deleted and the control limits recomputed, making the appropriate adjustments on the degrees of freedom, and re-testing the remaining points against the new limits.

6.5.4.3.4. Chart for Individual Observations -- Phase II

Control limits

In Phase II, each value of Q_j would be plotted against the UCL of

$$\frac{p(m+1)(m-1)}{m^2-mp}F_{\alpha,(p,m-p)}$$

with, as before, p denoting the number of characteristics.

Further Information

The control limit expressions given in this section and the immediately preceding sections are given in Ryan (2000, Chapter 9).

6.5.4.3.5. Charts for Controlling Multivariate Variability

No satisfactory charts for multivariate variability

Unfortunately, there are no charts for controlling multivariate variability, with either subgroups or individual observations, that are simple, easy-to-understand and implement, and statistically defensible. Methods based on the generalized variance have been proposed for subgroup data, but such methods have been criticized by Ryan (2000, Section 9.4) and some references cited therein. For individual observations, the multivariate analogue of a univariate moving range chart might be considered as an estimator of the variance-covariance matrix for Phase I, although the distribution of the estimator is unknown.

6.5.4.3.6. Constructing Multivariate Charts

Multivariate control charts not commonly available in statistical software

Although control charts were originally constructed and maintained by hand, it would be extremely impractical to try to do that with the chart procedures that were presented in Sections 6.5.4.3.1-6.5.4.3.4. Unfortunately, the well-known statistical software packages do not have capability for the four procedures just outlined. However, Dataplot, which is used for case studies and tutorials throughout this e-Handbook, does have that capability.

6.5.5. Principal Components

Dimension reduction tool

A Multivariate Analysis problem could start out with a substantial number of correlated variables. **Principal Component Analysis** is a dimension-reduction tool that can be used advantageously in such situations. Principal component analysis aims at reducing a large set of variables to a small set that still contains most of the information in the large set.

Principal factors

The technique of principal component analysis enables us to create and use a reduced set of variables, which are called *principal factors*. A reduced set is much easier to analyze and interpret. To study a data set that results in the estimation of roughly 500 parameters may be difficult, but if we could reduce these to 5 it would certainly make our day. We will show in what follows how to achieve substantial dimension reduction.

Inverse transformaion not possible

While these principal factors represent or replace one or more of the original variables, it should be noted that they are not just a one-to-one transformation, so inverse transformations are not possible.

Original data matrix

To shed a light on the structure of principal components analysis, let us consider a multivariate data matrix $\boldsymbol{X}$, with n rows and p columns. The p elements of each row are scores or measurements on a subject such as height, weight and age.

Linear function that maximizes variance

Next, *standardize* the $\boldsymbol{X}$ matrix so that each column mean is 0 and each column variance is 1. Call this matrix $\boldsymbol{Z}$. Each column is a vector variable, $\boldsymbol{z}_i$, $i = 1, \ldots, p$. The main idea behind principal component analysis is to derive a linear function $\boldsymbol{y}$ for each of the vector variables $\boldsymbol{z}_i$. This linear function possesses an extremely important property; namely, its variance is maximized.

Linear function is component of $\boldsymbol{z}$

This linear function is referred to as a component of $\boldsymbol{z}$. To illustrate the computation of a single element for the jth $\boldsymbol{y}$ vector, consider the product $\boldsymbol{y} = \boldsymbol{z}\,\boldsymbol{v}'$ where $\boldsymbol{v}'$ is a column vector of $\boldsymbol{V}$ and $\boldsymbol{V}$ is a p x p coefficient matrix that carries the p-element variable $\boldsymbol{z}$ into the derived n-element variable $\boldsymbol{y}$. $\boldsymbol{V}$ is known as the eigen vector matrix. The dimension of

$\boldsymbol{z}$ is 1 x p, the dimension of $\boldsymbol{v}'$ is p x 1. The scalar algebra for the component score for the ith individual of $\boldsymbol{y_j}$, j = 1, ...p is:

$$y_{ji} = v'_1 z_{1i} + v'_2 z_{2i} + ... + v'_p z_{pi}$$

This becomes in matrix notation for all of the y:

$$\boldsymbol{Y = ZV}$$

Mean and dispersion matrix of $\boldsymbol{y}$

The mean of $\boldsymbol{y}$ is $\boldsymbol{m}_y = \boldsymbol{V'm}_z = 0$, because $\boldsymbol{m}_z = 0$.

The dispersion matrix of $\boldsymbol{y}$ is

$$\boldsymbol{D_y = V'D_z V = V'RV}$$

***R** is correlation matrix*

Now, it can be shown that the dispersion matrix $\boldsymbol{D_z}$ of a standardized variable is a correlation matrix. Thus $\boldsymbol{R}$ is the correlation matrix for $\boldsymbol{z}$.

Number of parameters to estimate increases rapidly as p increases

At this juncture you may be tempted to say: "so what?". To answer this let us look at the intercorrelations among the elements of a vector variable. The number of parameters to be estimated for a p-element variable is

- p means
- p variances
- $(p^2 - p)/2$ covariances
- for a total of $2p + (p^2-p)/2$ parameters.

So

- If p = 2, there are 5 parameters
- If p = 10, there are 65 parameters
- If p = 30, there are 495 parameters

Uncorrelated variables require no covariance estimation

All these parameters must be estimated and interpreted. That is a herculean task, to say the least. Now, if we could transform the data so that we obtain a vector of uncorrelated variables, life becomes much more bearable, since there are no covariances.

6.5.5.1. Properties of Principal Components

Orthogonalizing Transformations

*Transformation from **z** to **y***

The equation ***y** = **V'z*** represents a transformation, where ***y*** is the transformed variable, **z** is the original standardized variable and ***V*** is the premultiplier to go from ***z*** to ***y***.

Orthogonal transformations simplify things

To produce a transformation vector for **y** for which the elements are uncorrelated is the same as saying that we want **V** such that $\mathbf{D_y}$ is a diagonal matrix. That is, all the off-diagonal elements of $\mathbf{D_y}$ must be zero. This is called an *orthogonalizing transformation*.

*Infinite number of values for **V***

There are an infinite number of values for **V** that will produce a diagonal $\mathbf{D_y}$ for any correlation matrix **R**. Thus the mathematical problem "find a unique **V** such that $\mathbf{D_y}$ is diagonal" cannot be solved as it stands. A number of famous statisticians such as Karl Pearson and Harold Hotelling pondered this problem and suggested a "variance maximizing" solution.

Principal components maximize variance of the transformed elements, one by one

Hotelling (1933) derived the "principal components" solution. It proceeds as follows: for the first principal component, which will be the first element of **y** and be defined by the coefficients in the first column of **V**, (denoted by $\mathbf{v}_1$), we want a solution such that the variance of $\mathbf{y}_1$ will be maximized.

*Constrain **v** to generate a unique solution*

The constraint on the numbers in $\mathbf{v}_1$ is that the sum of the squares of the coefficients equals 1. Expressed mathematically, we wish to maximize

$$\frac{1}{N}\sum_{i=1}^{N} y_{1i}^2$$

where

$$y_{1i} = \mathbf{v_1}' \mathbf{z_i}$$

and $\mathbf{v_1}'\mathbf{v_1} = 1$ (this is called "normalizing " $\mathbf{v}_1$).

*Computation of first principal component from **R** and $\mathbf{v_1}$*

Substituting the middle equation in the first yields

$$\frac{1}{N}\sum_{i=1}^{N} y_{1i}^2 = \mathbf{v_1'}\,\mathbf{R}\,\mathbf{v_1}$$

where **R** is the correlation matrix of **Z**, which, in turn, is the standardized matrix of **X**, the original data matrix. Therefore, we want to maximize $\mathbf{v_1'Rv_1}$ subject to $\mathbf{v_1'v_1} = 1$.

The eigenstructure

Lagrange multiplier approach

Let

$$\phi_1 = \mathbf{v_1'}\,\mathbf{R}\,\mathbf{v_1} - \lambda_1\left(\mathbf{v_1'v_1} - 1\right)$$

>

introducing the restriction on $\mathbf{v}_1$ via the Lagrange multiplier approach. It can be shown (T.W. Anderson, 1958, page 347, theorem 8) that the vector of partial derivatives is

$$\frac{\partial \phi_1}{\partial \mathbf{v}_1} = 2\mathbf{R}\mathbf{v}_1 - 2\lambda_1\mathbf{v}_1$$

and setting this equal to zero, dividing out 2 and factoring gives

$$\left(\mathbf{R} - \lambda_1\mathbf{I}\right)\mathbf{v}_1 = 0$$

This is known as "the problem of the eigenstructure of **R**".

Set of p homogeneous equations

The partial differentiation resulted in a set of *p* homogeneous equations, which may be written in matrix form as follows

$$\begin{bmatrix} (1-\lambda_i) & r_{12} & \cdots & r_{1p} \\ r_{21} & (1-\lambda_i) & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & (1-\lambda_i) \end{bmatrix} \begin{bmatrix} v_{1i} \\ v_{2i} \\ \vdots \\ v_{pi} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The characteristic equation

*Characterstic equation of **R** is a polynomial of*

The characteristic equation of **R** is a polynomial of degree *p*, which is obtained by expanding the determinant of

degree p

$$|\mathbf{R} - \lambda \mathbf{I}| = \begin{vmatrix} r_{11} - \lambda & & \cdots & r_{1p} \\ r_{21} & r_{22} - \lambda & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} - \lambda \end{vmatrix} = 0$$

and solving for the roots λ_j, $j = 1, 2, ..., p$.

Largest eigenvalue

Specifically, the largest eigenvalue, λ_1, and its associated vector, $\mathbf{v_1}$, are required. Solving for this eigenvalue and vector is another mammoth numerical task that can realistically only be performed by a computer. In general, software is involved and the algorithms are complex.

Remainig p eigenvalues

After obtaining the first eigenvalue, the process is repeated until all *p* eigenvalues are computed.

*Full eigenstructure of **R***

To succinctly define the full eigenstructure of **R**, we introduce another matrix **L**, which is a diagonal matrix with λ_j in the *j*th position on the diagonal. Then the full eigenstructure of **R** is given as

$$\mathbf{RV} = \mathbf{VL}$$

where

$$\mathbf{V'V} = \mathbf{VV'} = \mathbf{I}$$

and

$$\mathbf{V'RV} = \mathbf{L} = \mathbf{D}_y$$

Principal Factors

Scale to zero means and unit variances

It was mentioned before that it is helpful to scale any transformation **y** of a vector variable **z** so that its elements have zero means and unit variances. Such a standardized transformation is called a *factoring* of **z**, or of **R**, and each linear component of the transformation is called a factor.

Deriving unit variances for principal components

Now, the principal components already have zero means, but their variances are not 1; in fact, they are the eigenvalues, comprising the diagonal elements of **L**. It is possible to derive the principal factor with unit variance from the principal component as follows

$$f_i = \frac{y_i}{\sqrt{\lambda}}$$

or for all factors:

$$f = L^{-1/2} y$$

substituting ***V'z*** for ***y*** we have

$$f = L^{-1/2} V' z = B' z$$

where

$$\mathbf{B} = \mathbf{V}\mathbf{L}^{-1/2}$$

B matrix

The matrix **B** is then the matrix of *factor score coefficients* for principal factors.

How many Eigenvalues?

Dimensionality of the set of factor scores

The number of eigenvalues, *N*, used in the final set determines the dimensionality of the set of factor scores. For example, if the original test consisted of 8 measurements on 100 subjects, and we extract 2 eigenvalues, the set of factor scores is a matrix of 100 rows by 2 columns.

Eigenvalues greater than unity

Each column or principal factor should represent a number of original variables. Kaiser (1966) suggested a rule-of-thumb that takes as a value for *N*, the number of eigenvalues larger than unity.

Factor Structure

*Factor structure matrix **S***

The primary interpretative device in principal components is the factor structure, computed as

$$\mathbf{S} = \mathbf{V}\mathbf{L}^{1/2}$$

S is a matrix whose elements are the correlations between the principal components and the variables. If we retain, for example, two eigenvalues, meaning that there are two principal components, then the **S** matrix consists of two columns and *p* (number of variables) rows.

Table showing relation between variables and principal components

	Principal Component	
Variable	**1**	**2**
1	r_{11}	r_{12}
2	r_{21}	r_{22}

3	r_{31}	r_{32}
4	r_{41}	r_{42}

The r_{ij} are the correlation coefficients between variable *i* and principal component *j*, where *i* ranges from 1 to 4 and *j* from 1 to 2.

The communality

SS' is the source of the "explained" correlations among the variables. Its diagonal is called *"the communality"*.

Rotation

Factor analysis

If this correlation matrix, i.e., the factor structure matrix, does not help much in the interpretation, it is possible to rotate the axis of the principal components. This may result in the polarization of the correlation coefficients. Some practitioners refer to rotation after generating the factor structure as *factor analysis*.

Varimax rotation

A popular scheme for rotation was suggested by Henry Kaiser in 1958. He produced a method for orthogonal rotation of factors, called the varimax rotation, which cleans up the factors as follows:

> *for each factor, high loadings (correlations) will result for a few variables; the rest will be near zero.*

Example

The following computer output from a principal component analysis on a 4-variable data set, followed by varimax rotation of the factor structure, will illustrate his point.

	Before Rotation		After Rotation	
Variable	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
1	.853	-.989	.997	.058
2	.634	.762	.089	.987
3	.858	-.498	.989	.076
4	.633	.736	.103	.965

Communality

Formula for communality statistic

A measure of how well the selected factors (principal components) "explain" the variance of each of the variables is given by a statistic called *communality*. This is defined by

$$h_k^2 = \sum_{i=1}^{k} S_{ki}^2$$

Explanation of communality statistic

That is: the square of the correlation of variable *k* with factor *i* gives the part of the variance accounted for by that factor. The sum of these squares for *n* factors is the communality, or explained variable for that variable (row).

Roadmap to solve the V matrix

Main steps to obtaining eigenstructure for a correlation matrix

In summary, here are the main steps to obtain the eigenstructure for a correlation matrix.

1. Compute **R**, the correlation matrix of the original data. **R** is also the correlation matrix of the standardized data.

2. Obtain the characteristic equation of **R** which is a polynomial of degree *p* (the number of variables), obtained from expanding the determinant of $|\mathbf{R}-\lambda\mathbf{I}| = 0$ and solving for the roots λ_i, that is: $\lambda_1, \lambda_2, \ldots, \lambda_p$.

3. Then solve for the columns of the **V** matrix, ($\mathbf{v_1}, \mathbf{v_2}, \ldots \mathbf{v_p}$). The roots, λ_i, are called the *eigenvalue*s (or latent values). The columns of **V** are called the *eigenvectors*.

6.5.5.2. Numerical Example

Calculation of principal components example

A numerical example may clarify the mechanics of principal component analysis.

Sample data set

Let us analyze the following 3-variate dataset with 10 observations. Each observation consists of 3 measurements on a wafer: thickness, horizontal displacement and vertical displacement.

$$X = \begin{bmatrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 6 & 1 \\ 8 & 5 & 7 \\ 7 & 2 & 9 \\ 5 & 3 & 3 \\ 9 & 5 & 8 \\ 7 & 4 & 5 \\ 8 & 2 & 2 \end{bmatrix}$$

Compute the correlation matrix

First compute the correlation matrix

$$R = \begin{bmatrix} 1.00 & .67 & -.10 \\ .67 & 1.00 & -.29 \\ -.10 & -.29 & 1.00 \end{bmatrix}$$

*Solve for the roots of **R***

Next solve for the roots of **R,** using software

	λ value	**proportion**
1	1.769	.590
2	.927	.899
3	.304	1.000

Notice that

- Each eigenvalue satisfies $|\mathbf{R}-\lambda\mathbf{I}| = 0$.
- The sum of the eigenvalues = 3 = *p*, which is equal to the trace of **R** (i.e., the sum of the main diagonal elements).
- The determinant of **R** is the product of the eigenvalues.
- The product is $\lambda_1 \times \lambda_2 \times \lambda_3 = .499$.

*Compute the first column of the **V** matrix*

Substituting the first eigenvalue of 1.769 and **R** in the appropriate equation we obtain

$$\begin{bmatrix} -.769 & .67 & -.10 \\ .67 & -.769 & -.29 \\ -.10 & -.29 & -.769 \end{bmatrix} \cdot \begin{bmatrix} v_{11} \\ v_{21} \\ v_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

This is the matrix expression for 3 homogeneous equations with 3 unknowns and yields the first column of **V:** .64 .69 -.34 (again, a computerized solution is indispensable).

*Compute the remaining columns of the **V** matrix*

Repeating this procedure for the other 2 eigenvalues yields the matrix **V**

$$\mathbf{V} = \begin{bmatrix} .64 & .38 & -.66 \\ .69 & .10 & .72 \\ -.34 & .91 & .20 \end{bmatrix}$$

Notice that if you multiply **V** by its transpose, the result is an identity matrix, **V'V=I**.

Compute the $\mathbf{L}^{1/2}$ matrix

Now form the matrix $\mathbf{L}^{1/2}$, which is a diagonal matrix whose elements are the square roots of the eigenvalues of **R**. Then obtain **S**, the factor structure, using $\mathbf{S} = \mathbf{V}\,\mathbf{L}^{1/2}$

$$\begin{bmatrix} .64 & .38 & -.66 \\ .69 & .10 & .72 \\ -.34 & .91 & .20 \end{bmatrix} \begin{bmatrix} 1.33 & 0 & 0 \\ 0 & .96 & 0 \\ 0 & 0 & .55 \end{bmatrix} = \begin{bmatrix} .85 & .37 & -.37 \\ .91 & .10 & .40 \\ -.45 & .88 & .11 \end{bmatrix}$$

So, for example, .91 is the correlation between variable 2 and the first principal component.

Compute the communality

Next compute the communality, using the first two eigenvalues only

$$SS' = \begin{bmatrix} .85 & .37 \\ .91 & .09 \\ -.45 & .88 \end{bmatrix} \cdot \begin{bmatrix} .85 & .91 & -.45 \\ .37 & .09 & .88 \end{bmatrix} = \begin{bmatrix} ..8662 & .8140 & -.0606 \\ .8140 & .8420 & -.3321 \\ -.0606- & -.3321 & .9876 \end{bmatrix}$$

Diagonal elements report how much of the variability is explained

Communality consists of the diagonal elements.

	var
1	.8662
2	.8420
3	.9876

This means that the first two principal components "explain" 86.62% of the first variable, 84.20 % of the second variable, and 98.76% of the third.

Compute the coefficient matrix

The coefficient matrix, **B**, is formed using the reciprocals of the diagonals of $\mathbf{L}^{1/2}$

$$\mathbf{B} = \mathbf{V}\mathbf{L}^{-1/2} = \begin{bmatrix} .48 & .40 & -.18 \\ .52 & .10 & 1.31 \\ -.26 & .95 & .37 \end{bmatrix}$$

Compute the principal factors

Finally, we can compute the factor scores from **ZB**, where **Z** is **X** converted to standard score form. These columns are the *principal factors*.

$$\mathbf{F} = \mathbf{ZB} = \begin{bmatrix} .41 & -.69 & .06 \\ -2.11 & .07 & .63 \\ -.46 & -.32 & .30 \\ 1.62 & -1.00 & .70 \\ .70 & 1.09 & .65 \\ -.86 & 1.32 & -.85 \\ -.60 & -1.31 & .86 \\ .94 & 1.72 & -.04 \\ .22 & .03 & .34 \\ .15 & -.91 & -2.65 \end{bmatrix}$$

Principal factors control chart

These factors can be plotted against the indices, which could be times. If time is used, *the resulting plot is an example of a principal factors control chart.*

6. Process or Product Monitoring and Control

6.6. Case Studies in Process Monitoring

Detailed Examples

The general points of the first five sections are illustrated in this section using data from physical science and engineering applications. Each example is presented step-by-step in the text, and is often cross-linked with the relevant sections of the chapter describing the analysis in general.

Contents: Section 6

1. Lithography Process Example
2. Aerosol Particle Size Example

6.6.1. Lithography Process

Lithography Process This case study illustrates the use of control charts in analyzing a lithography process.

1. Background and Data
2. Graphical Representation of the Data
3. Subgroup Analysis
4. Shewhart Control Chart
5. Work This Example Yourself

6.6.1.1. Background and Data

Case Study for SPC in Batch Processing Environment

Semiconductor processing creates multiple sources of variability to monitor

One of the assumptions in using classical Shewhart SPC charts is that the only source of variation is from part to part (or within subgroup variation). This is the case for most continuous processing situations. However, many of today's processing situations have different sources of variation. The semiconductor industry is one of the areas where the processing creates multiple sources of variation.

In semiconductor processing, the basic experimental unit is a silicon wafer. Operations are performed on the wafer, but individual wafers can be grouped multiple ways. In the diffusion area, up to 150 wafers are processed in one time in a diffusion tube. In the etch area, single wafers are processed individually. In the lithography area, the light exposure is done on sub-areas of the wafer. There are many times during the production of a computer chip where the experimental unit varies and thus there are different sources of variation in this batch processing environment.

The following is a case study of a lithography process. Five sites are measured on each wafer, three wafers are measured in a cassette (typically a grouping of 24 - 25 wafers) and thirty cassettes of wafers are used in the study. The width of a line is the measurement under study. There are two line width variables. The first is the original data and the second has been cleaned up somewhat. This case study uses the raw data. The entire data table is 450 rows long with six columns.

Software

The analyses used in this case study can be generated using both Dataplot code and R code.

Case study data: wafer line width measurements

Cassette	Wafer	Site	Raw Line Width	Sequence	Cleaned Line Width
1	1	Top	3.199275	1	3.197275
1	1	Lef	2.253081	2	2.249081
1	1	Cen	2.074308	3	2.068308

```
 1         1     Rgt     2.418206      4
2.410206
 1         1     Bot     2.393732      5
2.383732
 1         2     Top     2.654947      6
2.642947
 1         2     Lef     2.003234      7
1.989234
 1         2     Cen     1.861268      8
1.845268
 1         2     Rgt     2.136102      9
2.118102
 1         2     Bot     1.976495     10
1.956495
 1         3     Top     2.887053     11
2.865053
 1         3     Lef     2.061239     12
2.037239
 1         3     Cen     1.625191     13
1.599191
 1         3     Rgt     2.304313     14
2.276313
 1         3     Bot     2.233187     15
2.203187
 2         1     Top     3.160233     16
3.128233
 2         1     Lef     2.518913     17
2.484913
 2         1     Cen     2.072211     18
2.036211
 2         1     Rgt     2.287210     19
2.249210
 2         1     Bot     2.120452     20
2.080452
 2         2     Top     2.063058     21
2.021058
 2         2     Lef     2.217220     22
2.173220
 2         2     Cen     1.472945     23
1.426945
 2         2     Rgt     1.684581     24
1.636581
 2         2     Bot     1.900688     25
1.850688
 2         3     Top     2.346254     26
2.294254
 2         3     Lef     2.172825     27
2.118825
 2         3     Cen     1.536538     28
1.480538
 2         3     Rgt     1.966630     29
1.908630
 2         3     Bot     2.251576     30
2.191576
 3         1     Top     2.198141     31
2.136141
 3         1     Lef     1.728784     32
1.664784
 3         1     Cen     1.357348     33
1.291348
 3         1     Rgt     1.673159     34
1.605159
 3         1     Bot     1.429586     35
1.359586
 3         2     Top     2.231291     36
2.159291
 3         2     Lef     1.561993     37
1.487993
 3         2     Cen     1.520104     38
1.444104
 3         2     Rgt     2.066068     39
1.988068
 3         2     Bot     1.777603     40
1.697603
 3         3     Top     2.244736     41
2.162736
 3         3     Lef     1.745877     42
1.661877
 3         3     Cen     1.366895     43
1.280895
 3         3     Rgt     1.615229     44
```

```
1.527229
 3         3     Bot     1.540863     45
1.450863
 4         1     Top     2.929037     46
2.837037
 4         1     Lef     2.035900     47
1.941900
 4         1     Cen     1.786147     48
1.690147
 4         1     Rgt     1.980323     49
1.882323
 4         1     Bot     2.162919     50
2.062919
 4         2     Top     2.855798     51
2.753798
 4         2     Lef     2.104193     52
2.000193
 4         2     Cen     1.919507     53
1.813507
 4         2     Rgt     2.019415     54
1.911415
 4         2     Bot     2.228705     55
2.118705
 4         3     Top     3.219292     56
3.107292
 4         3     Lef     2.900430     57
2.786430
 4         3     Cen     2.171262     58
2.055262
 4         3     Rgt     3.041250     59
2.923250
 4         3     Bot     3.188804     60
3.068804
 5         1     Top     3.051234     61
2.929234
 5         1     Lef     2.506230     62
2.382230
 5         1     Cen     1.950486     63
1.824486
 5         1     Rgt     2.467719     64
2.339719
 5         1     Bot     2.581881     65
2.451881
 5         2     Top     3.857221     66
3.725221
 5         2     Lef     3.347343     67
3.213343
 5         2     Cen     2.533870     68
2.397870
 5         2     Rgt     3.190375     69
3.052375
 5         2     Bot     3.362746     70
3.222746
 5         3     Top     3.690306     71
3.548306
 5         3     Lef     3.401584     72
3.257584
 5         3     Cen     2.963117     73
2.817117
 5         3     Rgt     2.945828     74
2.797828
 5         3     Bot     3.466115     75
3.316115
 6         1     Top     2.938241     76
2.786241
 6         1     Lef     2.526568     77
2.372568
 6         1     Cen     1.941370     78
1.785370
 6         1     Rgt     2.765849     79
2.607849
 6         1     Bot     2.382781     80
2.222781
 6         2     Top     3.219665     81
3.057665
 6         2     Lef     2.296011     82
2.132011
 6         2     Cen     2.256196     83
2.090196
 6         2     Rgt     2.645933     84
2.477933
```

6	2	Bot	2.422187	85	
2.252187					
6	3	Top	3.180348	86	
3.008348					
6	3	Lef	2.849264	87	
2.675264					
6	3	Cen	1.601288	88	
1.425288					
6	3	Rgt	2.810051	89	
2.632051					
6	3	Bot	2.902980	90	
2.722980					
7	1	Top	2.169679	91	
1.987679					
7	1	Lef	2.026506	92	
1.842506					
7	1	Cen	1.671804	93	
1.485804					
7	1	Rgt	1.660760	94	
1.472760					
7	1	Bot	2.314734	95	
2.124734					
7	2	Top	2.912838	96	
2.720838					
7	2	Lef	2.323665	97	
2.129665					
7	2	Cen	1.854223	98	
1.658223					
7	2	Rgt	2.391240	99	2.19324
7	2	Bot	2.196071	100	
1.996071					
7	3	Top	3.318517	101	
3.116517					
7	3	Lef	2.702735	102	
2.498735					
7	3	Cen	1.959008	103	
1.753008					
7	3	Rgt	2.512517	104	
2.304517					
7	3	Bot	2.827469	105	
2.617469					
8	1	Top	1.958022	106	
1.746022					
8	1	Lef	1.360106	107	
1.146106					
8	1	Cen	0.971193	108	
0.755193					
8	1	Rgt	1.947857	109	
1.729857					
8	1	Bot	1.643580	110	1.42358
8	2	Top	2.357633	111	
2.135633					
8	2	Lef	1.757725	112	
1.533725					
8	2	Cen	1.165886	113	
0.939886					
8	2	Rgt	2.231143	114	
2.003143					
8	2	Bot	1.311626	115	
1.081626					
8	3	Top	2.421686	116	
2.189686					
8	3	Lef	1.993855	117	
1.759855					
8	3	Cen	1.402543	118	
1.166543					
8	3	Rgt	2.008543	119	
1.770543					
8	3	Bot	2.139370	120	
1.899370					
9	1	Top	2.190676	121	
1.948676					
9	1	Lef	2.287483	122	
2.043483					
9	1	Cen	1.698943	123	
1.452943					
9	1	Rgt	1.925731	124	
1.677731					
9	1	Bot	2.057440	125	
1.807440					
9	2	Top	2.353597	126	

```
2.101597
 9          2     Lef     1.796236     127
1.542236
 9          2     Cen     1.241040     128
0.985040
 9          2     Rgt     1.677429     129
1.419429
 9          2     Bot     1.845041     130
1.585041
 9          3     Top     2.012669     131
1.750669
 9          3     Lef     1.523769     132
1.259769
 9          3     Cen     0.790789     133
0.524789
 9          3     Rgt     2.001942     134
1.733942
 9          3     Bot     1.350051     135
1.080051
10          1     Top     2.825749     136
2.553749
10          1     Lef     2.502445     137
2.228445
10          1     Cen     1.938239     138
1.662239
10          1     Rgt     2.349497     139
2.071497
10          1     Bot     2.310817     140
2.030817
10          2     Top     3.074576     141
2.792576
10          2     Lef     2.057821     142
1.773821
10          2     Cen     1.793617     143
1.507617
10          2     Rgt     1.862251     144
1.574251
10          2     Bot     1.956753     145
1.666753
10          3     Top     3.072840     146
2.780840
10          3     Lef     2.291035     147
1.997035
10          3     Cen     1.873878     148
1.577878
10          3     Rgt     2.475640     149
2.177640
10          3     Bot     2.021472     150
1.721472
11          1     Top     3.228835     151
2.926835
11          1     Lef     2.719495     152
2.415495
11          1     Cen     2.207198     153
1.901198
11          1     Rgt     2.391608     154
2.083608
11          1     Bot     2.525587     155
2.215587
11          2     Top     2.891103     156
2.579103
11          2     Lef     2.738007     157
2.424007
11          2     Cen     1.668337     158
1.352337
11          2     Rgt     2.496426     159
2.178426
11          2     Bot     2.417926     160
2.097926
11          3     Top     3.541799     161
3.219799
11          3     Lef     3.058768     162
2.734768
11          3     Cen     2.187061     163
1.861061
11          3     Rgt     2.790261     164
2.462261
11          3     Bot     3.279238     165
2.949238
12          1     Top     2.347662     166
2.015662
```

```
12          1     Lef     1.383336     167
1.049336
12          1     Cen     1.187168     168
0.851168
12          1     Rgt     1.693292     169
1.355292
12          1     Bot     1.664072     170
1.324072
12          2     Top     2.385320     171
2.043320
12          2     Lef     1.607784     172
1.263784
12          2     Cen     1.230307     173
0.884307
12          2     Rgt     1.945423     174
1.597423
12          2     Bot     1.907580     175
1.557580
12          3     Top     2.691576     176
2.339576
12          3     Lef     1.938755     177
1.584755
12          3     Cen     1.275409     178
0.919409
12          3     Rgt     1.777315     179
1.419315
12          3     Bot     2.146161     180
1.786161
13          1     Top     3.218655     181
2.856655
13          1     Lef     2.912180     182
2.548180
13          1     Cen     2.336436     183
1.970436
13          1     Rgt     2.956036     184
2.588036
13          1     Bot     2.423235     185
2.053235
13          2     Top     3.302224     186
2.930224
13          2     Lef     2.808816     187
2.434816
13          2     Cen     2.340386     188
1.964386
13          2     Rgt     2.795120     189
2.417120
13          2     Bot     2.865800     190
2.485800
13          3     Top     2.992217     191
2.610217
13          3     Lef     2.952106     192
2.568106
13          3     Cen     2.149299     193
1.763299
13          3     Rgt     2.448046     194
2.060046
13          3     Bot     2.507733     195
2.117733
14          1     Top     3.530112     196
3.138112
14          1     Lef     2.940489     197
2.546489
14          1     Cen     2.598357     198
2.202357
14          1     Rgt     2.905165     199
2.507165
14          1     Bot     2.692078     200
2.292078
14          2     Top     3.764270     201
3.362270
14          2     Lef     3.465960     202
3.061960
14          2     Cen     2.458628     203
2.052628
14          2     Rgt     3.141132     204
2.733132
14          2     Bot     2.816526     205
2.406526
14          3     Top     3.217614     206
2.805614
14          3     Lef     2.758171     207
```

```
2.344171
14          3     Cen     2.345921     208
1.929921
14          3     Rgt     2.773653     209
2.355653
14          3     Bot     3.109704     210
2.689704
15          1     Top     2.177593     211
1.755593
15          1     Lef     1.511781     212
1.087781
15          1     Cen     0.746546     213
0.320546
15          1     Rgt     1.491730     214
1.063730
15          1     Bot     1.268580     215
0.838580
15          2     Top     2.433994     216
2.001994
15          2     Lef     2.045667     217
1.611667
15          2     Cen     1.612699     218
1.176699
15          2     Rgt     2.082860     219
1.644860
15          2     Bot     1.887341     220
1.447341
15          3     Top     1.923003     221
1.481003
15          3     Lef     2.124461     222
1.680461
15          3     Cen     1.945048     223
1.499048
15          3     Rgt     2.210698     224
1.762698
15          3     Bot     1.985225     225
1.535225
16          1     Top     3.131536     226
2.679536
16          1     Lef     2.405975     227
1.951975
16          1     Cen     2.206320     228
1.750320
16          1     Rgt     3.012211     229
2.554211
16          1     Bot     2.628723     230
2.168723
16          2     Top     2.802486     231
2.340486
16          2     Lef     2.185010     232
1.721010
16          2     Cen     2.161802     233
1.695802
16          2     Rgt     2.102560     234
1.634560
16          2     Bot     1.961968     235
1.491968
16          3     Top     3.330183     236
2.858183
16          3     Lef     2.464046     237
1.990046
16          3     Cen     1.687408     238
1.211408
16          3     Rgt     2.043322     239
1.565322
16          3     Bot     2.570657     240
2.090657
17          1     Top     3.352633     241
2.870633
17          1     Lef     2.691645     242
2.207645
17          1     Cen     1.942410     243
1.456410
17          1     Rgt     2.366055     244
1.878055
17          1     Bot     2.500987     245
2.010987
17          2     Top     2.886284     246
2.394284
17          2     Lef     2.292503     247
1.798503
```

```
17          2     Cen     1.627562     248
1.131562
17          2     Rgt     2.415076     249
1.917076
17          2     Bot     2.086134     250
1.586134
17          3     Top     2.554848     251
2.052848
17          3     Lef     1.755843     252
1.251843
17          3     Cen     1.510124     253
1.004124
17          3     Rgt     2.257347     254
1.749347
17          3     Bot     1.958592     255
1.448592
18          1     Top     2.622733     256
2.110733
18          1     Lef     2.321079     257
1.807079
18          1     Cen     1.169269     258
0.653269
18          1     Rgt     1.921457     259
1.403457
18          1     Bot     2.176377     260
1.656377
18          2     Top     3.313367     261
2.791367
18          2     Lef     2.559725     262
2.035725
18          2     Cen     2.404662     263
1.878662
18          2     Rgt     2.405249     264
1.877249
18          2     Bot     2.535618     265
2.005618
18          3     Top     3.067851     266
2.535851
18          3     Lef     2.490359     267
1.956359
18          3     Cen     2.079477     268
1.543477
18          3     Rgt     2.669512     269
2.131512
18          3     Bot     2.105103     270
1.565103
19          1     Top     4.293889     271
3.751889
19          1     Lef     3.888826     272
3.344826
19          1     Cen     2.960655     273
2.414655
19          1     Rgt     3.618864     274
3.070864
19          1     Bot     3.562480     275
3.012480
19          2     Top     3.451872     276
2.899872
19          2     Lef     3.285934     277
2.731934
19          2     Cen     2.638294     278
2.082294
19          2     Rgt     2.918810     279
2.360810
19          2     Bot     3.076231     280
2.516231
19          3     Top     3.879683     281
3.317683
19          3     Lef     3.342026     282
2.778026
19          3     Cen     3.382833     283
2.816833
19          3     Rgt     3.491666     284
2.923666
19          3     Bot     3.617621     285
3.047621
20          1     Top     2.329987     286
1.757987
20          1     Lef     2.400277     287
1.826277
20          1     Cen     2.033941     288
```

```
1.457941
20          1     Rgt     2.544367     289
1.966367
20          1     Bot     2.493079     290
1.913079
20          2     Top     2.862084     291
2.280084
20          2     Lef     2.404703     292
1.820703
20          2     Cen     1.648662     293
1.062662
20          2     Rgt     2.115465     294
1.527465
20          2     Bot     2.633930     295
2.043930
20          3     Top     3.305211     296
2.713211
20          3     Lef     2.194991     297
1.600991
20          3     Cen     1.620963     298
1.024963
20          3     Rgt     2.322678     299
1.724678
20          3     Bot     2.818449     300
2.218449
21          1     Top     2.712915     301
2.110915
21          1     Lef     2.389121     302
1.785121
21          1     Cen     1.575833     303
0.969833
21          1     Rgt     1.870484     304
1.262484
21          1     Bot     2.203262     305
1.593262
21          2     Top     2.607972     306
1.995972
21          2     Lef     2.177747     307
1.563747
21          2     Cen     1.246016     308
0.630016
21          2     Rgt     1.663096     309
1.045096
21          2     Bot     1.843187     310
1.223187
21          3     Top     2.277813     311
1.655813
21          3     Lef     1.764940     312
1.140940
21          3     Cen     1.358137     313
0.732137
21          3     Rgt     2.065713     314
1.437713
21          3     Bot     1.885897     315
1.255897
22          1     Top     3.126184     316
2.494184
22          1     Lef     2.843505     317
2.209505
22          1     Cen     2.041466     318
1.405466
22          1     Rgt     2.816967     319
2.178967
22          1     Bot     2.635127     320
1.995127
22          2     Top     3.049442     321
2.407442
22          2     Lef     2.446904     322
1.802904
22          2     Cen     1.793442     323
1.147442
22          2     Rgt     2.676519     324
2.028519
22          2     Bot     2.187865     325
1.537865
22          3     Top     2.758416     326
2.106416
22          3     Lef     2.405744     327
1.751744
22          3     Cen     1.580387     328
0.924387
```

```
22          3     Rgt     2.508542     329
1.850542
22          3     Bot     2.574564     330
1.914564
23          1     Top     3.294288     331
2.632288
23          1     Lef     2.641762     332
1.977762
23          1     Cen     2.105774     333
1.439774
23          1     Rgt     2.655097     334
1.987097
23          1     Bot     2.622482     335
1.952482
23          2     Top     4.066631     336
3.394631
23          2     Lef     3.389733     337
2.715733
23          2     Cen     2.993666     338
2.317666
23          2     Rgt     3.613128     339
2.935128
23          2     Bot     3.213809     340
2.533809
23          3     Top     3.369665     341
2.687665
23          3     Lef     2.566891     342
1.882891
23          3     Cen     2.289899     343
1.603899
23          3     Rgt     2.517418     344
1.829418
23          3     Bot     2.862723     345
2.172723
24          1     Top     4.212664     346
3.520664
24          1     Lef     3.068342     347
2.374342
24          1     Cen     2.872188     348
2.176188
24          1     Rgt     3.040890     349
2.342890
24          1     Bot     3.376318     350
2.676318
24          2     Top     3.223384     351
2.521384
24          2     Lef     2.552726     352
1.848726
24          2     Cen     2.447344     353
1.741344
24          2     Rgt     3.011574     354
2.303574
24          2     Bot     2.711774     355
2.001774
24          3     Top     3.359505     356
2.647505
24          3     Lef     2.800742     357
2.086742
24          3     Cen     2.043396     358
1.327396
24          3     Rgt     2.929792     359
2.211792
24          3     Bot     2.935356     360
2.215356
25          1     Top     2.724871     361
2.002871
25          1     Lef     2.239013     362
1.515013
25          1     Cen     2.341512     363
1.615512
25          1     Rgt     2.263617     364
1.535617
25          1     Bot     2.062748     365
1.332748
25          2     Top     3.658082     366
2.926082
25          2     Lef     3.093268     367
2.359268
25          2     Cen     2.429341     368
1.693341
25          2     Rgt     2.538365     369
```

```
1.800365
25          2     Bot     3.161795     370
2.421795
25          3     Top     3.178246     371
2.436246
25          3     Lef     2.498102     372
1.754102
25          3     Cen     2.445810     373
1.699810
25          3     Rgt     2.231248     374
1.483248
25          3     Bot     2.302298     375
1.552298
26          1     Top     3.320688     376
2.568688
26          1     Lef     2.861800     377
2.107800
26          1     Cen     2.238258     378
1.482258
26          1     Rgt     3.122050     379
2.364050
26          1     Bot     3.160876     380
2.400876
26          2     Top     3.873888     381
3.111888
26          2     Lef     3.166345     382
2.402345
26          2     Cen     2.645267     383
1.879267
26          2     Rgt     3.309867     384
2.541867
26          2     Bot     3.542882     385
2.772882
26          3     Top     2.586453     386
1.814453
26          3     Lef     2.120604     387
1.346604
26          3     Cen     2.180847     388
1.404847
26          3     Rgt     2.480888     389
1.702888
26          3     Bot     1.938037     390
1.158037
27          1     Top     4.710718     391
3.928718
27          1     Lef     4.082083     392
3.298083
27          1     Cen     3.533026     393
2.747026
27          1     Rgt     4.269929     394
3.481929
27          1     Bot     4.038166     395
3.248166
27          2     Top     4.237233     396
3.445233
27          2     Lef     4.171702     397
3.377702
27          2     Cen     3.04394      398
2.247940
27          2     Rgt     3.91296      399
3.114960
27          2     Bot     3.714229     400
2.914229
27          3     Top     5.168668     401
4.366668
27          3     Lef     4.823275     402
4.019275
27          3     Cen     3.764272     403
2.958272
27          3     Rgt     4.396897     404
3.588897
27          3     Bot     4.442094     405
3.632094
28          1     Top     3.972279     406
3.160279
28          1     Lef     3.883295     407
3.069295
28          1     Cen     3.045145     408
2.229145
28          1     Rgt     3.51459      409
2.696590
```

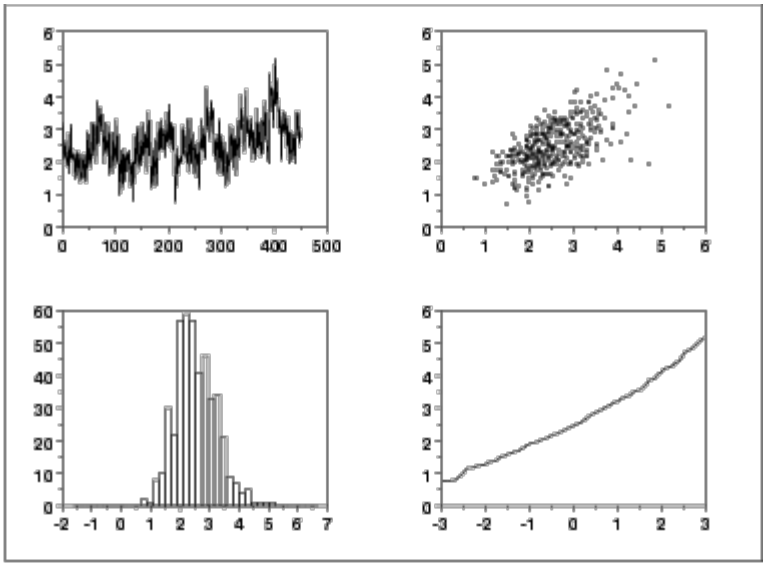
```
28          1     Bot     3.575446     410
2.755446
28          2     Top     3.024903     411
2.202903
28          2     Lef     3.099192     412
2.275192
28          2     Cen     2.048139     413
1.222139
28          2     Rgt     2.927978     414
2.099978
28          2     Bot     3.15257      415
2.322570
28          3     Top     3.55806      416
2.726060
28          3     Lef     3.176292     417
2.342292
28          3     Cen     2.852873     418
2.016873
28          3     Rgt     3.026064     419
2.188064
28          3     Bot     3.071975     420
2.231975
29          1     Top     3.496634     421
2.654634
29          1     Lef     3.087091     422
2.243091
29          1     Cen     2.517673     423
1.671673
29          1     Rgt     2.547344     424
1.699344
29          1     Bot     2.971948     425
2.121948
29          2     Top     3.371306     426
2.519306
29          2     Lef     2.175046     427
1.321046
29          2     Cen     1.940111     428
1.084111
29          2     Rgt     2.932408     429
2.074408
29          2     Bot     2.428069     430
1.568069
29          3     Top     2.941041     431
2.079041
29          3     Lef     2.294009     432
1.430009
29          3     Cen     2.025674     433
1.159674
29          3     Rgt     2.21154      434
1.343540
29          3     Bot     2.459684     435
1.589684
30          1     Top     2.86467      436
1.992670
30          1     Lef     2.695163     437
1.821163
30          1     Cen     2.229518     438
1.353518
30          1     Rgt     1.940917     439
1.062917
30          1     Bot     2.547318     440
1.667318
30          2     Top     3.537562     441
2.655562
30          2     Lef     3.311361     442
2.427361
30          2     Cen     2.767771     443
1.881771
30          2     Rgt     3.388622     444
2.500622
30          2     Bot     3.542701     445
2.652701
30          3     Top     3.184652     446
2.292652
30          3     Lef     2.620947     447
1.726947
30          3     Cen     2.697619     448
1.801619
30          3     Rgt     2.860684     449
1.962684
30          3     Bot     2.758571     450
```

```
1.858571
```

6.6.1.2. Graphical Representation of the Data

The first step in analyzing the data is to generate some simple plots of the response and then of the response versus the various factors.

4-Plot of Data



Interpretation

This 4-plot shows the following.

1. The run sequence plot (upper left) indicates that the location and scale are not constant over time. This indicates that the three factors do in fact have an effect of some kind.

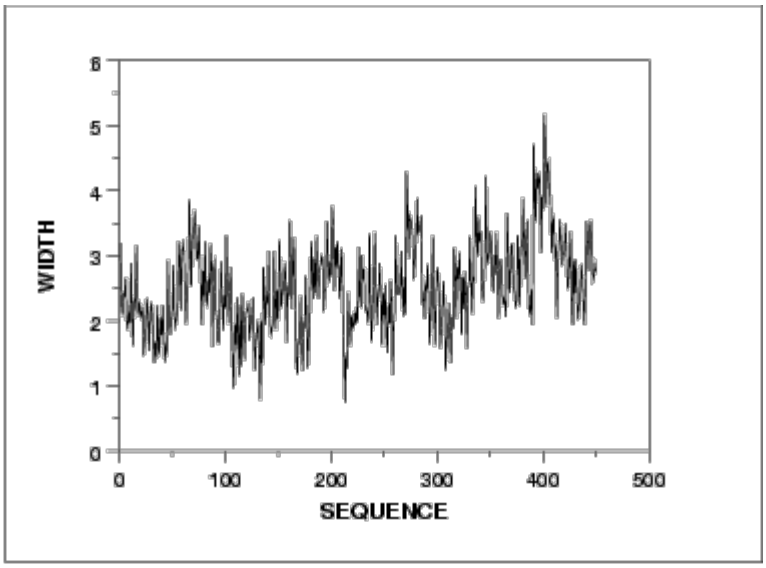
2. The lag plot (upper right) indicates that there is some mild autocorrelation in the data. This is not unexpected as the data are grouped in a logical order of the three factors (i.e., not randomly) and the run sequence plot indicates that there are factor effects.

3. The histogram (lower left) shows that most of the data fall between 1 and 5, with the center of the data at about 2.2.

4. Due to the non-constant location and scale and autocorrelation in the data, distributional inferences from the normal probability plot (lower right) are not meaningful.

The run sequence plot is shown at full size to show greater detail. In addition, a numerical summary of the data is generated.

Run Sequence Plot of Data



Numerical Summary

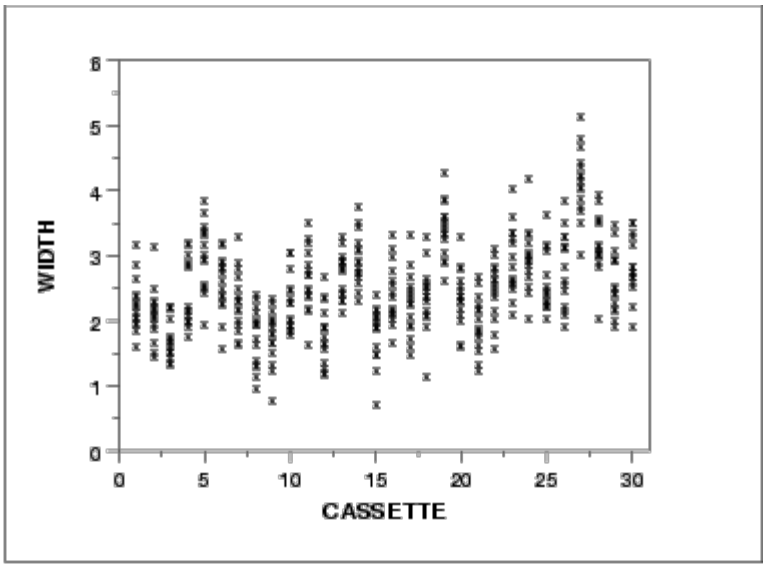
```
Sample size     = 450
Mean            =   2.53228
Median          =   2.45334
Minimum         =   0.74655
Maximum         =   5.16867
Range           =   4.42212
Stan. Dev.      =   0.69376
Autocorrelation =   0.60726
```

We are primarily interested in the mean and standard deviation. From the summary, we see that the mean is 2.53 and the standard deviation is 0.69.

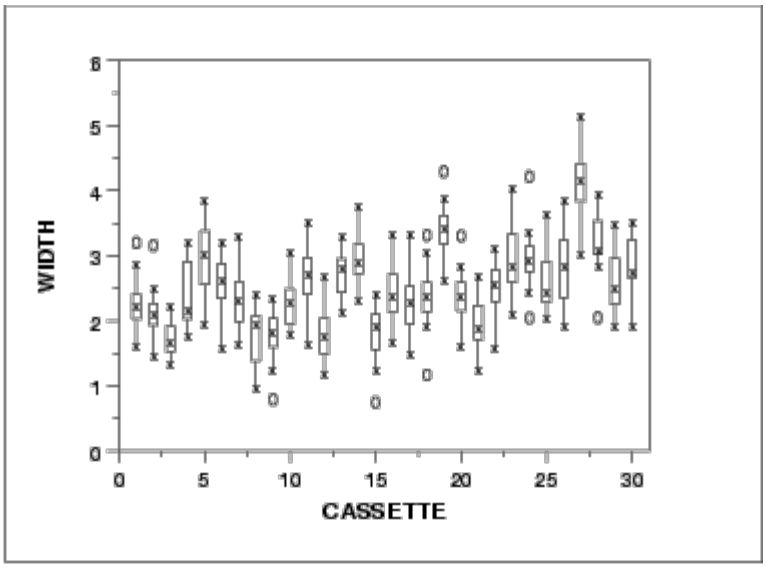
Plot response against individual factors

The next step is to plot the response against each individual factor. For comparison, we generate both a scatter plot and a box plot of the data. The scatter plot shows more detail. However, comparisons are usually easier to see with the box plot, particularly as the number of data points and groups become larger.

Scatter plot of width versus cassette



Box plot of width versus cassette

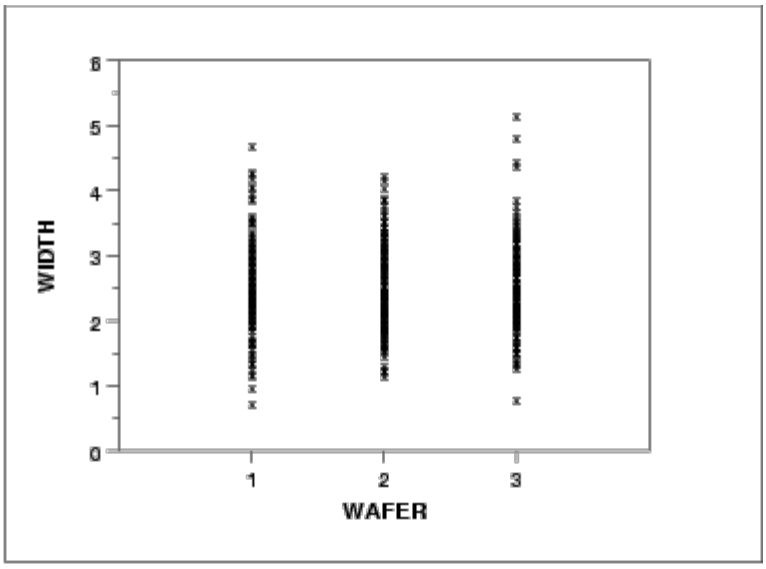


Interpretation

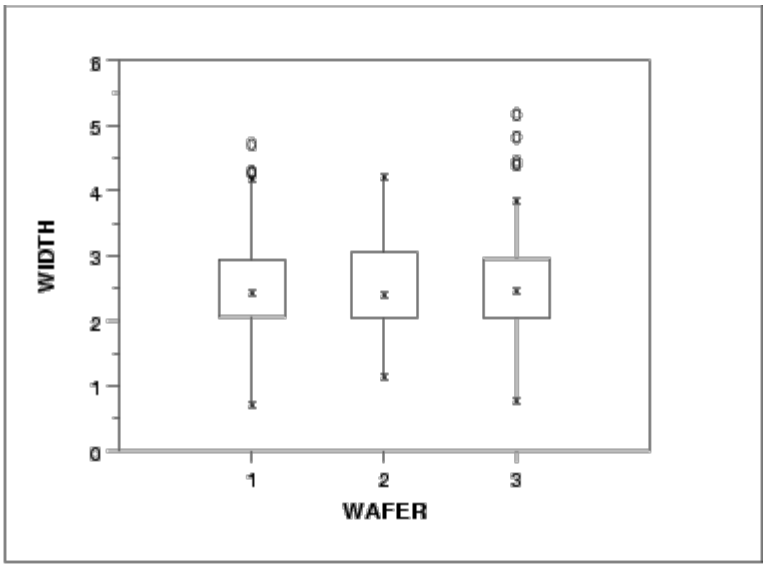
We can make the following conclusions based on the above scatter and box plots.

1. There is considerable variation in the location for the various cassettes. The medians vary from about 1.7 to 4.
2. There is also some variation in the scale.
3. There are a number of outliers.

Scatter plot of width versus wafer



Box plot of width versus wafer

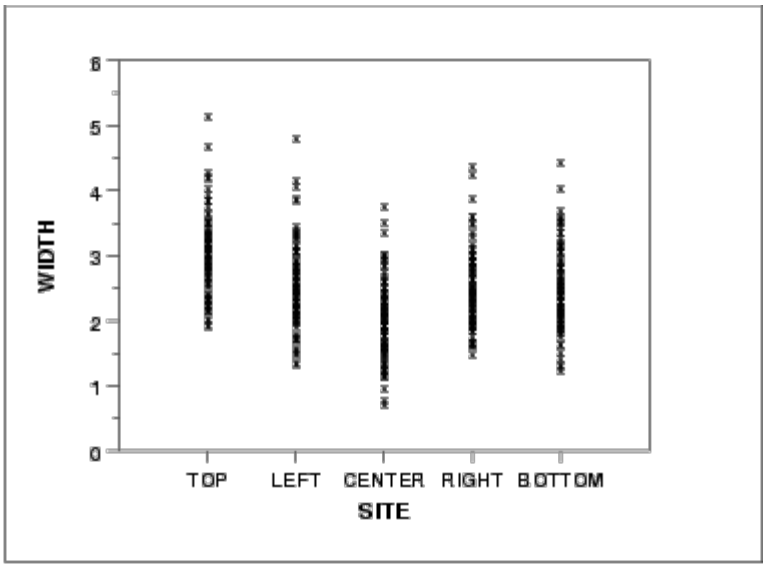


Interpretation

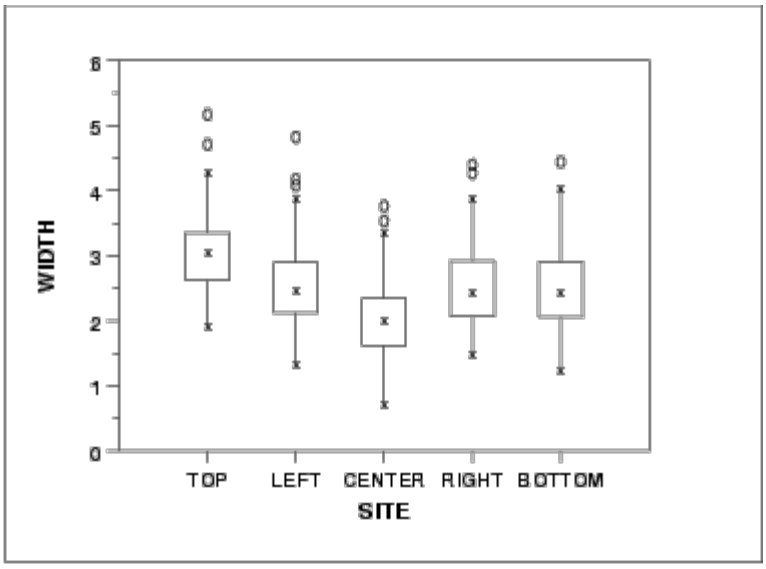
We can make the following conclusions based on the above scatter and box plots.

1. The locations for the three wafers are relatively constant.
2. The scales for the three wafers are relatively constant.
3. There are a few outliers on the high side.
4. It is reasonable to treat the wafer factor as homogeneous.

Scatter plot of width versus site



Box plot of width versus site



Interpretation

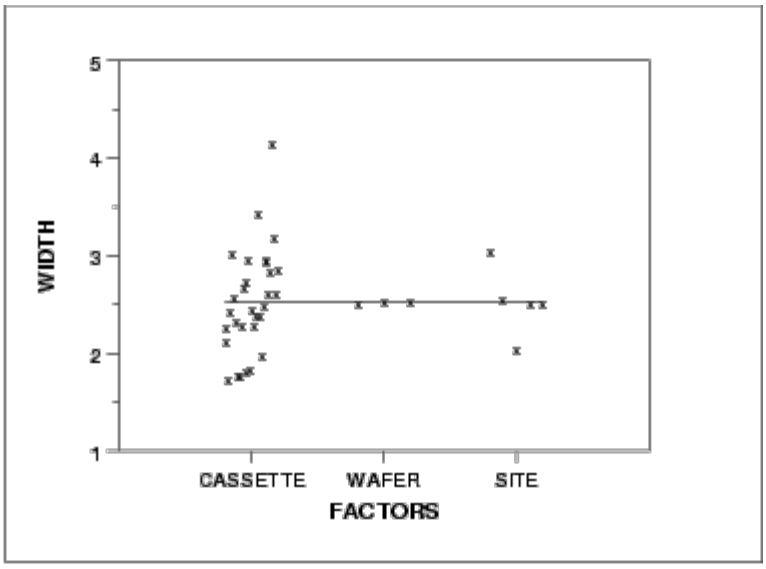
We can make the following conclusions based on the above scatter and box plots.

1. There is some variation in location based on site. The center site in particular has a lower median.
2. The scales are relatively constant across sites.
3. There are a few outliers.

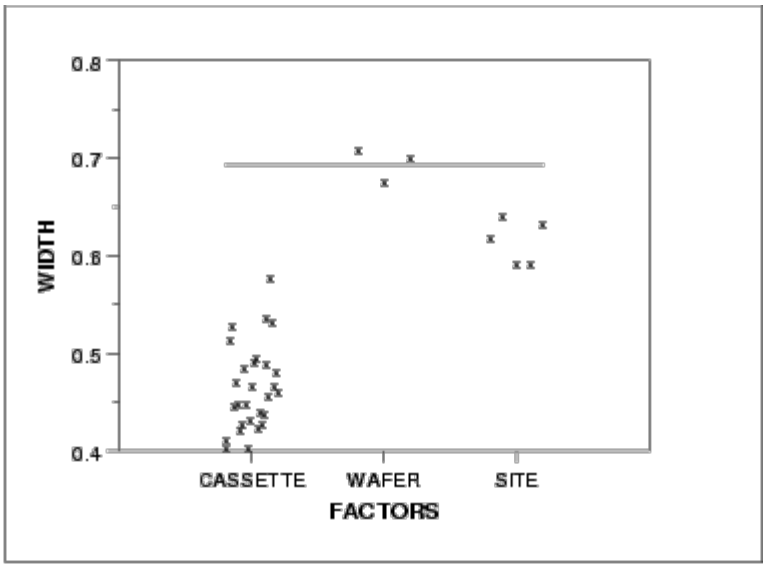
DOE mean and sd plots

We can use the DOE mean plot and the DOE standard deviation plot to show the factor means and standard deviations together for better comparison.

DOE mean plot



DOE sd plot



Summary

The above graphs show that there are differences between the lots and the sites.

There are various ways we can create subgroups of this dataset: each lot could be a subgroup, each wafer could be a subgroup, or each site measured could be a subgroup (with only one data value in each subgroup).

Recall that for a classical Shewhart means chart, the average within subgroup standard deviation is used to calculate the control limits for the means chart. However, with a means chart you are monitoring the subgroup mean-to-mean variation. There is no problem if you are in a continuous processing situation - this becomes an issue if you are operating in a batch processing environment.

We will look at various control charts based on different subgroupings in 6.6.1.3.

6.6.1.3. Subgroup Analysis

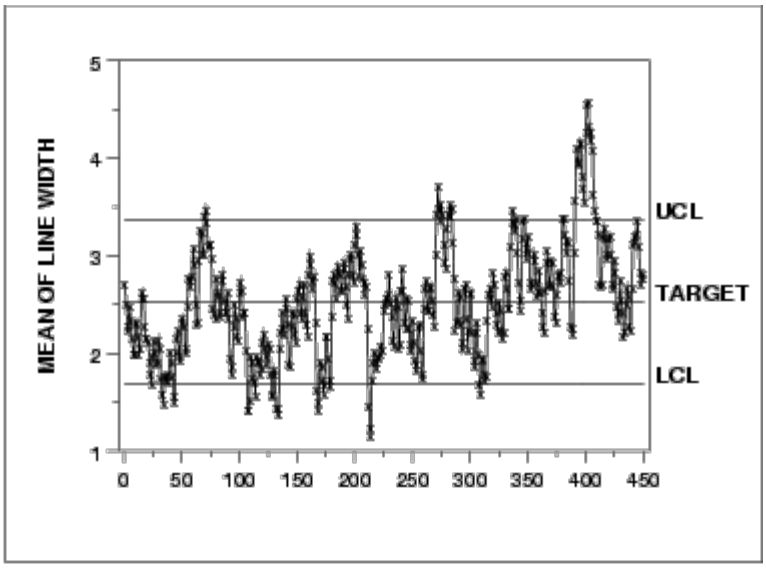
Control charts for subgroups

The resulting classical Shewhart control charts for each possible subgroup are shown below.

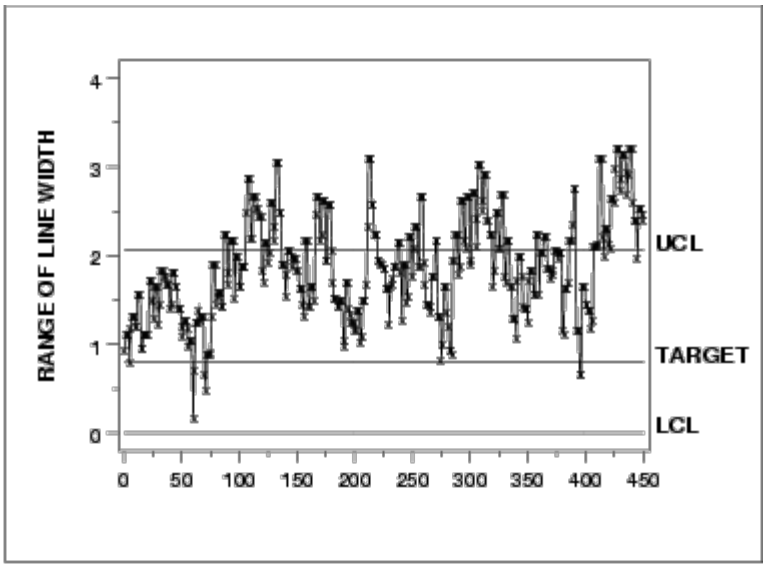
Site as subgroup

The first pair of control charts use the site as the subgroup. However, since site has a subgroup size of one we use the control charts for individual measurements. A moving average and a moving range chart are shown.

Moving average control chart



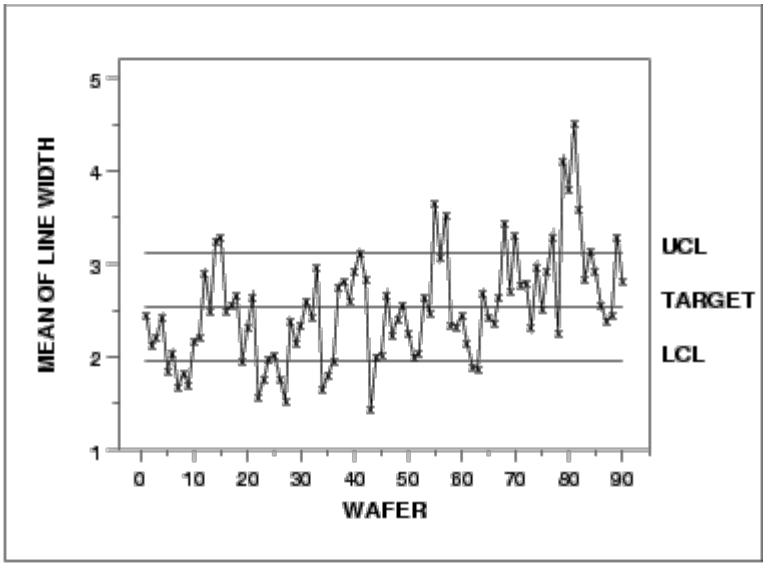
Moving range control chart



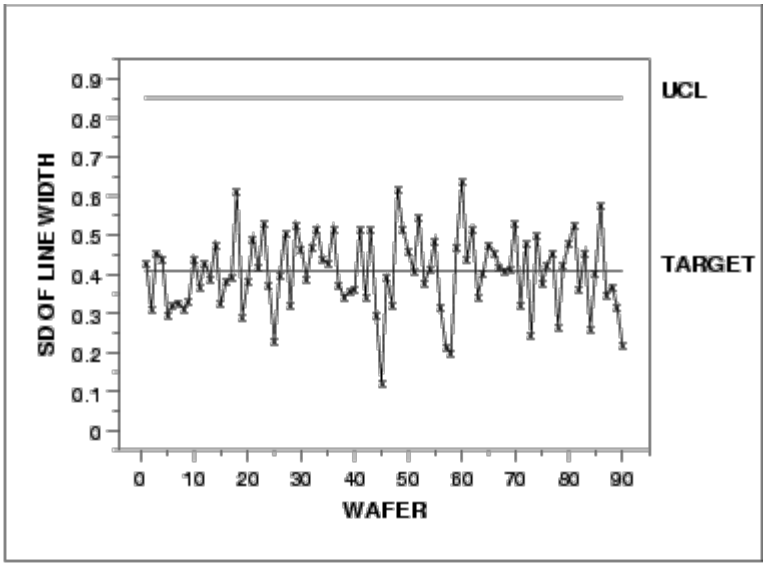
Wafer as subgroup

The next pair of control charts use the wafer as the subgroup. In this case, the subgroup size is five. A mean and a standard deviation control chart are shown.

Mean control chart



SD control chart

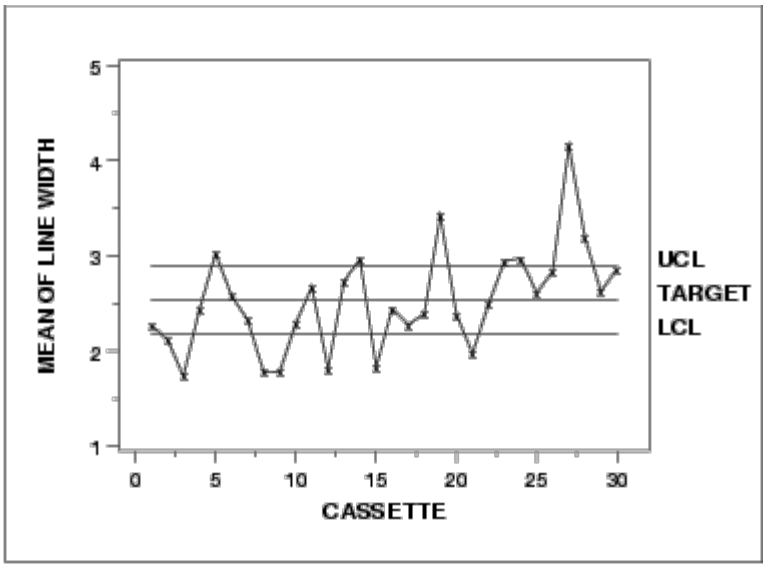


There is no LCL for the standard deviation chart because of the small subgroup size.

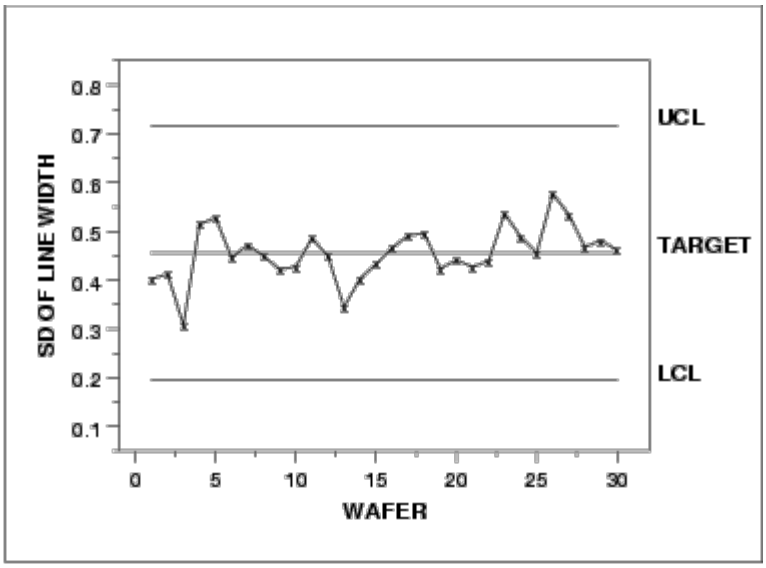
Cassette as subgroup

The next pair of control charts use the cassette as the subgroup. In this case, the subgroup size is 15. A mean and a standard deviation control chart are shown.

Mean control chart



SD control chart



Interpretation

Which of these subgroupings of the data is correct? As you can see, each sugrouping produces a different chart. Part of the answer lies in the manufacturing requirements for this process. Another aspect that can be statistically determined is the magnitude of each of the sources of variation. In order to understand our data structure and how much variation each of our sources contribute, we need to perform a variance component analysis. The variance component analysis for this data set is shown below.

Component	Variance Component Estimate
Cassette	0.2645
Wafer	0.0500
Site	0.1755

Variance Component Estimation

If your software does not generate the variance components directly, they can be computed from a standard analysis of variance output by equating mean squares (MS) to expected mean squares (EMS).

The sum of squares and mean squares for a nested, random effects model are shown below.

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Cassette	29	127.40293	4.3932
Wafer(Cassette)	60	25.52089	0.4253
Site(Cassette, Wafer)	360	63.17865	0.1755

The expected mean squares for cassette, wafer within cassette, and site within cassette and wafer, along with their associated mean squares, are the following.

```
4.3932 = (3*5)*Var(cassettes) + 5*Var(wafer) +
Var(site)
0.4253 = 5*Var(wafer) + Var(site)
0.1755 = Var(site)
```

Solving these equations, we obtain the variance component estimates 0.2645, 0.04997, and 0.1755 for cassettes, wafers, and sites, respectively.

All of the analyses in this section can be completed using R code.

6.6.1.4. Shewhart Control Chart

Choosing the right control charts to monitor the process

The largest source of variation in this data is the lot-to-lot variation. So, using classical Shewhart methods, if we specify our subgroup to be anything other than lot, we will be ignoring the known lot-to-lot variation and could get out-of-control points that already have a known, assignable cause - the data comes from different lots. However, in the lithography processing area the measurements of most interest are the site level measurements, not the lot means. How can we get around this seeming contradiction?

Chart sources of variation separately

One solution is to chart the important sources of variation separately. We would then be able to monitor the variation of our process and truly understand where the variation is coming from and if it changes. For this dataset, this approach would require having two sets of control charts, one for the individual site measurements and the other for the lot means. This would double the number of charts necessary for this process (we would have 4 charts for line width instead of 2).

Chart only most important source of variation

Another solution would be to have one chart on the largest source of variation. This would mean we would have one set of charts that monitor the lot-to-lot variation. From a manufacturing standpoint, this would be unacceptable.

Use boxplot type chart

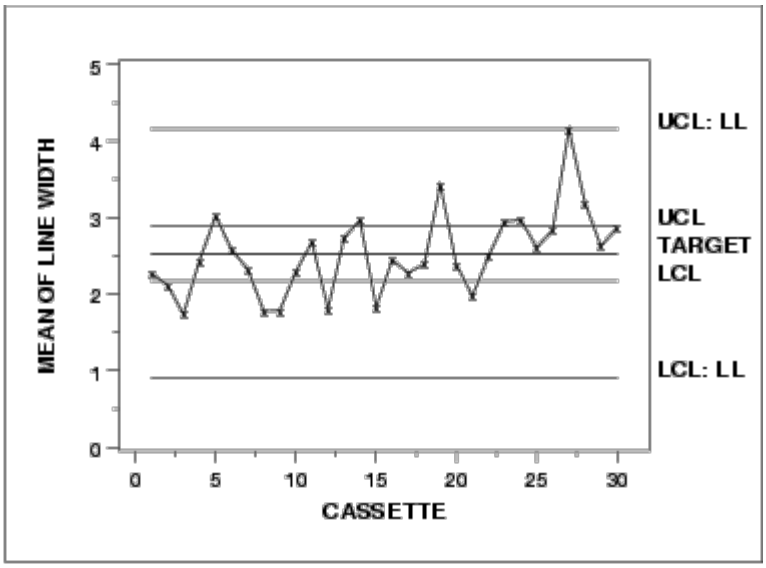
We could create a non-standard chart that would plot all the individual data values and group them together in a boxplot type format by lot. The control limits could be generated to monitor the individual data values while the lot-to-lot variation would be monitored by the patterns of the groupings. This would take special programming and management intervention to implement non-standard charts in most floor shop control systems.

Alternate form for mean control chart

A commonly applied solution is the first option; have multiple charts on this process. When creating the control limits for the lot means, care must be taken to use the lot-to-lot variation instead of the within lot variation. The resulting control charts are: the standard individuals/moving range charts (as seen previously), and a control chart on the lot means that is different from the previous lot means chart. This new chart

uses the lot-to-lot variation to calculate control limits instead of the average within-lot standard deviation. The accompanying standard deviation chart is the same as seen previously.

Mean control chart using lot-to-lot variation



The control limits labeled with "UCL" and "LCL" are the standard control limits. The control limits labeled with "UCL: LL" and "LCL: LL" are based on the lot-to-lot variation.

6.6.1.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output Window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 5 columns of numbers into Dataplot, variables CASSETTE, WAFER, SITE, WIDTH, and RUNSEQ.
2. Plot of the response variable 1. Numerical summary of WIDTH. 2. 4-Plot of WIDTH. 3. Run sequence plot of WIDTH.	1. The summary shows the mean line width is 2.53 and the standard deviation of the line width is 0.69. 2. The 4-plot shows non-constant location and scale and moderate

	autocorrelation. 3. The run sequence plot shows non-constant location and scale.
3. Generate scatter and box plots against individual factors. 1. Scatter plot of WIDTH versus CASSETTE. 2. Box plot of WIDTH versus CASSETTE. 3. Scatter plot of WIDTH versus WAFER. 4. Box plot of WIDTH versus WAFER. 5. Scatter plot of WIDTH versus SITE. 6. Box plot of WIDTH versus SITE. 7. DOE mean plot of WIDTH versus CASSETTE, WAFER, and SITE. 8. DOE sd plot of WIDTH versus CASSETTE, WAFER, and SITE.	1. The scatter plot shows considerable variation in location. 2. The box plot shows considerable variation in location and scale and the prescence of some outliers. 3. The scatter plot shows minimal variation in location and scale. 4. The box plot shows minimal variation in location and scale. It also show some outliers. 5. The scatter plot shows some variation in location. 6. The box plot shows some variation in location. Scale seems relatively constant. Some outliers. 7. The DOE mean plot shows effects for CASSETTE and SITE, no effect for WAFER. 8. The DOE sd plot shows effects for CASSETTE and SITE, no effect for WAFER.
4. Subgroup analysis. 1. Generate a moving mean control chart. 2. Generate a moving range control chart. 3. Generate a mean control chart for WAFER.	1. The moving mean plot shows a large number of out-of-control points. 2. The moving range plot shows a large number of out-of-

4. Generate a sd control chart for WAFER. 5. Generate a mean control chart for CASSETTE. 6. Generate a sd control chart for CASSETTE. 7. Generate an analysis of variance. This is not currently implemented in DATAPLOT for nested datasets. 8. Generate a mean control chart using lot-to-lot variation.	control points. 3. The mean control chart shows a large number of out-of-control points. 4. The sd control chart shows no out-of-control points. 5. The mean control chart shows a large number of out-of-control points. 6. The sd control chart shows no out-of-control points. 7. The analysis of variance and components of variance calculations show that cassette to cassette variation is 54% of the total and site to site variation is 36% of the total. 8. The mean control chart shows one point that is on the boundary of being out of control.

6.6.2. Aerosol Particle Size

Box-Jenkins Modeling of Aerosol Particle Size

This case study illustrates the use of Box-Jenkins modeling with aerosol particle size data.

1. Background and Data
2. Model Identification
3. Model Estimation
4. Model Validation
5. Work This Example Yourself

6.6.2.1. Background and Data

Data Source

The source of the data for this case study is Antuan Negiz who analyzed these data while he was a post-doc in the NIST Statistical Engineering Division from the Illinois Institute of Technology.

Data Collection

These data were collected from an aerosol mini-spray dryer device. The purpose of this device is to convert a slurry stream into deposited particles in a drying chamber. The device injects the slurry at high speed. The slurry is pulverized as it enters the drying chamber when it comes into contact with a hot gas stream at low humidity. The liquid contained in the pulverized slurry particles is vaporized, then transferred to the hot gas stream leaving behind dried small-sized particles.

The response variable is particle size, which is collected equidistant in time. There are a variety of associated variables that may affect the injection process itself and hence the size and quality of the deposited particles. For this case study, we restrict our analysis to the response variable.

Applications

Such deposition process operations have many applications from powdered laundry detergents at one extreme to ceramic molding at an important other extreme. In ceramic molding, the distribution and homogeneity of the particle sizes are particularly important because after the molds are baked and cured, the properties of the final molded ceramic product is strongly affected by the intermediate uniformity of the base ceramic particles, which in turn is directly reflective of the quality of the initial atomization process in the aerosol injection device.

Aerosol Particle Size Dynamic Modeling and Control

The data set consists of particle sizes collected over time. The basic distributional properties of this process are of interest in terms of distributional shape, constancy of size, and variation in size. In addition, this time series may be examined for autocorrelation structure to determine a prediction model of particle size as a function of time--such a model is frequently autoregressive in nature. Such a high-quality prediction equation would be essential as a first step in developing a predictor-corrective recursive feedback

mechanism which would serve as the core in developing and implementing real-time dynamic corrective algorithms. The net effect of such algorthms is, of course, a particle size distribution that is much less variable, much more stable in nature, and of much higher quality. All of this results in final ceramic mold products that are more uniform and predictable across a wide range of important performance characteristics.

For the purposes of this case study, we restrict the analysis to determining an appropriate Box-Jenkins model of the particle size.

Software

The analyses used in this case study can be generated using both Dataplot code and R code.

Case study data

```
115.36539
114.63150
114.63150
116.09940
116.34400
116.09940
116.34400
116.83331
116.34400
116.83331
117.32260
117.07800
117.32260
117.32260
117.81200
117.56730
118.30130
117.81200
118.30130
117.81200
118.30130
118.30130
118.54590
118.30130
117.07800
116.09940
118.30130
118.79060
118.05661
118.30130
118.54590
118.30130
118.54590
118.05661
118.30130
118.54590
118.30130
118.30130
118.30130
118.30130
118.30130
118.05661
118.30130
117.81200
118.30130
117.32260
117.32260
117.56730
117.81200
117.56730
117.81200
117.81200
117.32260
116.34400
116.58870
116.83331
116.58870
```

116.83331
116.83331
117.32260
116.34400
116.09940
115.61010
115.61010
115.61010
115.36539
115.12080
115.61010
115.85471
115.36539
115.36539
115.36539
115.12080
114.87611
114.87611
115.12080
114.87611
114.87611
114.63150
114.63150
114.14220
114.38680
114.14220
114.63150
114.87611
114.38680
114.87611
114.63150
114.14220
114.14220
113.89750
114.14220
113.89750
113.65289
113.65289
113.40820
113.40820
112.91890
113.40820
112.91890
113.40820
113.89750
113.40820
113.65289
113.89750
113.65289
113.65289
113.89750
113.65289
113.16360
114.14220
114.38680
113.65289
113.89750
113.89750
113.40820
113.65289
113.89750
113.65289
113.65289
114.14220
114.38680
114.63150
115.61010
115.12080
114.63150
114.38680
113.65289
113.40820
113.40820
113.16360
113.16360
113.16360
113.16360
113.16360
112.42960
113.40820
113.40820

113.16360
113.16360
113.16360
113.16360
111.20631
112.67420
112.91890
112.67420
112.91890
113.16360
112.91890
112.67420
112.91890
112.67420
112.91890
113.16360
112.67420
112.67420
112.91890
113.16360
112.67420
112.91890
111.20631
113.40820
112.91890
112.67420
113.16360
113.65289
113.40820
114.14220
114.87611
114.87611
116.09940
116.34400
116.58870
116.09940
116.34400
116.83331
117.07800
117.07800
116.58870
116.83331
116.58870
116.34400
116.83331
116.83331
117.07800
116.58870
116.58870
117.32260
116.83331
118.79060
116.83331
117.07800
116.58870
116.83331
116.34400
116.58870
116.34400
116.34400
116.34400
116.09940
116.09940
116.34400
115.85471
115.85471
115.85471
115.61010
115.61010
115.61010
115.36539
115.12080
115.61010
115.85471
115.12080
115.12080
114.87611
114.87611
114.38680
114.14220
114.14220

114.38680
114.14220
114.38680
114.38680
114.38680
114.38680
114.38680
114.14220
113.89750
114.14220
113.65289
113.16360
112.91890
112.67420
112.42960
112.42960
112.42960
112.18491
112.18491
112.42960
112.18491
112.42960
111.69560
112.42960
112.42960
111.69560
111.94030
112.18491
112.18491
112.18491
111.94030
111.69560
111.94030
111.94030
112.42960
112.18491
112.18491
111.94030
112.18491
112.18491
111.20631
111.69560
111.69560
111.69560
111.94030
111.94030
112.18491
111.69560
112.18491
111.94030
111.69560
112.18491
110.96170
111.69560
111.20631
111.20631
111.45100
110.22771
109.98310
110.22771
110.71700
110.22771
111.20631
111.45100
111.69560
112.18491
112.18491
112.18491
112.42960
112.67420
112.18491
112.42960
112.18491
112.91890
112.18491
112.42960
111.20631
112.42960
112.42960
112.42960
112.42960

```
113.16360
112.18491
112.91890
112.91890
112.67420
112.42960
112.42960
112.42960
112.91890
113.16360
112.67420
113.16360
112.91890
112.42960
112.67420
112.91890
112.18491
112.91890
113.16360
112.91890
112.91890
112.91890
112.67420
112.42960
112.42960
113.16360
112.91890
112.67420
113.16360
112.91890
113.16360
112.91890
112.67420
112.91890
112.67420
112.91890
112.91890
112.91890
113.16360
112.91890
112.91890
112.18491
112.42960
112.42960
112.18491
112.91890
112.67420
112.42960
112.42960
112.18491
112.42960
112.67420
112.42960
112.42960
112.18491
112.67420
112.42960
112.42960
112.67420
112.42960
112.42960
112.42960
112.67420
112.91890
113.40820
113.40820
113.40820
112.91890
112.67420
112.67420
112.91890
113.65289
113.89750
114.38680
114.87611
114.87611
115.12080
115.61010
115.36539
115.61010
115.85471
```

116.09940
116.83331
116.34400
116.58870
116.58870
116.34400
116.83331
116.83331
116.83331
117.32260
116.83331
117.32260
117.56730
117.32260
117.07800
117.32260
117.81200
117.81200
117.81200
118.54590
118.05661
118.05661
117.56730
117.32260
117.81200
118.30130
118.05661
118.54590
118.05661
118.30130
118.05661
118.30130
118.30130
118.30130
118.05661
117.81200
117.32260
118.30130
118.30130
117.81200
117.07800
118.05661
117.81200
117.56730
117.32260
117.32260
117.81200
117.32260
117.81200
117.07800
117.32260
116.83331
117.07800
116.83331
116.83331
117.07800
115.12080
116.58870
116.58870
116.34400
115.85471
116.34400
116.34400
115.85471
116.58870
116.34400
115.61010
115.85471
115.61010
115.85471
115.12080
115.61010
115.61010
115.85471
115.61010
115.36539
114.87611
114.87611
114.63150
114.87611
115.12080

```
114.63150
114.87611
115.12080
114.63150
114.38680
114.38680
114.87611
114.63150
114.63150
114.63150
114.63150
114.63150
114.14220
113.65289
113.65289
113.89750
113.65289
113.40820
113.40820
113.89750
113.89750
113.89750
113.65289
113.65289
113.89750
113.40820
113.40820
113.65289
113.89750
113.89750
114.14220
113.65289
113.40820
113.40820
113.65289
113.40820
114.14220
113.89750
114.14220
113.65289
113.65289
113.65289
113.89750
113.16360
113.16360
113.89750
113.65289
113.16360
113.65289
113.40820
112.91890
113.16360
113.16360
113.40820
113.40820
113.65289
113.16360
113.40820
113.16360
113.16360
112.91890
112.91890
112.91890
113.65289
113.65289
113.16360
112.91890
112.67420
113.16360
112.91890
112.67420
112.91890
112.91890
112.91890
111.20631
112.91890
113.16360
112.42960
112.67420
113.16360
112.42960
```

```
112.67420
112.91890
112.67420
111.20631
112.42960
112.67420
112.42960
113.16360
112.91890
112.67420
112.91890
112.42960
112.67420
112.18491
112.91890
112.42960
112.18491
```

6.6.2.2. Model Identification

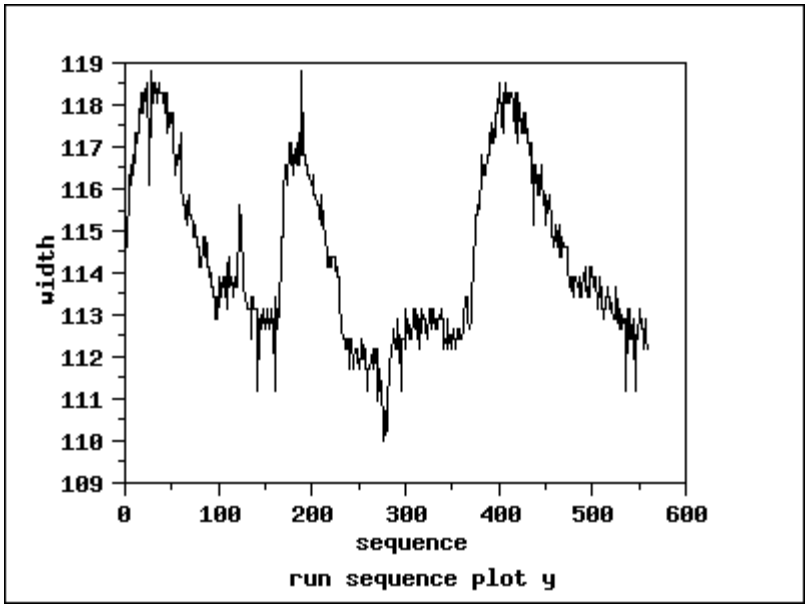
Check for Stationarity, Outliers, Seasonality

The first step in the analysis is to generate a run sequence plot of the response variable. A run sequence plot can indicate stationarity (i.e., constant location and scale), the presence of outliers, and seasonal patterns.

Non-stationarity can often be removed by differencing the data or fitting some type of trend curve. We would then attempt to fit a Box-Jenkins model to the differenced data or to the residuals after fitting a trend curve.

Although Box-Jenkins models can estimate seasonal components, the analyst needs to specify the seasonal period (for example, 12 for monthly data). Seasonal components are common for economic time series. They are less common for engineering and scientific data.

Run Sequence Plot



Interpretation of the Run Sequence Plot

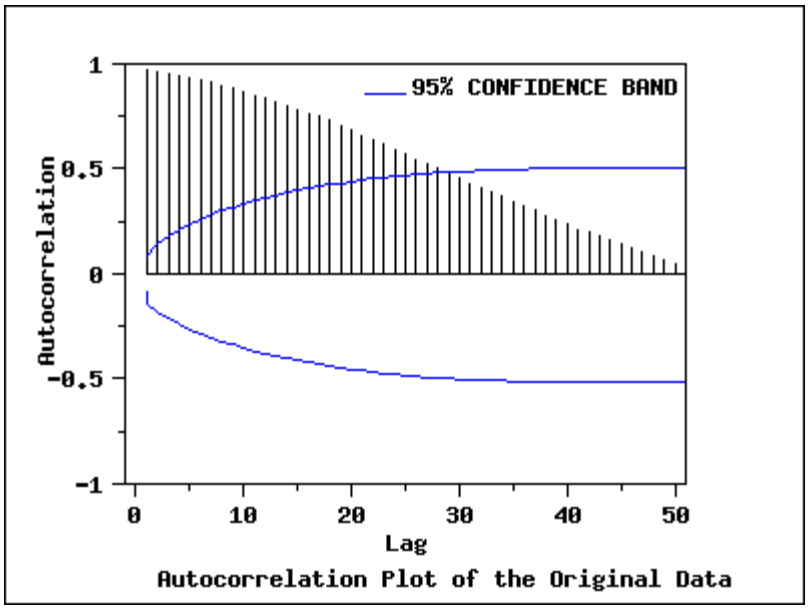
We can make the following conclusions from the run sequence plot.

1. The data show strong and positive autocorrelation.

2. There does not seem to be a significant trend or any obvious seasonal pattern in the data.

The next step is to examine the sample autocorrelations using the autocorrelation plot.

Autocorrelation Plot

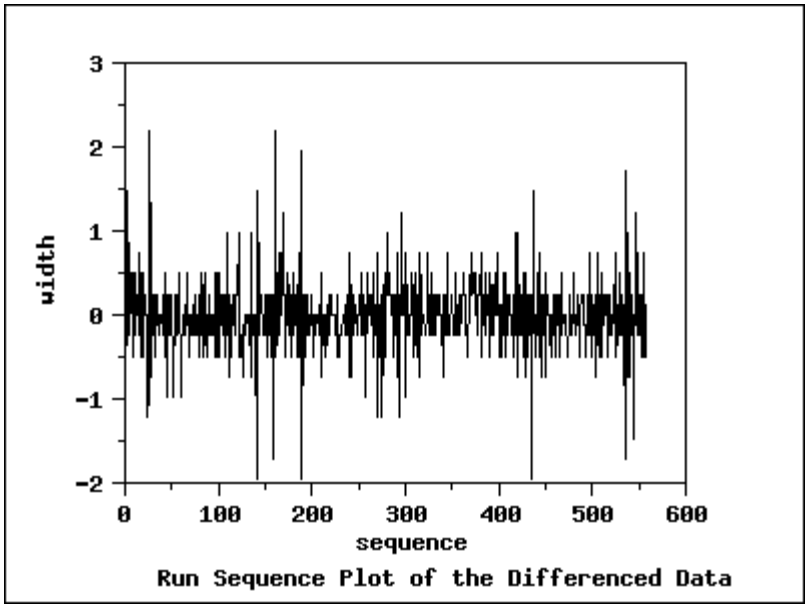


Interpretation of the Autocorrelation Plot

The autocorrelation plot has a 95% confidence band, which is constructed based on the assumption that the process is a moving average process. The autocorrelation plot shows that the sample autocorrelations are very strong and positive and decay very slowly.

The autocorrelation plot indicates that the process is non-stationary and suggests an ARIMA model. The next step is to difference the data.

Run Sequence Plot of Differenced Data



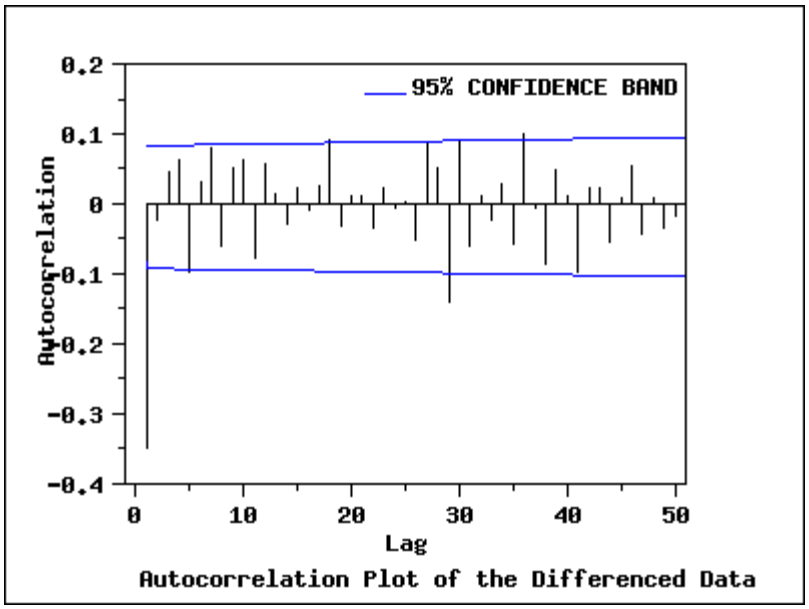
Interpretation of the Run Sequence Plot

The run sequence plot of the differenced data shows that the mean of the differenced data is around zero, with the differenced data less autocorrelated than the original data.

The next step is to examine the sample autocorrelations of

the differenced data.

Autocorrelation Plot of the Differenced Data

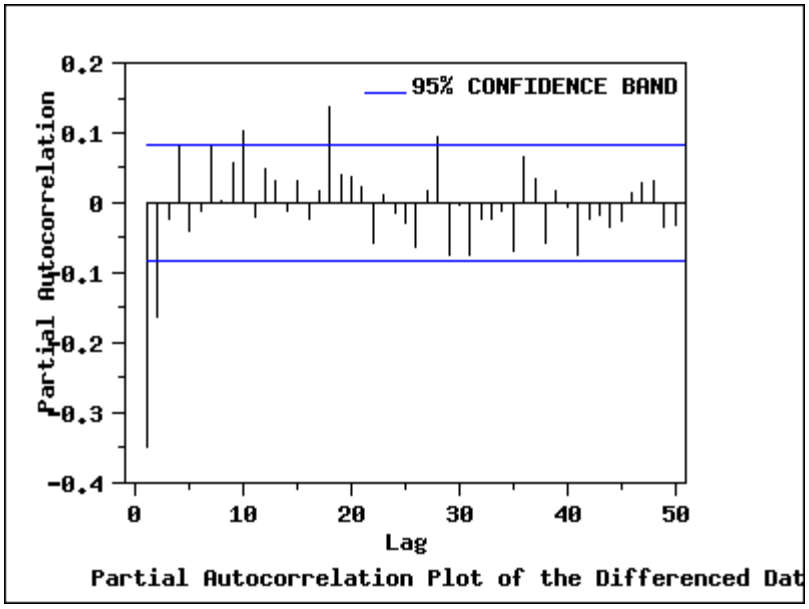


Interpretation of the Autocorrelation Plot of the Differenced Data

The autocorrelation plot of the differenced data with a 95% confidence band shows that only the autocorrelation at lag 1 is significant. The autocorrelation plot together with run sequence of the differenced data suggest that the differenced data are stationary. Based on the autocorrelation plot, an MA(1) model is suggested for the differenced data.

To examine other possible models, we produce the partial autocorrelation plot of the differenced data.

Partial Autocorrelation Plot of the Differenced Data



Interpretation of the Partial Autocorrelation Plot of the Differenced Data

The partial autocorrelation plot of the differenced data with 95% confidence bands shows that only the partial autocorrelations of the first and second lag are significant. This suggests an AR(2) model for the differenced data.

Akaike Information Criterion (AIC and AICC)	Information-based criteria, such as the AIC or AICC (see Brockwell and Davis (2002), pp. 171-174), can be used to automate the choice of an appropriate model. Many software programs for time series analysis will generate the AIC or AICC for a broad range of models.

Whatever method is used for model identification, model diagnostics should be performed on the selected model. Based on the plots in this section, we will examine the ARIMA(2,1,0) and ARIMA(0,1,1) models in detail.

6.6.2.3. Model Estimation

AR(2) Model Parameter Estimates

The following parameter estimates were computed for the AR(2) model based on the differenced data.

Source	Parameter Estimate	Standard Error	95 % Confidence Interval
Intercept	-0.0050	0.0119	
AR1	-0.4064	0.0419	(-0.4884, -0.3243)
AR2	-0.1649	0.0419	(-0.2469, -0.0829)

Number of Observations: 558
Degrees of Freedom: 558 - 3 = 555
Residual Standard Deviation: 0.4423

Both AR parameters are significant since the confidence intervals do not contain zero.

The model for the differenced data, Y_t, is an AR(2) model:

$$Y_t = -0.4064Y_{t-1} - 0.1649Y_{t-2} - 0.0050$$

with σ = 0.4423.

It is often more convenient to express the model in terms of the original data, X_t, rather than the differenced data. From the definition of the difference, $Y_t = X_t - X_{t-1}$, we can make the appropriate substitutions into the above equation:

$$X_t - X_{t-1} = -0.4064(X_{t-1} - X_{t-2}) - 0.1649(X_{t-2} - X_{t-3}) - 0.0050$$

to arrive at the model in terms of the original series:

$$X_t = 0.5936X_{t-1} + 0.2415X_{t-2} + 0.1649X_{t-3} - 0.0050$$

MA(1) Model Parameter Estimates

Alternatively, the parameter estimates for an MA(1) model based on the differenced data are the following.

Source	Parameter Estimate	Standard Error	95 % Confidence Interval
Intercept	-0.0051	0.0114	
MA1	-0.3921	0.0366	(-0.4638, -0.3205)

Number of Observations: 558
Degrees of Freedom: 558 - 2 = 556
Residual Standard Deviation: 0.4434

The model for the differenced data, Y_t, is an ARIMA(0,1,1) model:

$$Y_t = a_t - 0.3921a_{t-1} - 0.0051$$

with σ = 0.4434.

It is often more convenient to express the model in terms of the original data, X_t, rather than the differenced data. Making the appropriate substitutions into the above equation:

$$X_t - X_{t-1} = a_t - 0.3921a_{t-1} - 0.0051$$

we arrive at the model in terms of the original series:

$$X_t = X_{t-1} + a_t - 0.3921a_{t-1} - 0.0051$$

6.6.2.4. Model Validation

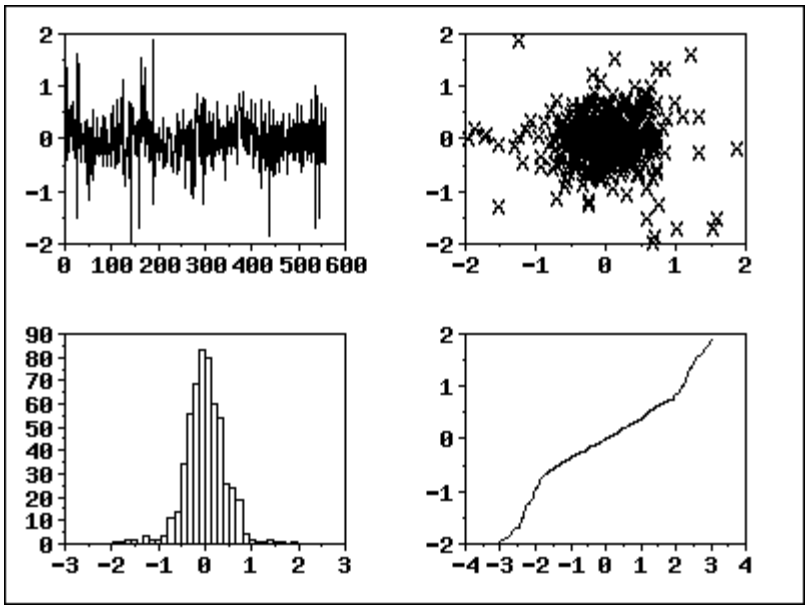
Residuals

After fitting the model, we should check whether the model is appropriate.

As with standard non-linear least squares fitting, the primary tool for model diagnostic checking is residual analysis.

4-Plot of Residuals from ARIMA(2,1,0) Model

The 4-plot is a convenient graphical technique for model validation in that it tests the assumptions for the residuals on a single graph.



Interpretation of the 4-Plot

We can make the following conclusions based on the above 4-plot.

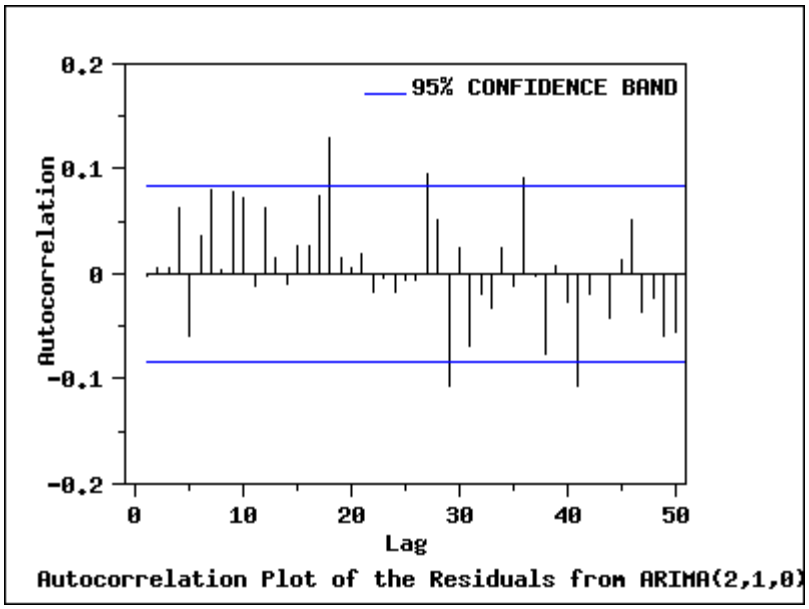
1. The run sequence plot shows that the residuals do not violate the assumption of constant location and scale. It also shows that most of the residuals are in the range (-1, 1).

2. The lag plot indicates that the residuals are not autocorrelated at lag 1.

3. The histogram and normal probability plot indicate that the normal distribution provides an adequate fit for this model.

Autocorrelation Plot of Residuals from ARIMA(2,1,0) Model

In addition, the autocorrelation plot of the residuals from the ARIMA(2,1,0) model was generated.



Interpretation of the Autocorrelation Plot

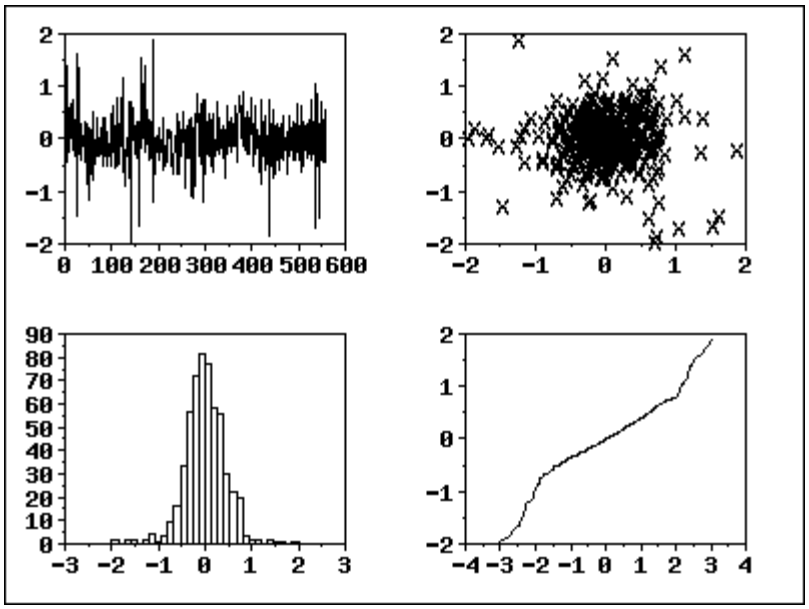
The autocorrelation plot shows that for the first 25 lags, all sample autocorrelations except those at lags 7 and 18 fall inside the 95 % confidence bounds indicating the residuals appear to be random.

Test the Randomness of Residuals From the ARIMA(2,1,0) Model Fit

We apply the Box-Ljung test to the residuals from the ARIMA(2,1,0) model fit to determine whether residuals are random. In this example, the Box-Ljung test shows that the first 24 lag autocorrelations among the residuals are zero (*p*-value = 0.080), indicating that the residuals are random and that the model provides an adequate fit to the data.

4-Plot of Residuals from ARIMA(0,1,1) Model

The 4-plot is a convenient graphical technique for model validation in that it tests the assumptions for the residuals on a single graph.



Interpretation of the 4-Plot from the ARIMA(0,1,1) Model

We can make the following conclusions based on the above 4-plot.

1. The run sequence plot shows that the residuals do not violate the assumption of constant location and scale. It also shows that most of the residuals are in the range (-1, 1).

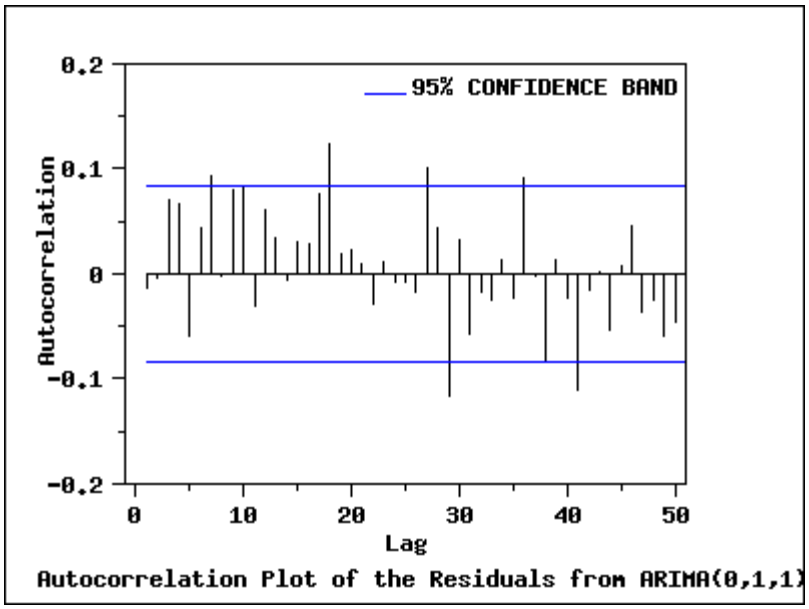
2. The lag plot indicates that the residuals are not autocorrelated at lag 1.

3. The histogram and normal probability plot indicate that the normal distribution provides an adequate fit for this model.

This 4-plot of the residuals indicates that the fitted model is adequate for the data.

Autocorrelation Plot of Residuals from ARIMA(0,1,1) Model

The autocorrelation plot of the residuals from ARIMA(0,1,1) was generated.



Interpretation of the Autocorrelation Plot	Similar to the result for the ARIMA(2,1,0) model, it shows that for the first 25 lags, all sample autocorrelations expect those at lags 7 and 18 fall inside the 95% confidence bounds indicating the residuals appear to be random.
Test the Randomness of Residuals From the ARIMA(0,1,1) Model Fit	The Box-Ljung test is also applied to the residuals from the ARIMA(0,1,1) model. The test indicates that there is at least one non-zero autocorrelation amont the first 24 lags. We conclude that there is not enough evidence to claim that the residuals are random (p-value = 0.026).
Summary	Overall, the ARIMA(0,1,1) is an adequate model. However, the ARIMA(2,1,0) is a little better than the ARIMA(0,1,1).

6.6.2.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output Window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read one column of numbers into Dataplot, variable Y.
2. Model identification plots 1. Run sequence plot of Y. 2. Autocorrelation plot of Y. 3. Run sequence plot of the differenced data of Y.	1. The run sequence plot shows that the data show strong and positive autocorrelation. 2. The autocorrelation plot indicates significant autocorrelation and that the data are not stationary.

4. Autocorrelation plot of the differenced data of Y. 5. Partial autocorrelation plot of the differenced data of Y.	3. The run sequence plot shows that the differenced data appear to be stationary and do not exhibit seasonality. 4. The autocorrelation plot of the differenced data suggests an ARIMA(0,1,1) model may be appropriate. 5. The partial autocorrelation plot suggests an ARIMA(2,1,0) model may be appropriate.
3. Estimate the model. 1. ARIMA(2,1,0) fit of Y. 2. ARIMA(0,1,1) fit of Y.	1. The ARMA fit generates parameter estimates for the ARIMA(2,1,0) model. 2. The ARMA fit generates parameter estimates for the ARIMA(0,1,1) model.
4. Model validation. 1. Generate a 4-plot of the residuals from the ARIMA(2,1,0) model. 2. Generate an autocorrelation plot of the residuals from the ARIMA(2,1,0) model. 3. Perform a Ljung-Box test of randomness for the residuals from the ARIMA(2,1,0) model. 4. Generate a 4-plot of the residuals from the ARIMA(0,1,1) model. 5. Generate an autocorrelation plot of the residuals from the ARIMA(0,1,1) model. 6. Perform a Ljung-Box test of randomness for the residuals from the ARIMA(0,1,1) model.	1. The 4-plot shows that the assumptions for the residuals are satisfied. 2. The autocorrelation plot of the residuals indicates that the residuals are random. 3. The Ljung-Box test indicates that the residuals are random. 4. The 4-plot shows that the assumptions for the residuals are satisfied. 5. The

	autocorrelation plot of the residuals indicates that the residuals are random. 6. The Ljung-Box test indicates that the residuals are not random at the 95% level, but are random at the 99% level.

6.7. References

Selected References

Time Series Analysis

Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*, Wiley, New York, NY.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis, Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Clifs, NJ.

Box, G. E. P. and McGregor, J. F. (1974). "The Analysis of Closed-Loop Dynamic Stochastic Systems", *Technometrics*, Vol. 16-3.

Brockwell, Peter J. and Davis, Richard A. (1987). *Time Series: Theory and Methods*, Springer-Verlang.

Brockwell, Peter J. and Davis, Richard A. (2002). *Introduction to Time Series and Forecasting*, 2nd. ed., Springer-Verlang.

Chatfield, C. (1996). *The Analysis of Time Series*, 5th ed., Chapman & Hall, New York, NY.

DeLurgio, S. A. (1998). *Forecasting Principles and Applications*, Irwin McGraw-Hill, Boston, MA.

Ljung, G. and Box, G. (1978). "On a Measure of Lack of Fit in Time Series Models", *Biometrika*, 65, 297-303.

Nelson, C. R. (1973). *Applied Time Series Analysis for Managerial Forecasting*, Holden-Day, Boca-Raton, FL.

Makradakis, S., Wheelwright, S. C. and McGhee, V. E. (1983). *Forecasting: Methods and Applications*, 2nd ed., Wiley, New York, NY.

Statistical Process and Quality Control

Army Chemical Corps (1953). *Master Sampling Plans for Single, Duplicate, Double and Multiple Sampling*, Manual No. 2.

Bissell, A. F. (1990). "How Reliable is Your Capability Index?", Applied Statistics, 39, 331-340.

Champ, C.W., and Woodall, W.H. (1987). "Exact Results for Shewhart

Control Charts with Supplementary Runs Rules", *Technometrics*, 29, 393-399.

Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 5th ed., Irwin, Homewood, IL.

Hotelling, H. (1947). Multivariate Quality Control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. *Techniques of Statistical Analysis*. New York: McGraw-Hill.

Juran, J. M. (1997). "Early SQC: A Historical Supplement", *Quality Progress*, 30(9) 73-81.

Montgomery, D. C. (2000). *Introduction to Statistical Quality Control*, 4th ed., Wiley, New York, NY.

Kotz, S. and Johnson, N. L. (1992). *Process Capability Indices*, Chapman & Hall, London.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). "A Multivariate Exponentially Weighted Moving Average Chart", *Technometrics*, 34, 46-53.

Lucas, J. M. and Saccucci, M. S. (1990). "Exponentially weighted moving average control schemes: Properties and enhancements", *Technometrics 32*, 1-29.

Ott, E. R. and Schilling, E. G. (1990). *Process Quality Control*, 2nd ed., McGraw-Hill, New York, NY.

Quesenberry, C. P. (1993). "The effect of sample size on estimated limits for $\overline{\mathbf{X}}$ and **X** control charts", *Journal of Quality Technology*, 25(4) 237-247.

Ryan, T.P. (2000). *Statistical Methods for Quality Improvement*, 2nd ed., Wiley, New York, NY.

Ryan, T. P. and Schwertman, N. C. (1997). "Optimal limits for attributes control charts", *Journal of Quality Technology*, 29 (1), 86-98.

Schilling, E. G. (1982). *Acceptance Sampling in Quality Control*, Marcel Dekker, New York, NY.

Tracy, N. D., Young, J. C. and Mason, R. L. (1992). "Multivariate Control Charts for Individual Observations", *Journal of Quality Technology*, 24(2), 88-95.

Woodall, W. H. (1997). "Control Charting Based on Attribute Data: Bibliography and Review", *Journal of Quality Technology*, 29, 172-183.

Woodall, W. H., and Adams, B. M. (1993); "The Statistical Design of CUSUM Charts", *Quality Engineering*, 5(4), 559-570.

Zhang, Stenback, and Wardrop (1990). "Interval Estimation of the Process Capability Index", *Communications in Statistics: Theory and Methods,* 19(21), 4455-4470.

Statistical Analysis

Anderson, T. W. (1984). *Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley New York, NY.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*, Fourth Ed., Prentice Hall, Upper Saddle River, NJ.

7. Product and Process Comparisons

This chapter presents the background and specific analysis techniques needed to compare the performance of one or more processes against known standards or one another.

1. Introduction

1. Scope
2. Assumptions
3. Statistical Tests
4. Confidence Intervals
5. Equivalence of Tests and Intervals
6. Outliers
7. Trends

2. Comparisons: One Process

1. Comparing to a Distribution
2. Comparing to a Nominal Mean
3. Comparing to Nominal Variability
4. Fraction Defective
5. Defect Density
6. Location of Population Values

3. Comparisons: Two Processes

1. Means: Normal Data
2. Variability: Normal Data
3. Fraction Defective
4. Failure Rates
5. Means: General Case

4. Comparisons: Three + Processes

1. Comparing Populations
2. Comparing Variances
3. Comparing Means
4. Variance Components
5. Comparing Categorical Datasets
6. Comparing Fraction Defectives
7. Multiple Comparisons

Detailed table of contents
References for Chapter 7

7. Product and Process Comparisons - Detailed Table of Contents [7.]

1. Introduction [7.1.]
    1. What is the scope? [7.1.1.]
    2. What assumptions are typically made? [7.1.2.]
    3. What are statistical tests? [7.1.3.]
        1. Critical values and p values [7.1.3.1.]
    4. What are confidence intervals? [7.1.4.]
    5. What is the relationship between a test and a confidence interval? [7.1.5.]
    6. What are outliers in the data? [7.1.6.]
    7. What are trends in sequential process or product data? [7.1.7.]

2. Comparisons based on data from one process [7.2.]
    1. Do the observations come from a particular distribution? [7.2.1.]
        1. Chi-square goodness-of-fit test [7.2.1.1.]
        2. Kolmogorov- Smirnov test [7.2.1.2.]
        3. Anderson-Darling and Shapiro-Wilk tests [7.2.1.3.]
    2. Are the data consistent with the assumed process mean? [7.2.2.]
        1. Confidence interval approach [7.2.2.1.]
        2. Sample sizes required [7.2.2.2.]
    3. Are the data consistent with a nominal standard deviation? [7.2.3.]
        1. Confidence interval approach [7.2.3.1.]
        2. Sample sizes required [7.2.3.2.]
    4. Does the proportion of defectives meet requirements? [7.2.4.]
        1. Confidence intervals [7.2.4.1.]
        2. Sample sizes required [7.2.4.2.]
    5. Does the defect density meet requirements? [7.2.5.]
    6. What intervals contain a fixed percentage of the population values? [7.2.6.]
        1. Approximate intervals that contain most of the population values [7.2.6.1.]
        2. Percentiles [7.2.6.2.]
        3. Tolerance intervals for a normal distribution [7.2.6.3.]
        4. Tolerance intervals based on the largest and smallest observations [7.2.6.4.]

3. Comparisons based on data from two processes [7.3.]
    1. Do two processes have the same mean? [7.3.1.]
        1. Analysis of paired observations [7.3.1.1.]
        2. Confidence intervals for differences between means [7.3.1.2.]
    2. Do two processes have the same standard deviation? [7.3.2.]
    3. How can we determine whether two processes produce the same proportion of defectives? [7.3.3.]
    4. Assuming the observations are failure times, are the failure rates (or Mean Times To Failure) for two distributions the same? [7.3.4.]
    5. Do two arbitrary processes have the same central tendency? [7.3.5.]

4. Comparisons based on data from more than two processes [7.4.]

1. How can we compare several populations with unknown distributions (the Kruskal-Wallis test)? [7.4.1.]
2. Assuming the observations are normal, do the processes have the same variance? [7.4.2.]
3. Are the means equal? [7.4.3.]
    1. 1-Way ANOVA overview [7.4.3.1.]
    2. The 1-way ANOVA model and assumptions [7.4.3.2.]
    3. The ANOVA table and tests of hypotheses about means [7.4.3.3.]
    4. 1-Way ANOVA calculations [7.4.3.4.]
    5. Confidence intervals for the difference of treatment means [7.4.3.5.]
    6. Assessing the response from any factor combination [7.4.3.6.]
    7. The two-way ANOVA [7.4.3.7.]
    8. Models and calculations for the two-way ANOVA [7.4.3.8.]
4. What are variance components? [7.4.4.]
5. How can we compare the results of classifying according to several categories? [7.4.5.]
6. Do all the processes have the same proportion of defects? [7.4.6.]
7. How can we make multiple comparisons? [7.4.7.]
    1. Tukey's method [7.4.7.1.]
    2. Scheffe's method [7.4.7.2.]
    3. Bonferroni's method [7.4.7.3.]
    4. Comparing multiple proportions: The Marascuillo procedure [7.4.7.4.]

5. References [7.5.]

7. Product and Process Comparisons

7.1. Introduction

Goals of this section

The primary goal of this section is to lay a foundation for understanding statistical tests and confidence intervals that are useful for making decisions about processes and comparisons among processes. The materials covered are:

- Scope
- Assumptions
- Introduction to hypothesis testing
- Introduction to confidence intervals
- Relationship between hypothesis testing and confidence intervals
- Outlier detection
- Detection of sequential trends in data or processes

Hypothesis testing and confidence intervals

This chapter explores the types of comparisons which can be made from data and explains hypothesis testing, confidence intervals, and the interpretation of each.

7. Product and Process Comparisons
7.1. Introduction

7.1.1. What is the scope?

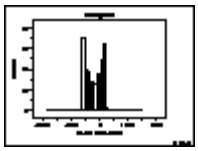
Data from one process	This section deals with introductory material related to comparisons that can be made on data from one process for cases where the process standard deviation may be known or unknown.

7.1.2. What assumptions are typically made?

Validity of tests

The validity of the tests described in this chapter depend on the following assumptions:

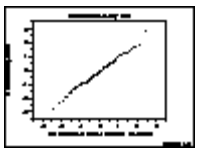
1. The data come from a single process that can be represented by a single statistical distribution.
2. The distribution is a normal distribution.
3. The data are uncorrelated over time.



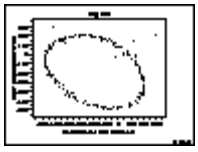
An easy method for checking the assumption of a single normal distribution is to construct a histogram of the data.

Clarification

The tests described in this chapter depend on the assumption of normality, and the data should be examined for departures from normality before the tests are applied. However, the tests are robust to small departures from normality; i.e., they work fairly well as long as the data are bell-shaped and the tails are not heavy. Quantitative methods for checking the normality assumption are discussed in the next section.



Another graphical method for testing the normality assumption is the normal probability plot.



A graphical method for testing for correlation among measurements is a time-lag plot. Correlation may not be a problem if measurements are properly structured over time. Correlation problems often occur when measurements are made close together in time.

7.1.3. What are statistical tests?

What is meant by a statistical test?

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. Not rejecting may be a good result if we want to continue to act as if we "believe" the null hypothesis is true. Or it may be a disappointing result, possibly indicating we may not yet have enough data to "prove" something by rejecting the null hypothesis.

For more discussion about the meaning of a statistical hypothesis test, see Chapter 1.

Concept of null hypothesis

A classic use of a statistical test occurs in process control studies. For example, suppose that we are interested in ensuring that photomasks in a production process have mean linewidths of 500 micrometers. The null hypothesis, in this case, is that the mean linewidth is 500 micrometers. Implicit in this statement is the need to flag photomasks which have mean linewidths that are either much greater or much less than 500 micrometers. This translates into the alternative hypothesis that the mean linewidths are not equal to 500 micrometers. This is a two-sided alternative because it guards against alternatives in opposite directions; namely, that the linewidths are too small or too large.

The testing procedure works this way. Linewidths at random positions on the photomask are measured using a scanning electron microscope. A test statistic is computed from the data and tested against pre-determined upper and lower critical values. If the test statistic is greater than the upper critical value or less than the lower critical value, the null hypothesis is rejected because there is evidence that the mean linewidth is not 500 micrometers.

One-sided tests of hypothesis

Null and alternative hypotheses can also be one-sided. For example, to ensure that a lot of light bulbs has a mean lifetime of at least 500 hours, a testing program is implemented. The null hypothesis, in this case, is that the mean lifetime is greater than or equal to 500 hours. The complement or alternative hypothesis that is being guarded against is that the mean lifetime is less than 500 hours. The test statistic is compared with a lower critical value, and if it

is less than this limit, the null hypothesis is rejected.

Thus, a statistical test requires a pair of hypotheses; namely,

- H_0: a null hypothesis
- H_a: an alternative hypothesis.

Significance levels

The null hypothesis is a statement about a belief. We may doubt that the null hypothesis is true, which might be why we are "testing" it. The alternative hypothesis might, in fact, be what we believe to be true. The test procedure is constructed so that the risk of rejecting the null hypothesis, when it is in fact true, is small. This risk, α, is often referred to as the *significance level* of the test. By having a test with a small value of α, we feel that we have actually "proved" something when we reject the null hypothesis.

Errors of the second kind

The risk of failing to reject the null hypothesis when it is in fact false is not chosen by the user but is determined, as one might expect, by the magnitude of the real discrepancy. This risk, β, is usually referred to as the *error of the second kind*. Large discrepancies between reality and the null hypothesis are easier to detect and lead to small errors of the second kind; while small discrepancies are more difficult to detect and lead to large errors of the second kind. Also the risk β increases as the risk α decreases. The risks of errors of the second kind are usually summarized by an *operating characteristic curve (OC)* for the test. OC curves for several types of tests are shown in (Natrella, 1962).

Guidance in this chapter

This chapter gives methods for constructing test statistics and their corresponding critical values for both one-sided and two-sided tests for the specific situations outlined under the scope. It also provides guidance on the sample sizes required for these tests.

Further guidance on statistical hypothesis testing, significance levels and critical regions, is given in Chapter 1.

7.1.3.1. Critical values and p values

Determination of critical values

Critical values for a test of hypothesis depend upon a test statistic, which is specific to the type of test, and the significance level, α, which defines the sensitivity of the test. A value of α = 0.05 implies that the null hypothesis is rejected 5% of the time when it is in fact true. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are common. Critical values are essentially cut-off values that define regions where the test statistic is unlikely to lie; for example, a region where the critical value is exceeded with probability α if the null hypothesis is true. The null hypothesis is rejected if the test statistic lies within this region which is often referred to as the rejection region(s). Critical values for specific tests of hypothesis are tabled in chapter 1.

Information in this chapter

This chapter gives formulas for the test statistics and points to the appropriate tables of critical values for tests of hypothesis regarding means, standard deviations, and proportion defectives.

P values

Another quantitative measure for reporting the result of a test of hypothesis is the *p*-value. The *p*-value is the probability of the test statistic being at least as extreme as the one observed given that the null hypothesis is true. A small *p*-value is an indication that the null hypothesis is false.

Good practice

It is good practice to decide in advance of the test how small a *p*-value is required to reject the test. This is exactly analagous to choosing a significance level, α for test. For example, we decide either to reject the null hypothesis if the test statistic exceeds the critical value (for α = 0.05) or analagously to reject the null hypothesis if the *p*-value is smaller than 0.05. It is important to understand the relationship between the two concepts because some statistical software packages report *p*-values rather than critical values.

7.1.4. What are confidence intervals?

How do we form a confidence interval?

The purpose of taking a random sample from a lot or population and computing a statistic, such as the mean from the data, is to approximate the mean of the population. How well the sample statistic estimates the underlying population value is always an issue. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

Confidence levels

Confidence intervals are constructed at a *confidence level*, such as 95%, selected by the user. What does this mean? It means that if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95% of the cases. A confidence stated at a $1 - \alpha$ level can be thought of as the inverse of a significance level, α .

One and two-sided confidence intervals

In the same way that statistical tests can be one or two-sided, confidence intervals can be one or two-sided. A two-sided confidence interval brackets the population parameter from above and below. A one-sided confidence interval brackets the population parameter either from above or below and furnishes an upper or lower bound to its magnitude.

Example of a two-sided confidence interval

For example, a 100($1 - \alpha$)% confidence interval for the mean of a normal population is;

$$\bar{Y} \pm \frac{z_{1-\alpha/2}\sigma}{\sqrt{N}}$$

where $\bar{Y}$ is the sample mean, $z_{1-\alpha/2}$ is the 1-α/2 critical value of the standard normal distribution which is found in the table of the standard normal distribution, σ is the known population standard deviation, and N is the sample size.

Guidance in this chapter

This chapter provides methods for estimating the population parameters and confidence intervals for the situations described under the scope.

Problem with unknown standard deviation

In the normal course of events, population standard deviations are not known, and must be estimated from the data. Confidence intervals, given the same confidence level, are by necessity wider if the standard deviation is estimated from limited data because of the uncertainty in this estimate. Procedures for creating confidence intervals in this situation are described fully in this chapter.

More information on confidence intervals can also be found in Chapter 1.

7.1.5. What is the relationship between a test and a confidence interval?

There is a correspondence between hypothesis testing and confidence intervals

In general, for every test of hypothesis there is an equivalent statement about whether the hypothesized parameter value is included in a confidence interval. For example, consider the previous example of linewidths where photomasks are tested to ensure that their linewidths have a mean of 500 micrometers. The null and alternative hypotheses are:

H_0: mean linewidth = 500 micrometers

H_a: mean linewidth $\neq$ 500 micrometers

Hypothesis test for the mean

For the test, the sample mean, $\bar{Y}$, is calculated from N linewidths chosen at random positions on each photomask. For the purpose of the test, it is assumed that the standard deviation, σ, is known from a long history of this process. A test statistic is calculated from these sample statistics, and the null hypothesis is rejected if:

$$\frac{\bar{Y} - 500}{\sigma/\sqrt{N}} \leq z_{\alpha/2} \quad \text{or} \quad \frac{\bar{Y} - 500}{\sigma/\sqrt{N}} \geq z_{1-\alpha/2}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are tabled values from the normal distribution.

Equivalent confidence interval

With some algebra, it can be seen that the null hypothesis is rejected if and only if the value 500 micrometers is not in the confidence interval

$$\bar{Y} \pm \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{N}}$$

Equivalent confidence interval

In fact, all values bracketed by this interval would be accepted as null values for a given set of test data.

7.1.6. What are outliers in the data?

Definition of outliers

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Ways to describe data

Two activities are essential for characterizing a set of data:

1. Examination of the overall shape of the graphed data for important features, including symmetry and departures from assumptions. The chapter on Exploratory Data Analysis (EDA) discusses assumptions and summarization of data in detail.

2. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box plots, along with an analytic procedure for detecting outliers when the distribution is normal (Grubbs' Test), are also discussed in detail in the EDA chapter.

Box plot construction

The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q2, then the difference (Q2 - Q1) is called the interquartile range or IQ.

Box plots with fences

A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called *fences*) are needed for identifying extreme values in the tails of the distribution:

1. lower inner fence: Q1 - 1.5*IQ
2. upper inner fence: Q2 + 1.5*IQ
3. lower outer fence: Q1 - 3*IQ
4. upper outer fence: Q2 + 3*IQ

Outlier detection criteria

A point beyond an inner fence on either side is considered a **mild outlier**. A point beyond an outer fence is considered an **extreme outlier**.

Example of an outlier

The data set of $N = 90$ ordered observations as shown below is examined for outliers:

box plot

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441

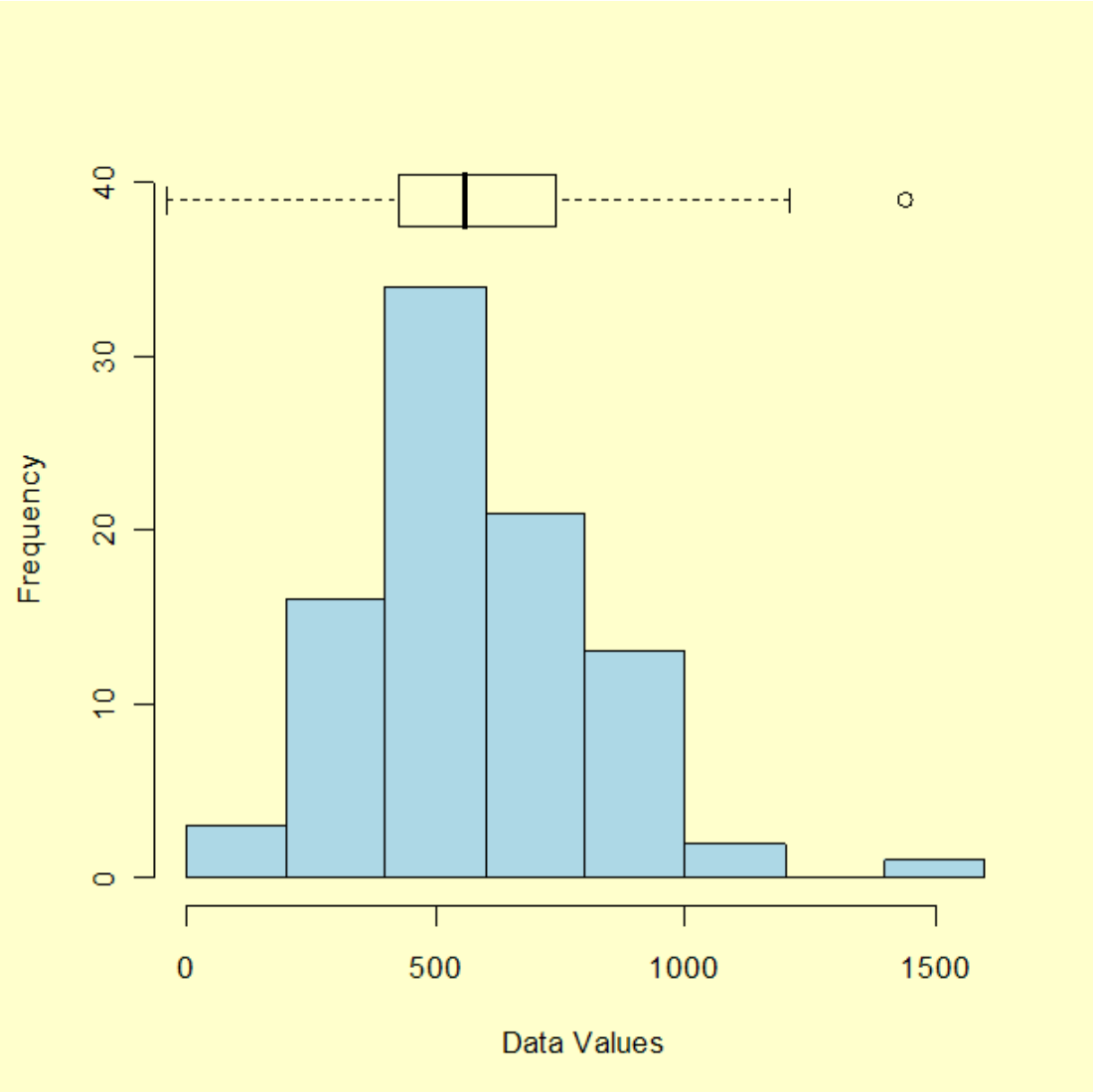
The computations are as follows:

- Median = (n+1)/2 largest data point = the average of the 45th and 46th ordered points = (559 + 560)/2 = 559.5
- Lower quartile = .25(N+1)th ordered point = 22.75th ordered point = 411 + .75(436-411) = 429.75
- Upper quartile = .75(N+1)th ordered point = 68.25th ordered point = 739 +.25(752-739) = 742.25
- Interquartile range = 742.25 - 429.75 = 312.5
- Lower inner fence = 429.75 - 1.5 (312.5) = -39.0
- Upper inner fence = 742.25 + 1.5 (312.5) = 1211.0
- Lower outer fence = 429.75 - 3.0 (312.5) = -507.75
- Upper outer fence = 742.25 + 3.0 (312.5) = 1679.75

From an examination of the fence points and the data, one point (1441) exceeds the upper inner fence and stands out as a mild outlier; there are no extreme outliers.

Histogram with box plot

A histogram with an overlaid box plot are shown below.



The outlier is identified as the largest value in the data set, 1441, and appears as the circle to the right of the box plot.

Outliers may contain important information

Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points.

7.1.7. What are trends in sequential process or product data?

Detecting trends by plotting the data points to see if a line with an obviously non-zero slope fits the points

Detecting trends is equivalent to comparing the process values to what we would expect a series of numbers to look like if there were no trends. If we see a significant departure from a model where the next observation is equally likely to go up or down, then we would reject the hypothesis of "no trend".

A common way of investigating for trends is to fit a straight line to the data and observe the line's direction (or slope). If the line looks horizontal, then there is no evidence of a trend; otherwise there is. Formally, this is done by testing whether the slope of the line is significantly different from zero. The methodology for this is covered in Chapter 4.

Other trend tests

A non-parametric approach for detecting significant trends known as the Reverse Arrangement Test is described in Chapter 8.

7.2. Comparisons based on data from one process

Questions answered in this section

For a single process, the current state of the process can be compared with a nominal or hypothesized state. This section outlines techniques for answering the following questions from data gathered from a single process:

1. Do the observations come from a particular distribution?
    1. Chi-Square Goodness-of-Fit test for a continuous or discrete distribution
    2. Kolmogorov- Smirnov test for a continuous distribution
    3. Anderson-Darling and Shapiro-Wilk tests for a continuous distribution
2. Are the data consistent with the assumed process mean?
    1. Confidence interval approach
    2. Sample sizes required
3. Are the data consistent with a nominal standard deviation?
    1. Confidence interval approach
    2. Sample sizes required
4. Does the proportion of defectives meet requirements?
    1. Confidence intervals
    2. Sample sizes required
5. Does the defect density meet requirements?
6. What intervals contain a fixed percentage of the data?
    1. Approximate intervals that contain most of the population values
    2. Percentiles
    3. Tolerance intervals
    4. Tolerance intervals based on the smallest and largest observations

General forms of testing

These questions are addressed either by an hypothesis test or by a confidence interval.

Parametric vs. non-parametric

All hypothesis-testing procedures can be broadly described as either parametric or non-parametric/distribution-free. Parametric test procedures are those that:

testing

1. Involve hypothesis testing of specified parameters (such as "the population mean=50 grams"...).
2. Require a stringent set of assumptions about the underlying sampling distributions.

When to use nonparametric methods?

When do we require non-parametric or distribution-free methods? Here are a few circumstances that may be candidates:

1. The measurements are only categorical; i.e., they are nominally scaled, or ordinally (in ranks) scaled.
2. The assumptions underlying the use of parametric methods cannot be met.
3. The situation at hand requires an investigation of such features as randomness, independence, symmetry, or goodness of fit rather than the testing of hypotheses about specific values of particular population parameters.

Difference between non-parametric and distribution-free

Some authors distinguish between non-parametric and distribution-free procedures.

Distribution-free test procedures are broadly defined as:

1. Those whose test statistic does not depend on the form of the underlying population distribution from which the sample data were drawn, or
2. Those for which the data are nominally or ordinally scaled.

Nonparametric test procedures are defined as those that are not concerned with the parameters of a distribution.

Advantages of nonparametric methods.

Distribution-free or nonparametric methods have several advantages, or benefits:

1. They may be used on all types of data-categorical data, which are nominally scaled or are in rank form, called ordinally scaled, as well as interval or ratio-scaled data.
2. For small sample sizes they are easy to apply.
3. They make fewer and less stringent assumptions than their parametric counterparts.
4. Depending on the particular procedure they may be *almost* as powerful as the corresponding parametric procedure when the assumptions of the latter are

met, and when this is not the case, they are generally more powerful.

Disadvantages of nonparametric methods

Of course there are also disadvantages:

1. If the assumptions of the parametric methods can be met, it is generally more efficient to use them.

2. For large sample sizes, data manipulations tend to become more laborious, unless computer software is available.

3. Often special tables of critical values are needed for the test statistic, and these values cannot always be generated by computer software. On the other hand, the critical values for the parametric tests are readily available and generally easy to incorporate in computer programs.

7.2.1. Do the observations come from a particular distribution?

Data are often assumed to come from a particular distribution.

Goodness-of-fit tests indicate whether or not it is reasonable to assume that a random sample comes from a specific distribution. Statistical techniques often rely on observations having come from a population that has a distribution of a specific form (e.g., normal, lognormal, Poisson, etc.). Standard control charts for continuous measurements, for instance, require that the data come from a normal distribution. Accurate lifetime modeling requires specifying the correct distributional model. There may be historical or theoretical reasons to assume that a sample comes from a particular population, as well. Past data may have consistently fit a known distribution, for example, or theory may predict that the underlying population should be of a specific form.

Hypothesis Test model for Goodness-of-fit

Goodness-of-fit tests are a form of hypothesis testing where the null and alternative hypotheses are

H_0: Sample data come from the stated distribution.
H_A: Sample data **do not** come from the stated distribution.

Parameters may be assumed or estimated from the data

One needs to consider whether a simple or composite hypothesis is being tested. For a simple hypothesis, values of the distribution's parameters are specified prior to drawing the sample. For a composite hypothesis, one or more of the parameters is unknown. Often, these parameters are estimated using the sample observations.

A simple hypothesis would be:

H_0: Data are from a normal distribution, $\mu = 0$ and $\sigma = 1$.

A composite hypothesis would be:

H_0: Data are from a normal distribution, unknown μ and σ.

Composite hypotheses are more common because they allow us to decide whether a sample comes from any distribution of a specific type. In this situation, the form of the distribution is of interest, regardless of the values of the parameters. Unfortunately, composite hypotheses are more difficult to

work with because the critical values are often hard to compute.

Problems with censored data

A second issue that affects a test is whether the data are censored. When data are censored, sample values are in some way restricted. Censoring occurs if the range of potential values are limited such that values from one or both tails of the distribution are unavailable (e.g., right and/or left censoring - where high and/or low values are missing). Censoring frequently occurs in reliability testing, when either the testing time or the number of failures to be observed is fixed in advance. A thorough treatment of goodness-of-fit testing under censoring is beyond the scope of this document. See D'Agostino & Stephens (1986) for more details.

Three types of tests will be covered

Three goodness-of-fit tests are examined in detail:

1. Chi-square test for continuous and discrete distributions;
2. Kolmogorov-Smirnov test for continuous distributions based on the empirical distribution function (EDF);
3. Anderson-Darling test for continuous distributions.

A more extensive treatment of goodness-of-fit techniques is presented in D'Agostino & Stephens (1986). Along with the tests mentioned above, other general and specific tests are examined, including tests based on regression and graphical techniques.

7.2.1.1. Chi-square goodness-of-fit test

Choice of number of groups for "Goodness of Fit" tests is important - but only useful rules of thumb can be given	The test requires that the data first be grouped. The actual number of observations in each group is compared to the expected number of observations and the test statistic is calculated as a function of this difference. The number of groups and how group membership is defined will affect the power of the test (i.e., how sensitive it is to detecting departures from the null hypothesis). Power will not only be affected by the number of groups and how they are defined, but by the sample size and shape of the null and underlying (true) distributions. Despite the lack of a clear "best method", some useful rules of thumb can be given.
Group Membership	When data are discrete, group membership is unambiguous. Tabulation or cross tabulation can be used to categorize the data. Continuous data present a more difficult challenge. One defines groups by segmenting the range of possible values into non-overlapping intervals. Group membership can then be defined by the endpoints of the intervals. In general, power is maximized by choosing endpoints such that group membership is equiprobable (i.e., the probabilities associated with an observation falling into a given group are divided as evenly as possible across the intervals). Many commercial software packages follow this procedure.
Rule-of-thumb for number of groups	One rule-of-thumb suggests using the value $\mathbf{2}\boldsymbol{n}^{\mathbf{2/5}}$ as a good starting point for choosing the number of groups. Another well known rule-of-thumb requires every group to have at least 5 data points.
Computation of the chi-square goodness-of-fit test	The formulas for the computation of the chi-square goodnes-of-fit test are given in the EDA chapter.

7.2.1.2. Kolmogorov- Smirnov test

The K-S test is a good alternative to the chi-square test.

The Kolmogorov-Smirnov (K-S) test was originally proposed in the 1930's in papers by Kolmogorov (1933) and Smirnov (1936). Unlike the Chi-Square test, which can be used for testing against both continuous and discrete distributions, the K-S test is only appropriate for testing data against a continuous distribution, such as the normal or Weibull distribution. It is one of a number of tests that are based on the empirical cumulative distribution function (ECDF).

K-S procedure

Details on the construction and interpretation of the K-S test statistic, *D*, and examples for several distributions are outlined in Chapter 1.

The probability associated with the test statistic is difficult to compute.

Critical values associated with the test statistic, *D*, are difficult to compute for finite sample sizes, often requiring Monte Carlo simulation. However, some general purpose statistical software programs support the Kolmogorov-Smirnov test at least for some of the more common distributions. Tabled values can be found in Birnbaum (1952). A correction factor can be applied if the parameters of the distribution are estimated with the same data that are being tested. See D'Agostino and Stephens (1986) for details.

7.2.1.3. Anderson-Darling and Shapiro-Wilk tests

Purpose: Test for distributional adequacy

The Anderson-Darling Test

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data comes from a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails of the distribution than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested.

Requires critical values for each distribution

The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Tables of critical values are not given in this handbook (see Stephens 1974, 1976, 1977, and 1979) because this test is usually applied with a statistical software program that produces the relevant critical values. Currently, Dataplot computes critical values for the Anderson-Darling test for the following distributions:

- normal
- lognormal
- Weibull
- extreme value type I.

Anderson-Darling procedure

Details on the construction and interpretation of the Anderson-Darling test statistic, A^2, and examples for several distributions are outlined in Chapter 1.

Shapiro-Wilk test for normality

The Shapiro-Wilk Test For Normality

The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, $x_1, x_2, ..., x_n$ comes from (specifically) a normal distribution . Small values of W are evidence of departure from normality and percentage points for the W statistic, obtained via Monte Carlo simulations, were reproduced by Pearson and Hartley (1972, Table 16). This test has done very well in

comparison studies with other goodness of fit tests.

The *W* statistic is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

where the $x_{(i)}$ are the ordered sample values ($x_{(1)}$ is the smallest) and the a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size *n* from a normal distribution (see Pearson and Hartley (1972, Table 15)).

For more information about the Shapiro-Wilk test the reader is referred to the original Shapiro and Wilk (1965) paper and the tables in Pearson and Hartley (1972).

7.2.2. Are the data consistent with the assumed process mean?

The testing of H_0 for a single population mean

Given a random sample of measurements, Y_1, ..., Y_N, there are three types of questions regarding the true mean of the population that can be addressed with the sample data. They are:

1. Does the true mean agree with a known standard or assumed mean?
2. Is the true mean of the population less than a given standard?
3. Is the true mean of the population at least as large as a given standard?

Typical null hypotheses

The corresponding null hypotheses that test the true mean, μ, against the standard or assumed mean, μ_0 are:

1. $H0 : \mu = \mu0$
2. $H0 : \mu \leq \mu0$
3. $H0 : \mu \geq \mu0$

Test statistic where the standard deviation is not known

The basic statistics for the test are the sample mean and the standard deviation. The form of the test statistic depends on whether the poulation standard deviation, σ, is known or is estimated from the data at hand. The more typical case is where the standard deviation must be estimated from the data, and the test statistic is

$$t = \frac{\bar{Y} - \mu_0}{s / \sqrt{N}}$$

where the sample mean is

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

and the sample standard deviation is

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

with N - 1 degrees of freedom.

Comparison with critical values

For a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10, the hypothesis associated with each case enumerated above is rejected if:

1. $|t| \geq t_{1-\alpha/2,\, N-1}$
2. $t \geq t_{1-\alpha,\, N-1}$
3. $t \leq t_{\alpha,\, N-1}$

where $t_{1-\alpha/2,\, N-1}$ is the 1-α/2 critical value from the t distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the *t*-table in Chapter 1.

Test statistic where the standard deviation is known

If the standard deviation is known, the form of the test statistic is

$$z = \frac{\bar{Y} - \mu_0}{\sigma / \sqrt{N}}$$

For case (1), the test statistic is compared with $z_{1-\alpha/2}$, which is the 1-α/2 critical value from the standard normal distribution, and similarly for cases (2) and (3).

Caution

If the standard deviation is assumed known for the purpose of this test, this assumption should be checked by a test of hypothesis for the standard deviation.

An illustrative example of the t-test

The following numbers are particle (contamination) counts for a sample of 10 semiconductor silicon wafers:

```
50   48   44   56   61   52   53   55   67   51
```

The mean = 53.7 counts and the standard deviation = 6.567 counts.

The test is two-sided

Over a long run the process average for wafer particle counts has been 50 counts per wafer, and on the basis of the sample, we want to test whether a change has occurred. The null hypothesis that the process mean is 50 counts is tested against the alternative hypothesis that the process mean is not equal to 50 counts. The purpose of the two-sided alternative is to rule out a possible process change in either direction.

Critical values

For a significance level of α = 0.05, the chances of erroneously rejecting the null hypothesis when it is true are 5

% or less. (For a review of hypothesis testing basics, see Chapter 1).

Even though there is a history on this process, it has not been stable enough to justify the assumption that the standard deviation is known. Therefore, the appropriate test statistic is the *t*-statistic. Substituting the sample mean, sample standard deviation, and sample size into the formula for the test statistic gives a value of

$$t = 1.782$$

with degrees of freedom $N - 1 = 9$. This value is tested against the critical value

$$t_{1-0.025;9} = 2.262$$

from the *t*-table where the critical value is found under the column labeled 0.975 for the probability of exceeding the critical value and in the row for 9 degrees of freedom. The critical value is based on $\alpha/2$ instead of α because of the two-sided alternative (two-tailed test) which requires equal probabilities in each tail of the distribution that add to α.

Conclusion Because the value of the test statistic falls in the interval (-2.262, 2.262), we cannot reject the null hypothesis and, therefore, we may continue to assume the process mean is 50 counts.

7.2.2.1. Confidence interval approach

Testing using a confidence interval

The hypothesis test results in a "yes" or "no" answer. The null hypothesis is either rejected or not rejected. There is another way of testing a mean and that is by constructing a confidence interval about the true but unknown mean.

General form of confidence intervals where the standard deviation is unknown

Tests of hypotheses that can be made from a single sample of data were discussed on the foregoing page. As with null hypotheses, confidence intervals can be two-sided or one-sided, depending on the question at hand. The general form of confidence intervals, for the three cases discussed earlier, where the standard deviation is unknown are:

1. Two-sided confidence interval for μ:

$$\bar{Y} + \frac{s}{\sqrt{N}} t_{\alpha/2,N-1} \leq \mu \leq \bar{Y} + \frac{s}{\sqrt{N}} t_{1-\alpha/2,N-1}$$

2. Lower one-sided confidence interval for μ:

$$\mu \geq \bar{Y} + \frac{s}{\sqrt{N}} t_{\alpha,N-1}$$

3. Upper one-sided confidence interval for μ:

$$\mu \leq \bar{Y} + \frac{s}{\sqrt{N}} t_{1-\alpha,N-1}$$

where $t_{\alpha/2,\, N-1}$ is the $\alpha/2$ critical value from the t distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the t table in Chapter 1.

Confidence level

The confidence intervals are constructed so that the probability of the interval containing the mean is 1 - α. Such intervals are referred to as 100(1-α)% confidence intervals.

A 95% confidence interval for

The corresponding confidence interval for the test of hypothesis example on the foregoing page is shown below. A 95 % confidence interval for the population mean of particle counts per wafer is given by

the example

$$\bar{Y} + \frac{s}{\sqrt{N}}(t_{0.025,9}) \leq \mu \leq \bar{Y} + \frac{s}{\sqrt{N}}(t_{0.975,9})$$

$$53.7 + \frac{6.567}{\sqrt{10}}(-2.262) \leq \mu \leq 53.7 + \frac{6.567}{\sqrt{10}}(2.262)$$

$$49.0 \leq \mu \leq 58.4$$

Interpretation The 95 % confidence interval includes the null hypothesis if, and only if, it would be accepted at the 5 % level. This interval includes the null hypothesis of 50 counts so we cannot reject the hypothesis that the process mean for particle counts is 50. The confidence interval includes all null hypothesis values for the population mean that would be accepted by an hypothesis test at the 5 % significance level. This assumes, of course, a two-sided alternative.

7.2.2.2. Sample sizes required

The computation of sample sizes depends on many things, some of which have to be assumed in advance

Perhaps one of the most frequent questions asked of a statistician is,

> "How many measurements should be included in the sample? "

Unfortunately, there is no correct answer without additional information (or assumptions). The sample size required for an experiment designed to investigate the behavior of an unknown population mean will be influenced by the following:

- value selected for α, the risk of rejecting a true hypothesis
- value of β, the risk of accepting a false null hypothesis when a particular value of the alternative hypothesis is true.
- value of the population standard deviation.

Application - estimating a minimum sample size, N, for limiting the error in the estimate of the mean

For example, suppose that we wish to estimate the average daily yield, μ, of a chemical process by the mean of a sample, Y_1, ..., Y_N, such that the error of estimation is less than δ with a probability of 95%. This means that a 95% confidence interval centered at the sample mean should be

$$\bar{Y} - \delta \le \mu \le \bar{Y} + \delta$$

and if the standard deviation is known,

$$\delta = \frac{\sigma}{\sqrt{N}} z_{1-0.025}$$

The critical value from the normal distribution for 1-α/2 = 0.975 is 1.96. Therefore,

$$N \ge \left(\frac{1.96}{\delta}\right)^2 \sigma^2$$

Limitation and interpretation

A restriction is that the standard deviation must be known. Lacking an exact value for the standard deviation requires some accommodation, perhaps the best estimate available from a previous experiment.

Controlling the risk of accepting a false hypothesis

To control the risk of accepting a false hypothesis, we set not only α, the probability of rejecting the null hypothesis when it is true, but also β, the probability of accepting the null hypothesis when in fact the population mean is $\mu+\delta$ where δ is the difference or shift we want to detect.

Standard deviation assumed to be known

The minimum sample size, N, is shown below for two- and one-sided tests of hypotheses with σ assumed to be known.

$$N = (z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{\sigma}{\delta}\right)^2 \rightarrow two-sided\ test$$

$$N = (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma}{\delta}\right)^2 \rightarrow one-sided\ test$$

The quantities $z_{1-\alpha/2}$ and $z_{1-\beta}$ are critical values from the normal distribution.

Note that it is usual to state the shift, δ, in units of the standard deviation, thereby simplifying the calculation.

Example where the shift is stated in terms of the standard deviation

For a one-sided hypothesis test where we wish to detect an increase in the population mean of one standard deviation, the following information is required: α, the significance level of the test, and β, the probability of failing to detect a shift of one standard deviation. For a test with α = 0.05 and β = 0.10, the minimum sample size required for the test is

$$N = (1.645 + 1.282)^2 = 8.567 \sim 9.$$

More often we must compute the sample size with the population standard deviation being unknown

The procedures for computing sample sizes when the standard deviation is not known are similar to, but more complex, than when the standard deviation is known. The formulation depends on the *t* distribution where the minimum sample size is given by

$$N = (t_{1-\alpha/2} + t_{1-\beta})^2 \left(\frac{s}{\delta}\right)^2 \rightarrow two-sided\ test$$

$$N = (t_{1-\alpha} + t_{1-\beta})^2 \left(\frac{s}{\delta}\right)^2 \rightarrow one-sided\ test$$

The drawback is that critical values of the *t* distribution depend on known degrees of freedom, which in turn depend upon the sample size which we are trying to estimate.

Iterate on the initial estimate using critical values from

Therefore, the best procedure is to start with an intial estimate based on a sample standard deviation and iterate. Take the example discussed above where the the minimum sample size is computed to be N = 9. This estimate is low. Now use the formula above with degrees of freedom N - 1 = 8 which gives a second estimate of

the t table

$$N = (1.860 + 1.397)^2 = 10.6 \sim 11.$$

It is possible to apply another iteration using degrees of freedom 10, but in practice one iteration is usually sufficient. For the purpose of this example, results have been rounded to the closest integer; however, computer programs for finding critical values from the *t* distribution allow non-integer degrees of freedom.

Table showing minimum sample sizes for a two-sided test

The table below gives sample sizes for a two-sided test of hypothesis that the mean is a given value, with the shift to be detected a multiple of the standard deviation. For a one-sided test at significance level α, look under the value of 2α in column 1. Note that this table is based on the normal approximation (i.e., the standard deviation is known).

Sample Size Table for Two-Sided Tests

α	β	$\delta = .5\sigma$	$\delta = 1.0\sigma$	$\delta = 1.5\sigma$
.01	.01	98	25	11
.01	.05	73	18	8
.01	.10	61	15	7
.01	.20	47	12	6
.01	.50	27	7	3
.05	.01	75	19	9
.05	.05	53	13	6
.05	.10	43	11	5
.05	.20	33	8	4
.05	.50	16	4	3
.10	.01	65	16	8
.10	.05	45	11	5
.10	.10	35	9	4
.10	.20	25	7	3
.10	.50	11	3	3
.20	.01	53	14	6
.20	.05	35	9	4
.20	.10	27	7	3
.20	.20	19	5	3
.20	.50	7	3	3

7.2.3. Are the data consistent with a nominal standard deviation?

The testing of H_0 for a single population mean

Given a random sample of measurements, Y_1, ..., Y_N, there are three types of questions regarding the true standard deviation of the population that can be addressed with the sample data. They are:

1. Does the true standard deviation agree with a nominal value?
2. Is the true standard deviation of the population less than or equal to a nominal value?
3. Is the true stanard deviation of the population at least as large as a nominal value?

Corresponding null hypotheses

The corresponding null hypotheses that test the true standard deviation, σ, against the nominal value, σ_0 are:

1. H_0: $\sigma = \sigma_0$
2. H_0: $\sigma <= \sigma_0$
3. H_0: $\sigma >= \sigma_0$

Test statistic

The basic test statistic is the chi-square statistic

$$\chi^2 = \frac{(N-1)s^2}{\sigma_0^2}$$

with N - 1 degrees of freedom where s is the sample standard deviation; i.e.,

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

.

Comparison with critical values

For a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10, the hypothesis associated with each case enumerated above is rejected if:

1. $\chi^2 \geq \chi^2_{1-\alpha/2}$ or $\chi^2 \leq \chi^2_{\alpha/2}$
2. $\chi^2 \geq \chi^2_{1-\alpha}$
3. $\chi^2 \leq \chi^2_{\alpha}$

where $X^2_{\alpha/2}$ is the $\alpha/2$ critical value from the chi-square distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the chi-square table in Chapter 1.

Warning

Because the chi-square distribution is a non-negative, asymmetrical distribution, care must be taken in looking up critical values from tables. For two-sided tests, critical values are required for both tails of the distribution.

Example

A supplier of 100 ohm·cm silicon wafers claims that his fabrication process can produce wafers with sufficient consistency so that the standard deviation of resistivity for the lot does not exceed 10 ohm·cm. A sample of N = 10 wafers taken from the lot has a standard deviation of 13.97 ohm.cm. Is the suppliers claim reasonable? This question falls under null hypothesis (2) above. For a test at significance level, α = 0.05, the test statistic,

$$\chi^2 = \frac{(N-1)s^2}{\sigma_0^2} = \frac{9(13.97)^2}{100} = 17.56$$

is compared with the critical value, $X^2_{0.95,\,9}$ = 16.92.

Since the test statistic (17.56) exceeds the critical value (16.92) of the chi-square distribution with 9 degrees of freedom, the manufacturer's claim is rejected.

7.2.3.1. Confidence interval approach

Confidence intervals for the standard deviation

Confidence intervals for the true standard deviation can be constructed using the chi-square distribution. The 100(1- α)% confidence intervals that correspond to the tests of hypothesis on the previous page are given by

1. Two-sided confidence interval for σ

$$\frac{s\sqrt{N-1}}{\sqrt{\chi^2_{1-\alpha/2,N-1}}} \leq \sigma \leq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{\alpha/2,N-1}}}$$

2. Lower one-sided confidence interval for σ

$$\sigma \geq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{1-\alpha,N-1}}}$$

3. Upper one-sided confidence interval for σ

$$0 \leq \sigma \leq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{\alpha,N-1}}}$$

where for case (1), $X^2_{\alpha/2}$ is the $\alpha/2$ critical value from the chi-square distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the chi-square table in Chapter 1.

Choice of risk level α can change the conclusion

Confidence interval (1) is equivalent to a two-sided test for the standard deviation. That is, if the hypothesized or nominal value, σ_0, is not contained within these limits, then the hypothesis that the standard deviation is equal to the nominal value is rejected.

A dilemma of hypothesis testing

A change in α can lead to a change in the conclusion. This poses a dilemma. What should α be? Unfortunately, there is no clear-cut answer that will work in all situations. The usual strategy is to set α small so as to guarantee that the null hypothesis is *wrongly* rejected in only a small number of

cases. The risk, β, of failing to reject the null hypothesis when it is false depends on the size of the discrepancy, and also depends on α. The discussion on the next page shows how to choose the sample size so that this risk is kept small for specific discrepancies.

7.2.3.2. Sample sizes required

Sample sizes to minimize risk of false acceptance

The following procedure for computing sample sizes for tests involving standard deviations follows W. Diamond (1989). The idea is to find a sample size that is large enough to guarantee that the risk, β, of accepting a false hypothesis is small.

Alternatives are specific departures from the null hypothesis

This procedure is stated in terms of changes in the variance, not the standard deviation, which makes it somewhat difficult to interpret. Tests that are generally of interest are stated in terms of δ, a discrepancy from the hypothesized variance. For example:

1. Is the true variance larger than its hypothesized value by δ ?
2. Is the true variance smaller than its hypothesized value by δ ?

That is, the tests of interest are:

1. $H_0: \sigma^2 \geq \sigma_0^2 + \delta; \quad \delta \geq 0$
2. $H_0: \sigma^2 \leq \sigma_0^2 - \delta; \quad \delta \geq 0$

Interpretation

The experimenter wants to assure that the probability of erroneously accepting the null hypothesis of unchanged variance is at most β. The sample size, N, required for this type of detection depends on the factor, δ; the significance level, α; and the risk, β.

First choose the level of significance and beta risk

The sample size is determined by first choosing appropriate values of α and β and then following the directions below to find the degrees of freedom, ν, from the chi-square distribution.

The calculations should be done by creating a table or

First compute

$$R = 1 + \frac{\delta}{\sigma_0^2}$$

Then generate a table of degrees of freedom, ν, say

spreadsheet

between 1 and 200. For case (1) or (2) above, calculate β_ν and the corresponding value of C_ν for each value of degrees of freedom in the table where

$$\begin{aligned} 1.\quad \beta_\nu &= \chi^2_{1-\alpha,\nu}/R \\ C_\nu &= \Pr(\chi^2_\nu < \beta_\nu) \\ 2.\quad \beta_\nu &= \chi^2_{\alpha,\nu}/R \\ C_\nu &= \Pr(\chi^2_\nu > \beta_\nu) \end{aligned}$$

The value of ν where C_ν is closest to β is the correct degrees of freedom and

$$N = \nu + 1$$

Hints on using software packages to do the calculations

The quantity $X^2_{1-\alpha,\,\nu}$ is the critical value from the chi-square distribution with ν degrees of freedom which is exceeded with probability α. It is sometimes referred to as the percent point function (PPF) or the inverse chi-square function. The probability that is evaluated to get C_ν is called the cumulative density function (CDF).

Example

Consider the case where the variance for resistivity measurements on a lot of silicon wafers is claimed to be 100 $(\text{ohm}\cdot\text{cm})^2$. A buyer is unwilling to accept a shipment if δ is greater than 55 ohm·cm for a particular lot. This problem falls under case (1) above. How many samples are needed to assure risks of $\alpha = 0.05$ and $\beta = 0.01$?

Calculations

If software is available to compute the roots (or zero values) of a univariate function, then we can determine the sample size by finding the roots of a function that calculates C_ν for a given value of ν. The procedure is:

1. Define constants.
   $\alpha = 0.05$
   $\beta = 0.01$
   $\delta = 55$
   $\sigma_0^2 = 100$
   $R = 1 + \delta/\sigma_0^2$

2. Create a function, Cnu.
   $Cnu = F(\ F^{-1}(\alpha,\ \nu)/R,\ \nu\ ) - \beta$

   $F(x,\ \nu)$ returns the probability of a chi-square random variable with ν degrees of freedom that is less than or equal to x and $F^{-1}(\alpha,\ \nu)$ returns x such that $F(x,\ \nu) = \alpha$.

3. Find the value of ν for which the function, Cnu, is zero.

Using this procedure, Cnu is zero when ν is 169.3. Therefore, the minimum sample size needed to guarantee the risk level is N = 170.

Alternatively, we can determine the sample size by simply printing computed values of Cnu for various values of ν.

1. Define constants.
   α = 0.05
   δ = 55
   σ_0^2 = 100
   R = 1 + δ/σ_0^2

2. Generate Cnu for values of ν from 1 to 200.
   Bnu = $F^{-1}(\alpha, \nu)$ / R
   Cnu = F(Bnu, ν)

The values of Cnu generated for ν between 165 and 175 degrees of freedom are shown below.

ν	Bnu	Cnu
165	126.4344	0.0114
166	127.1380	0.0110
167	127.8414	0.0107
168	128.5446	0.0104
169	129.2477	0.0101
170	129.9506	0.0098
171	130.6533	0.0095
172	131.3558	0.0092
173	132.0582	0.0090
174	132.7604	0.0087
175	133.4625	0.0085

The value of Cnu closest to 0.01 is 0.0101, which is associated with ν = 169 degrees of freedom. Therefore, the minimum sample size needed to guarantee the risk level is N = 170.

The calculations used in this section can be performed using both Dataplot code and R code.

7.2.4. Does the proportion of defectives meet requirements?

Testing proportion defective is based on the binomial distribution

The proportion of defective items in a manufacturing process can be monitored using statistics based on the observed number of defectives in a random sample of size N from a continuous manufacturing process, or from a large population or lot. The proportion defective in a sample follows the binomial distribution where p is the probability of an individual item being found defective. Questions of interest for quality control are:

1. Is the proportion of defective items within prescribed limits?
2. Is the proportion of defective items less than a prescribed limit?
3. Is the proportion of defective items greater than a prescribed limit?

Hypotheses regarding proportion defective

The corresponding hypotheses that can be tested are:

1. $p \neq p_0$
2. $p \leq p_0$
3. $p \geq p_0$

where p_0 is the prescribed proportion defective.

Test statistic based on a normal approximation

Given a random sample of measurements Y_1, ..., Y_N from a population, the proportion of items that are judged defective from these N measurements is denoted $\hat{p}$. The test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$$

depends on a normal approximation to the binomial distribution that is valid for large N, ($N > 30$). This approximation simplifies the calculations using critical values from the table of the normal distribution as shown below.

Restriction on sample size

Because the test is approximate, N needs to be large for the test to be valid. One criterion is that N should be chosen so that

$$\min\{Np_0, N(1 - p_0)\} \geq 5$$

For example, if $p_0 = 0.1$, then N should be at least 50 and if $p_0 = 0.01$, then N should be at least 500. Criteria for choosing a sample size in order to guarantee detecting a change of size δ are discussed on another page.

One and two-sided tests for proportion defective

Tests at the 1 - α confidence level corresponding to hypotheses (1), (2), and (3) are shown below. For hypothesis (1), the test statistic, z, is compared with $z_{1-\alpha/2}$, the critical value from the normal distribution that is exceeded with probability $\alpha/2$ and similarly for (2) and (3). If

1. $|z| \geq z_{1-\alpha/2}$
2. $z \leq z_{\alpha}$
3. $z \geq z_{1-\alpha}$

the null hypothesis is rejected.

Example of a one-sided test for proportion defective

After a new method of processing wafers was introduced into a fabrication process, two hundred wafers were tested, and twenty-six showed some type of defect. Thus, for N= 200, the proportion defective is estimated to be $\hat{p}$ = 26/200 = 0.13. In the past, the fabrication process was capable of producing wafers with a proportion defective of at most 0.10. The issue is whether the new process has degraded the quality of the wafers. The relevant test is the one-sided test (3) which guards against an increase in proportion defective from its historical level.

Calculations for a one-sided test of proportion defective

For a test at significance level α = 0.05, the hypothesis of no degradation is validated if the test statistic z is less than the critical value, $z_{0.95}$ = 1.645. The test statistic is computed to be

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} = \frac{0.13 - 0.10}{\sqrt{\frac{0.10(0.90)}{200}}} = 1.414$$

Interpretation

Because the test statistic is less than the critical value (1.645), we cannot reject hypothesis (3) and, therefore, we cannot conclude that the new fabrication method is degrading the quality of the wafers. The new process may, indeed, be worse, but more evidence would be needed to reach that conclusion at the 95% confidence level.

7.2.4.1. Confidence intervals

Confidence intervals using the method of Agresti and Coull

The method recommended by Agresti and Coull (1998) and also by Brown, Cai and DasGupta (2001) (the methodology was originally developed by Wilson in 1927) is to use the form of the confidence interval that corresponds to the hypothesis test given in Section 7.2.4. That is, solve for the two values of p_0 (say, p_{upper} and p_{lower}) that result from setting $z = z_{1-\alpha/2}$ and solving for $p_0 = p_{upper}$, and then setting $z = z_{\alpha/2}$ and solving for $p_0 = p_{lower}$. (Here, as in Section 7.2.4, $z_{\alpha/2}$ denotes the variate value from the standard normal distribution such that the area to the left of the value is $\alpha/2$.) Although solving for the two values of p_0 might sound complicated, the appropriate expressions can be obtained by straightforward but slightly tedious algebra. Such algebraic manipulation isn't necessary, however, as the appropriate expressions are given in various sources. Specifically, we have

Formulas for the confidence intervals

$$\text{U.L.} = \frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}}$$

$$\text{L.L.} = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{z^2_{\alpha/2}}{n}}$$

Procedure does not strongly depend on values of p and n

This approach can be substantiated on the grounds that it is the exact algebraic counterpart to the (large-sample) hypothesis test given in section 7.2.4 and is also supported by the research of Agresti and Coull. One advantage of this procedure is that its worth does not strongly depend upon the value of *n* and/or *p*, and indeed was recommended by Agresti and Coull for virtually all combinations of *n* and *p*.

Another advantage is that the lower limit cannot be negative

Another advantage is that the lower limit cannot be negative. That is not true for the confidence expression most frequently used:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A confidence limit approach that produces a lower limit which is an impossible value for the parameter for which the interval is constructed is an inferior approach. This also applies to limits for the control charts that are discussed in Chapter 6.

One-sided confidence intervals

A one-sided confidence interval can also be constructed simply by replacing each $z_{\alpha/2}$ by z_{α} in the expression for the lower or upper limit, whichever is desired. The 95% one-sided interval for *p* for the example in the preceding section is:

Example

$$p \geq \text{lower limit}$$

$$p \geq \frac{\hat{p} + \frac{z_\alpha^2}{2n} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_\alpha^2}{4n^2}}}{1 + \frac{z_\alpha^2}{n}}$$

$$p \geq \frac{0.013 + \frac{(-1.645)^2}{2(200)} - 1.645\sqrt{\frac{0.013(1-0.013)}{200} + \frac{(-1.645)^2}{4(200)^2}}}{1 + \frac{(-1.645)^2}{200}}$$

$$p \geq 0.09577$$

Conclusion from the example

Since the lower bound does not exceed 0.10, in which case it would exceed the hypothesized value, the null hypothesis that the proportion defective is at most 0.10, which was given in the preceding section, would not be rejected if we used the confidence interval to test the hypothesis. Of course a confidence interval has value in its own right and does not have to be used for hypothesis testing.

Exact Intervals for Small Numbers of Failures and/or Small Sample Sizes

Construction of exact two-sided confidence intervals based on the binomial distribution

If the number of failures is very small or if the sample size *N* is very small, symmetical confidence limits that are approximated using the normal distribution may not be accurate enough for some applications. An *exact method* based on the binomial distribution is shown next. To construct a two-sided confidence interval at the 100(1-α)% confidence level for the true proportion defective *p* where N_d defects are found in a sample of size *N* follow the steps below.

1. Solve the equation

$$\sum_{k=0}^{N_d}\binom{N}{k}p_U^k(1-p_U)^{N-k} = \alpha/2$$

   for p_U to obtain the upper 100(1-α)% limit for *p*.

2. Next solve the equation

$$\sum_{k=0}^{N_d-1} \binom{N}{k} p_L^k (1-p_L)^{N-k} = 1-\alpha/2$$

for p_L to obtain the lower 100(1-α)% limit for p.

Note

The interval (p_L, p_U) is an exact 100(1-α)% confidence interval for p. However, it is not symmetric about the observed proportion defective, $\hat{p}=N_d/N$.

Binomial confidence interval example

The equations above that determine p_L and p_U can be solved using readily available functions. Take as an example the situation where twenty units are sampled from a continuous production line and four items are found to be defective. The proportion defective is estimated to be $\hat{p}$ = 4/20 = 0.20. The steps for calculating a 90 % confidence interval for the true proportion defective, p follow.

1. Initalize constants.
   ```
   alpha = 0.10
   Nd = 4
   N = 20
   ```
2. Define a function for upper limit (fu) and a function for the lower limit (fl).
   ```
   fu = F(Nd,pu,20) - alpha/2
   fl = F(Nd-1,pl,20) - (1-alpha/2)
   ```
   F is the cumulative density function for the binominal distribution.
3. Find the value of pu that corresponds to fu = 0 and the value of pl that corresponds to fl = 0 using software to find the roots of a function.

The values of pu and pl for our example are:

```
pu = 0.401029
pl = 0.071354
```

Thus, a 90 % confidence interval for the proportion defective, p, is (0.071, 0.400). Whether or not the interval is truly "exact" depends on the software.

The calculations used in this example can be performed using both Dataplot code and R code.

7.2.4.2. Sample sizes required

Derivation of formula for required sample size when testing proportions

The method of determining sample sizes for testing proportions is similar to the method for determining sample sizes for testing the mean. Although the sampling distribution for proportions actually follows a binomial distribution, the normal approximation is used for this derivation.

Minimum sample size

If we are interested in detecting a change in the proportion defective of size δ in either direction, the minimum sample size is

1. For a two-sided test

$$N \geq \frac{p(1-p)}{\delta^2} z^2_{1-\alpha/2}$$

2. For a one-sided test

$$N \geq \frac{p(1-p)}{\delta^2} z^2_{1-\alpha}$$

Interpretation and sample size for high probability of detecting a change

This requirement on the sample size only guarantees that a change of size δ is detected with 50% probability. The derivation of the sample size when we are interested in protecting against a change δ with probability 1 -β (where β is small) is

1. For a two-sided test

$$N \geq (z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{p(1-p)}{\delta^2}\right)$$

2. For a one-sided test

$$N \geq (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{p(1-p)}{\delta^2}\right)$$

where $z_{1-\beta}$ is the critical value from the normal distribution that is

exceeded with probability β.

Value for the true proportion defective

The equations above require that p be known. Usually, this is not the case. If we are interested in detecting a change relative to an historical or hypothesized value, this value is taken as the value of p for this purpose. Note that taking the value of the proportion defective to be 0.5 leads to the largest possible sample size.

Example of calculating sample size for testing proportion defective

Suppose that a department manager needs to be able to detect any change above 0.10 in the current proportion defective of his product line, which is running at approximately 10% defective. He is interested in a one-sided test and does not want to stop the line except when the process has clearly degraded and, therefore, he chooses a significance level for the test of 5%. Suppose, also, that he is willing to take a risk of 10% of failing to detect a change of this magnitude. With these criteria:

1. $z_{0.95} = 1.645$; $z_{0.90}=1.282$
2. $\delta = 0.10$
3. $p = 0.10$

and the minimum sample size for a one-sided test procedure is

$$N \geq \frac{p(1-p)}{\delta^2}(z_{0.95} + z_{0.90})^2 = \frac{(0.10)(0.90)(2.927)^2}{(0.10)^2} \cong 77$$

7. Product and Process Comparisons
7.2. Comparisons based on data from one process

7.2.5. Does the defect density meet requirements?

Testing defect densities is based on the Poisson distribution

The number of defects observed in an area of size A units is often assumed to have a Poisson distribution with parameter A x D, where D is the actual process defect density (D is defects per unit area). In other words:

$$P\{\#\text{Defects} = n\} = \frac{(AD)^n}{n!}e^{-AD}.$$

The questions of primary interest for quality control are:

1. Is the defect density within prescribed limits?
2. Is the defect density less than a prescribed limit?
3. Is the defect density greater than a prescribed limit?

Normal approximation to the Poisson

We assume that AD is large enough so that the normal approximation to the Poisson applies (in other words, $AD > 10$ for a reasonable approximation and $AD > 20$ for a good one). That translates to

$$P\{\#\text{Defects} < n\} = \Phi\left(\frac{n - AD}{\sqrt{AD}}\right)$$

where Φ is the standard normal distribution function.

Test statistic based on a normal approximation

If, for a sample of area A with a defect density target of D_0, a defect count of C is observed, then the test statistic

$$Z = \frac{C - AD_0}{\sqrt{AD_0}}$$

can be used exactly as shown in the discussion of the test statistic for fraction defectives in the preceding section.

Testing the hypothesis that the process defect density is less than or equal to D_0

For example, after choosing a sample size of area A (see below for sample size calculation) we can reject that the process defect density is less than or equal to the target D_0 if the number of defects C in the sample is greater than C_A, where

$$C_A = z_{1-\alpha}\sqrt{AD_0} + AD_0$$

and $z_{1-\alpha}$ is the 100(1-α) percentile of the standard normal distribution. The test significance level is 100(1-α). For a 90% significance level use $z_{0.90}$ = 1.282 and for a 95% test use $z_{0.95}$ = 1.645. α is the maximum risk that an acceptable process with a defect density at least as low as D_0 "fails" the test.

Choice of sample size (or area) to examine for defects

In order to determine a suitable area A to examine for defects, you first need to choose an unacceptable defect density level. Call this unacceptable defect density $D_1 = kD_0$, where $k > 1$.

We want to have a probability of less than or equal to β is of "passing" the test (and not rejecting the hypothesis that the true level is D_0 or better) when, in fact, the true defect level is D_1 or worse. Typically β will be 0.2, 0.1 or 0.05. Then we need to count defects in a sample size of area A, where A is equal to

$$A = \frac{k}{D_0}\left(\frac{\frac{z_{1-\alpha}}{\sqrt{k}} - z_\beta}{k-1}\right)^2$$

Example

Suppose the target is D_0 = 4 defects per wafer and we want to verify a new process meets that target. We choose α = 0.1 to be the chance of failing the test if the new process is as good as D_0 (α = the Type I error probability or the "producer's risk") and we choose β = 0.1 for the chance of passing the test if the new process is as bad as 6 defects per wafer (β = the Type II error probability or the "consumer's risk"). That means $z_{1-\alpha}$ = 1.282 and z_β = -1.282.

The sample size needed is A wafers, where

$$A = \frac{1.5}{4}\left(\frac{\frac{1.282}{\sqrt{1.5}} - (-1.282)}{k-1}\right)^2 = 8.1$$

which we round up to 9.

The test criteria is to "accept" that the new process meets target unless the number of defects in the sample of 9 wafers exceeds

$$C_A = z_{1-\alpha}\sqrt{AD_0} + AD_0 = 1.282\sqrt{36} + 36 = 43.7$$

In other words, the reject criteria for the test of the new process is 44 or more defects in the sample of 9 wafers.

Note: Technically, all we can say if we run this test and end up *not rejecting* is that we do not have statistically significant evidence that the new process exceeds target. However, the way we chose the sample size for this test assures us we most likely would have had statistically significant evidence for rejection if

the process had been as bad as 1.5 times the target.

7.2.6. What intervals contain a fixed percentage of the population values?

Observations tend to cluster around the median or mean

Empirical studies have demonstrated that it is typical for a large number of the observations in any study to cluster near the median. In right-skewed data this clustering takes place to the left of (i.e., below) the median and in left-skewed data the observations tend to cluster to the right (i.e., above) the median. In symmetrical data, where the median and the mean are the same, the observations tend to distribute equally around these measures of central tendency.

Various methods

Several types of intervals about the mean that contain a large percentage of the population values are discussed in this section.

- Approximate intervals that contain most of the population values
- Percentiles
- Tolerance intervals for a normal distribution
- Tolerance intervals based on the smallest and largest observations

7.2.6.1. Approximate intervals that contain most of the population values

Empirical intervals

A rule of thumb is that where there is no evidence of significant skewness or clustering, two out of every three observations (67%) should be contained within a distance of one standard deviation of the mean; 90% to 95% of the observations should be contained within a distance of two standard deviations of the mean; 99-100% should be contained within a distance of three standard deviations. This rule can help identify outliers in the data.

Intervals that apply to any distribution

The **Bienayme-Chebyshev** rule states that regardless of how the data are distributed, the percentage of observations that are contained within a distance of k tandard deviations of the mean is at least $(1 - 1/k^2)100\%$.

Exact intervals for the normal distribution

The Bienayme-Chebyshev rule is conservative because it applies to any distribution. For a *normal* distribution, a higher percentage of the observations are contained within k standard deviations of the mean as shown in the following table.

Percentage of observations contained between the mean and k standard deviations

k, No. of Standard Deviations	Empircal Rule	Bienayme-Chebychev	Normal Distribution
1	67%	N/A	68.26%
2	90-95%	at least 75%	95.44%
3	99-100%	at least 88.89%	99.73%
4	N/A	at least 93.75%	99.99%

7.2.6.2. Percentiles

Definitions of order statistics and ranks

For a series of measurements Y_1, ..., Y_N, denote the data ordered in increasing order of magnitude by $Y_{[1]}$, ..., $Y_{[N]}$. These ordered data are called order statistics. If $Y_{[j]}$ is the order statistic that corresponds to the measurement Y_i, then the rank for Y_i is j; i.e.,

$$Y_{[j]} \sim Y_i \Rightarrow r_i = j$$

Definition of percentiles

Order statistics provide a way of estimating proportions of the data that should fall above and below a given value, called a *percentile*. The pth percentile is a value, $Y_{(p)}$, such that at most $(100p)$ % of the measurements are less than this value and at most $100(1-p)$ % are greater. The 50th percentile is called the *median*.

Percentiles split a set of ordered data into hundredths. (Deciles split ordered data into tenths). For example, 70 % of the data should fall below the 70th percentile.

Estimation of percentiles

Percentiles can be estimated from N measurements as follows: for the pth percentile, set $p(N+1)$ equal to $k + d$ for k an integer, and d, a fraction greater than or equal to 0 and less than 1.

1. For $0 < k < N$, $Y(p) = Y_{[k]} + d\left(Y_{[k+1]} - Y_{[k]}\right)$
2. For $k = 0$, $Y(p) = Y_{[1]}$
3. For $k = N$, $Y(p) = Y_{[N]}$

Example and interpretation

For the purpose of illustration, twelve measurements from a gage study are shown below. The measurements are resistivities of silicon wafers measured in ohm·cm.

i	Measurements	Order stats	Ranks
1	95.1772	95.0610	9
2	95.1567	95.0925	6
3	95.1937	95.1065	10

4	95.1959	95.1195	11
5	95.1442	95.1442	5
6	95.0610	95.1567	1
7	95.1591	95.1591	7
8	95.1195	95.1682	4
9	95.1065	95.1772	3
10	95.0925	95.1937	2
11	95.1990	95.1959	12
12	95.1682	95.1990	8

To find the 90th percentile, $p(N+1) = 0.9(13) = 11.7$; $k = 11$, and $d = 0.7$. From condition (1) above, $Y(0.90)$ is estimated to be 95.1981 ohm·cm. This percentile, although it is an estimate from a small sample of resistivities measurements, gives an indication of the percentile for a population of resistivity measurements.

Note that there are other ways of calculating percentiles in common use

Some software packages set $1+p(N-1)$ equal to $k + d$, then proceed as above. The two methods give fairly similar results.

A third way of calculating percentiles (given in some elementary textbooks) starts by calculating pN. If that is not an integer, round up to the next highest integer k and use $Y_{[k]}$ as the percentile estimate. If pN is an integer k, use $0.5(Y_{[k]} + Y_{[k+1]})$.

Definition of Tolerance Interval

An interval covering population percentiles can be interpreted as "covering a proportion p of the population with a level of confidence, say, 90 %." This is known as a tolerance interval.

7. Product and Process Comparisons
7.2. Comparisons based on data from one process
7.2.6. What intervals contain a fixed percentage of the population values?

7.2.6.3. Tolerance intervals for a normal distribution

Definition of a tolerance interval

A confidence interval covers a population parameter with a stated confidence, that is, a certain proportion of the time. There is also a way to cover a fixed proportion of the population with a stated confidence. Such an interval is called a *tolerance interval*. The endpoints of a tolerance interval are called *tolerance limits*. An application of tolerance intervals to manufacturing involves comparing specification limits prescribed by the client with tolerance limits that cover a specified proportion of the population.

Difference between confidence and tolerance intervals

Confidence limits are limits within which we expect a given population parameter, such as the mean, to lie. Statistical tolerance limits are limits within which we expect a stated proportion of the population to lie.

Not related to engineering tolerances

Statistical tolerance intervals have a probabilistic interpretation. *Engineering tolerances* are specified outer limits of acceptability which are usually prescribed by a design engineer and do not necessarily reflect a characteristic of the actual measurements.

Three types of tolerance intervals

Three types of questions can be addressed by tolerance intervals. Question (1) leads to a two-sided interval; questions (2) and (3) lead to one-sided intervals.

1. What interval will contain p percent of the population measurements?
2. What interval guarantees that p percent of population measurements will not fall below a lower limit?
3. What interval guarantees that p percent of population measurements will not exceed an upper limit?

Tolerance intervals for measurements from a normal distribution

For the questions above, the corresponding tolerance intervals are defined by lower (L) and upper (U) tolerance limits which are computed from a series of measurements Y_1, ..., Y_N :

1. $Y_L = \bar{Y} - k_2 s; \quad Y_U = \bar{Y} + k_2 s$
2. $Y_L = \bar{Y} - k_1 s$
3. $Y_U = \bar{Y} + k_1 s$

where the k factors are determined so that the intervals cover at least a proportion p of the population with confidence, γ.

Calculation

If the data are from a normally distributed population, an approximate value for the

of k factor for a two-sided tolerance limit for a normal distribution

factor as a function of p and γ for a two-sided tolerance interval (Howe, 1969) is

$$k_2 = \sqrt{\frac{(N-1)\left(1+\frac{1}{N}\right) z^2_{1-(1-p)/2}}{\chi^2_{1-\gamma,N-1}}}$$

where $X^2_{1-\gamma,\, N-1}$, is the critical value of the chi-square distribution with degrees of freedom, N - 1, that is exceeded with probability γ and $z_{1-(1-p)/2}$ is the critical value of the normal distribution which is exceeded with probability (1-p)/2.

Example of calculation

For example, suppose that we take a sample of N = 43 silicon wafers from a lot and measure their thicknesses in order to find tolerance limits within which a proportion p = 0.90 of the wafers in the lot fall with probability γ = 0.99.

Use of tables in calculating two-sided tolerance intervals

Values of the k factor as a function of p and γ are tabulated in some textbooks, such as Dixon and Massey (1969). To use the tables in this handbook, follow the steps outlined below:

1. Calculate $\alpha = (1 - p)/2 = 0.05$
2. Go to the page describing critical values of the normal distribution and in the summary table under the column labeled 0.95 find $z_{1-(1-p)/2} = z_{0.95} = 1.645$.
3. Go to the table of lower critical values of the chi-square distribution and under the column labeled 0.01 in the row labeled degrees of freedom = 42, find
   $X^2_{1-\gamma,\, N-1} = X^2_{0.01,\, 42} = 23.650$.
4. Calculate

$$k_2 = \sqrt{\frac{(N-1)\left(1+\frac{1}{N}\right) z^2_{1-(1-p)/2}}{\chi^2_{1-\gamma,N-1}}} = \sqrt{\frac{42\left(\frac{44}{43}\right)(1.645)^2}{23.650}} = 2.217$$

The tolerance limits are then computed from the sample mean, $\bar{Y}$, and standard deviation, s, according to case(1).

Important notes

The notation for the critical value of the chi-square distribution can be confusing. Values as tabulated are, in a sense, already squared; whereas the critical value for the normal distribution must be squared in the formula above.

Some software is capable of computing a tolerance intervals for a given set of data so that the user does not need to perform all the calculations. All the tolerance intervals shown in this section can be computed using both Dataplot code and R code. R and Dataplot examples include the case where a tolerance interval is computed automatically from a data set.

Calculation of a one-sided tolerance interval for a normal

The calculation of an approximate k factor for one-sided tolerance intervals comes directly from the following set of formulas (Natrella, 1963):

distribution

$$k_1 = \frac{z_p + \sqrt{z_p^2 - ab}}{a}$$

$$a = 1 - \frac{z_\gamma^2}{2(N-1)}$$

$$b = z_p^2 - \frac{z_\gamma^2}{N}$$

A one-sided tolerance interval example

For the example above, it may also be of interest to guarantee with 0.99 probability (or 99% confidence) that 90% of the wafers have thicknesses less than an upper tolerance limit. This problem falls under case (3). The calculations for the k_1 factor for a one-sided tolerance interval are:

$$a = 1 - \frac{1}{2(43-1)}(2.3263)^2 = 0.9356$$

$$b = (1.2816)^2 - \frac{1}{43}(2.3263)^2 = 1.5165$$

$$k_1 = \frac{1.2816 + \sqrt{(1.2816)^2 - (0.9356)(1.5165)}}{0.9356} = 1.8752$$

Tolerance factor based on the non-central t distribution

The value of k_1 can also be computed using the inverse cumulative distribution function for the non-central *t* distribution. This method may give more accurate results for small values of *N*. The value of k_1 using the non-central *t* distribution (using the same example as above) is:

$$\delta = z_p\sqrt{N} = 1.2816\sqrt{43} = 8.4037$$

$$k_1 = \frac{t_{\gamma, N-1, \delta}}{\sqrt{N}} = \frac{12.28834}{\sqrt{43}} = 1.8740$$

where δ is the non-centrality parameter.

In this case, the difference between the two computations is negligble (1.8752 versus 1.8740). However, the difference becomes more pronounced as the value of *N* gets smaller (in particular, for $N \leq 10$). For example, if $N = 43$ is replaced with $N = 6$, the non-central *t* method returns a value of 4.4111 for k_1 while the method based on the Natrella formuals returns a value of 5.2808.

The disadvantage of the non-central *t* method is that it depends on the inverse cumulative distribution function for the non-central *t* distribution. This function is not available in many statistical and spreadsheet software programs, but it is available in Dataplot and R (see Dataplot code and R code). The Natrella formulas only depend on the inverse cumulative distribution function for the normal distribution (which is available in just about all statistical and spreadsheet software programs). Unless you have small samples (say $N \leq 10$), the difference in the

methods should not have much practical effect.

7.2.6.4. Tolerance intervals based on the largest and smallest observations

Tolerance intervals can be constructed for a distribution of any form

The methods on the previous pages for computing tolerance limits are based on the assumption that the measurements come from a normal distribution. If the distribution is not normal, tolerance intervals based on this assumption will not provide coverage for the intended proportion *p* of the population. However, there are methods for achieving the intended coverage if the form of the distribution is not known, but these methods may produce substantially wider tolerance intervals.

Risks associated with making assumptions about the distribution

There are situations where it would be particularly dangerous to make unwarranted assumptions about the exact shape of the distribution, for example, when testing the strength of glass for airplane windshields where it is imperative that a very large proportion of the population fall within acceptable limits.

Tolerance intervals based on largest and smallest observations

One obvious choice for a two-sided tolerance interval for an unknown distribution is the interval between the smallest and largest observations from a sample of Y_1, ..., Y_N measurements. Given the sample size *N* and coverage *p*, an equation from Hahn and Meeker (p. 91),

$$\gamma = 1 - Np^{N-1} + (N-1)p^N$$

allows us to calculate the confidence γ of the tolerance interval. For example, the confidence levels for selected coverages between 0.5 and 0.9999 are shown below for *N* = 25.

Confidence	Coverage
1.000	0.5000
0.993	0.7500
0.729	0.9000
0.358	0.9500
0.129	0.9750
0.026	0.9900
0.007	0.9950
0.0	0.9990
0.0	0.9995
0.0	0.9999

Note that if 99 % confidence is required, the interval that covers the entire sample data set is guaranteed to achieve a coverage of only 75 % of the population values.

What is the optimal sample size?

Another question of interest is, "How large should a sample be so that one can be assured with probability γ that the tolerance interval will contain at least a proportion *p* of the population?"

Approximation for N

A rather good approximation for the required sample size is given by

$$N \cong \frac{1}{4}\frac{(1+p)}{(1-p)}\chi^2_{\gamma,4} + \frac{1}{2}$$

where $X^2_{\gamma,\,4}$ is the critical value of the chi-square distribution with 4 degrees of freedom that is exceeded with probability γ.

Example of the effect of p on the sample size

Suppose we want to know how many measurements to make in order to guarantee that the interval between the smallest and largest observations covers a proportion p of the population with probability γ = 0.95. From the table for the upper critical value of the chi-square distribution, look under the column labeled 0.95 in the row for 4 degrees of freedom. The value is found to be $X^2_{0.95,\,4}$ = 9.488 and calculations are shown below for p equal to 0.90 and 0.99.

For p = 0.90, γ = 0.95:

$$N \cong \frac{1}{4}\frac{(1+0.90)}{(1-0.90)}\chi^2_{0.95,4} + \frac{1}{2} = 0.25(19)(9.488) + 0.5 = 45.57 = 46$$

For p = 0.99, γ = 0.95:

$$N \cong \frac{1}{4}\frac{(1+0.99)}{(1-0.99)}\chi^2_{0.95,4} + \frac{1}{2} = 0.25(199)(9.488) + 0.5 = 472.5 = 473$$

These calculations demonstrate that requiring the tolerance interval to cover a very large proportion of the population may lead to an unacceptably large sample size.

7.3. Comparisons based on data from two processes

Outline for this section

In many manufacturing environments it is common to have two or more processes performing the same task or generating similar products. The following pages describe tests covering several of the most common and useful cases for two processes.

1. Do two processes have the same mean?
    1. Tests when the standard deviations are equal
    2. Tests when the standard deviations are unequal
    3. Tests for paired data
2. Do two processes have the same standard deviation?
3. Do two processes produce the same proportion of defectives?
4. If the observations are failure times, are the failure rates (or mean times to failure) the same?
5. Do two arbitrary processes have the same central tendency?

Example of a dual track process

For example, in an automobile manufacturing plant, there may exist several assembly lines producing the same part. If one line goes down for some reason, parts can still be produced and production will not be stopped. For example, if the parts are piston rings for a particular model car, the rings produced by either line should conform to a given set of specifications.

How does one confirm that the two processes are in fact producing rings that are similar? That is, how does one determine if the two processes are similar?

The goal is to determine if the two processes are similar

In order to answer this question, data on piston rings are collected for each process. For example, on a particular day, data on the diameters of ten piston rings from each process are measured over a one-hour time frame.

To determine if the two processes are similar, we are interested in answering the following questions:

1. Do the two processes produce piston rings with the same diameter?
2. Do the two processes have similar variability in the

diameters of the rings produced?

Unknown standard deviation

The second question assumes that one does not know the standard deviation of either process and therefore it must be estimated from the data. This is usually the case, and the tests in this section assume that the population standard deviations are unknown.

Assumption of a normal distribution

The statistical methodology used (i.e., the specific test to be used) to answer these two questions depends on the underlying distribution of the measurements. The tests in this section assume that the data are normally distributed.

7.3.1. Do two processes have the same mean?

Testing hypotheses related to the means of two processes

Given two random samples of measurements,

$$Y_1, ..., Y_N \text{ and } Z_1, ..., Z_N$$

from two independent processes (the Y's are sampled from process 1 and the Z's are sampled from process 2), there are three types of questions regarding the true means of the processes that are often asked. They are:

1. Are the means from the two processes the same?
2. Is the mean of process 1 less than or equal to the mean of process 2?
3. Is the mean of process 1 greater than or equal to the mean of process 2?

Typical null hypotheses

The corresponding null hypotheses that test the true mean of the first process, μ_1, against the true mean of the second process, μ_2 are:

1. H_0: $\mu_1 = \mu_2$
2. H_0: μ_1 < or equal to μ_2
3. H_0: μ_1 > or equal to μ_2

Note that as previously discussed, our choice of which null hypothesis to use is typically made based on one of the following considerations:

1. When we are hoping to prove something new with the sample data, we make that the alternative hypothesis, whenever possible.
2. When we want to continue to assume a reasonable or traditional hypothesis still applies, unless very strong contradictory evidence is present, we make that the null hypothesis, whenever possible.

Basic statistics from the two processes

The basic statistics for the test are the sample means

$$\overline{Y} = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i \; ; \; \overline{Z} = \frac{1}{N_2}\sum_{i=1}^{N_2} Z_i$$

and the sample standard deviations

$$s_1 = \sqrt{\frac{\sum_{i=1}^{N_1}(Y_i - \bar{Y})^2}{N_1 - 1}}$$

$$s_2 = \sqrt{\frac{\sum_{i=1}^{N_2}(Z_i - \bar{Z})^2}{N_2 - 1}}$$

with degrees of freedom $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$ respectively.

Form of the test statistic where the two processes have equivalent standard deviations

If the standard deviations from the two processes are equivalent, and this should be tested before this assumption is made, the test statistic is

$$t = \frac{\bar{Y} - \bar{Z}}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

where the pooled standard deviation is estimated as

$$s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)}}$$

with degrees of freedom $\nu = N_1 + N_2 - 2$.

Form of the test statistic where the two processes do NOT have equivalent standard deviations

If it cannot be assumed that the standard deviations from the two processes are equivalent, the test statistic is

$$t = \frac{\bar{Y} - \bar{Z}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

The degrees of freedom are not known exactly but can be estimated using the Welch-Satterthwaite approximation

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

Test strategies

The strategy for testing the hypotheses under (1), (2) or (3) above is to calculate the appropriate *t* statistic from one of the formulas above, and then perform a test at significance level α, where α is chosen to be small, typically .01, .05 or .10. The hypothesis associated with each case enumerated above is rejected if:

1. $|t| \geq t_{1-\alpha/2,\,\nu}$
2. $t \geq t_{1-\alpha,\,\nu}$
3. $t \leq t_{\alpha,\,\nu}$

Explanation of critical values

The critical values from the *t* table depend on the significance level and the degrees of freedom in the standard deviation. For hypothesis (1) $t_{1-\alpha/2,\,\nu}$ is the 1-α/2 critical value from the *t* table with ν degrees of freedom and similarly for hypotheses (2) and (3).

Example of unequal number of data points

A new procedure (process 2) to assemble a device is introduced and tested for possible improvement in time of assembly. The question being addressed is whether the mean, μ_2, of the new assembly process is smaller than the mean, μ_1, for the old assembly process (process 1). We choose to test hypothesis (2) in the hope that we will reject this null hypothesis and thereby feel we have a strong degree of confidence that the new process is an improvement worth implementing. Data (in minutes required to assemble a device) for both the new and old processes are listed below along with their relevant statistics.

Device	Process 1 (Old)	Process 2 (New)
1	32	36
2	37	31
3	35	30
4	28	31
5	41	34
6	44	36
7	35	29
8	31	32
9	34	31
10	38	
11	42	
Mean	36.0909	32.2222
Standard deviation	4.9082	2.5386
No. measurements	11	9
Degrees freedom	10	8

Computation of the test statistic

From this table we generate the test statistic

$$t = \frac{\bar{Y} - \bar{Z}}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} = \frac{36.0909 - 32.2222}{\sqrt{4.9082^2/11 + 2.5386^2/9}} = 2.2694$$

with the degrees of freedom approximated by

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}} = \frac{\left(\frac{4.9082^2}{11} + \frac{2.5386^2}{9}\right)^2}{\frac{4.9082^4}{1210} + \frac{2.5386^4}{648}} = 15.5$$

Decision process

For a one-sided test at the 5% significance level, go to the *t* table for 0.95 signficance level, and look up the critical value for degrees of

freedom v = 16. The critical value is 1.746. Thus, hypothesis (2) is rejected because the test statistic (t = 2.269) is greater than 1.746 and, therefore, we conclude that process 2 has improved assembly time (smaller mean) over process 1.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

7.3.1.1. Analysis of paired observations

Definition of paired comparisons

Given two random samples,

$$\boldsymbol{Y_1}, ..., \boldsymbol{Y_N} \quad \text{and} \quad \boldsymbol{Z_1}, ..., \boldsymbol{Z_N}$$

from two populations, the data are said to be paired if the *i*th measurement on the first sample is naturally paired with the *i*th measurement on the second sample. For example, if *N* supposedly identical products are chosen from a production line, and each one, in turn, is tested with first one measuring device and then with a second measuring device, it is possible to decide whether the measuring devices are compatible; i.e., whether there is a difference between the two measurement systems. Similarly, if "before" and "after" measurements are made with the same device on *N* objects, it is possible to decide if there is a difference between "before" and "after"; for example, whether a cleaning process changes an important characteristic of an object. Each "before" measurement is paired with the corresponding "after" measurement, and the differences

$$d_i = Y_i - X_i \quad (i = 1, \cdots, N)$$

are calculated.

Basic statistics for the test

The mean and standard deviation for the differences are calculated as

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

and

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}$$

with $v = N - 1$ degrees of freedom.

Test statistic based on the t

The paired-sample *t* test is used to test for the difference of two means before and after a treatment. The test statistic is:

distribution

$$t = \frac{\bar{d}}{s_d / \sqrt{N}}$$

The hypotheses described on the foregoing page are rejected if:

1. $|t| \geq t_{1-\alpha/2, \nu}$
2. $t \geq t_{1-\alpha, \nu}$
3. $t \leq t_{\alpha, \nu}$

where for hypothesis (1) $t_{1-\alpha/2, \nu}$ is the 1-α/2 critical value from the t distribution with ν degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the t table in Chapter 1.

7.3.1.2. Confidence intervals for differences between means

Definition of confidence interval for difference between population means

Given two random samples,

$$Y_1, ..., Y_N \quad \text{and} \quad Z_1, ..., Z_N$$

from two populations, two-sided confidence intervals with 100 (1- α)% coverage for the difference between the unknown population means, μ_1 and μ_2, are shown in the table below. Relevant statistics for paired observations and for unpaired observations are shown elsewhere.

Two-sided confidence intervals with 100(1- α)% coverage for μ_1 - μ_2:

Paired observations

$\mu_1 - \mu_2$ (where $\sigma_1 = \sigma_2$)	$\bar{d} \pm t_{1-\alpha/2,N-1} \frac{s_d}{\sqrt{N}}$

Unpaired observations

$\mu_1 - \mu_2$ (where $\sigma_1 = \sigma_2$)	$\bar{Y} - \bar{Z} \pm t_{1-\alpha/2,N_1+N_2-2}\ s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$
$\mu_1 - \mu_2$ (where $\sigma_1 \neq \sigma_2$)	$\bar{Y} - \bar{Z} \pm t_{1-\alpha/2,effective\ df}\ s\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$

Interpretation of confidence interval

One interpretation of the confidence interval for means is that if zero is contained within the confidence interval, the two population means are equivalent.

7.3.2. Do two processes have the same standard deviation?

Testing hypotheses related to standard deviations from two processes

Given two random samples of measurements,

$$Y_1, ..., Y_N \quad \text{and} \quad Z_1, ..., Z_N$$

from two independent processes, there are three types of questions regarding the true standard deviations of the processes that can be addressed with the sample data. They are:

1. Are the standard deviations from the two processes the same?
2. Is the standard deviation of one process less than the standard deviation of the other process?
3. Is the standard deviation of one process greater than the standard deviation of the other process?

Typical null hypotheses

The corresponding null hypotheses that test the true standard deviation of the first process, σ_1, against the true standard deviation of the second process, σ_2 are:

1. H_0: $\sigma_1 = \sigma_2$
2. H_0: $\sigma_1 \leq \sigma_2$
3. H_0: $\sigma_1 \geq \sigma_2$

Basic statistics from the two processes

The basic statistics for the test are the sample variances

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N1} (Y_i - \bar{Y})^2$$

$$s_2^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N2} (Z_i - \bar{Z})^2$$

and degrees of freedom $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$, respectively.

Form of the test statistic

The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

Test strategies

The strategy for testing the hypotheses under (1), (2) or (3) above is to calculate the *F* statistic from the formula above, and then perform a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10. The hypothesis associated with each case enumerated above is rejected if:

1. $F \leq \frac{1}{F_{\alpha/2;\nu 2;\nu 1}}$ or $F \geq F_{\alpha/2;\nu 1;\nu 2}$
2. $F \geq F_{\alpha;\nu 1;\nu 2}$
3. $F \leq \frac{1}{F_{\alpha;\nu 2;\nu 1}}$

Explanation of critical values

The critical values from the *F* table depend on the significance level and the degrees of freedom in the standard deviations from the two processes. For hypothesis (1):

- $F_{\alpha/2;\nu_2;\nu_1}$ is the upper critical value from the F table with
- $\nu_2 = N_2 - 1$ degrees of freedom for the numerator and
- $\nu_1 = N_1 - 1$ degrees of freedom for the denominator

and

- $F_{\alpha/2;\nu_1;\nu_2}$ is the upper critical value from the F table with
- $\nu_1 = N_1 - 1$ degrees of freedom for the numerator and
- $\nu_2 = N_2 - 1$ degrees of freedom for the denominator.

***Caution** on looking up critical values*

The *F* distribution has the property that

$$F_{1-\alpha/2;\nu_1;\nu_2} = \frac{1}{F_{\alpha/2;\nu_2;\nu_1}}$$

which means that only upper critical values are required for two-sided tests. However, note that the degrees of freedom are interchanged in the ratio. For example, for a two-sided test at significance level 0.05, go to the *F* table labeled "2.5% significance level".

- For $F_{\alpha/2;\nu_2;\nu_1}$, reverse the order of the degrees of freedom; i.e., look across the top of the table for and down the table for .

$\nu_2 = N_2 - 1$ $\nu_1 = N_1 - 1$

- For $F_{\alpha/2;\nu_1;\nu_2}$, look across the top of the table for $\nu_1 = N_1 - 1$ and down the table for $\nu_2 = N_2 - 1$.

Critical values for cases (2) and (3) are defined similarly, except that the critical values for the one-sided tests are based on α rather than on $\alpha/2$.

Two-sided confidence interval

The two-sided confidence interval for the ratio of the two unknown variances (squares of the standard deviations) is shown below.

Two-sided confidence interval with 100(1- α)% coverage for:

$$\frac{\sigma_1^2}{\sigma_2^2} \quad \left[\frac{1}{F_{\alpha/2;N1-1;N2-1}} \left(\frac{s_1^2}{s_2^2} \right), F_{\alpha/2;N2-1;N1-1} \left(\frac{s_1^2}{s_2^2} \right) \right]$$

One interpretation of the confidence interval is that if the quantity "one" is contained within the interval, the standard deviations are equivalent.

Example of unequal number of data points

A new procedure to assemble a device is introduced and tested for possible improvement in time of assembly. The question being addressed is whether the standard deviation, σ_2, of the new assembly process is better (i.e., smaller) than the standard deviation, σ_1, for the old assembly process. Therefore, we test the null hypothesis that $\sigma_1 \leq \sigma_2$. We form the hypothesis in this way because we hope to reject it, and therefore accept the alternative that σ_2 is less than σ_1. This is hypothesis (2). Data (in minutes required to assemble a device) for both the old and new processes are listed on an earlier page. Relevant statistics are shown below:

	Process 1	Process 2
Mean	36.0909	32.2222
Standard deviation	4.9082	2.5874
No. measurements	11	9
Degrees freedom	10	8

Computation of the test statistic

From this table we generate the test statistic

$$F = \frac{s_1^2}{s_2^2} = \left(\frac{4.9082}{2.5874} \right)^2 = 3.60$$

Decision process

For a test at the 5% significance level, go to the *F* table for 5% signficance level, and look up the critical value for numerator degrees of freedom $\nu_1 = N_1 - 1$ = 10 and denominator degrees of freedom $\nu_2 = N_2 - 1$ = 8. The critical value is 3.35. Thus, hypothesis (2) can be rejected because

the test statistic (F = 3.60) is greater than 3.35. Therefore, we accept the alternative hypothesis that process 2 has better precision (smaller standard deviation) than process 1.

7.3.3. How can we determine whether two processes produce the same proportion of defectives?

Case 1: Large Samples (Normal Approximation to Binomial)

The hypothesis of equal proportions can be tested using a z statistic

If the samples are reasonably large we can use the normal approximation to the binomial to develop a test similar to testing whether two normal means are equal.

Let sample 1 have x_1 defects out of n_1 and sample 2 have x_2 defects out of n_2. Calculate the proportion of defects for each sample and the z statistic below:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

Compare $|z|$ to the normal $z_{1-\alpha/2}$ table value for a two-sided test. For a one-sided test, assuming the alternative hypothesis is $p_1 > p_2$, compare z to the normal $z_{1-\alpha}$ table value. If the alternative hypothesis is $p_1 < p_2$, compare z to z_{α}.

Case 2: An Exact Test for Small Samples

The Fisher Exact Probability test is an excellent choice for small samples

The **Fisher Exact Probability Test** is an excellent nonparametric technique for analyzing discrete data (either nominal or ordinal), when the two independent samples are small in size. It is used when the results from two independent random samples fall into one or the other of two mutually exclusive classes (i.e., defect versus good, or successes vs failures).

Example of a 2x2

In other words, every subject in each group has one of two possible scores. These scores are represented by frequencies

contingency table

in a 2x2 contingency table. The following discussion, using a 2x2 contingency table, illustrates how the test operates.

We are working with two independent groups, such as experiments and controls, males and females, the Chicago Bulls and the New York Knicks, etc.

	-	+	Total
Group I	A	B	A+B
Group II	C	D	C+D
Total	A+C	B+D	N

The column headings, here arbitrarily indicated as plus and minus, may be of any two classifications, such as: above and below the median, passed and failed, Democrat and Republican, agree and disagree, etc.

Determine whether two groups differ in the proportion with which they fall into two classifications

Fisher's test determines whether the two groups differ in the proportion with which they fall into the two classifications. For the table above, the test would determine whether Group I and Group II differ significantly in the proportion of plusses and minuses attributed to them.

The method proceeds as follows:

The exact probability of observing a particular set of frequencies in a 2 × 2 table, when the marginal totals are regarded as fixed, is given by the hypergeometric distribution

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$$

$$= \frac{\frac{(A+C)!}{A!C!}\frac{(B+D)!}{B!D!}}{\frac{N!}{(A+B)!(C+D)!}}$$

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

But the test does not just look at the observed case. If needed, it also computes the probability of more extreme outcomes, *with the same marginal totals.* By "more extreme", we mean relative to the null hypothesis of equal proportions.

Example of Fisher's test

This will become clear in the next illustrative example. Consider the following set of 2 x 2 contingency tables:

Observed Data			More extreme outcomes with same marginals					
(a)			(b)			(c)		
2	5	7	1	6	7	0	7	7
3	2	5	4	1	5	5	0	5
5	7	12	5	7	12	5	7	12

Table (a) shows the observed frequencies and tables (b) and (c) show the two more extreme distributions of frequencies that could occur with the same marginal totals 7, 5. Given the observed data in table (a) , we wish to test the null hypothesis at, say, $\alpha = 0.05$.

Applying the previous formula to tables (a), (b), and (c), we obtain

$$p_a = \frac{7!5!5!7!}{12!2!5!3!2!} = .26515$$

$$p_b = \frac{7!5!5!7!}{12!1!6!4!1!} = .04419$$

$$p_c = \frac{7!5!5!7!}{12!0!7!5!0!} = .00126$$

The probability associated with the occurrence of values as extreme as the observed results under H_0 is given by adding these three p's:

$$.26515 + .04419 + .00126 = .31060$$

So $p = 0.31060$ is the probability that we get from Fisher's test. Since 0.31060 is larger than α, we cannot reject the null hypothesis.

Tocher's Modification

Tocher's modification makes Fisher's test less conservative

Tocher (1950) showed that a slight modification of the Fisher test makes it a more useful test. Tocher starts by isolating the probability of all cases more extreme than the observed one. In this example that is

$$p_b + p_c = .04419 + .00126 = .04545$$

Now, if this probability is larger than α, we cannot reject H_0. But if this probability is less than α, while the probability that we got from Fisher's test is greater than α

(as is the case in our example) then Tocher advises to compute the following ratio:

$$\frac{\alpha - p_{\text{more extreme cases}}}{p_{\text{observed alone}}}$$

For the data in the example, that would be

$$\frac{\alpha - (p_b + p_c)}{p_a} = \frac{.05 - .04545}{.2615} = .0172$$

Now we go to a table of random numbers and at random draw a number between 0 and 1. If this random number is *smaller* than the ratio above of 0.0172, we reject H_0. If it is larger we cannot reject H_0. This added small probability of rejecting H_0 brings the test procedure Type I error (i.e., α value) to exactly 0.05 and makes the Fisher test less conservative.

The test is a one-tailed test. For a two-tailed test, the value of *p* obtained from the formula must be doubled.

A difficulty with the Tocher procedure is that someone else analyzing the same data would draw a different random number and possibly make a different decision about the validity of H_0.

7.3.4. Assuming the observations are failure times, are the failure rates (or Mean Times To Failure) for two distributions the same?

Comparing two exponential distributions is to compare the means or hazard rates

The comparison of two (or more) life distributions is a common objective when performing statistical analyses of lifetime data. Here we look at the one-parameter exponential distribution case.

In this case, comparing two exponential distributions is equivalent to comparing their means (or the reciprocal of their means, known as their hazard rates).

Type II Censored data

Definition of Type II censored data

Definition: Type II censored data occur when a life test is terminated exactly when a pre-specified number of failures have occurred. The remaining units have not yet failed. If n units were on test, and the pre-specified number of failures is r (where r is less than or equal to n), then the test ends at t_r = the time of the r-th failure.

Two exponential samples oredered by time

Suppose we have Type II censored data from two exponential distributions with means θ_1 and θ_2. We have two samples from these distributions, of sizes n_1 on test with r_1 failures and n_2 on test with r_2 failures, respectively. The observations are time to failure and are therefore ordered by time.

$$t_{1(1)} < \cdots < t_{1(r_1)} \quad (r_1 \le n_1)$$
$$t_{2(1)} < \cdots < t_{2(r_2)} \quad (r_2 \le n_2)$$

Test of equality of θ_1 and θ_2 and confidence

Letting

$$T_i = \sum_{j=1}^{r_i} t_{i(j)} + (n_i - r_i)t_{i(r_i)} \qquad i = 1, 2$$

interval for θ_1/θ_2

Then

$$2T_1/\theta_1 \approx \chi^2_{2r1}$$

and

$$2T_2/\theta_2 \approx \chi^2_{2r2}$$

with T_1 and T_2 independent. Thus

$$U = \frac{2T_1/(2r_1\theta_1)}{2T_2/(2r_2\theta_2)} = \frac{\hat{\theta}_1\theta_2}{\hat{\theta}_2\theta_1},$$

where

$$\hat{\theta}_1 = \frac{T_1}{r_1} \text{ and } \hat{\theta}_2 = \frac{T_2}{r_2}$$

has an F distribution with $(2r_1, 2r_2)$ degrees of freedom. Tests of equality of θ_1 and θ_2 can be performed using tables of the F distribution or computer programs. Confidence intervals for θ_1 / θ_2, which is the ratio of the means or the hazard rates for the two distributions, are also readily obtained.

Numerical example

A numerical application will illustrate the concepts outlined above.

For this example,

H_0: $\theta_1/\theta_2 = 1$

H_a: $\theta_1/\theta_2 \neq 1$

Two samples of size 10 from exponential distributions were put on life test. The first sample was censored after 7 failures and the second sample was censored after 5 failures. The times to failure were:

Sample 1: 125 189 210 356 468 550 610
Sample 2: 170 234 280 350 467

So $r_1 = 7$, $r_2 = 5$ and $t_{1,(r1)} = 610$, $t_{2,(r2)}=467$.

Then $T_1 = 4338$ and $T_2 = 3836$.

The estimator for θ_1 is 4338 / 7 = 619.71 and the estimator for θ_2 is 3836 / 5 = 767.20.

The ratio of the estimators = U = 619.71 / 767.20 = .808.

If the means are the same, the ratio of the estimators, U, follows an F distribution with $2r_1$, $2r_2$ degrees of freedom. The $P(F < .808) = .348$. The associated p-value is $2(.348) = .696$. Based on this p-value, we find no evidence to reject the null hypothesis (that the true but unknown ratio = 1). Note that this is a two-sided test, and we would reject the null hyposthesis if the p-value is either too small (i.e., less or equal to .025) or too large (i.e., greater than or equal to .975) for a 95% significance level test.

We can also put a 95% confidence interval around the ratio of the two means. Since the .025 and .975 quantiles of $F_{(14,10)}$ are 0.3178 and 3.5504, respectively, we have

$$\Pr(U/3.5504 < \theta_1/\theta_2 < U/.3178) = .95$$

and (.228, 2.542) is a 95% confidence interval for the ratio of the unknown means. The value of 1 is within this range, which is another way of showing that we cannot reject the null hypothesis at the 95% significance level.

7. Product and Process Comparisons
7.3. Comparisons based on data from two processes

7.3.5. Do two arbitrary processes have the same central tendency?

The nonparametric equivalent of the t test is due to Mann and Whitney, called the U test

By "arbitrary" we mean that we make no underlying assumptions about normality or any other distribution. The test is called the **Mann-Whitney U Test**, which is the nonparametric equivalent of the *t* test for means.

The *U*-test (as the majority of nonparametric tests) uses the rank sums of the two samples.

Procedure

The test is implemented as follows.

1. Rank all (n_1 + n_2) observations in ascending order. Ties receive the average of their observations.
2. Calculate the sum of the ranks, call these T_a and T_b
3. Calculate the U statistic,

$$U_a = n_1(n_2) + 0.5(n_1)(n_1 + 1) - T_a$$

or

$$U_b = n_1(n_2) + 0.5(n_2)(n_2 + 1) - T_b$$

where $U_a + U_b = n_1(n_2)$.

Null Hypothesis

The null hypothesis is: the two populations have the same central tendency. The alternative hypothesis is: The central tendencies are **NOT** the same.

Test statistic

The test statistic, *U*, is the smaller of U_a and U_b. For sample sizes larger than 20, we can use the normal *z* as follows:

$$z = [\ U - E(U)] / \sigma$$

where

$$E(U) = .5(n_1)(n_2) \text{ and } \sigma^2 = \left[n_1(n_2)(n_1 + n_2 + 1)\right] / 12$$

The critical value is the normal tabled *z* for $\alpha/2$ for a two-tailed test or *z* at α

level, for a one-tail test.

For small samples, tables are readily available in most textbooks on nonparametric statistics.

Example

An illustrative example of the U test

Two processing systems were used to clean wafers. The following data represent the (coded) particle counts. The null hypothesis is that there is no difference between the central tendencies of the particle counts; the alternative hypothesis is that there is a difference. The solution shows the typical kind of output software for this procedure would generate, based on the large sample approximation.

Group A	Rank	Group B	Rank
.55	8	.49	5
.67	15.5	.68	17
.43	1	.59	9.5
.51	6	.72	19
.48	3.5	.67	15.5
.60	11	.75	20.5
.71	18	.65	13.5
.53	7	.77	22
.44	2	.62	12
.65	13.5	.48	3.5
.75	20.5	.59	9.5

	N	Sum of Ranks	U	Std. Dev of U	Median
A	11	106.000	81.000	15.229	0.540
B	11	147.000	40.000	15.229	0.635

For U = 40.0 and E[U] = 0.5(n_1)(n_2) = 60.5, the test statistic is

$$z = \frac{U - \mathrm{E}(U)}{\sigma} = \frac{40.0 - 60.5}{15.23} = -1.346$$

where

$$\sigma = \sqrt{\frac{n_1(n_2)(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{11(11)(11 + 11 + 1)}{12}} = 15.23$$

For a two-sided test with significance level α = 0.05, the critical value is $z_{1-\alpha/2}$ = 1.96. Since |z| is less than the critical value, we do not reject the null

hypothesis and conclude that there is not enough evidence to claim that two groups have different central tendencies.

7. Product and Process Comparisons

7.4. Comparisons based on data from more than two processes

Introduction This section begins with a nonparametric procedure for comparing several populations with unknown distributions. Then the following topics are discussed:

- Comparing variances
- Comparing means (ANOVA technique)
- Estimating variance components
- Comparing categorical data
- Comparing population proportion defectives
- Making multiple comparisons

7.4.1. How can we compare several populations with unknown distributions (the Kruskal-Wallis test)?

The Kruskal-Wallis (KW) Test for Comparing Populations with Unknown Distributions

A nonparametric test for comparing population medians by Kruskal and Wallis

The KW procedure tests the null hypothesis that k samples from possibly different populations actually originate from similar populations, at least as far as their central tendencies, or medians, are concerned. The test assumes that the variables under consideration have underlying continuous distributions.

In what follows assume we have k samples, and the sample size of the i-th sample is n_i, $i = 1, 2, \ldots, k$.

Test based on ranks of combined data

In the computation of the KW statistic, each observation is replaced by its rank in an ordered combination of all the k samples. By this we mean that the data from the k samples combined are ranked in a single series. The minimum observation is replaced by a rank of 1, the next-to-the-smallest by a rank of 2, and the largest or maximum observation is replaced by the rank of N, where N is the total number of observations in all the samples (N is the sum of the n_i).

Compute the sum of the ranks for each sample

The next step is to compute the sum of the ranks for each of the original samples. The KW test determines whether these sums of ranks are so different by sample that they are not likely to have all come from the same population.

Test statistic follows a χ^2 distribution

It can be shown that if the k samples come from the same population, that is, if the null hypothesis is true, then the test statistic, H, used in the KW procedure is distributed approximately as a chi-square statistic with $df = k - 1$, provided that the sample sizes of the k samples are not too small (say, $n_i > 4$, for all i). H is defined as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where

- k = number of samples (groups)
- n_i = number of observations for the i-th sample or group
- N = total number of observations (sum of all the n_i)
- R_i = sum of ranks for group i

Example

An illustrative example

The following data are from a comparison of four investment firms. The observations represent percentage of growth during a three month period.for recommended funds.

A	B	C	D
4.2	3.3	1.9	3.5
4.6	2.4	2.4	3.1
3.9	2.6	2.1	3.7
4.0	3.8	2.7	4.1
	2.8	1.8	4.4

Step 1: Express the data in terms of their ranks

	A	B	C	D
	17	10	2	11
	19	4.5	4.5	9
	14	6	3	12
	15	13	7	16
		8	1	18
SUM	65	41.5	17.5	66

Compute the test statistic

The corresponding H test statistic is

$$H = \frac{12}{(19)(20)}\left[\frac{65^2}{4} + \frac{41.5^2}{5} + \frac{17.5^2}{5} + \frac{66^2}{5}\right] - 3(20) = 13.678$$

From the chi-square table in Chapter 1, the critical value for $1-\alpha = 0.95$ with $df = k-1 = 3$ is 7.812. Since 13.678 > 7.812, we reject the null hypothesis.

Note that the rejection region for the KW procedure is one-sided, since we only reject the null hypothesis when the H statistic is too large.

7.4.2. Assuming the observations are normal, do the processes have the same variance?

Before comparing means, test whether the variances are equal

Techniques for comparing means of normal populations generally assume the populations have the same variance. Before using these ANOVA techniques, it is advisable to test whether this assumption of homogeneity of variance is reasonable. The following procedure is widely used for this purpose.

Bartlett's Test for Homogeneity of Variances

Null hypothesis

Bartlett's test is a commonly used test for equal variances. Let's examine the null and alternative hypotheses.

$$H_0 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

against

$$H_a = \text{the } \sigma_i^2 \text{ are not all equal}$$

Test statistic

Assume we have samples of size n_i from the i-th population, $i = 1, 2, \ldots, k$, and the usual variance estimates from each sample:

$$s_1^2, s_2^2, \ldots, s_k^2$$

where

$$s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2/(n_j - 1)$$

Now introduce the following notation: $\nu_j = n_j - 1$ (the ν_j are the degrees of freedom) and

$$\nu = \sum_{i=1}^{k} \nu_i$$

$$s^2 = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{\nu}$$

The Bartlett's test statistic M is defined by

$$M = \nu \log s^2 - \sum_{i=1}^{k} \nu_i \log s_i^2$$

Distribution of the test statistic

When none of the degrees of freedom is small, Bartlett showed that M is distributed approximately as χ^2_{k-1}. The chi-square approximation is generally acceptable if all the n_i are at least 5.

Bias correction

This is a slightly biased test, according to Bartlett. It can be improved by dividing M by the factor

$$C = 1 + \frac{1}{3(k-1)}\left(\left[\sum_{i=1}^{k} \frac{1}{\nu_i}\right] - \frac{1}{\nu}\right)$$

Instead of M, it is suggested to use M/C for the test statistic.

Bartlett's test is not robust

This test is not robust, it is very sensitive to departures from normality.

An alternative description of Bartlett's test appears in Chapter 1.

Gear Data Example (from Chapter 1):

An illustrative example of Bartlett's test

Gear diameter measurements were made on 10 batches of product. The complete set of measurements appears in Chapter 1. Bartlett's test was applied to this dataset leading to a rejection of the assumption of equal batch variances at the .05 critical value level. applied to this dataset

The Levene Test for Homogeneity of Variances

The Levene test for equality of variances

Levene's test offers a more robust alternative to Bartlett's procedure. That means it will be less likely to reject a true hypothesis of equality of variances just because the distributions of the sampled populations are not normal. When non-normality is suspected, Levene's procedure is a better choice than Bartlett's.

Levene's test is described in Chapter 1. This description also includes an example where the test is applied to the gear data. Levene's test does not reject the assumption of equality of batch variances for these data. This differs from the conclusion drawn from Bartlett's test and is a better answer if, indeed, the batch population distributions are non-normal.

7.4.3. Are the means equal?

Test equality of means

The procedure known as the *Analysis of Variance* or *ANOVA* is used to test hypotheses concerning means when we have several populations.

The Analysis of Variance (ANOVA)

The ANOVA procedure is one of the most powerful statistical techniques

ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed.

A couple of questions come immediately to mind: **what** means? and why analyze **variances** in order to derive conclusions about the **means**?

Both questions will be answered as we delve further into the subject.

Introduction to ANOVA

To begin, let us study the effect of temperature on a passive component such as a resistor. We select three different temperatures and observe their effect on the resistors. This experiment can be conducted by measuring all the participating resistors before placing *n* resistors each in three different ovens.

Each oven is heated to a selected temperature. Then we measure the resistors again after, say, 24 hours and analyze the responses, which are the differences between before and after being subjected to the temperatures. The temperature is called a *factor*. The different temperature settings are called *levels*. In this example there are three levels or settings of the factor Temperature.

What is a factor?

A factor is an independent treatment variable whose settings (values) are controlled and varied by the experimenter. The intensity setting of a factor is the level.

- **Levels may be quantitative numbers or, in many cases, simply "present" or "not present" ("0" or "1").**

The 1-way

In the experiment above, there is only one factor,

ANOVA

temperature, and the analysis of variance that we will be using to analyze the effect of temperature is called a *one-way* or *one-factor ANOVA*.

The 2-way or 3-way ANOVA

We could have opted to also study the effect of positions in the oven. In this case there would be two factors, temperature and oven position. Here we speak of a *two-way* or *two-factor ANOVA*. Furthermore, we may be interested in a third factor, the effect of time. Now we deal with a *three-way* or *three-factor* ANOVA. In each of these ANOVA's we test a variety of hypotheses of equality of means (or average responses when the factors are varied).

Hypotheses that can be tested in an ANOVA

First consider the one-way ANOVA. The null hypothesis is: there is no difference in the population means of the different levels of factor A (the only factor).

The alternative hypothesis is: the means are not the same.

For the 2-way ANOVA, the possible null hypotheses are:

1. There is no difference in the means of factor A
2. There is no difference in means of factor B
3. There is no interaction between factors A and B

The alternative hypothesis for cases 1 and 2 is: the means are not equal.

The alternative hypothesis for case 3 is: there is an interaction between A and B.

For the 3-way ANOVA: The main effects are factors A, B and C. The 2-factor interactions are: AB, AC, and BC. There is also a three-factor interaction: ABC.

For each of the seven cases the null hypothesis is the same: there is no difference in means, and the alternative hypothesis is the means are not equal.

The n-way ANOVA

In general, the number of main effects and interactions can be found by the following expression:

$$N = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}$$

The first term is for the overall mean, and is always 1. The second term is for the number of main effects. The third term is for the number of 2-factor interactions, and so on. The last term is for the *n*-factor interaction and is always 1.

In what follows, we will discuss only the 1-way and 2-way ANOVA.

7.4.3.1. 1-Way ANOVA overview

Overview and principles

This section gives an overview of the one-way ANOVA. First we explain the principles involved in the 1-way ANOVA.

Partition response into components

In an analysis of variance the variation in the response measurements is partitoned into components that correspond to different sources of variation.

The goal in this procedure is to split the total variation in the data into a portion due to random error and portions due to changes in the values of the independent variable(s).

Variance of n measurements

The variance of *n* measurements is given by

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

where $\bar{y}$ is the mean of the *n* measurements.

Sums of squares and degrees of freedom

The numerator part is called the *sum of squares* of deviations from the mean, and the denominator is called the *degrees of freedom*.

The variance, after some algebra, can be rewritten as:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n}{n-1}$$

The first term in the numerator is called the "*raw sum of squares*" and the second term is called the "*correction term for the mean*". Another name for the numerator is the "*corrected sum of squares*", and this is usually abbreviated by *Total SS or SS(Total)*.

The SS in a 1-way ANOVA can be split into two components, called the *"sum of squares of treatments*" and *"sum of squares of error*", abbreviated as SST and SSE, respectively.

The guiding principle behind ANOVA is the decomposition of the sums of squares, or Total SS

Algebraically, this is expressed by

$$\text{Total SS} = \text{SST} + \text{SSE}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$$

where k is the number of *treatments* and the bar over the y.. denotes the "grand" or "overall" mean. Each n_i is the number of observations for treatment i. The total number of observations is N (the sum of the n_i).

Note on subscripting

Don't be alarmed by the double subscripting. The total SS can be written single or double subscripted. The double subscript stems from the way the data are arranged in the data table. The table is usually a rectangular array with k columns and each column consists of n_i rows (however, the lengths of the rows, or the n_i, may be unequal).

Definition of "Treatment"

We introduced the concept of treatment. The definition is:
A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.

7. Product and Process Comparisons
7.4. Comparisons based on data from more than two processes
7.4.3. Are the means equal?

7.4.3.2. The 1-way ANOVA model and assumptions

A model that describes the relationship between the response and the treatment (between the dependent and independent variables)

The mathematical model that describes the relationship between the response and treatment for the one-way ANOVA is given by

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where Y_{ij} represents the j-th observation (j = 1, 2, ...n_i) on the i-th treatment (i = 1, 2, ..., k levels). So, Y_{23} represents the third observation using level 2 of the factor. μ is the common effect for the whole experiment, τ_i represents the i-th treatment effect and ϵ_{ij} represents the random error present in the j-th observation on the i-th treatment.

Fixed effects model

The errors ϵ_{ij} are assumed to be normally and independently (NID) distributed, with mean zero and variance σ_ϵ^2. μ is always a fixed parameter and $\tau_1, \tau_2, ..., \tau_k$ are considered to be fixed parameters *if the levels of the treatment are fixed*, and not a random sample from a population of possible levels. It is also assumed that μ is chosen so that

$$\sum \tau_i = 0 \qquad i = 1, ..., k$$

holds. This is the *fixed effects model*.

Random effects model

If the k levels of treatment are chosen at random, the model equation remains the same. However, now the τ_i's are random variables assumed to be NID(0, σ_τ). This is the *random effects model*.

Whether the levels are fixed or random depends on how these levels are chosen in a given experiment.

7.4.3.3. The ANOVA table and tests of hypotheses about means

Sums of Squares help us compute the variance estimates displayed in ANOVA Tables

The sums of squares **SST** and **SSE** previously computed for the one-way ANOVA are used to form two mean squares, one for *treatments* and the second for *error*. These mean squares are denoted by *MST* and *MSE*, respectively. These are typically displayed in a tabular form, known as an *ANOVA Table*. The ANOVA table also shows the statistics used to test hypotheses about the population means.

Ratio of MST and MSE

When the null hypothesis of equal means is true, the two mean squares estimate the same quantity (error variance), and should be of approximately equal magnitude. In other words, their ratio should be close to 1. If the null hypothesis is false, MST should be larger than MSE.

Divide sum of squares by degrees of freedom to obtain mean squares

The mean squares are formed by dividing the sum of squares by the associated degrees of freedom.

Let $N = \sum n_i$. Then, the degrees of freedom for treatment, DFT = k - 1, and the degrees of freedom for error, DFE = N - k.

The corresponding *mean squares* are:

$$MST = SST / DFT$$
$$MSE = SSE / DFE$$

The F-test

The test statistic, used in testing the equality of treatment means is: F = MST / MSE.

The critical value is the tabular value of the F distribution, based on the chosen α level and the degrees of freedom DFT and DFE.

The calculations are displayed in an ANOVA table, as follows:

ANOVA table

Source	SS	DF	MS	F

Treatments	SST	k-1	SST / (k-1)	MST/MSE
Error	SSE	N-k	SSE / (N-k)	
Total (corrected)	SS	N-1		

The word "source" stands for source of variation. Some authors prefer to use "between" and "within" instead of "treatments" and "error", respectively.

ANOVA Table Example

A numerical example

The data below resulted from measuring the difference in resistance resulting from subjecting identical resistors to three different temperatures for a period of 24 hours. The sample size of each group was 5. In the language of Design of Experiments, we have an experiment in which each of three treatments was replicated 5 times.

	Level 1	Level 2	Level 3
	6.9	8.3	8.0
	5.4	6.8	10.5
	5.8	7.8	8.1
	4.6	9.2	6.9
	4.0	6.5	9.3
means	5.34	7.72	8.56

The resulting ANOVA table is

Example ANOVA table

Source	SS	DF	MS	F
Treatments	27.897	2	13.949	9.59
Error	17.452	12	1.454	
Total (corrected)	45.349	14		
Correction Factor	779.041	1		

Interpretation of the ANOVA table

The test statistic is the F value of 9.59. Using an α of .05, we have that $F_{.05;\,2,\,12} = 3.89$ (see the F distribution table in Chapter 1). Since the test statistic is much larger than the critical value, we reject the null hypothesis of equal population means and conclude that there is a (statistically)

significant difference among the population means. The *p*-value for 9.59 is .00325, so the test statistic is significant at that level.

Techniques for further analysis

The populations here are resistor readings while operating under the three different temperatures. What we do **not** know at this point is whether the three means are all different or which of the three means is different from the other two, and by how much.

There are several techniques we might use to further analyze the differences. These are:

- constructing confidence intervals around the difference of two means,
- estimating combinations of factor levels with confidence bounds
- multiple comparisons of combinations of factor levels tested simultaneously.

7.4.3.4. 1-Way ANOVA calculations

Formulas for 1-way ANOVA hand calculations

Although computer programs that do ANOVA calculations now are common, for reference purposes this page describes how to calculate the various entries in an ANOVA table. Remember, the goal is to produce two variances (of treatments and error) and their ratio. The various computational formulas will be shown and applied to the data from the previous example.

Step 1: compute CM

STEP 1 Compute CM, the correction for the mean.

$$CM = \frac{\left(\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}\right)^2}{N_{total}} = \frac{(\text{Total of all observations})^2}{N_{total}}$$

$$= \frac{(108.1)^2}{15} = 779.041$$

Step 2: compute total SS

STEP 2 Compute the total SS.

The total SS = sum of squares of all observations - CM

$$SS_{total} = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^2 - CM$$

$$= (6.9)^2 + (5.4)^2 + \cdots + (6.9)^2 + 9.3^2 - CM$$

$$= 829.390 - 779.041 = 45.349$$

The 829.390 SS is called the "raw" or "uncorrected " sum of squares.

Step 3: compute SST

STEP 3 Compute SST, the treatment sum of squares.

First we compute the total (sum) for each treatment.

T_1 = (6.9) + (5.4) + ... + (4.0) = 26.7

$T_2 = (8.3) + (6.8) + ... + (6.5) = 38.6$
$T_1 = (8.0) + (10.5) + ... + (9.3) = 42.8$

Then

$$SST = \sum_{i=1}^{3} \frac{T_i^2}{n_i} - CM$$
$$= \frac{(26.7)^2}{5} + \frac{(38.6)^2}{5} + \frac{(42.8)^2}{5} - 779.041 = 27.897$$

Step 4: compute SSE

STEP 4 Compute SSE, the error sum of squares.

Here we utilize the property that the treatment sum of squares plus the error sum of squares equals the total sum of squares.

Hence, SSE = SS Total - SST = 45.349 - 27.897 = 17.45.

Step 5: Compute MST, MSE, and F

STEP 5 Compute MST, MSE and their ratio, *F*.

MST is the mean square of treatments, MSE is the mean square of error (MSE is also frequently denoted by $\hat{\sigma}_\varepsilon^2$).

MST = SST / (k-1) = 27.897 / 2 = 13.949
MSE = SSE / (N-k) = 17.452/ 12 = 1.454

where *N* is the total number of observations and *k* is the number of treatments. Finally, compute *F* as

F = MST / MSE = 9.59

That is it. These numbers are the quantities that are assembled in the ANOVA table that was shown previously.

7.4.3.5. Confidence intervals for the difference of treatment means

Confidence intervals for the difference between two means

This page shows how to construct a confidence interval around (μ_i- μ_j) for the one-way ANOVA by continuing the example shown on a previous page.

Formula for the confidence interval

The formula for a (1- α) 100% confidence interval for the difference between two treatment means is:

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm t_{1-\alpha/2,N-k} \sqrt{\hat{\sigma}_\epsilon^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\hat{\sigma}_\epsilon^2$ = *MSE*.

Computation of the confidence interval for μ_3 - μ_1

For the example, we have the following quantities for the formula:

- $\bar{y}_3$ = 8.56
- $\bar{y}_1$ = 5.34
- $\sqrt{1.454(1/5 + 1/5)} = 0.763$
- $t_{0.975,\ 12}$ = 2.179

Substituting these values yields (8.56 - 5.34) $\pm$ 2.179(0.763) or 3.22 $\pm$ 1.616.

That is, the confidence interval is from 1.604 to 4.836.

Additional 95% confidence intervals

A 95% confidence interval for μ_3 - μ_2 is: from -1.787 to 3.467.

A 95% confidence interval for μ_2 - μ_1 is: from -0.247 to 5.007.

Contrasts

Later on the topic of estimating more general linear

discussed later	combinations of means (primarily contrasts) will be discussed, including how to put confidence bounds around contrasts.

7.4.3.6. Assessing the response from any factor combination

Contrasts

This page treats how to estimate and put confidence bounds around the response to different combinations of factors. Primary focus is on the combinations that are known as *contrasts*. We begin, however, with the simple case of a single factor-level mean.

Estimation of a Factor Level Mean With Confidence Bounds

Estimating factor level means

An unbiased estimator of the factor level mean μ_i in the 1-way ANOVA model is given by:

$$\hat{\mu}_i = \bar{Y}_{i.}$$

where

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} = \frac{Y_{i.}}{n_i}$$

Variance of the factor level means

The variance of this sample mean estimator is

$$s^2_{\bar{Y}_{i.}} = \frac{MSE}{n_i} = \frac{\hat{\sigma}^2_e}{n_i}$$

Confidence intervals for the factor level means

It can be shown that:

$$t = \frac{\bar{Y}_{i.} - \mu_i}{s_{\bar{Y}_{i.}}}$$

has a t distribution with (N - k) degrees of freedom for the ANOVA model under consideration, where N is the total number of observations and k is the number of factor levels or groups. The degrees of freedom are the same as were used to calculate the MSE in the ANOVA table. That is: *dfe* (degrees of freedom for error) = N - k. From this we can calculate (1- α)100% confidence limits for each μ_i. These are given by:

$$\bar{Y}_{i\cdot} \pm t_{1-\alpha/2,N-k}\sqrt{\frac{\hat{\sigma}_\epsilon^2}{n_i}}$$

Example 1

Example for a 4-level treatment (or 4 different treatments)

The data in the accompanying table resulted from an experiment run in a completely randomized design in which each of four treatments was replicated five times.

						Total	Mean
Group 1	6.9	5.4	5.8	4.6	4.0	26.70	5.34
Group 2	8.3	6.8	7.8	9.2	6.5	38.60	7.72
Group 3	8.0	10.5	8.1	6.9	9.3	42.80	8.56
Group 4	5.8	3.8	6.1	5.6	6.2	27.50	5.50
All Groups						135.60	6.78

1-Way ANOVA table layout

This experiment can be illustrated by the table layout for this 1-way ANOVA experiment shown below:

Level	Sample j						
i	1	2	...	5	Sum	Mean	N
1	Y_{11}	Y_{12}	...	Y_{15}	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$	n_1
2	Y_{21}	Y_{22}	...	Y_{25}	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$	n_2
3	Y_{31}	Y_{32}	...	Y_{35}	$Y_{3\cdot}$	$\bar{Y}_{3\cdot}$	n_3
4	Y_{41}	Y_{42}	...	Y_{45}	$Y_{4\cdot}$	$\bar{Y}_{4\cdot}$	n_4
All					$Y_{\cdot}$	$\bar{Y}_{\cdot\cdot}$	n_t

ANOVA table

The resulting ANOVA table is

Source	SS	DF	MS	F
Treatments	38.820	3	12.940	9.724
Error	21.292	16	1.331	
Total (Corrected)	60.112	19		
Mean	919.368	1		
Total (Raw)	979.480	20		

The estimate for the mean of group 1 is 5.34, and the sample size is $n_1 = 5$.

Computing the confidence interval

Since the confidence interval is two-sided, the entry (1 - $\alpha/2$) value for the t table is (1 - 0.05/2) = 0.975, and the associated degrees of freedom is N - 4, or 20 - 4 = 16.

From the t table in Chapter 1, we obtain $t_{0.975;16} = 2.120$.

Next we need the standard error of the mean for group 1:

$$s^2_{Y_{1.}} = \frac{\text{MSE}}{n_1} = \frac{1.331}{5} = 0.2662$$

$$s_{Y_{1.}} = \sqrt{0.2662} = 0.5159$$

Hence, we obtain confidence limits 5.34 ± 2.120 (0.5159) and the confidence interval is

$$4.246 \leq \mu_1 \leq 6.434$$

Definition and Estimation of Contrasts

Definition of contrasts and orthogonal contrasts

Definitions

A contrast is a linear combination of 2 or more factor level means with coefficients that sum to zero.

Two contrasts are orthogonal if the sum of the products of corresponding coefficients (i.e., coefficients for the same means) adds to zero.

Formally, the definition of a contrast is expressed below, using the notation μ_i for the i-th treatment mean:

$$C = c_1\mu_1 + c_2\mu_2 + \ldots + c_j\mu_j + \ldots + c_k\mu_k$$

where

$$c_1 + c_2 + \ldots + c_j + \ldots + c_k = \sum_{j=1}^{k} c_j = 0$$

Simple contrasts include the case of the difference between two factor means, such as $\mu_1 - \mu_2$. If one wishes to compare treatments 1 and 2 with treatment 3, one way of expressing this is by: $\mu_1 + \mu_2 - 2\mu_3$. Note that

$\mu_1 - \mu_2$ has coefficients +1, -1
$\mu_1 + \mu_2 - 2\mu_3$ has coefficients +1, +1, -2.
These coefficients sum to zero.

An example of orthogonal contrasts

As an example of *orthogonal contrasts*, note the three contrasts defined by the table below, where the rows denote coefficients for the column treatment means.

	μ_1	μ_2	μ_3	μ_4
c_1	+1	0	0	-1
c_2	0	+1	-1	0
c_3	+1	-1	-1	+1

Some properties of orthogonal contrasts

The following is true:

1. The sum of the coefficients for each contrast is zero.
2. The sum of the products of coefficients of each pair of contrasts is also 0 (orthogonality property).
3. The first two contrasts are simply pairwise comparisons, the third one involves all the treatments.

Estimation of contrasts

As might be expected, *contrasts are estimated by taking the same linear combination of treatment mean estimators*. In other words:

$$\hat{C} = \sum_{i=1}^{r} c_i \overline{Y}_{i\cdot}$$

and

$$Var\left(\hat{C}\right) = \sum_{i=1}^{r} c_i^2 Var\left(\overline{Y}_{i\cdot}\right) = \sum_{i=1}^{r} c_i^2 \left(\frac{\sigma^2}{n_i}\right) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

Note: These formulas hold for any linear combination of treatment means, not just for contrasts.

Confidence Interval for a Contrast

Confidence intervals for contrasts

An unbiased estimator for a contrast *C* is given by

$$\hat{C} = \sum_{i=1}^{r} c_i \overline{Y}_{i\cdot}$$

The estimator of $Var(\hat{C})$ is

$$s_{\hat{C}} = \hat{\sigma}_e \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

The estimator $\hat{C}$ is normally distributed because it is a linear combination of independent normal random variables. It can be shown that:

$$\frac{\hat{C} - C}{s_{\hat{C}}}$$

is distributed as t_{N-r} for the one-way ANOVA model under discussion.

Therefore, the 1- α confidence limits for *C* are:

$$\hat{C} \pm t_{1-\alpha/2,N-r}\ s_{\hat{C}}$$

Example 2 (estimating contrast)

Contrast to estimate

We wish to estimate, in our previous example, the following contrast:

$$C = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

and construct a 95 % confidence interval for *C*.

Computing the point estimate and standard error

The point estimate is:

$$\hat{C} = \frac{\bar{Y}_1 + \bar{Y}_2}{2} - \frac{\bar{Y}_3 + \bar{Y}_4}{2} = -0.5$$

Applying the formulas above we obtain

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = 0.2$$

and

$$s_{\hat{C}}^2 = MSE \sum_{i=1}^{4} \frac{c_i^2}{n_i} = 1.331(0.2) = 0.2662$$

and the standard error is $\sqrt{0.2661}$ = 0.5159.

Confidence interval

For a confidence coefficient of 95 % and *df* = 20 - 4 = 16, $t_{0.975,16}$ = 2.12. Therefore, the desired 95 % confidence interval is -0.5 ± 2.12(0.5159) or

(-1.594, 0.594).

Estimation of Linear Combinations

Estimating linear combinations

Sometimes we are interested in a linear combination of the factor-level means that is not a contrast. Assume that in our sample experiment certain costs are associated with each group. For example, there might be costs associated with each factor as follows:

Factor	Cost in $
1	3
2	5
3	2
4	1

The following linear combination might then be of interest:

$$C = 3\mu_1 + 5\mu_2 + 2\mu_3 + 1\mu_4$$

Coefficients do not have to sum to zero for linear combinations

This resembles a contrast, **but the coefficients c_i do not sum to zero**. A linear combination is given by the definition:

$$C = \sum_{i=1}^{r} c_i \mu_i$$

with no restrictions on the coefficients c_i.

Confidence interval identical to contrast

Confidence limits for a linear combination C are obtained in precisely the same way as those for a contrast, using the same calculation for the point estimator and estimated variance.

7.4.3.7. The two-way ANOVA

Definition of a factorial experiment

The 2-way ANOVA is probably the most popular layout in the Design of Experiments. To begin with, let us define a *factorial experiment*:

An experiment that utilizes every combination of factor levels as treatments is called a factorial experiment.

Model for the two-way factorial experiment

In a factorial experiment with factor A at *a* levels and factor B at *b* levels, the model for the general layout can be written as

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, r$$

where μ is the overall mean response, τ_i is the effect due to the *i*-th level of factor A, β_j is the effect due to the *j*-th level of factor B and γ_{ij} is the effect due to any interaction between the *i*-th level of A and the *j*-th level of B.

Fixed factors and fixed effects models

At this point, consider the levels of factor A and of factor B chosen for the experiment to be the only levels of interest to the experimenter such as predetermined levels for temperature settings or the length of time for process step. The factors A and B are said to be *fixed factors* and the model is a *fixed-effects model*. Random actors will be discussed later.

When an *a* x *b* factorial experiment is conducted with an equal number of observations per treatment combination, the total (corrected) sum of squares is partitioned as:

SS(total) = SS(A) + SS(B) + SS(AB) + SSE

where AB represents the interaction between A and B.

For reference, the formulas for the sums of squares are:

$$SS(A) = rb\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SS(B) = ra\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS(AB) = r\sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SSE = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{ij.})^2$$

$$SS(Total) = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{...})^2$$

The breakdown of the total (corrected for the mean) sums of squares

The resulting ANOVA table for an *a* x *b* factorial experiment is

Source	*SS*	*df*	*MS*
Factor A	*SS*(A)	(*a* - 1)	*MS*(A) = *SS*(A)/(*a*-1)
Factor B	*SS*(B)	(*b* - 1)	*MS*(B) = *SS*(B)/(*b*-1)
Interaction AB	*SS*(AB)	(*a*-1)(*b*-1)	*MS*(AB)= *SS*(AB)/(*a*-1)(*b*-1)
Error	*SSE*	(*N* - *ab*)	*SSE*/(*N* - *ab*)
Total (Corrected)	SS(Total)	(*N* - 1)	

The ANOVA table can be used to test hypotheses about the effects and interactions

The various hypotheses that can be tested using this ANOVA table concern whether the different levels of Factor A, or Factor B, really make a difference in the response, and whether the AB interaction is significant (see previous discussion of ANOVA hypotheses).

7.4.3.8. Models and calculations for the two-way ANOVA

Basic Layout

The balanced 2-way factorial layout

Factor A has 1, 2, ..., *a* levels. Factor B has 1, 2, ..., *b* levels. There are *ab* treatment combinations (or cells) in a complete factorial layout. Assume that each treatment cell has *r* independent obsevations (known as replications). When each cell has the same number of replications, the design is a *balanced factorial*. In this case, the *abr* data points $\{y_{ijk}\}$ can be shown pictorially as follows:

		Factor B			
		1	2	...	b
Factor A	1	$y_{111}, y_{112}, ..., y_{11r}$	$y_{121}, y_{122}, ..., y_{12r}$	...	$y_{1b1}, y_{1b2}, ..., y_{1br}$
	2	$y_{211}, y_{212}, ..., y_{21r}$	$y_{221}, y_{222}, ..., y_{22r}$	...	$y_{2b1}, y_{2b2}, ..., y_{2br}$
	.	...	....		...
	.				
	a	$y_{a11}, y_{a12}, ..., y_{a1r}$	$y_{a21}, y_{a22}, ..., y_{a2r}$	...	$y_{ab1}, y_{ab2}, ..., y_{abr}$

How to obtain sums of squares for the balanced factorial layout

Next, we will calculate the sums of squares needed for the ANOVA table.

- Let A_i be the sum of all observations of level *i* of factor A, *i* = 1, ... ,*a*. The A_i are the row sums.
- Let B_j be the sum of all observations of level *j* of factor B, *j* = 1, ...,*b*. The B_j are the column sums.
- Let $(AB)_{ij}$ be the sum of all observations of level *i* of A and level *j* of B. These are cell sums.
- Let *r* be the number of replicates in the experiment; that is: the number of times each factorial treatment combination appears in the experiment.

Then the total number of observations for each level of factor A is *rb* and the total number of observations for each level of factor B is *ra* and the total number of observations for each interaction is *r*.

Finally, the total number of observations *n* in the experiment is *abr*.

With the help of these expressions we arrive (omitting derivations) at

$$CM = \frac{(\text{Sum of all observations})^2}{rab}$$

$$SS_{total} = \sum(\text{each observation})^2 - CM$$

$$SS(A) = \frac{\sum_{i=1}^{a} A_i^2}{rb} - CM$$

$$SS(B) = \frac{\sum_{i=1}^{b} B_i^2}{ra} - CM$$

$$SS(AB) = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(AB)_{ij}^2}{r} - CM - SS(A) - SS(B)$$

$$SSE = SS_{total} - SS(A) - S(B) - SS(AB)$$

These expressions are used to calculate the ANOVA table entries for the (fixed effects) 2-way ANOVA.

Two-Way ANOVA Example:

Data An evaluation of a new coating applied to 3 different materials was conducted at 2 different laboratories. Each laboratory tested 3 samples from each of the treated materials. The results are given in the next table:

	Materials (B)		
LABS (A)	1	2	3
1	4.1	3.1	3.5
	3.9	2.8	3.2
	4.3	3.3	3.6
2	2.7	1.9	2.7
	3.1	2.2	2.3
	2.6	2.3	2.5

Row and column sums The preliminary part of the analysis yields a table of row and column sums.

Material (B)

Lab (A)	1	2	3	Total (A_i)
1	12.3	9.2	10.3	31.8
2	8.4	6.4	7.5	22.3
Total (B_j)	20.7	15.6	17.8	54.1

ANOVA table From this table we generate the ANOVA table.

Source	*SS*	*df*	*MS*	*F*	*p*-value
A	5.0139	1	5.0139	100.28	0
B	2.1811	2	1.0906	21.81	.0001
AB	0.1344	2	0.0672	1.34	.298
Error	0.6000	12	0.0500		
Total (Corr)	7.9294	17			

7.4.4. What are variance components?

Fixed and Random Factors and Components of Variance

A fixed level of a factor or variable means that the levels in the experiment are the only ones we are interested in

In the previous example, the levels of the factor temperature were considered as *fixed*; that is, the three temperatures were the only ones that we were interested in (this may sound somewhat unlikely, but let us accept it without opposition). The model employed for fixed levels is called a *fixed model*. When the levels of a factor are random, such as operators, days, lots or batches, where the levels in the experiment might have been chosen at *random* from a large number of possible levels, the model is called a *random model*, and inferences are to be extended to all levels of the population.

Random levels are chosen at random from a large or infinite set of levels

In a random model the experimenter is often interested in estimating *components of variance*. Let us run an example that analyzes and interprets a component of variance or random model.

Components of Variance Example for Random Factors

Data for the example

A company supplies a customer with a larger number of batches of raw materials. The customer makes three sample determinations from each of 5 randomly selected batches to control the quality of the incoming material. The model is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

and the k levels (e.g., the batches) are chosen at random from a population with variance σ_τ. The data are shown below

Batch				
1	2	3	4	5
74	68	75	72	79

76	71	77	74	81
75	72	77	73	79

ANOVA table for example

A 1-way ANOVA is performed on the data with the following results:

ANOVA

Source	*SS*	*df*	*MS*	*EMS*
Treatment (batches)	147.74	4	36.935	$\sigma_\epsilon^2 + 3\sigma_\tau^2$
Error	17.99	10	1.799	σ_ϵ^2
Total (corrected)	165.73	14		

Interpretation of the ANOVA table

The computations that produce the SS are the same for both the fixed and the random effects model. For the random model, however, the treatment sum of squares, SST, is an estimate of $\{\sigma_\epsilon^2 + 3\sigma_\tau^2\}$. This is shown in the EMS (Expected Mean Squares) column of the ANOVA table.

The test statistic from the ANOVA table is F = 36.94 / 1.80 = 20.5.

If we had chosen an α value of .01, then the F value from the table in Chapter 1 for a *df* of 4 in the numerator and 10 in the denominator is 5.99.

Method of moments

Since the test statistic is larger than the critical value, we reject the hypothesis of equal means. Since these batches were chosen via a random selection process, it may be of interest to find out how much of the variance in the experiment might be attributed to batch diferences and how much to random error. In order to answer these questions, we can use the EMS column. The estimate of σ_ϵ^2 is 1.80 and the computed treatment mean square of 36.94 is an estimate of $\sigma_\epsilon^2 + 3\sigma_\tau^2$. Setting the MS values equal to the EMS values (this is called the *Method of Moments*), we obtain

$$s_\epsilon^2 = 1.80 \quad \text{and} \quad s_\epsilon^2 + 3s_\tau^2 = 36.94$$

where we use s^2 since these are estimators of the corresponding σ^2's.

Computation of the

Solving these expressions

components of variance

$$s_\tau^2 = \frac{36.94 - 1.80}{3} = 11.71$$

The total variance can be estimated as

$$s_{total} = s_\tau^2 + s_\epsilon^2 = 11.71 + 1.80 = 13.51$$

Interpretation

In terms of percentages, we see that 11.71/13.51 = 86.7 percent of the total variance is attributable to batch differences and 13.3 percent to error variability within the batches.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

7.4.5. How can we compare the results of classifying according to several categories?

Contingency Table approach

When items are classified according to two or more criteria, it is often of interest to decide whether these criteria act independently of one another.

For example, suppose we wish to classify defects found in wafers produced in a manufacturing plant, first according to the type of defect and, second, according to the production shift during which the wafers were produced. If the proportions of the various types of defects are constant from shift to shift, then classification by defects is independent of the classification by production shift. On the other hand, if the proportions of the various defects vary from shift to shift, then the classification by defects depends upon or is *contingent* upon the shift classification and the classifications are dependent.

In the process of investigating whether one method of classification is contingent upon another, it is customary to display the data by using a cross classification in an array consisting of r rows and c columns called a ***contingency table***. A contingency table consists of r x c cells representing the r x c possible outcomes in the classification process. Let us construct an industrial case:

Industrial example

A total of 309 wafer defects were recorded and the defects were classified as being one of four types, *A*, *B*, *C*, or *D*. At the same time each wafer was identified according to the production shift in which it was manufactured, 1, 2, or 3.

Contingency table classifying defects in wafers according to type and production shift

These counts are presented in the following table.

	Type of Defects				
Shift	A	B	C	D	Total
1	15(22.51)	21(20.99)	45(38.94)	13(11.56)	94
2	26(22.9)	31(21.44)	34(39.77)	5(11.81)	96
3	33(28.50)	17(26.57)	49(49.29)	20(14.63)	119
Total	74	69	128	38	309

(Note: the numbers in parentheses are the expected cell frequencies).

Column probabilities

Let p_A be the probability that a defect will be of type A. Likewise, define p_B, p_C, and p_D as the probabilities of observing the other three types of defects. These probabilities, which are called the ***column probabilities***, will satisfy the requirement

$$p_A + p_B + p_C + p_D = 1$$

Row probabilities

By the same token, let p_i (i=1, 2, or 3) be the **row probability** that a defect will have occurred during shift i, where

$$p_1 + p_2 + p_3 = 1$$

Multiplicative Law of Probability

Then if the two classifications are independent of each other, a cell probability will equal the product of its respective row and column probabilities in accordance with the Multiplicative Law of Probability.

Example of obtaining column and row probabilities

For example, the probability that a particular defect will occur in shift 1 and is of type A is (p_1) (p_A). While the numerical values of the cell probabilities are unspecified, the null hypothesis states that each cell probability will equal the product of its respective row and column probabilities. This condition implies independence of the two classifications. The alternative hypothesis is that this equality does not hold for at least one cell.

In other words, we state the null hypothesis as H_0: the two classifications are independent, while the alternative hypothesis is H_a: the classifications are dependent.

To obtain the observed column probability, divide the column total by the grand total, n. Denoting the total of column j as c_j, we get

$$\hat{p}_A = \frac{c_1}{n} = \frac{74}{309} \qquad \hat{p}_C = \frac{c_3}{n} = \frac{128}{309}$$
$$\hat{p}_B = \frac{c_2}{n} = \frac{69}{309} \qquad \hat{p}_D = \frac{c_4}{n} = \frac{38}{309}$$

Similarly, the row probabilities p_1, p_2, and p_3 are estimated by dividing the row totals r_1, r_2, and r_3 by the grand total n, respectively

$$\hat{p}_1 = \frac{r_1}{n} = \frac{94}{309} \quad \hat{p}_2 = \frac{r_2}{n} = \frac{96}{309} \quad \hat{p}_3 = \frac{r_3}{n} = \frac{119}{309}$$

Expected cell frequencies

Denote the observed frequency of the cell in row i and column $j$of the contingency table by n_{ij}. Then we have

$$\hat{E}(n_{ij}) = n(\hat{p}_i \hat{p}_j) = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right) = \frac{r_i c_j}{n}$$

Estimated expected cell frequency when H_0 is true.

In other words, when the row and column classifications are independent, the estimated expected value of the observed cell frequency n_{ij} in an r x c contingency table is equal to its respective row and column totals divided by the total frequency.

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$$

The estimated cell frequencies are shown in parentheses in the contingency table above.

Test statistic

From here we use the expected and observed frequencies shown in the table to calculate the value of the test statistic

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

$$\chi^2 = \frac{(15 - 22.51)^2}{22.51} + \frac{(26 - 22.99)^2}{22.99} + \cdots + \frac{(20 - 14.63)^2}{14.63} = 19.18$$

df = (r-1)(c-1)

The next step is to find the appropriate number of degrees of freedom associated with the test statistic. Leaving out the details of the derivation, we state the result:

> *The number of degrees of freedom associated with a contingency table consisting of r rows and c columns is (r-1) (c-1).*

So for our example we have (3-1) (4-1) = 6 d.f.

Testing the null hypothesis

In order to test the null hypothesis, we compare the test statistic with the critical value of $X^2_{1-\alpha/2}$ at a selected value of α. Let us use α = 0.05. Then the critical value is $X^2_{0.95,6}$ = 12.5916 (see the chi square table in Chapter 1). Since the test statistic of 19.18 exceeds the critical value, we reject the null hypothesis and conclude that there is significant evidence that the proportions of the different defect types vary from shift to shift. In this case, the *p*-value of the test statistic is 0.00387.

7.4.6. Do all the processes have the same proportion of defects?

The contingency table approach

Testing for homogeneity of proportions using the chi-square distribution via contingency tables

When we have samples from n populations (i.e., lots, vendors, production runs, etc.), we can test whether there are significant differences in the proportion defectives for these populations using a contingency table approach. The contingency table we construct has two rows and n columns.

To test the null hypothesis of no difference in the proportions among the n populations

$$H_0: \ p_1 = p_2 = ... = p_n$$

against the alternative that not all n population proportions are equal

$$H_1: \text{Not all } p_i \text{ are equal } (i = 1, 2, ..., n)$$

The chi-square test statistic

we use the following test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_c)^2}{f_c}$$

where f_o is the observed frequency in a given cell of a 2 x n contingency table, and f_c is the theoretical count or expected frequency in a given cell if the null hypothesis were true.

The critical value

The critical value is obtained from the χ^2 distribution table with degrees of freedom (2-1)(n-1) = n-1, at a given level of significance.

An illustrative example

Data for the example

Diodes used on a printed circuit board are produced in lots of size 4000. To study the homogeneity of lots with

respect to a demanding specification, we take random samples of size 300 from 5 consecutive lots and test the diodes. The results are:

	Lot					
Results	1	2	3	4	5	Totals
Nonconforming	36	46	42	63	38	225
Conforming	264	254	258	237	262	1275
Totals	300	300	300	300	300	1500

Computation of the overall proportion of nonconforming units

Assuming the null hypothesis is true, we can estimate the single overall proportion of nonconforming diodes by pooling the results of all the samples as

$$\bar{p} = \frac{(36+46+42+63+38)}{(5 \text{ x } 300)} = 225/1500 = .15$$

Computation of the overall proportion of conforming units

We estimate the proportion of conforming ("good") diodes by the complement 1 - 0.15 = 0.85. Multiplying these two proportions by the sample sizes used for each lot results in the expected frequencies of nonconforming and conforming diodes. These are presented below:

Table of expected frequencies

	Lot					
Results	1	2	3	4	5	Totals
Nonconforming	45	45	45	45	45	225
Conforming	255	255	255	255	255	1275
Totals	300	300	300	300	300	1500

Null and alternate hypotheses

To test the null hypothesis of homogeneity or equality of proportions

H_0: $p_1 = p_2 = ... = p_5$

against the alternative that not all 5 population proportions are equal

H_1: Not all p_i are equal (i = 1, 2, ...,5)

Table for computing the test statistic

we use the observed and expected values from the tables above to compute the χ^2 test statistic. The calculations are presented below:

2 2

f_o	f_c	$(f_o - f_c)$	$(f_o - f_c)$	$(f_o - f_c)\ /\ f_c$
36	45	-9	81	1.800
46	45	1	1	0.022
42	45	-3	9	0.200
63	45	18	324	7.200
38	45	-7	49	1.089
264	225	9	81	0.318
254	255	-1	1	0.004
258	255	3	9	0.035
237	255	-18	324	1.271
262	255	7	49	0.192
				12.131

Conclusions If we choose a .05 level of significance, the critical value of χ^2 with 4 degrees of freedom is 9.488 (see the chi square distribution table in Chapter 1). Since the test statistic (12.131) exceeds this critical value, we reject the null hypothesis.

7.4.7. How can we make multiple comparisons?

What to do after equality of means is rejected

When processes are compared and the null hypothesis of equality (or homogeneity) is rejected, all we know at that point is that there is no equality amongst them. But we do not know the form of the inequality.

Typical questions

Questions concerning the reason for the rejection of the null hypothesis arise in the form of:

- "Which mean(s) or proportion (s) differ from a standard or from each other?"
- "Does the mean of treatment 1 differ from that of treatment 2?"
- "Does the average of treatments 1 and 2 differ from the average of treatments 3 and 4?"

Multiple Comparison test procedures are needed

One popular way to investigate the cause of rejection of the null hypothesis is a *Multiple Comparison Procedure.* These are methods which examine or compare more than one pair of means or proportions at the same time.

Note: Doing pairwise comparison procedures over and over again for all possible pairs will not, in general, work. This is because the overall significance level is not as specified for a single pair comparison.

ANOVA F test is a preliminary test

The ANOVA uses the F test to determine whether there exists a significant difference among treatment means or interactions. In this sense it is a preliminary test that informs us if we should continue the investigation of the data at hand.

If the null hypothesis (no difference among treatments or interactions) is accepted, there is an implication that no relation exists between the factor levels and the response. There is not much we can learn, and we are finished with the analysis.

When the F test rejects the null hypothesis, we usually want to undertake a thorough analysis of the nature of the factor-

level effects.

Procedures for examining factor-level effects

Previously, we discussed several procedures for examining particular factor-level effects. These were

- Estimation of the Difference Between Two Factor Means
- Estimation of Factor Level Effects
- Confidence Intervals For A Contrast

Determine contrasts in advance of observing the experimental results

These types of investigations should be done on combinations of factors that were determined in advance of observing the experimental results, or else the confidence levels are not as specified by the procedure. Also, doing several comparisons might change the overall confidence level (see note above). This can be avoided by carefully selecting contrasts to investigate in advance and making sure that:

- the number of such contrasts does not exceed the number of degrees of freedom between the treatments
- only orthogonal contrasts are chosen.

However, there are also several powerful multiple comparison procedures we can use after observing the experimental results.

Tests on Means after Experimentation

Procedures for performing multiple comparisons

If the decision on what comparisons to make is withheld until after the data are examined, the following procedures can be used:

- Tukey's Method *to test all possible pairwise differences of means to determine if at least one difference is significantly different from 0.*
- Scheffé's Method *to test all possible contrasts at the same time, to see if at least one is significantly different from 0.*
- Bonferroni Method *to test, or put simultaneous confidence intervals around, a pre-selected group of contrasts*

Multiple Comparisons Between Proportions

Procedure for proportion defective data

When we are dealing with population proportion defective data, the Marascuilo procedure can be used to simultaneously examine comparisons between all groups after the data have been collected.

7.4.7.1. Tukey's method

Tukey's method considers all possible pairwise differences of means at the same time

The Tukey method applies simultaneously to the set of all pairwise comparisons

$$\{\mu_i - \mu_j\}$$

The confidence coefficient for the set, when all sample sizes are equal, is exactly 1-α. For unequal sample sizes, the confidence coefficient is greater than 1-α. In other words, the Tukey method is conservative when there are unequal sample sizes.

Studentized Range Distribution

The studentized range q

The Tukey method uses the *studentized range distribution*. Suppose we have r independent observations $y_1, ..., y_r$ from a normal distribution with mean μ and variance σ^2. Let w be the range for this set , i.e., the maximum minus the minimum. Now suppose that we have an estimate s^2 of the variance σ^2 which is based on ν degrees of freedom and is independent of the y_i. The studentized range is defined as

$$q_{r,\nu} = w/s$$

The distribution of q is tabulated in many textbooks and can be calculated using Dataplot

The distribution of q has been tabulated and appears in many textbooks on statistics. In addition, Dataplot has a CDF function (SRACDF) and a percentile function (SRAPPF) for q.

As an example, let r = 5 and ν = 10. The 95th percentile is $q_{.05;5,10}$ = 4.65. This means:

$$P\left\{\frac{w}{s} \le 4.65\right\} = .95$$

So, if we have five observations from a normal distribution, the probability is .95 that their range is not more than 4.65 times as great as an independent sample standard deviation estimate for which the estimator has 10 degrees of freedom.

Tukey's Method

Confidence limits for Tukey's method

The Tukey confidence limits for all pairwise comparisons with confidence coefficient of at least 1- α are:

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \frac{1}{\sqrt{2}} q_{\alpha;r,N-r} \hat{\sigma}_\epsilon \sqrt{\frac{2}{n}} \qquad i,j = 1,\ldots,r; i \neq j$$

Notice that the point estimator and the estimated variance are the same as those for a single pairwise comparison that was illustrated previously. The only difference between the confidence limits for simultaneous comparisons and those for a single comparison is the multiple of the estimated standard deviation.

Also note that the sample sizes must be equal when using the studentized range approach.

Example

Data

We use the data from a previous example.

Set of all pairwise comparisons

The set of all pairwise comparisons consists of:

μ_2 - μ_1, μ_3 - μ_1, μ_1 - μ_4,
μ_2 - μ_3, μ_2 - μ_4, μ_3 - μ_4

Confidence intervals for each pair

Assume we want a confidence coefficient of 95 percent, or .95. Since r = 4 and n_t = 20, the required percentile of the studentized range distribution is $q_{.05;\,4,16}$. Using the Tukey method for each of the six comparisons yields:

$$0.29 \leq \mu_2 - \mu_1 \leq 4.47$$

$$1.13 \leq \mu_3 - \mu_1 \leq 5.31$$

$$-2.25 \leq \mu_1 - \mu_4 \leq 1.93$$

$$-2.93 \leq \mu_2 - \mu_3 \leq 1.25$$

$$0.13 \leq \mu_2 - \mu_4 \leq 4.31$$

$$0.97 \leq \mu_3 - \mu_4 \leq 5.15$$

Conclusions

The simultaneous pairwise comparisons indicate that the differences μ_1 - μ_4 and μ_2 - μ_3 are not significantly different from 0 (their confidence intervals include 0), and all the other pairs are significantly different.

Unequal sample sizes

It is possible to work with unequal sample sizes. In this case, one has to calculate the estimated standard deviation for each

pairwise comparison. The Tukey procedure for unequal sample sizes is sometimes referred to as the *Tukey-Kramer Method*.

7.4.7.2. Scheffe's method

Scheffe's method tests all possible contrasts at the same time

Scheffé's method applies to the set of estimates of all possible contrasts among the factor level means, not just the pairwise differences considered by Tukey's method.

Definition of contrast

An arbitrary contrast is defined by

$$C = \sum_{i=1}^{r} c_i \mu_i$$

where

$$\sum_{i=1}^{r} c_i = 0$$

Infinite number of contrasts

Technically there is an infinite number of contrasts. The simultaneous confidence coefficient is exactly 1- α, whether the factor level sample sizes are equal or unequal.

Estimate and variance for C

As was described earlier, we estimate C by:

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{Y}_{i\cdot}$$

for which the estimated variance is:

$$s^2_{\hat{C}} = \hat{\sigma}^2_e \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

Simultaneous confidence interval

It can be shown that the probability is 1 - α that all confidence limits of the type

$$\hat{C} \pm \sqrt{(r-1) F_{\alpha;r-1,N-r}} \; s_{\hat{C}}$$

are correct simultaneously.

Scheffe method example

Contrasts to estimate

We wish to estimate, in our previous experiment, the following contrasts

$$C_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$C_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

and construct 95 percent confidence intervals for them.

Compute the point estimates of the individual contrasts

The point estimates are:

$$\hat{C}_1 = \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} = -0.5$$
$$\hat{C}_2 = \frac{\overline{Y}_1 + \overline{Y}_3}{2} - \frac{\overline{Y}_2 + \overline{Y}_4}{2} = .34$$

Compute the point estimate and variance of C

Applying the formulas above we obtain in both cases:

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = .2$$

and

$$s^2_{\hat{C}} = \hat{\sigma}^2_\epsilon \sum_{i=1}^{4} \frac{c_i^2}{4} = 1.331(.2) = .2661$$

where σ^2_ϵ = 1.331 was computed in our previous example. The standard error = .5158 (square root of .2661).

Scheffe confidence interval

For a confidence coefficient of 95 percent and degrees of freedom in the numerator of r - 1 = 4 - 1 = 3, and in the denominator of 20 - 4 = 16, we have:

$$\sqrt{(r-1)F_{\alpha;r-1;N-r}} = \sqrt{3F_{.05;3;16}} = 3.12$$

The confidence limits for C_1 are -.5 ± 3.12(.5158) = -.5 ± 1.608, and for C_2 they are .34 ± 1.608.

The desired simultaneous 95 percent confidence intervals are

$$-2.108 \leq C_1 \leq 1.108$$

$$-1.268 \leq C_2 \leq 1.948$$

Comparison to confidence interval for a single contrast

Recall that when we constructed a confidence interval for a single contrast, we found the 95 percent confidence interval:

$$-1.594 \leq C \leq 0.594$$

As expected, the Scheffé confidence interval procedure that generates simultaneous intervals for all contrasts is considerabley wider.

Comparison of Scheffé's Method with Tukey's Method

Tukey preferred when only pairwise comparisons are of interest

If only pairwise comparisons are to be made, the Tukey method will result in a narrower confidence limit, which is preferable.

Consider for example the comparison between μ_3 and μ_1.

Tukey: $1.13 < \mu_3 - \mu_1 < 5.31$

Scheffé: $0.95 < \mu_3 - \mu_1 < 5.49$

which gives Tukey's method the edge.

The normalized contrast, using sums, for the Scheffé method is 4.413, which is close to the maximum contrast.

Scheffe preferred when many contrasts are of interest

In the general case when many or all contrasts might be of interest, the Scheffé method tends to give narrower confidence limits and is therefore the preferred method.

7.4.7.3. Bonferroni's method

Simple method

The Bonferroni method is a simple method that allows many comparison statements to be made (or confidence intervals to be constructed) while still assuring an overall confidence coefficient is maintained.

Applies for a finite number of contrasts

This method applies to an ANOVA situation when the analyst has picked out a particular set of pairwise comparisons or contrasts or linear combinations in advance. This set is not infinite, as in the Scheffé case, but may exceed the set of pairwise comparisons specified in the Tukey procedure.

Valid for both equal and unequal sample sizes

The Bonferroni method is valid for equal and unequal sample sizes. We restrict ourselves to only linear combinations or comparisons of treatment level means (pairwise comparisons and contrasts are special cases of linear combinations). We denote the number of statements or comparisons in the finite set by *g*.

Bonferroni general inequality

Formally, the Bonferroni general inequality is presented by:

$$P\left(\bigcap_{i=1}^{g} A_i\right) \geq 1 - \sum_{i=1}^{g} P[\bar{A}_i]$$

where A_i and its complement $\bar{A}_i$ are any events.

Interpretation of Bonferroni inequality

In particular, if each A_i is the event that a calculated confidence interval for a particular linear combination of treatments includes the true value of that combination, then the left-hand side of the inequality is the probability that all the confidence intervals simultaneously cover their respective true values. The right-hand side is one minus the sum of the probabilities of each of the intervals missing their true values. Therefore, if simultaneous multiple interval estimates are desired with an overall confidence coefficient 1-α, *one can construct each interval with confidence coefficient (1-*α*/g), and the Bonferroni inequality insures that the overall confidence coefficient is at least 1-*α.

Formula for Bonferroni confidence interval

In summary, the Bonferroni method states that the confidence coefficient is at least 1- α that simultaneously all the following confidence limits for the *g* linear combinations C_i are "correct" (or capture their respective true values):

$$\hat{C}_i \pm t_{1-\alpha/(2g),N-r}\ s_{\hat{C}_i}$$

where

$$s_{\hat{C}_i} = \hat{\sigma}_\epsilon \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

Example using Bonferroni method

Contrasts to estimate

We wish to estimate, as we did using the Scheffe method, the following linear combinations (contrasts):

$$C_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$C_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

and construct 95 % confidence intervals around the estimates.

Compute the point estimates of the individual contrasts

The point estimates are:

$$\hat{C}_1 = \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} = -0.5$$
$$\hat{C}_2 = \frac{\overline{Y}_1 + \overline{Y}_3}{2} - \frac{\overline{Y}_2 + \overline{Y}_4}{2} = .34$$

Compute the point estimate and variance of C

As before, for both contrasts, we have

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = .2$$

and

$$s_{\hat{C}}^2 = \hat{\sigma}_\epsilon^2 \sum_{i=1}^{4} \frac{c_i^2}{4} = 1.331(.2) = .2661$$

where σ_ϵ^2 = 1.331 was computed in our previous example. The standard error is .5158 (the square root of .2661).

Compute the Bonferroni simultaneous confidence interval

For a 95 % overall confidence coefficient using the Bonferroni method, the t value is $t_{1-0.05/(2*2),16} = t_{0.9875,16}$ = 2.473 (from the t table in Chapter 1). Now we can calculate the confidence intervals for the two contrasts. For C_1 we have confidence limits -0.5 ± 2.473 (.5158) and for C_2 we have confidence limits 0.34 ± 2.473 (0.5158).

Thus, the confidence intervals are:

$$-1.776 \leq C_1 \leq 0.776$$
$$-0.936 \leq C_2 \leq 1.616$$

Comparison to Scheffe interval

Notice that the Scheffé interval for C_1 is:

$$-2.108 \leq C_1 \leq 1.108$$

which is wider and therefore less attractive.

Comparison of Bonferroni Method with Scheffé and Tukey Methods

No one comparison method is uniformly best - each has its uses

1. If all pairwise comparisons are of interest, Tukey has the edge. If only a subset of pairwise comparisons are required, Bonferroni may sometimes be better.
2. When the number of contrasts to be estimated is small, (about as many as there are factors) Bonferroni is better than Scheffé. Actually, unless the number of desired contrasts is at least twice the number of factors, Scheffé will always show wider confidence bands than Bonferroni.
3. Many computer packages include all three methods. So, study the output and select the method with the smallest confidence band.
4. No single method of multiple comparisons is uniformly best among all the methods.

7. Product and Process Comparisons
7.4. Comparisons based on data from more than two processes
7.4.7. How can we make multiple comparisons?

7.4.7.4. Comparing multiple proportions: The Marascuillo procedure

Testing for equal proportions of defects

Earlier, we discussed how to test whether several populations have the same proportion of defects. The example given there led to rejection of the null hypothesis of equality.

Marascuilo procedure allows comparison of all possible pairs of proportions

Rejecting the null hypothesis only allows us to conclude that not (in this case) all lots are equal with respect to the proportion of defectives. However, it does not tell us which lot or lots caused the rejection.

The Marascuilo procedure enables us to simultaneously test the differences of all pairs of proportions when there are several populations under investigation.

The Marascuillo Procedure

Step 1: compute differences p_i - p_j

Assume we have samples of size n_i (i = 1, 2, ..., k) from k populations. The first step of this procedure is to compute the differences p_i - p_j, (where i is not equal to j) among all $k(k\text{-}1)/2$ pairs of proportions.

The absolute values of these differences are the test-statistics.

Step 2: compute test statistics

Step 2 is to pick a significance level and compute the corresponding critical values for the Marascuilo procedure from

$$r_{ij} = \sqrt{\chi^2_{1-\alpha,k-1}} \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}$$

Step 3: compare test statistics against corresponding critical values

The third and last step is to compare each of the $k(k\text{-}1)/2$ test statistics against its corresponding critical r_{ij} value. Those pairs that have a test statistic that exceeds the critical value are significant at the α level.

Example

Sample proportions

To illustrate the Marascuillo procedure, we use the data from the previous example. Since there were 5 lots, there are (5 x 4)/2 = 10 possible pairwise comparisons to be made and ten critical ranges to compute. The five sample proportions are:

$$p_1 = 36/300 = .120$$
$$p_2 = 46/300 = .153$$
$$p_3 = 42/300 = .140$$
$$p_4 = 63/300 = .210$$
$$p_5 = 38/300 = .127$$

Table of critical values

For an overall level of significance of 0.05, the critical value of the chi-square distribution having four degrees of freedom is $X^2_{0.95,4} = 9.488$ and the square root of 9.488 is 3.080. Calculating the 10 absolute differences and the 10 critical values leads to the following summary table.

contrast	value	critical range	significant
$\|p_1 - p_2\|$	.033	0.086	no
$\|p_1 - p_3\|$	.020	0.085	no
$\|p_1 - p_4\|$	.090	0.093	no
$\|p_1 - p_5\|$	.007	0.083	no
$\|p_2 - p_3\|$	.013	0.089	no
$\|p_2 - p_4\|$	.057	0.097	no
$\|p_2 - p_5\|$	.026	0.087	no
$\|p_3 - p_4\|$	.070	0.095	no
$\|p_3 - p_5\|$	.013	0.086	no
$\|p_4 - p_5\|$	.083	0.094	no

The table of critical values can be generated using both Dataplot code and R code.

No individual contrast is statistically significant

A difference is statistically significant if its value exceeds the critical range value. In this example, even though the null hypothesis of equality was rejected earlier, there is not enough data to conclude any particular difference is significant. Note, however, that all the comparisons involving population 4 come the closest to significance - leading us to suspect that more data might actually show that population 4 does have a significantly higher proportion of defects.

7.5. References

Primary References

Agresti, A. and Coull, B. A. (1998). *Approximate is better than "exact" for interval estimation of binomial proportions*", The American Statistician, 52(2), 119-126.

Berenson M.L. and Levine D.M. (1996) *Basic Business Statistics*, Prentice-Hall, Englewood Cliffs, New Jersey.

Bhattacharyya, G. K., and R. A. Johnson, (1997). *Statistical Concepts and Methods*, John Wiley and Sons, New York.

Birnbaum, Z. W. (1952). "*Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size*", Journal of the American Statistical Association, 47, page 425.

Brown, L. D. Cai, T. T. and DasGupta, A. (2001). *Interval estimation for a binomial proportion*", Statistical Science, 16(2), 101-133.

Diamond, W. J. (1989). *Practical Experiment Designs*, Van-Nostrand Reinhold, New York.

Dixon, W. J. and Massey, F.J. (1969). *Introduction to Statistical Analysis*, McGraw-Hill, New York.

Draper, N. and Smith, H., (1981). *Applied Regression Analysis*, John Wiley & Sons, New York.

Fliess, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, Third Edition, John Wiley & Sons, New York.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, John Wiley & Sons, New York.

Hicks, C. R. (1973). *Fundamental Concepts in the Design of Experiments*, Holt, Rinehart and Winston, New York.

Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*, John Wiley & Sons, New York.

Howe, W. G. (1969). *"Two-sided Tolerance Limits for Normal Populations - Some Improvements"*, Journal of the Americal Statistical Association, 64 , pages 610-620.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of*

Statistics, Volume 2: Inference and Relationship. Charles Griffin & Co. Limited, London.

Mendenhall, W., Reinmuth, J. E. and Beaver, R. J. *Statistics for Management and Economics*, Duxbury Press, Belmont, CA.

Montgomery, D. C. (1991). *Design and Analysis of Experiments*, John Wiley & Sons, New York.

Moore, D. S. (1986). "*Tests of Chi-Square Type*". From Goodness-of-Fit Techniques (D'Agostino & Stephens eds.).

Myers, R. H., (1990). *Classical and Modern Regression with Applications*, PWS-Kent, Boston, MA.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models*, 3rd Edition, Irwin, Boston, MA.

Lawless, J. F., (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Pearson, A. V., and Hartley, H. O. (1972). *Biometrica Tables for Statisticians, Vol 2*, Cambridge, England, Cambridge University Press.

Sarhan, A. E. and Greenberg, B. G. (1956). "*Estimation of location and scale parameters by order statistics from singly and double censored samples*," Part I, Annals of Mathematical Statistics, 27, 427-451.

Searle, S. S., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

Siegel, S. (1956). *Nonparametric Statistics*, McGraw-Hill, New York.

Shapiro, S. S. and Wilk, M. B. (1965). "*An analysis of variance test for normality (complete samples)*", Biometrika, 52, 3 and 4, pages 591-611.

Some Additional References and Bibliography

Books

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-FitTechniques*, Marcel Dekker, Inc., New York.

Hicks, C. R. 1973. *Fundamental Concepts in the Design of Experiments*. Holt, Rhinehart and Winston,New-York

Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, Springer-Verlag, New York.

Neter, Wasserman, and Whitmore (1993). *Applied Statistics*, 4th Edition, Allyn and Bacon, Boston, MA.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models,* 3rd Edition, Irwin, Boston, MA.

Scheffe, H. (1959). *The Analysis of Variance*, John Wiley, New-York.

Articles

Begun, J. M. and Gabriel, K. R. (1981). "*Closure of the Newman-Keuls Multiple Comparisons Procedure*", Journal of the American Statistical Association, 76, page 374.

Carmer, S. G. and Swanson, M. R. (1973. *"Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte-Carlo Methods"*, Journal of the American Statistical Association, 68, pages 66-74.

Duncan, D. B. (1975). "*t-Tests and Intervals for Comparisons suggested by the Data*" Biometrics, 31, pages 339-359.

Dunnett, C. W. (1980). "*Pairwise Multiple Comparisons in the Homogeneous Variance for Unequal Sample Size Case*", Journal of the American Statistical Association, 75, page 789.

Einot, I. and Gabriel, K. R. (1975). "*A Study of the Powers of Several Methods of Multiple Comparison*", Journal of the American Statistical Association, 70, page 351.

Gabriel, K. R. (1978). "*A Simple Method of Multiple Comparisons of Means*", Journal of the American Statistical Association, 73, page 364.

Hochburg, Y. (1974). "*Some Conservative Generalizations of the T-Method in Simultaneous Inference*", Journal of Multivariate Analysis, 4, pages 224-234.

Kramer, C. Y. (1956). "*Extension of Multiple Range Tests to Group Means with Unequal Sample Sizes*", Biometrics, 12, pages 307-310.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). "*On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance*", Biometrics, 63, pages 655-660.

Ryan, T. A. (1959). "*Multiple Comparisons in Psychological Research*", Psychological Bulletin, 56, pages 26-47.

Ryan, T. A. (1960). "*Significance Tests for Multiple Comparisons of Proportions, Variances, and Other Statistics*", Psychological Bulletin, 57, pages 318-328.

Scheffe, H. (1953). "*A Method for Judging All Contrasts in the Analysis of Variance*", Biometrika,40, pages 87-104.

Sidak, Z., (1967). "*Rectangular Confidence Regions for the Means of Multivariate Normal Distributions*", Journal of the American Statistical Association, 62, pages 626-633.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*, Unpublished Manuscript.

Waller, R. A. and Duncan, D. B. (1969). "*A Bayes Rule for the Symmetric Multiple Comparison Problem*", Journal of the American Statistical Association 64, pages 1484-1504.

Waller, R. A. and Kemp, K. E. (1976). "*Computations of Bayesian t-Values for Multiple Comparisons*", Journal of Statistical Computation and Simulation, 75, pages 169-172.

Welsch, R. E. (1977). "*Stepwise Multiple Comparison Procedure*", Journal of the American Statistical Association, 72, page 359.

8. Assessing Product Reliability

This chapter describes the terms, models and techniques used to evaluate and predict product reliability.

1. Introduction

1. Why important?
2. Basic terms and models
3. Common difficulties
4. Modeling "physical acceleration"
5. Common acceleration models
6. Basic non-repairable lifetime distributions
7. Basic models for repairable systems
8. Evaluate reliability "bottom-up"
9. Modeling reliability growth
10. Bayesian methodology

2. Assumptions/Prerequisites

1. Choosing appropriate life distribution
2. Plotting reliability data
3. Testing assumptions
4. Choosing a physical acceleration model
5. Models and assumptions for Bayesian methods

3. Reliability Data Collection

1. Planning reliability assessment tests

4. Reliability Data Analysis

1. Estimating parameters from censored data
2. Fitting an acceleration model
3. Projecting reliability at use conditions
4. Comparing reliability between two or more populations
5. Fitting system repair rate models
6. Estimating reliability using a Bayesian gamma prior

Click here for a detailed table of contents
References for Chapter 8

8. Assessing Product Reliability - Detailed Table of Contents [8.]

1. Introduction [8.1.]
    1. Why is the assessment and control of product reliability important? [8.1.1.]
        1. Quality versus reliability [8.1.1.1.]
        2. Competitive driving factors [8.1.1.2.]
        3. Safety and health considerations [8.1.1.3.]
    2. What are the basic terms and models used for reliability evaluation? [8.1.2.]
        1. Repairable systems, non-repairable populations and lifetime distribution models [8.1.2.1.]
        2. Reliability or survival function [8.1.2.2.]
        3. Failure (or hazard) rate [8.1.2.3.]
        4. "Bathtub" curve [8.1.2.4.]
        5. Repair rate or ROCOF [8.1.2.5.]
    3. What are some common difficulties with reliability data and how are they overcome? [8.1.3.]
        1. Censoring [8.1.3.1.]
        2. Lack of failures [8.1.3.2.]
    4. What is "physical acceleration" and how do we model it? [8.1.4.]
    5. What are some common acceleration models? [8.1.5.]
        1. Arrhenius [8.1.5.1.]
        2. Eyring [8.1.5.2.]
        3. Other models [8.1.5.3.]
    6. What are the basic lifetime distribution models used for non-repairable populations? [8.1.6.]
        1. Exponential [8.1.6.1.]
        2. Weibull [8.1.6.2.]
        3. Extreme value distributions [8.1.6.3.]
        4. Lognormal [8.1.6.4.]
        5. Gamma [8.1.6.5.]
        6. Fatigue life (Birnbaum-Saunders) [8.1.6.6.]
        7. Proportional hazards model [8.1.6.7.]
    7. What are some basic repair rate models used for repairable systems? [8.1.7.]
        1. Homogeneous Poisson Process (HPP) [8.1.7.1.]
        2. Non-Homogeneous Poisson Process (NHPP) - power law [8.1.7.2.]
        3. Exponential law [8.1.7.3.]
    8. How can you evaluate reliability from the "bottom-up" (component failure mode to system failure rate)? [8.1.8.]
        1. Competing risk model [8.1.8.1.]
        2. Series model [8.1.8.2.]
        3. Parallel or redundant model [8.1.8.3.]
        4. R out of N model [8.1.8.4.]
        5. Standby model [8.1.8.5.]
        6. Complex systems [8.1.8.6.]
    9. How can you model reliability growth? [8.1.9.]
        1. NHPP power law [8.1.9.1.]
        2. Duane plots [8.1.9.2.]

3. NHPP exponential law [8.1.9.3.]
10. How can Bayesian methodology be used for reliability evaluation? [8.1.10.]

2. Assumptions/Prerequisites [8.2.]
   1. How do you choose an appropriate life distribution model? [8.2.1.]
      1. Based on failure mode [8.2.1.1.]
      2. Extreme value argument [8.2.1.2.]
      3. Multiplicative degradation argument [8.2.1.3.]
      4. Fatigue life (Birnbaum-Saunders) model [8.2.1.4.]
      5. Empirical model fitting - distribution free (Kaplan-Meier) approach [8.2.1.5.]
   2. How do you plot reliability data? [8.2.2.]
      1. Probability plotting [8.2.2.1.]
      2. Hazard and cum hazard plotting [8.2.2.2.]
      3. Trend and growth plotting (Duane plots) [8.2.2.3.]
   3. How can you test reliability model assumptions? [8.2.3.]
      1. Visual tests [8.2.3.1.]
      2. Goodness of fit tests [8.2.3.2.]
      3. Likelihood ratio tests [8.2.3.3.]
      4. Trend tests [8.2.3.4.]
   4. How do you choose an appropriate physical acceleration model? [8.2.4.]
   5. What models and assumptions are typically made when Bayesian methods are used for reliability evaluation? [8.2.5.]

3. Reliability Data Collection [8.3.]
   1. How do you plan a reliability assessment test? [8.3.1.]
      1. Exponential life distribution (or HPP model) tests [8.3.1.1.]
      2. Lognormal or Weibull tests [8.3.1.2.]
      3. Reliability growth (Duane model) [8.3.1.3.]
      4. Accelerated life tests [8.3.1.4.]
      5. Bayesian gamma prior model [8.3.1.5.]

4. Reliability Data Analysis [8.4.]
   1. How do you estimate life distribution parameters from censored data? [8.4.1.]
      1. Graphical estimation [8.4.1.1.]
      2. Maximum likelihood estimation [8.4.1.2.]
      3. A Weibull maximum likelihood estimation example [8.4.1.3.]
   2. How do you fit an acceleration model? [8.4.2.]
      1. Graphical estimation [8.4.2.1.]
      2. Maximum likelihood [8.4.2.2.]
      3. Fitting models using degradation data instead of failures [8.4.2.3.]
   3. How do you project reliability at use conditions? [8.4.3.]
   4. How do you compare reliability between two or more populations? [8.4.4.]
   5. How do you fit system repair rate models? [8.4.5.]
      1. Constant repair rate (HPP/exponential) model [8.4.5.1.]
      2. Power law (Duane) model [8.4.5.2.]
      3. Exponential law model [8.4.5.3.]
   6. How do you estimate reliability using the Bayesian gamma prior model? [8.4.6.]
   7. References For Chapter 8: Assessing Product Reliability [8.4.7.]

8.1. Introduction

This section introduces the terminology and models that will be used to describe and quantify product reliability. The terminology, probability distributions and models used for reliability analysis differ in many cases from those used in other statistical applications.

Detailed contents of Section 1

1. Introduction
    1. Why is the assessment and control of product reliability important?
        1. Quality versus reliability
        2. Competitive driving factors
        3. Safety and health considerations
    2. What are the basic terms and models used for reliability evaluation?
        1. Repairable systems, non-repairable populations and lifetime distribution models
        2. Reliability or survival function
        3. Failure (or hazard) rate
        4. "Bathtub" curve
        5. Repair rate or ROCOF
    3. What are some common difficulties with reliability data and how are they overcome?
        1. Censoring
        2. Lack of failures
    4. What is "physical acceleration" and how do we model it?
    5. What are some common acceleration models?
        1. Arrhenius
        2. Eyring
        3. Other models
    6. What are the basic lifetime distribution models used for non-repairable populations?
        1. Exponential
        2. Weibull
        3. Extreme value distributions
        4. Lognormal
        5. Gamma
        6. Fatigue life (Birnbaum-Saunders)
        7. Proportional hazards model
    7. What are some basic repair rate models used for repairable systems?
        1. Homogeneous Poisson Process (HPP)

2. Non-Homogeneous Poisson Process (NHPP) with power law
3. Exponential law

8. How can you evaluate reliability from the "bottom- up" (component failure mode to system failure rates)?
    1. Competing risk model
    2. Series model
    3. Parallel or redundant model
    4. R out of N model
    5. Standby model
    6. Complex systems
9. How can you model reliability growth?
    1. NHPP power law
    2. Duane plots
    3. NHPP exponential law
10. How can Bayesian methodology be used for reliability evaluation?

8.1.1. Why is the assessment and control of product reliability important?

We depend on, demand, and expect reliable products

In today's technological world nearly everyone depends upon the continued functioning of a wide array of complex machinery and equipment for their everyday health, safety, mobility and economic welfare. We expect our cars, computers, electrical appliances, lights, televisions, etc. to function whenever we need them - day after day, year after year. When they fail the results can be catastrophic: injury, loss of life and/or costly lawsuits can occur. More often, repeated failure leads to annoyance, inconvenience and a lasting customer dissatisfaction that can play havoc with the responsible company's marketplace position.

Shipping unreliable products can destroy a company's reputation

It takes a long time for a company to build up a reputation for reliability, and only a short time to be branded as "unreliable" after shipping a flawed product. Continual assessment of new product reliability and ongoing control of the reliability of everything shipped are critical necessities in today's competitive business arena.

8. Assessing Product Reliability
8.1. Introduction
8.1.1. Why is the assessment and control of product reliability important?

8.1.1.1. Quality versus reliability

Reliability is "quality changing over time"

The everyday usage term "quality of a product" is loosely taken to mean its inherent degree of excellence. In industry, this is made more precise by defining quality to be "conformance to requirements at the start of use". Assuming the product specifications adequately capture customer requirements, the quality level can now be precisely measured by the fraction of units shipped that meet specifications.

A motion picture instead of a snapshot

But how many of these units still meet specifications after a week of operation? Or after a month, or at the end of a one year warranty period? That is where "reliability" comes in. Quality is a snapshot at the start of life and reliability is a motion picture of the day-by-day operation. Time zero defects are manufacturing mistakes that escaped final test. The additional defects that appear over time are "reliability defects" or reliability fallout.

Life distributions model fraction fallout over time

The quality level might be described by a single fraction defective. To describe reliability fallout a probability model that describes the fraction fallout over time is needed. This is known as the life distribution model.

8. Assessing Product Reliability
8.1. Introduction
8.1.1. Why is the assessment and control of product reliability important?

8.1.1.2. Competitive driving factors

Reliability is a major economic factor in determining a product's success

Accurate prediction and control of reliability plays an important role in the profitability of a product. Service costs for products within the warranty period or under a service contract are a major expense and a significant pricing factor. Proper spare part stocking and support personnel hiring and training also depend upon good reliability fallout predictions. On the other hand, missing reliability targets may invoke contractual penalties and cost future business.

Companies that can economically design and market products that meet their customers' reliability expectations have a strong competitive advantage in today's marketplace.

8. Assessing Product Reliability
8.1. Introduction
8.1.1. Why is the assessment and control of product reliability important?

8.1.1.3. Safety and health considerations

Some failures have serious social consequences and this should be taken into account when planning reliability studies

Sometimes equipment failure can have a major impact on human safety and/or health. Automobiles, planes, life support equipment, and power generating plants are a few examples.

From the point of view of "assessing product reliability", we treat these kinds of catastrophic failures no differently from the failure that occurs when a key parameter measured on a manufacturing tool drifts slightly out of specification, calling for an unscheduled maintenance action.

It is up to the reliability engineer (and the relevant customer) to define what constitutes a failure in any reliability study. More resource (test time and test units) should be planned for when an incorrect reliability assessment could negatively impact safety and/or health.

8.1.2. What are the basic terms and models used for reliability evaluation?

Reliability methods and terminology began with 19th century insurance companies

Reliability theory developed apart from the mainstream of probability and statistics, and was used primarily as a tool to help nineteenth century maritime and life insurance companies compute profitable rates to charge their customers. Even today, the terms "failure rate" and "hazard rate" are often used interchangeably.

The following sections will define some of the concepts, terms, and models we need to describe, estimate and predict reliability.

8. Assessing Product Reliability
8.1. Introduction
8.1.2. What are the basic terms and models used for reliability evaluation?

8.1.2.1. Repairable systems, non-repairable populations and lifetime distribution models

Life distribution models describe how non-repairable populations fail over time

A repairable system is one which can be restored to satisfactory operation by any action, including parts replacements or changes to adjustable settings. When discussing the rate at which failures occur during system operation time (and are then repaired) we will define a Rate Of Occurrence Of Failure (ROCF) or "repair rate". It would be incorrect to talk about failure rates or hazard rates for repairable systems, as these terms apply only to the first failure times for a population of non repairable components.

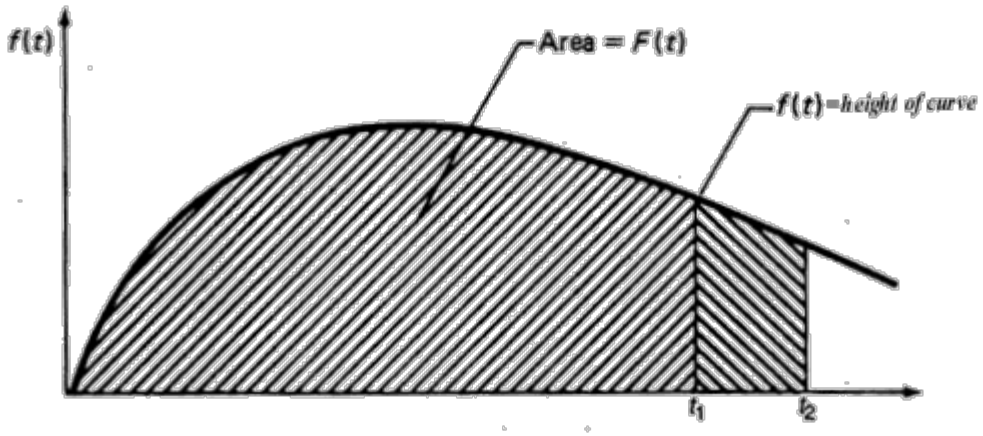
A non-repairable population is one for which individual items that fail are removed permanently from the population. While the system may be repaired by replacing failed units from either a similar or a different population, the members of the original population dwindle over time until all have eventually failed.

We begin with models and definitions for non-repairable populations. Repair rates for repairable populations will be defined in a later section.

The theoretical population models used to describe unit lifetimes are known as **Lifetime Distribution Models**. The population is generally considered to be all of the possible unit lifetimes for all of the units that could be manufactured based on a particular design and choice of materials and manufacturing process. A random sample of size *n* from this population is the collection of failure times observed for a randomly selected group of *n* units.

Any continuous PDF defined only for non-negative values can be a lifetime distribution model

A lifetime distribution model can be any *probability density function* (or PDF) *f(t)* defined over the range of time from $t = 0$ to $t =$ infinity. The corresponding *cumulative distribution function* (or CDF) $F(t)$ is a very useful function, as it gives the probability that a randomly selected unit will fail by time t. The figure below shows the relationship between *f(t)* and *F(t)* and gives three descriptions of *F(t)*.



1. *F(t)* = the area under the PDF *f(t)* to the left of *t*.

2. *F(t)* = the probability that a single randomly chosen new unit will fail by time *t*.

3. *F(t)* = the proportion of the entire population that fails by time *t*.

The figure above also shows a shaded area under *f(t)* between the two times t_1 and t_2. This area is $[F(t_2) - F(t_1)]$ and represents the proportion of the population that fails between times t_1 and t_2 (or the probability that a brand new randomly chosen unit will survive to time t_1 but fail before time t_2).

Note that the PDF *f(t)* has only non-negative values and eventually either becomes 0 as *t* increases, or decreases towards 0. The CDF *F(t)* is monotonically increasing and goes from 0 to 1 as *t* approaches infinity. In other words, the total area under the curve is always 1.

The Weibull model is a good example of a life distribution

The 2-parameter Weibull distribution is an example of a popular *F(t)*. It has the CDF and PDF equations given by:

$$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}, \quad f(t) = \frac{\gamma}{t}\left(\frac{t}{\alpha}\right)^{\gamma} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$$

where γ is the "shape" parameter and α is a scale parameter called the **characteristic life**.

Example: A company produces automotive fuel pumps that fail according to a Weibull life distribution model with shape parameter γ = 1.5 and scale parameter 8,000 (time measured in use hours). If a typical pump is used 800 hours a year, what proportion are likely to fail within 5 years?

Solution: The probability associated with the 800*5 quantile of a Weibull distribution with γ = 1.5 and α = 8000 is 0.298. Thus about 30% of the pumps will fail in the first 5 years.

Functions for computing PDF values and CDF values, are available in both

Dataplot code and R code.

8. Assessing Product Reliability
8.1. Introduction
8.1.2. What are the basic terms and models used for reliability evaluation?

8.1.2.2. Reliability or survival function

Survival is the complementary event to failure

The **Reliability Function** $R(t)$, also known as the **Survival Function** $S(t)$, is defined by:

$R(t) = S(t)$ = the probability a unit survives beyond time t.

Since a unit either fails, or survives, and one of these two mutually exclusive alternatives must occur, we have

$$R(t) = 1 - F(t), \quad F(t) = 1 - R(t)$$

Calculations using $R(t)$ often occur when building up from single components to subsystems with many components. For example, if one microprocessor comes from a population with reliability function $R_m(t)$ and two of them are used for the CPU in a system, then the system CPU has a reliability function given by

$$R_{cpu}(t) = R_m^2(t)$$

The reliability of the system is the product of the reliability functions of the components

since both must survive in order for the system to survive. This building up to the system from the individual components will be discussed in detail when we look at the "Bottom-Up" method. The general rule is: to calculate the reliability of a system of independent components, multiply the reliability functions of all the components together.

8. Assessing Product Reliability
8.1. Introduction
8.1.2. What are the basic terms and models used for reliability evaluation?

8.1.2.3. Failure (or hazard) rate

The failure rate is the rate at which the population survivors at any given instant are "falling over the cliff"

The failure rate is defined for non repairable populations as the (instantaneous) rate of failure for the survivors to time *t* during the next instant of time. It is a rate per unit of time similar in meaning to reading a car speedometer at a particular instant and seeing 45 mph. The next instant the failure rate may change and the units that have already failed play no further role since only the survivors count.

The failure rate (or hazard rate) is denoted by *h(t)* and calculated from

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)} = \text{the instantaneous (conditional) failure rate.}$$

The failure rate is sometimes called a "conditional failure rate" since the denominator 1 - *F(t)* (i.e., the population survivors) converts the expression into a conditional rate, given survival past time *t*.

Since *h(t)* is also equal to the negative of the derivative of ln{*R(t)*}, we have the useful identity:

$$F(t) = 1 - \exp\left\{ -\int_0^t h(t)dt \right\}$$

If we let

$$H(t) = \int_0^t h(t)dt$$

be the **Cumulative Hazard Function**, we then have $F(t) = 1 - e^{-H(t)}$. Two other useful identities that follow from these formulas are:

$$h(t) = -\frac{d \ln R(t)}{dt}$$

$$H(t) = -\ln R(t).$$

It is also sometimes useful to define an average failure rate over any interval (T_1, T_2) that "averages" the failure rate over that interval. This rate, denoted by $AFR(T_1,T_2)$, is a single number that can be used as a specification or target for the population failure rate over that interval. If T_1 is 0, it is dropped from the expression. Thus, for example, $AFR(40,000)$ would be the average failure rate for the population over the first 40,000 hours of operation.

The formulas for calculating AFR's are:

$$AFR(T_2 - T_1) = \frac{\left(\int_{T_1}^{T_2} h(t)dt\right)}{T_2 - T_1} = \frac{H(T_2) - H(T_1)}{T_2 - T_1} = \frac{\ln R(T_1) - \ln R(T_2)}{T_2 - T_1}$$

$$AFR(0,T) = AFR(T) = \frac{H(T)}{T} = \frac{-\ln R(T)}{T}$$

8.1.2.4. "Bathtub" curve

A plot of the failure rate over time for most products yields a curve that looks like a drawing of a bathtub

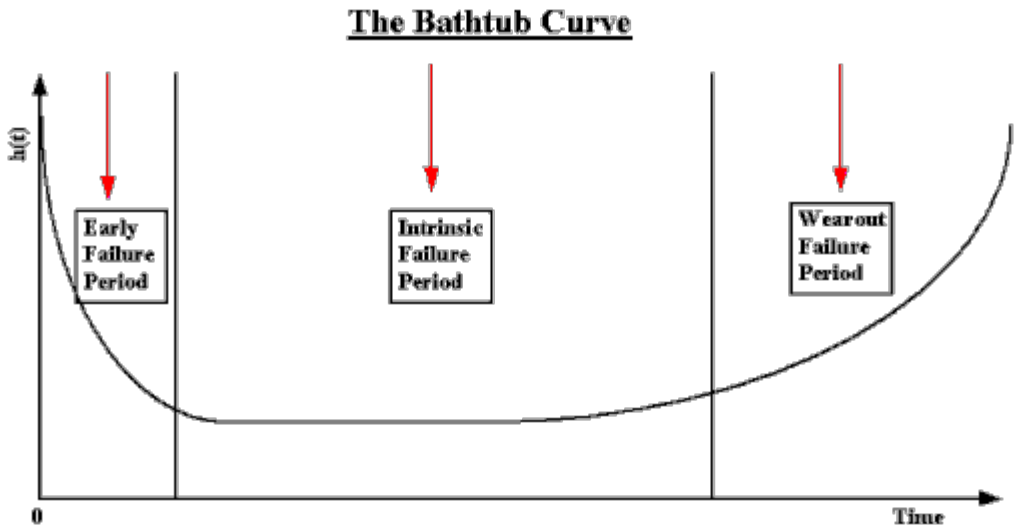
If enough units from a given population are observed operating and failing over time, it is relatively easy to compute week-by-week (or month-by-month) estimates of the failure rate *h(t)*. For example, if N_{12} units survive to start the 13th month of life and r_{13} of them fail during the next month (or 720 hours) of life, then a simple empirical estimate of *h(t)* averaged across the 13th month of life (or between 8640 hours and 9360 hours of age), is given by (r_{13} / N_{12} * 720). Similar estimates are discussed in detail in the section on Empirical Model Fitting.

Over many years, and across a wide variety of mechanical and electronic components and systems, people have calculated empirical population failure rates as units age over time and repeatedly obtained a graph such as shown below. Because of the shape of this failure rate curve, it has become widely known as the "Bathtub" curve.

The initial region that begins at time zero when a customer first begins to use the product is characterized by a high but rapidly decreasing failure rate. This region is known as the **Early Failure Period** (also referred to as **Infant Mortality Period**, from the actuarial origins of the first bathtub curve plots). This decreasing failure rate typically lasts several weeks to a few months.

Next, the failure rate levels off and remains roughly constant for (hopefully) the majority of the useful life of the product. This long period of a level failure rate is known as the **Intrinsic Failure Period** (also called the **Stable Failure Period**) and the constant failure rate level is called the **Intrinsic Failure Rate**. Note that most systems spend most of their lifetimes operating in this flat portion of the bathtub curve

Finally, if units from the population remain in use long enough, the failure rate begins to increase as materials wear out and degradation failures occur at an ever increasing rate. This is the **Wearout Failure Period**.



NOTE: The Bathtub Curve also applies (based on much empirical evidence) to Repairable Systems. In this case, the vertical axis is the Repair Rate or the Rate of Occurrence of Failures (ROCOF).

8. Assessing Product Reliability
8.1. Introduction
8.1.2. What are the basic terms and models used for reliability evaluation?

8.1.2.5. Repair rate or ROCOF

Repair Rate models are based on counting the cumulative number of failures over time

A different approach is used for modeling the rate of occurrence of failure incidences for a repairable system. In this chapter, these rates are called *repair rates* (not to be confused with the length of time for a repair, which is not discussed in this chapter). Time is measured by system power-on-hours from initial turn-on at time zero, to the end of system life. Failures occur as given system ages and the system is repaired to a state that may be the same as new, or better, or worse. The frequency of repairs may be increasing, decreasing, or staying at a roughly constant rate.

Let *N(t)* be a counting function that keeps track of the cumulative number of failures a given system has had from time zero to time *t*. *N(t)* is a step function that jumps up one every time a failure occurs and stays at the new level until the next failure.

Every system will have its own observed *N(t)* function over time. If we observed the *N(t)* curves for a large number of similar systems and "averaged" these curves, we would have an estimate of *M(t)* = the expected number (average number) of cumulative failures by time *t* for these systems.

The Repair Rate (or ROCOF) is the mean rate of failures per unit time

The derivative of *M(t)*, denoted *m(t)*, is defined to be the **Repair Rate** or the **Rate Of Occurrence Of Failures at Time *t*** or **ROCOF**.

Models for *N(t)*, *M(t)* and *m(t)* will be described in the section on Repair Rate Models.

8.1.3. What are some common difficulties with reliability data and how are they overcome?

The Paradox of Reliability Analysis: The more reliable a product is, the harder it is to get the failure data needed to "prove" it is reliable!

There are two closely related problems that are typical with reliability data and not common with most other forms of statistical data. These are:

- Censoring (when the observation period ends, not all units have failed - some are survivors)
- Lack of Failures (if there is too much censoring, even though a large number of units may be under observation, the information in the data is limited due to the lack of actual failures)

These problems cause considerable practical difficulty when planning reliability assessment tests and analyzing failure data. Some solutions are discussed in the next two sections. Typically, the solutions involve making additional assumptions and using complicated models.

8.1.3.1. Censoring

When not all units on test fail we have censored data

Consider a situation in which we are reliability testing n (non repairable) units taken randomly from a population. We are investigating the population to determine if its failure rate is acceptable. In the typical test scenario, we have a fixed time T to run the units to see if they survive or fail. The data obtained are called **Censored Type I** data.

Censored Type I Data

During the T hours of test we observe r failures (where r can be any number from 0 to n). The (exact) failure times are t_1, t_2, ..., t_r and there are $(n - r)$ units that survived the entire T-hour test without failing. Note that T is fixed in advance and r is random, since we don't know how many failures will occur until the test is run. Note also that we assume the exact times of failure are recorded when there are failures.

This type of censoring is also called "right censored" data since the times of failure to the right (i.e., larger than T) are missing.

Another (much less common) way to test is to decide in advance that you want to see exactly r failure times and then test until they occur. For example, you might put 100 units on test and decide you want to see at least half of them fail. Then r = 50, but T is unknown until the 50th fail occurs. This is called **Censored Type II** data.

Censored Type II Data

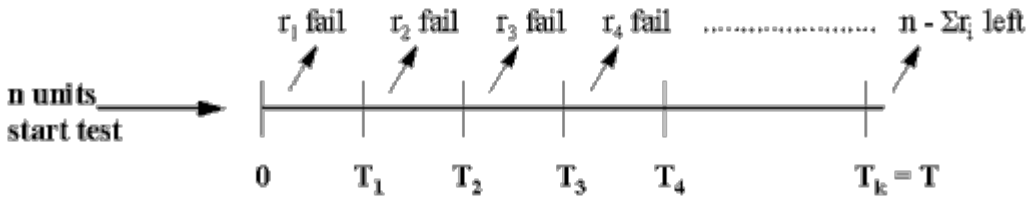
We observe t_1, t_2, ..., t_r, where r is specified in advance. The test ends at time T = t_r, and $(n-r)$ units have survived. Again we assume it is possible to observe the exact time of failure for failed units.

Type II censoring has the significant advantage that you know in advance how many failure times your test will yield - this helps enormously when planning adequate tests. However, an open-ended random test time is generally impractical from a management point of view and this type of testing is rarely seen.

Sometimes we don't even know the exact time of failure

Readout or Interval Data

Sometimes exact times of failure are not known; only an interval of time in which the failure occurred is recorded. This kind of data is called **Readout** or **Interval** data and the situation is shown in the figure below:



.

Multicensored Data

In the most general case, every unit observed yields exactly one of the following three types of information:

- a run-time if the unit did not fail while under observation
- an exact failure time
- an interval of time during which the unit failed.

The units may all have different run-times and/or readout intervals.

Many special methods have been developed to handle censored data

How do we handle censored data?

Many statistical methods can be used to fit models and estimate failure rates, even with censored data. In later sections we will discuss the Kaplan-Meier approach, Probability Plotting, Hazard Plotting, Graphical Estimation, and Maximum Likelihood Estimation.

Separating out Failure Modes

Note that when a data set consists of failure times that can be sorted into several different failure modes, it is possible (and often necessary) to analyze and model each mode separately. Consider all failures due to modes other than the one being analyzed as censoring times, with the censored run-time equal to the time it failed due to the different (independent) failure mode. This is discussed further in the competing risk section and later analysis sections.

8.1.3.2. Lack of failures

Failure data is needed to accurately assess and improve reliability - this poses problems when testing highly reliable parts

When fitting models and estimating failure rates from reliability data, the precision of the estimates (as measured by the width of the confidence intervals) tends to vary inversely with the square root of the number of failures observed - not the number of units on test or the length of the test. In other words, a test where 5 fail out of a total of 10 on test gives more information than a test with 1000 units but only 2 failures.

Since the number of failures *r* is critical, and not the sample size *n* on test, it becomes increasingly difficult to assess the failure rates of highly reliable components. Parts like memory chips, that in typical use have failure rates measured in parts per million per thousand hours, will have few or no failures when tested for reasonable time periods with affordable sample sizes. This gives little or no information for accomplishing the two primary purposes of reliability testing, namely:

- accurately assessing population failure rates
- obtaining failure mode information to feedback for product improvement.

Testing at much higher than typical stresses can yield failures but models are then needed to relate these back to use stress

How can tests be designed to overcome an expected lack of failures?

The answer is to make failures occur by testing at much higher stresses than the units would normally see in their intended application. This creates a new problem: how can these failures at higher-than-normal stresses be related to what would be expected to happen over the course of many years at normal use stresses? The models that relate high stress reliability to normal use reliability are called acceleration models.

8.1.4. What is "physical acceleration" and how do we model it?

When changing stress is equivalent to multiplying time to fail by a constant, we have true (physical) acceleration

Physical Acceleration (sometimes called **True Acceleration** or just **Acceleration)** means that operating a unit at high stress (i.e., higher temperature or voltage or humidity or duty cycle, etc.) produces the same failures that would occur at typical-use stresses, except that they happen much quicker.

Failure may be due to mechanical fatigue, corrosion, chemical reaction, diffusion, migration, etc. These are the same causes of failure under normal stress; the time scale is simply different.

An Acceleration Factor is the constant multiplier between the two stress levels

When there is true acceleration, changing stress is equivalent to transforming the time scale used to record when failures occur. The transformations commonly used are *linear*, which means that time-to-fail at high stress just has to be multiplied by a constant (the **acceleration factor**) to obtain the equivalent time-to-fail at use stress.

We use the following notation:

t_s = time-to-fail at stress	t_u = corresponding time-to-fail at use
$F_s(t)$ = CDF at stress	$F_u(t)$ = CDF at use
$f_s(t)$ = PDF at stress	$f_u(t)$ = PDF at use
$h_s(t)$ = failure rate at stress	$h_u(t)$ = failure rate at use

Then, an acceleration factor AF between stress and use means the following relationships hold:

Linear Acceleration Relationships

Time-to-Fail	$t_u = AF \times t_s$
Failure Probability	$F_u(t) = F_s(t/AF)$
Reliability	$R_u(t) = R_s(t/AF)$
PDF or Density Function	$f_u(t) = (1/AF)f_s(t/AF)$
Failure Rate	$h_u(t) = (1/AF)\, h_s(t/AF)$

Each failure mode has its own acceleration factor

Failure data should be separated by failure mode when analyzed, if acceleration is relevant

Probability plots of data from different stress cells have the same slope (if there is acceleration)

Note: Acceleration requires that there be a stress dependent physical process causing change or degradation that leads to failure. In general, different failure modes will be affected differently by stress and have different acceleration factors. Therefore, it is unlikely that a single acceleration factor will apply to more than one failure mechanism. In general, different failure modes will be affected differently by stress and have different acceleration factors. Separate out different types of failure when analyzing failure data.

Also, a consequence of the linear acceleration relationships shown above (which follows directly from "true acceleration") is the following:

> The Shape Parameter for the key life distribution models (Weibull, Lognormal) does not change for units operating under different stresses. Probability plots of data from different stress cells will line up roughly parallel.

These distributions and probability plotting will be discussed in later sections.

8.1.5. What are some common acceleration models?

Acceleration models predict time to fail as a function of stress

Acceleration factors show how time-to-fail at a particular operating stress level (for one failure mode or mechanism) can be used to predict the equivalent time to fail at a different operating stress level.

A model that predicts time-to-fail as a function of stress would be even better than a collection of acceleration factors. If we write $t_f = G(S)$, with $G(S)$ denoting the model equation for an arbitrary stress level S, then the acceleration factor between two stress levels S_1 and S_2 can be evaluated simply by $AF = G(S_1)/G(S_2)$. Now we can test at the higher stress S_2, obtain a sufficient number of failures to fit life distribution models and evaluate failure rates, and use the Linear Acceleration Relationships Table to predict what will occur at the lower use stress S_1.

A model that predicts time-to-fail as a function of operating stresses is known as an **acceleration model**.

Acceleration models are often derived from physics or kinetics models related to the failure mechanism

Acceleration models are usually based on the physics or chemistry underlying a particular failure mechanism. Successful empirical models often turn out to be approximations of complicated physics or kinetics models, when the theory of the failure mechanism is better understood. The following sections will consider a variety of powerful and useful models:

- Arrhenius
- Eyring
- Other Models

8.1.5.1. Arrhenius

The Arrhenius model predicts failure acceleration due to temperature increase

One of the earliest and most successful acceleration models predicts how time-to-fail varies with temperature. This empirically based model is known as the Arrhenius equation. It takes the form

$$t_f = A \exp\left\{\frac{\Delta H}{kT}\right\}$$

with T denoting temperature measured in degrees Kelvin (273.16 + degrees Celsius) at the point when the failure process takes place and k is Boltzmann's constant (8.617 x 10^{-5} in ev/K). The constant A is a scaling factor that drops out when calculating acceleration factors, with ΔH (pronounced "Delta H") denoting the activation energy, which is the critical parameter in the model.

The Arrhenius activation energy, ΔH, is all you need to know to calculate temperature acceleration

The value of ΔH depends on the failure mechanism and the materials involved, and typically ranges from .3 or .4 up to 1.5, or even higher. Acceleration factors between two temperatures increase exponentially as ΔH increases.

The acceleration factor between a higher temperature T_2 and a lower temperature T_1 is given by

$$AF = \exp\left\{\frac{\Delta H}{k}\left[\frac{1}{T_1} - \frac{1}{T_2}\right]\right\}$$

Using the value of k given above, this can be written in terms of T in degrees Celsius as

$$AF = \exp\left\{\Delta H \times 11605 \times \left[\frac{1}{(T_1 + 273.16)} - \frac{1}{(T_2 + 273.16)}\right]\right\}$$

Note that the only unknown parameter in this formula is ΔH.

Example: The acceleration factor between 25°C and 125°C is 133 if ΔH = .5 and 17,597 if ΔH = 1.0.

The Arrhenius model has been used successfully for failure

mechanisms that depend on chemical reactions, diffusion processes or migration processes. This covers many of the non mechanical (or non material fatigue) failure modes that cause electronic equipment failure.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

8.1.5.2. Eyring

The Eyring model has a theoretical basis in chemistry and quantum mechanics and can be used to model acceleration when many stresses are involved

Henry Eyring's contributions to chemical reaction rate theory have led to a very general and powerful model for acceleration known as the Eyring Model. This model has several key features:

- It has a theoretical basis from chemistry and quantum mechanics.
- If a chemical process (chemical reaction, diffusion, corrosion, migration, etc.) is causing degradation leading to failure, the Eyring model describes how the rate of degradation varies with stress or, equivalently, how time to failure varies with stress.
- The model includes temperature and can be expanded to include other relevant stresses.
- The temperature term by itself is very similar to the Arrhenius empirical model, explaining why that model has been so successful in establishing the connection between the ΔH parameter and the quantum theory concept of "activation energy needed to cross an energy barrier and initiate a reaction".

The model for temperature and one additional stress takes the general form:

$$t_f = AT^{\alpha} \exp\left\{\frac{\Delta H}{kT} + \left(B + \frac{C}{T}\right)S_1\right\}$$

for which S_1 could be some function of voltage or current or any other relevant stress and the parameters α, ΔH, B, and C determine acceleration between stress combinations. As with the Arrhenius Model, k is Boltzmann's constant and temperature is in degrees Kelvin.

If we want to add an additional non-thermal stress term, the model becomes

$$t_f = AT^{\alpha} \exp\left\{\frac{\Delta H}{kT} + \left(B + \frac{C}{T}\right)S_1 + \left(D + \frac{E}{T}\right)S_2\right\}$$

and as many stresses as are relevant can be included by adding similar terms.

Models with multiple stresses generally have no interaction terms - which means you can multiply acceleration factors due to different stresses

Note that the general Eyring model includes terms that have stress and temperature interactions (in other words, the effect of changing temperature varies, depending on the levels of other stresses). Most models in actual use do not include any interaction terms, so that the relative change in acceleration factors when only one stress changes does not depend on the level of the other stresses.

In models with no interaction, you can compute acceleration factors for each stress and multiply them together. This would not be true if the physical mechanism required interaction terms - but, at least to first approximations, it seems to work for most examples in the literature.

The Eyring model can also be used to model rate of degradation leading to failure as a function of stress

Advantages of the Eyring Model

- Can handle many stresses.
- Can be used to model degradation data as well as failure data.
- The ΔH parameter has a physical meaning and has been studied and estimated for many well known failure mechanisms and materials.

In practice, the Eyring Model is usually too complicated to use in its most general form and must be "customized" or simplified for any particular failure mechanism

Disadvantages of the Eyring Model

- Even with just two stresses, there are 5 parameters to estimate. Each additional stress adds 2 more unknown parameters.
- Many of the parameters may have only a second-order effect. For example, setting α = 0 works quite well since the temperature term then becomes the same as in the Arrhenius model. Also, the constants C and E are only needed if there is a significant temperature interaction effect with respect to the other stresses.
- The form in which the other stresses appear is not specified by the general model and may vary according to the particular failure mechanism. In other words, S_1 may be voltage or ln (voltage) or some other function of voltage.

Many well-known models are simplified versions of the Eyring model with appropriate functions of relevant stresses chosen for S_1 and S_2. Some of these will be shown in the Other Models section. The trick is to find the right simplification to use for a particular failure mechanism.

8.1.5.3. Other models

Many useful 1, 2 and 3 stress models are simple Eyring models. Six are described

This section will discuss several acceleration models whose successful use has been described in the literature.

- The (Inverse) Power Rule for Voltage
- The Exponential Voltage Model
- Two Temperature/Voltage Models
- The Electromigration Model
- Three Stress Models (Temperature, Voltage and Humidity)
- The Coffin-Manson Mechanical Crack Growth Model

The (Inverse) Power Rule for Voltage

This model, used for capacitors, has only voltage dependency and takes the form:

$$t_f = AV^{-\beta}$$

This is a very simplified Eyring model with α, ΔH, and C all 0, and $S = \ln V$, and $\beta = -B$.

The Exponential Voltage Model

In some cases, voltage dependence is modeled better with an exponential model:

$$t_f = Ae^{-BV}$$

Two Temperature/Voltage Models

Temperature/Voltage models are common in the literature and take one of the two forms given below:

$$t_f = Ae^{\frac{\Delta H}{kT}}V^{-\beta}$$

or

$$t_f = Ae^{\frac{\Delta H}{kT}}e^{-BV}$$

Again, these are just simplified two stress Eyring models with the appropriate choice of constants and functions of voltage.

The Electromigration Model

Electromigration is a semiconductor failure mechanism where open failures occur in metal thin film conductors due to the movement of ions toward the anode. This ionic movement is accelerated high temperatures and high current density. The (modified Eyring) model takes the form

$$t_f = AJ^{-n}e^{\frac{\Delta H}{kT}}$$

with J denoting the current density. ΔH is typically between .5 and 1.2 electron volts, while an n around 2 is common.

Three-Stress Models (Temperature, Voltage and Humidity)

Humidity plays an important role in many failure mechanisms that depend on corrosion or ionic movement. A common 3-stress model takes the form

$$t_f = Ae^{\frac{\Delta H}{kT}}V^{-\beta}RH^{-\gamma}$$

Here RH is percent relative humidity. Other obvious variations on this model would be to use an exponential voltage term and/or an exponential RH term.

Even this simplified Eyring 3-stress model has 4 unknown parameters and an extensive experimental setup would be required to fit the model and calculate acceleration factors.

The Coffin-Manson Model is a useful non-Eyring model for crack growth or material fatigue

The Coffin-Manson Mechanical Crack Growth Model

Models for mechanical failure, material fatigue or material deformation are not forms of the Eyring model. These models typically have terms relating to cycles of stress or frequency of use or change in temperatures. A model of this type known as the (modified) Coffin-Manson model has been used successfully to model crack growth in solder and other metals due to repeated temperature cycling as equipment is turned on and off. This model takes the form

$$N_f = Af^{-\alpha}\Delta T^{-\beta}G(T_{max})$$

with

- N_f = the number of cycles to fail

- f = the cycling frequency
- ΔT = the temperature range during a cycle

and $G(T_{max})$ is an Arrhenius term evaluated at the maximum temperature reached in each cycle.

Typical values for the cycling frequency exponent α and the temperature range exponent β are around -1/3 and 2, respectively (note that reducing the cycling frequency reduces the number of cycles to failure). The ΔH activation energy term in $G(T_{max})$ is around 1.25.

8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

A handful of lifetime distribution models have enjoyed great practical success

There are a handful of parametric models that have successfully served as population models for failure times arising from a wide range of products and failure mechanisms. Sometimes there are probabilistic arguments based on the physics of the failure mode that tend to justify the choice of model. Other times the model is used solely because of its empirical success in fitting actual failure data.

Seven models will be described in this section:

1. Exponential
2. Weibull
3. Extreme Value
4. Lognormal
5. Gamma
6. Birnbaum-Saunders
7. Proportional hazards

8. Assessing Product Reliability
8.1. Introduction
8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

8.1.6.1. Exponential

All the key formulas for using the exponential model

Formulas and Plots

The exponential model, with only one unknown parameter, is the simplest of all life distribution models. The key equations for the exponential are shown below:

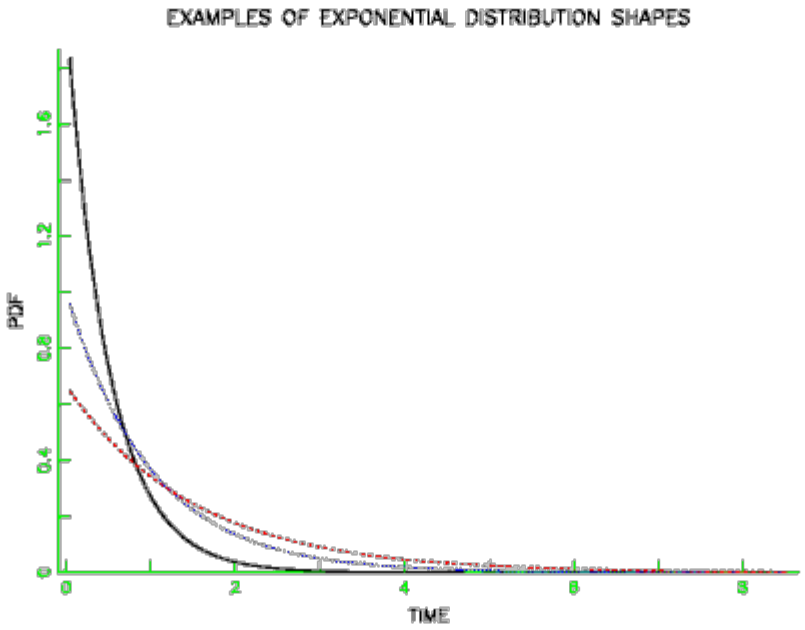
PDF:	$f(t,\lambda) = \lambda e^{-\lambda t}$
CDF:	$F(t) = 1 - e^{-\lambda t}$
Reliability:	$R(t) = e^{-\lambda t}$
Failure Rate:	$h(t) = \lambda$
Mean:	$\frac{1}{\lambda}$
Median:	$\frac{\ln 2}{\lambda} \cong \frac{0.693}{\lambda}$
Variance:	$\frac{1}{\lambda^2}$

Note that the failure rate reduces to the constant λ for any time. The exponential distribution is the only distribution to have a constant failure rate. Also, another name for the exponential mean is the **Mean Time To Fail** or **MTTF** and we have MTTF = $1/\lambda$.

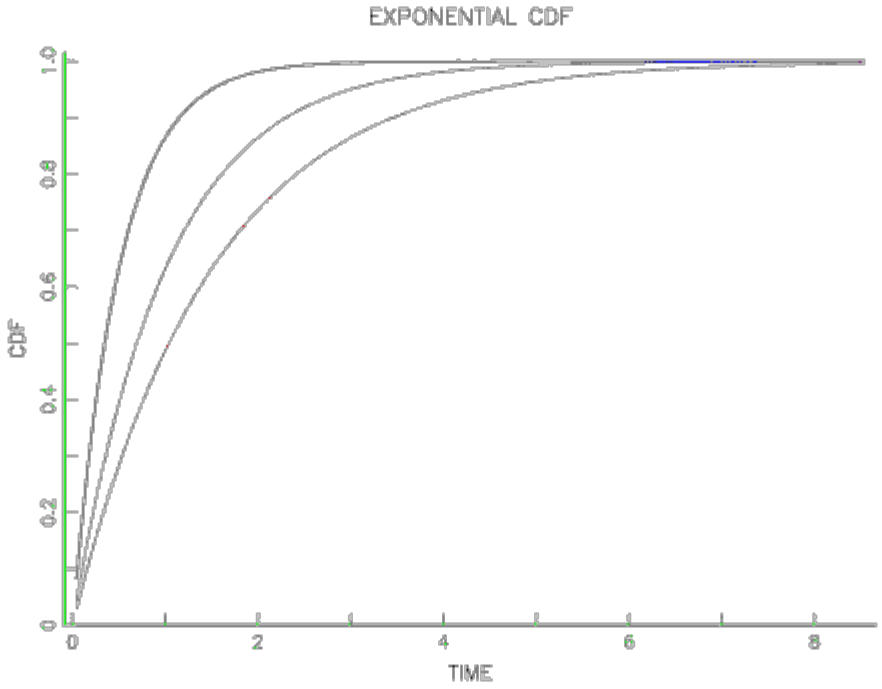
The cumulative hazard function for the exponential is just the integral of the failure rate or $H(t) = \lambda t$.

The PDF for the exponential has the familiar shape shown below.

The Exponential distribution 'shape'

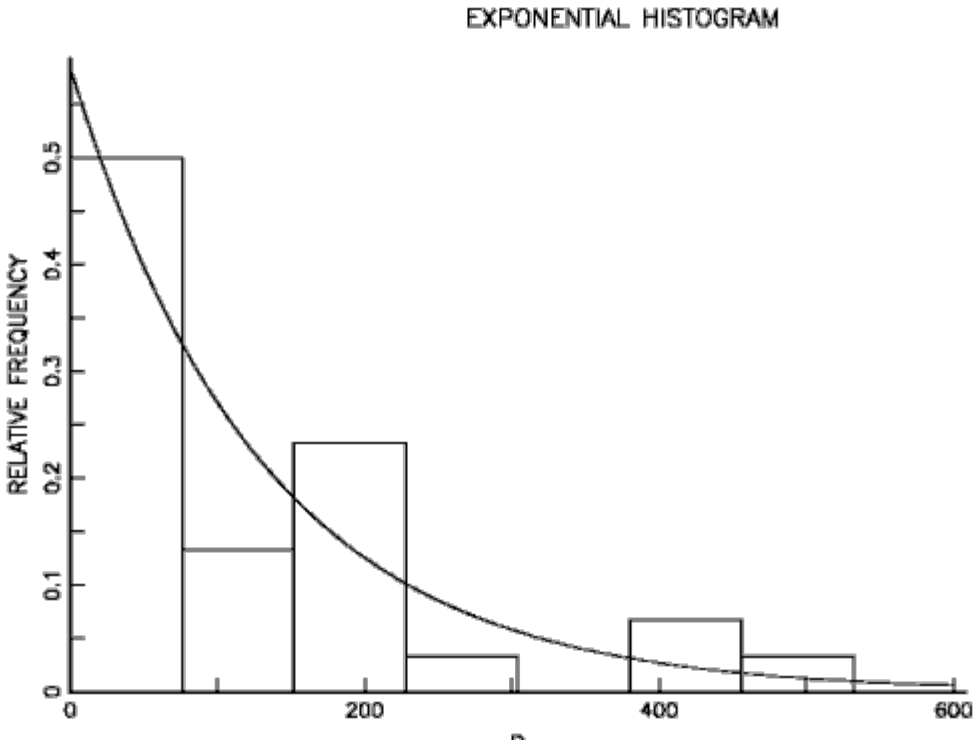


The Exponential CDF



Below is an example of typical exponential lifetime data displayed in Histogram form with corresponding exponential PDF drawn through the histogram.

Histogram of Exponential Data



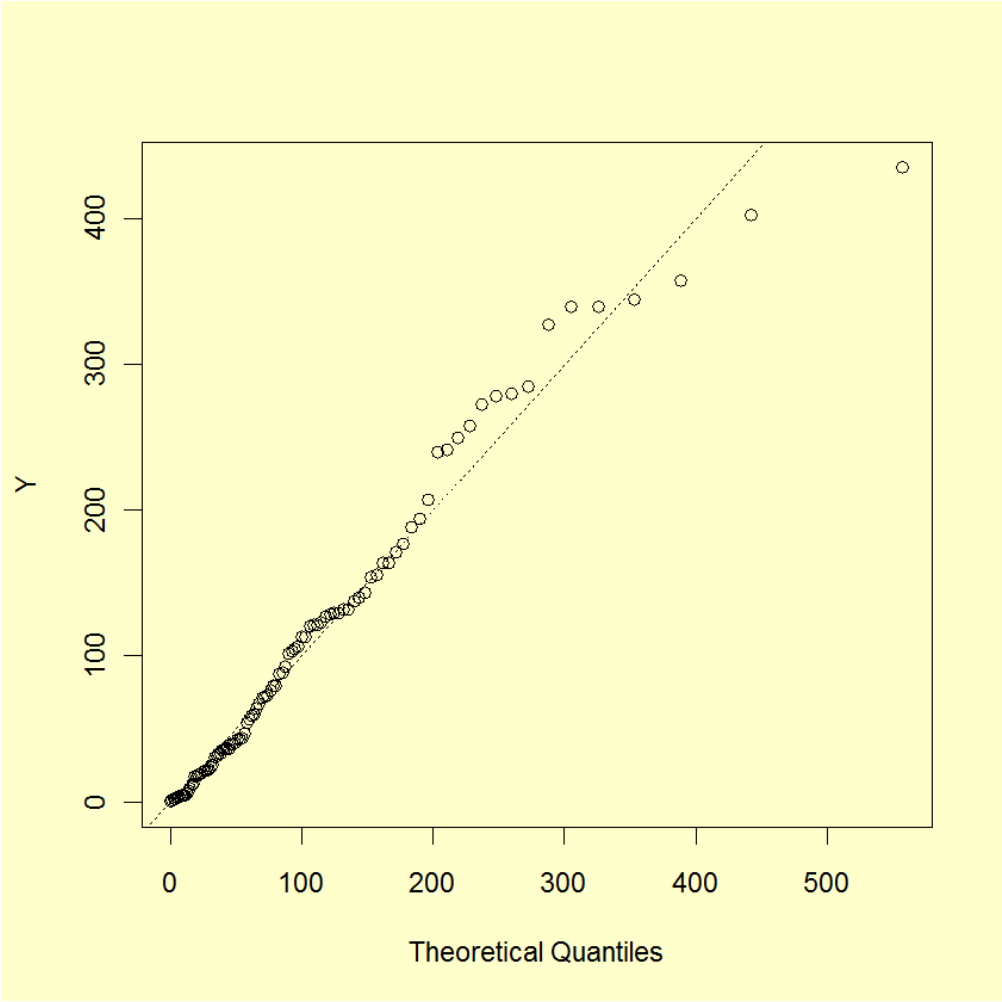
The Exponential models the flat portion of the "bathtub" curve - where most systems spend most of their 'lives'

Uses of the Exponential Distribution Model

1. Because of its constant failure rate property, the exponential distribution is an excellent model for the long flat "intrinsic failure" portion of the Bathtub Curve. Since most components and systems spend most of their lifetimes in this portion of the Bathtub Curve, this justifies frequent use of the exponential distribution (when early failures or wear out is not a concern).
2. Just as it is often useful to approximate a curve by piecewise straight line segments, we can approximate any failure rate curve by week-by-week or month-by-month constant rates that are the average of the actual changing rate during the respective time durations. That way we can approximate any model by piecewise exponential distribution segments patched together.
3. Some natural phenomena have a constant failure rate (or occurrence rate) property; for example, the arrival rate of cosmic ray alpha particles or Geiger counter tics. The exponential model works well for inter arrival times (while the Poisson distribution describes the total number of events in a given period). When these events trigger failures, the exponential life distribution model will naturally apply.

Exponential probability plot

We can generate a probability plot of normalized exponential data, so that a perfect exponential fit is a diagonal line with slope 1. The probability plot for 100 normalized random exponential observations ($\lambda = 0.01$) is shown below.



We can calculate the exponential PDF and CDF at 100 hours for the case where $\lambda = 0.01$. The PDF value is 0.0037 and the CDF value is 0.6321.

Functions for computing exponential PDF values, CDF values, and for producing probability plots, are found in both Dataplot code and R code.

8. Assessing Product Reliability
8.1. Introduction
8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

8.1.6.2. Weibull

Weibull Formulas

Formulas and Plots

The Weibull is a very flexible life distribution model with two parameters. It has CDF and PDF and other key formulas given by:

PDF:	$f(t,\gamma,\alpha) = \frac{\gamma}{t}\left(\frac{t}{\alpha}\right)^{\gamma} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$
CDF:	$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$
Reliability:	$R(t) = e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$
Failure Rate:	$h(t) = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1}$
Mean:	$\alpha\Gamma\left(1+\frac{1}{\gamma}\right)$
Median:	$\alpha(\ln 2)^{\frac{1}{\gamma}}$
Variance:	$\alpha^2\Gamma\left(1+\frac{2}{\gamma}\right) - \left[\alpha\Gamma\left(1+\frac{1}{\gamma}\right)\right]^2$

with α the scale parameter (the **Characteristic Life**), γ (gamma) the **Shape Parameter,** and Γ is the Gamma function with $\Gamma(N) = (N\text{-}1)!$ for integer N.

The cumulative hazard function for the Weibull is the integral of the failure rate or

$$H(t) = \left(\frac{t}{\alpha}\right)^{\gamma}$$

A more general three-parameter form of the Weibull includes an additional **waiting time** parameter μ (sometimes called a **shift** or **location** parameter). The formulas for the 3-parameter Weibull are easily obtained from the above formulas by replacing t by $(t - \mu)$ wherever t appears. No failure can occur before μ hours, so the time scale starts at μ, and not 0. If a shift parameter μ is known (based, perhaps, on the physics of the failure mode), then all you have to do is subtract μ from all the observed failure times and/or readout times and

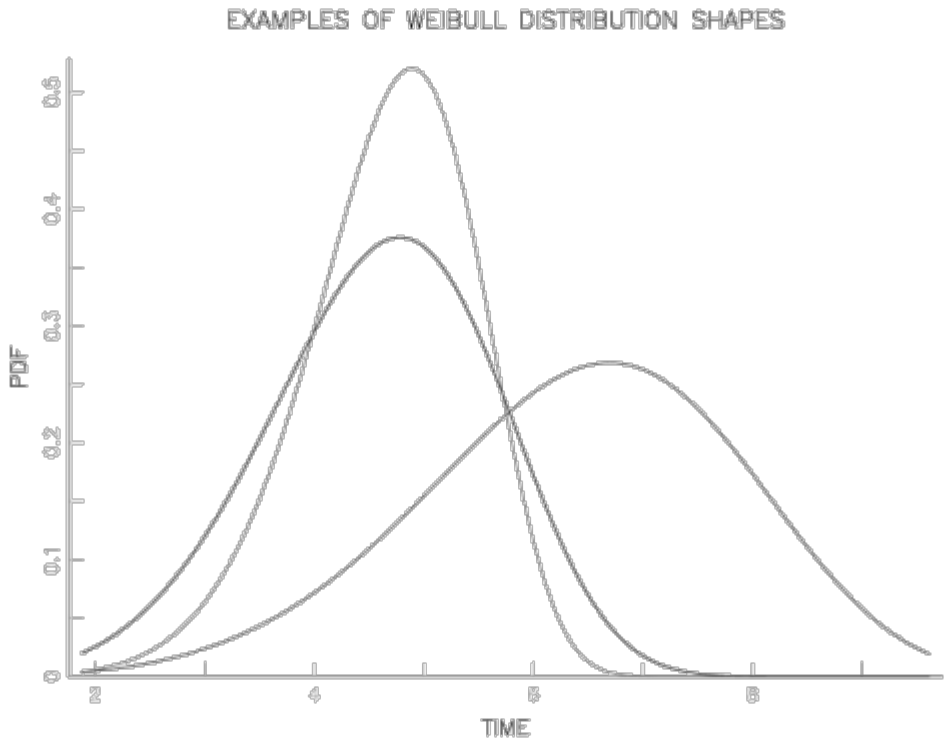
analyze the resulting shifted data with a two-parameter Weibull.

NOTE: Various texts and articles in the literature use a variety of different symbols for the same Weibull parameters. For example, the characteristic life is sometimes called c (v = nu or η = eta) and the shape parameter is also called m (or β = beta). To add to the confusion, some software uses β as the characteristic life parameter and α as the shape parameter. Some authors even parameterize the density function differently, using a scale parameter $\theta = \alpha^{\gamma}$.

Special Case: When γ = 1, the Weibull reduces to the Exponential Model, with α = 1/λ = the **mean time to fail (MTTF)**.

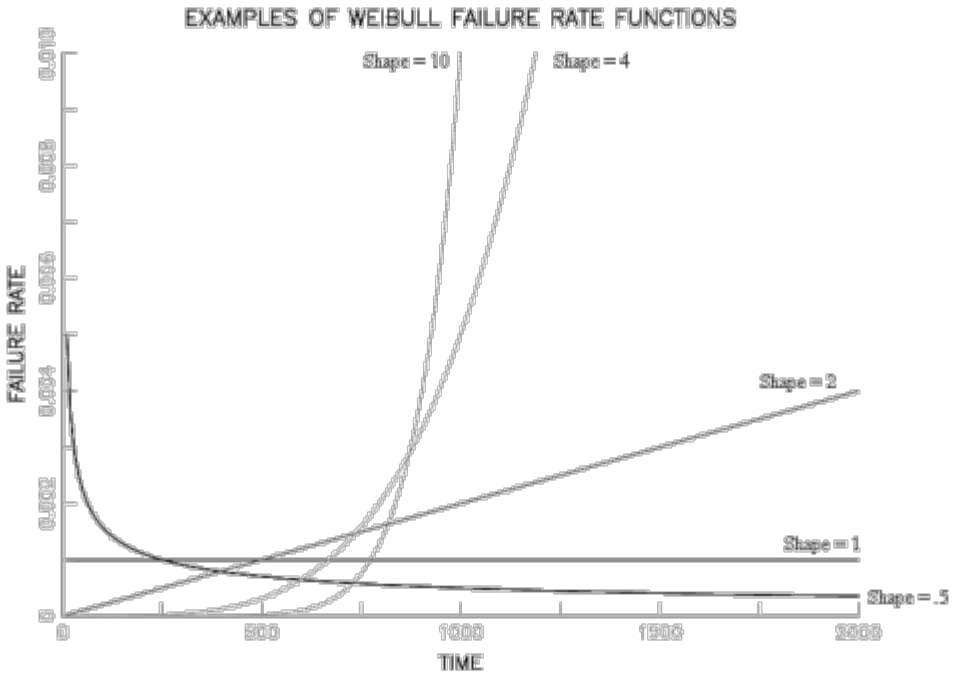
Depending on the value of the shape parameter γ, the Weibull model can empirically fit a wide range of data histogram shapes. This is shown by the PDF example curves below.

Weibull data 'shapes'



From a failure rate model viewpoint, the Weibull is a natural extension of the constant failure rate exponential model since the Weibull has a polynomial failure rate with exponent {γ - 1}. This makes all the failure rate curves shown in the following plot possible.

Weibull failure rate 'shapes'



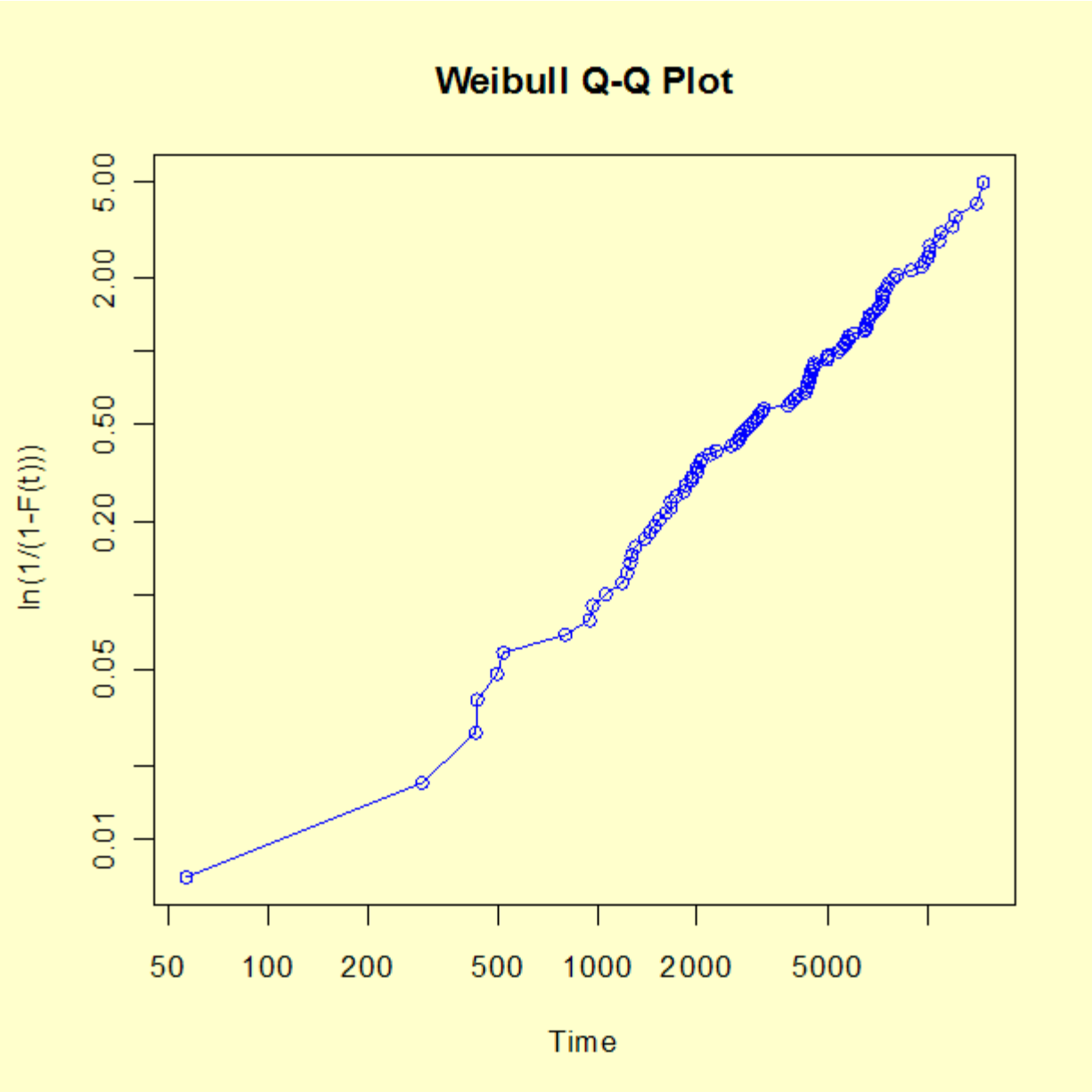
The Weibull is very flexible and also has theoretical justification in many applications

Uses of the Weibull Distribution Model

1. Because of its flexible shape and ability to model a wide range of failure rates, the Weibull has been used successfully in many applications as a purely empirical model.
2. The Weibull model can be derived theoretically as a form of Extreme Value Distribution, governing the time to occurrence of the "weakest link" of many competing failure processes. This may explain why it has been so successful in applications such as capacitor, ball bearing, relay and material strength failures.
3. Another special case of the Weibull occurs when the shape parameter is 2. The distribution is called the Rayleigh Distribution and it turns out to be the theoretical probability model for the magnitude of radial error when the x and y coordinate errors are independent normals with 0 mean and the same standard deviation.

Weibull probability plot

We generated 100 Weibull random variables using T = 1000, γ = 1.5 and α = 5000. To see how well these random Weibull data points are actually fit by a Weibull distribution, we generated the probability plot shown below. Note the log scale used is base 10.



If the data follow a Weibull distribution, the points should follow a straight line.

We can comput the PDF and CDF values for failure time T = 1000, using the example Weibull distribution with γ = 1.5 and α = 5000. The PDF value is 0.000123 and the CDF value is 0.08556.

Functions for computing Weibull PDF values, CDF values, and for producing probability plots, are found in both Dataplot code and R code.

8. Assessing Product Reliability
8.1. Introduction
8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

8.1.6.3. Extreme value distributions

The Extreme Value Distribution usually refers to the distribution of the minimum of a large number of unbounded random observations

Description, Formulas and Plots

We have already referred to **Extreme Value Distributions** when describing the uses of the Weibull distribution. Extreme value distributions are the limiting distributions for the minimum or the maximum of a very large collection of random observations from the same arbitrary distribution. Gumbel (1958) showed that for any well-behaved initial distribution (i.e., *F(x)* is continuous and has an inverse), only a few models are needed, depending on whether you are interested in the maximum or the minimum, and also if the observations are bounded above or below.

In the context of reliability modeling, extreme value distributions for the minimum are frequently encountered. For example, if a system consists of *n* identical components in series, and the system fails when the first of these components fails, then system failure times are the minimum of *n* random component failure times. Extreme value theory says that, independent of the choice of component model, the system model will approach a Weibull as *n* becomes large. The same reasoning can also be applied at a component level, if the component failure occurs when the first of many similar competing failure processes reaches a critical level.

The distribution often referred to as the **Extreme Value Distribution (Type I)** is the limiting distribution of the minimum of a large number of unbounded identically distributed random variables. The PDF and CDF are given by:

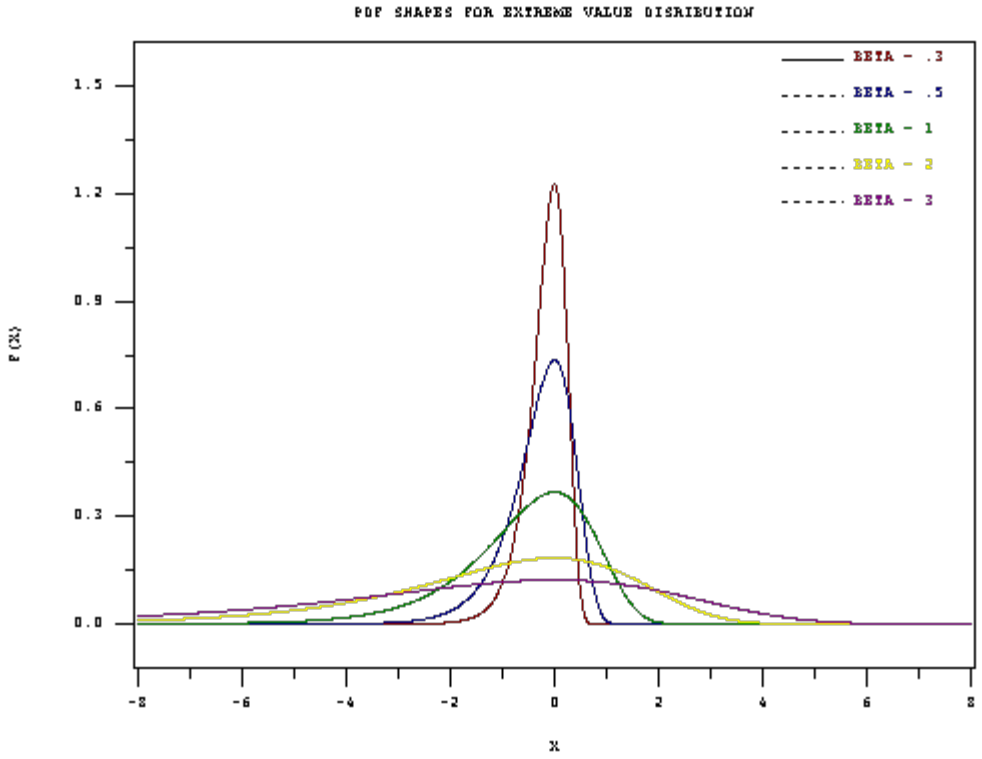
Extreme Value Distribution formulas and PDF shapes

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}}, \quad -\infty < x < \infty, \beta > 0$$

$$F(x) = 1 - e^{-e^{\frac{x-\mu}{\beta}}}, \quad -\infty < x < \infty, \beta > 0$$

If the *x* values are bounded below (as is the case with times of failure) then the limiting distribution is the Weibull. Formulas and uses of the Weibull have already been discussed.

PDF Shapes for the (minimum) Extreme Value Distribution (Type I) are shown in the following figure.



The natural log of Weibull data is extreme value data

Uses of the Extreme Value Distribution Model

1. In any modeling application for which the variable of interest is the minimum of many random factors, all of which can take positive or negative values, try the extreme value distribution as a likely candidate model. For lifetime distribution modeling, since failure times are bounded below by zero, the Weibull distribution is a better choice.
2. The Weibull distribution and the extreme value distribution have a useful mathematical relationship. If $t_1, t_2, ...,t_n$ are a sample of random times of fail from a Weibull distribution, then ln t_1, ln t_2, ...,ln t_n are random observations from the extreme value distribution. In other words, the natural log of a Weibull random time is an extreme value random observation.

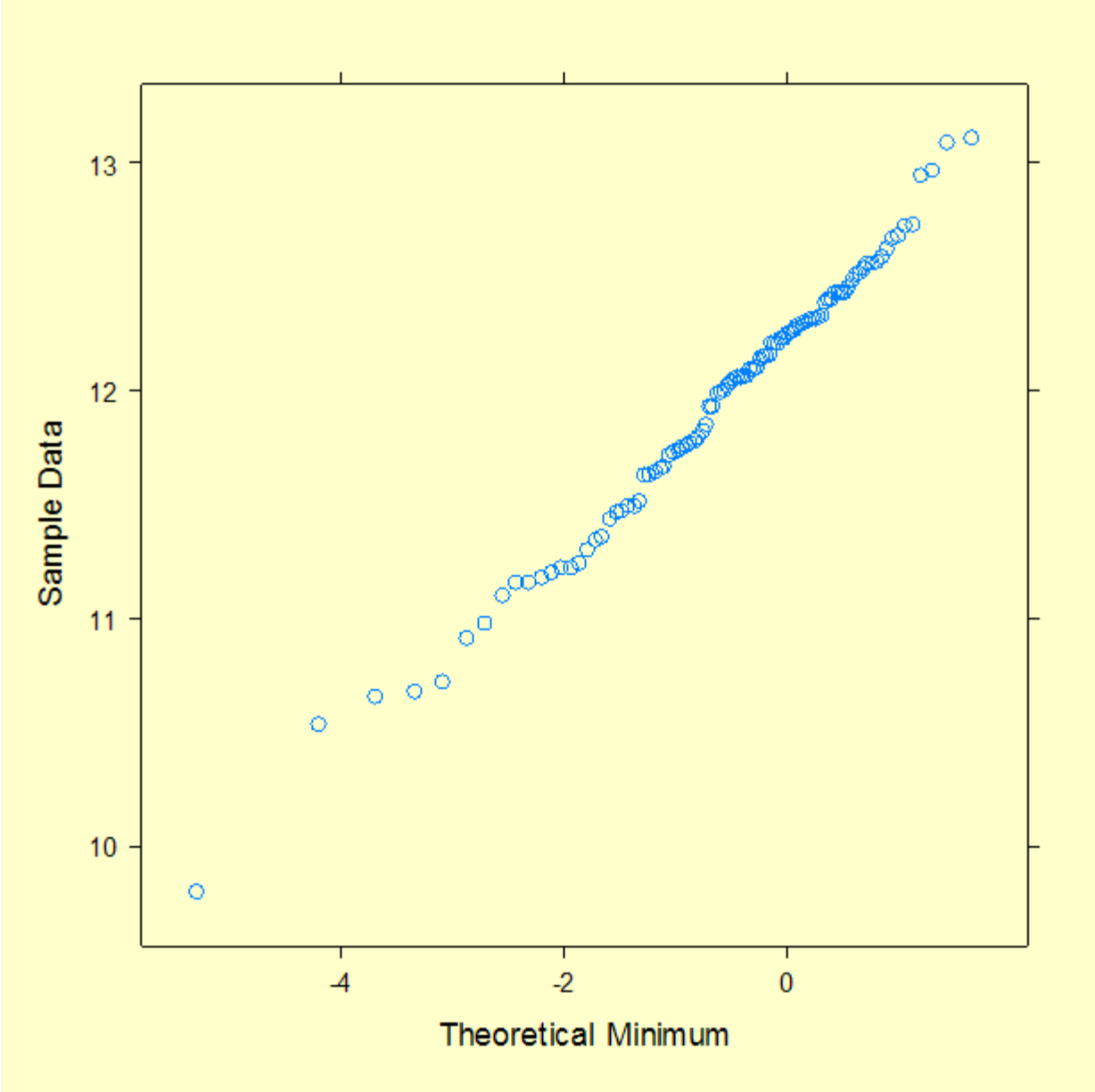
If the Weibull has shape parameter γ and characteristic life α, then the extreme value distribution (after taking natural logarithms) has $\mu = \ln \alpha$, $\beta = 1/\gamma$.

Because of this relationship, computer programs designed for the extreme value distribution can be used to analyze Weibull data. The situation exactly parallels using normal distribution programs to analyze lognormal data, after first taking natural logarithms of the data points.

Probability plot for the extreme value distribution

Assume μ = ln 200,000 = 12.206 and β = 1/2 = 0.5. The extreme value distribution associated with these parameters could be obtained by taking natural logarithms of data from a Weibull population with characteristic life α = 200,000 and shape γ = 2.

We generate 100 random numbers from this extreme value distribution and construct the following probability plot.



Data from an extreme value distribution will line up approximately along a straight line when this kind of plot is constructed. The slope of the line is an estimate of β, "y-axis" value on the line corresponding to the "x-axis" 0 point is an estimate of μ. For the graph above, these turn out to be very close to the actual values of β and μ.

For the example extreme value distribution with μ = ln 200,000 = 12.206 and β = 1/2 = 0.5, the PDF values corresponding to the points 5, 8, 10, 12, 12.8. are 0.110E-5, 0.444E-3, 0.024, 0.683 and 0.247. and the CDF values corresponding to the same points are 0.551E-6, 0.222E-3, 0.012, 0.484 and 0.962.

Functions for computing extreme value distribution PDF values, CDF values, and for producing probability plots, are found in both Dataplot code and R code.

8.1.6.4. Lognormal

Lognormal Formulas and relationship to the normal distribution

Formulas and Plots

The lognormal life distribution, like the Weibull, is a very flexible model that can empirically fit many types of failure data. The two-parameter form has parameters σ is the **shape** parameter and T_{50} is the **median** (a **scale** parameter).

Note: If time to failure, t_f, has a lognormal distribution, then the (natural) logarithm of time to failure has a normal distribution with mean $\mu = \ln T_{50}$ and standard deviation σ. This makes lognormal data convenient to work with; just take natural logarithms of all the failure times and censoring times and analyze the resulting normal data. Later on, convert back to real time and lognormal parameters using σ as the lognormal shape and $T_{50} = e^{\mu}$ as the (median) scale parameter.

Below is a summary of the key formulas for the lognormal.

PDF: $$f(t) = \frac{1}{\sigma t \sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln T_{50})^2}$$

CDF: $$F(t) = \int_0^T \frac{1}{\sigma t \sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln T_{50})^2} dt$$

$$F(t) = \Phi\left(\frac{\ln t - \ln T_{50}}{\sigma}\right)$$

$\Phi(z)$ denotes the standard normal CDF.

Reliability: $$R(t) = 1 - F(t)$$

Failure Rate: $$h(t) = \frac{f(t)}{R(t)}$$

Mean: $$T_{50}\, e^{\frac{\sigma^2}{2}}$$

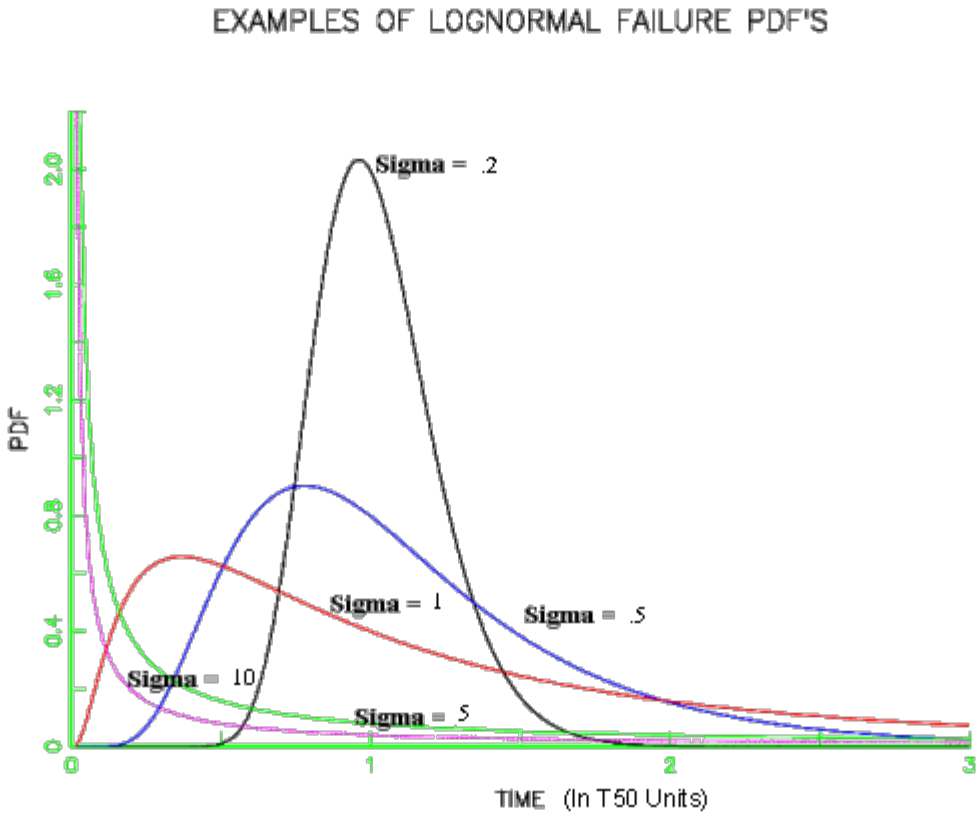
Median: $$T_{50}$$

Variance: $$T_{50}^2\, e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$$

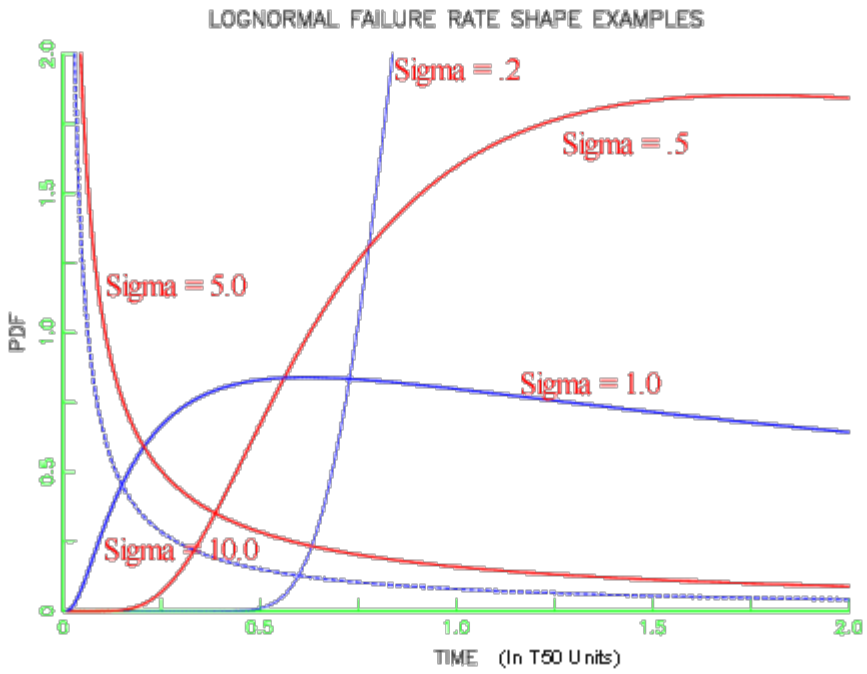
Note: A more general three-parameter form of the lognormal includes an additional **waiting time** parameter θ (sometimes called a **shift** or **location** parameter). The formulas for the three-parameter lognormal are easily obtained from the above formulas by replacing *t* by (*t* - θ) wherever *t* appears. No failure can occur before θ hours, so the time scale starts at θ and not 0. If a shift parameter θ is known (based, perhaps, on the physics of the failure mode), then all you have to do is subtract θ from all the observed failure times and/or readout times and analyze the resulting shifted data with a two-parameter lognormal.

Examples of lognormal PDF and failure rate plots are shown below. Note that lognormal shapes for small sigmas are very similar to Weibull shapes when the shape parameter γ is large and large sigmas give plots similar to small Weibull γ's. Both distributions are very flexible and it is often difficult to choose which to use based on empirical fits to small samples of (possibly censored) data.

Lognormal data 'shapes'



Lognormal failure rate 'shapes'



A very flexible model that also can apply (theoretically) to many degradation process

Uses of the Lognormal Distribution Model

1. As shown in the preceding plots, the lognormal PDF and failure rate shapes are flexible enough to make the lognormal a very useful empirical model. In addition, the relationship to the normal (just take natural logarithms of all the data and time points and you have "normal" data) makes it easy to work with mathematically, with many good software analysis programs available to treat normal data.
2. The lognormal model can be theoretically derived under assumptions matching many

failure modes

failure degradation processes common to electronic (semiconductor) failure mechanisms. Some of these are: corrosion, diffusion, migration, crack growth, electromigration, and, in general, failures resulting from chemical reactions or processes. That does not mean that the lognormal is always the correct model for these mechanisms, but it does perhaps explain why it has been empirically successful in so many of these cases.

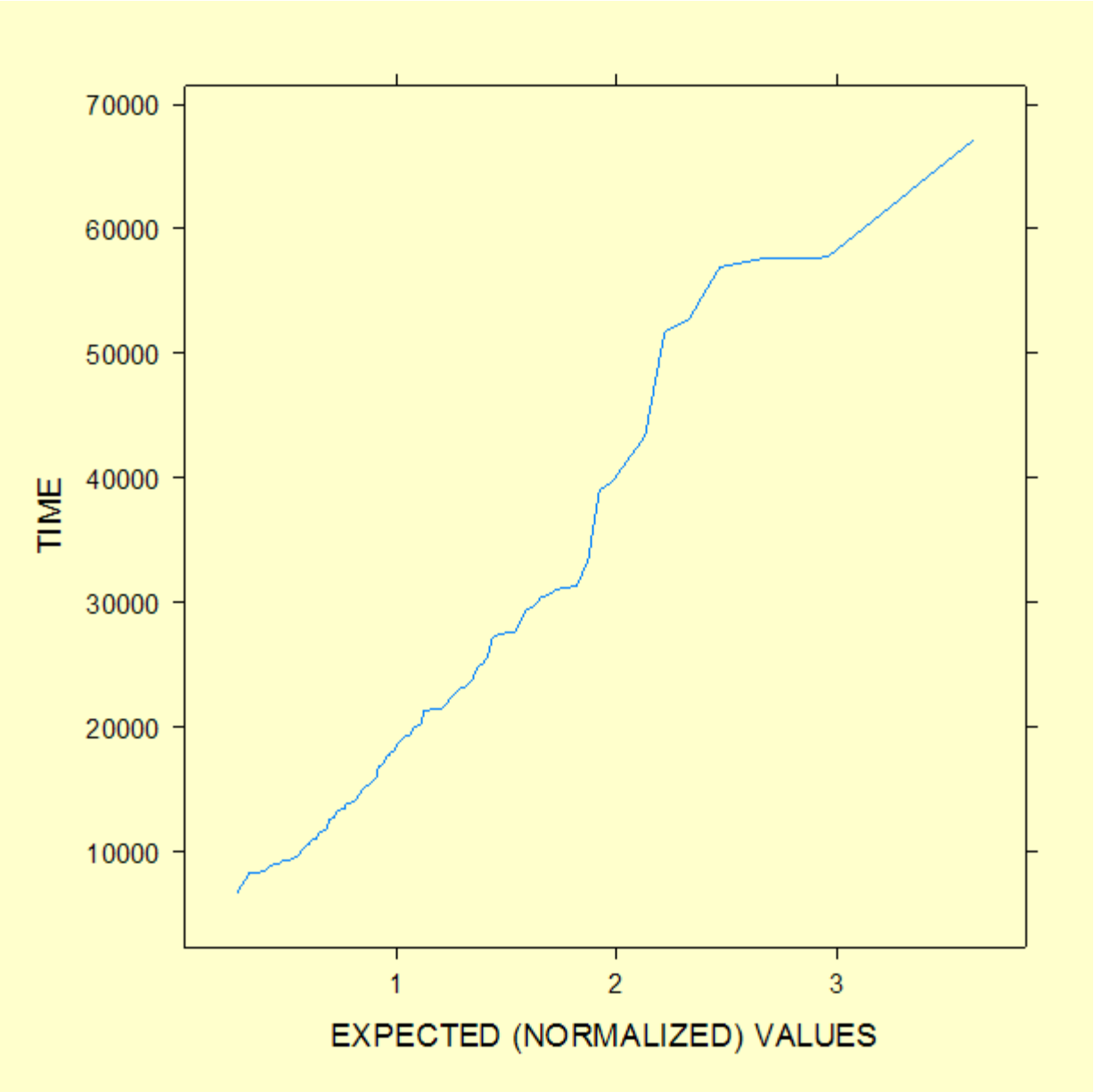
A brief sketch of the theoretical arguments leading to a lognormal model follows.

> Applying the Central Limit Theorem to small additive errors in the log domain and justifying a normal model is equivalent to justifying the lognormal model in real time when a process moves towards failure based on the cumulative effect of many small "multiplicative" shocks. More precisely, if at any instant in time a degradation process undergoes a small increase in the total amount of degradation that is proportional to the current total amount of degradation, then it is reasonable to expect the time to failure (i.e., reaching a critical amount of degradation) to follow a lognormal distribution (Kolmogorov, 1941).

A more detailed description of the multiplicative degradation argument appears in a later section.

Lognormal probability plot

We generated 100 random numbers from a lognormal distribution with shape 0.5 and median life 20,000. To see how well these random lognormal data points are fit by a lognormal distribution, we generate the lognormal probability plot shown below. Points that line up approximately on a straight line indicates a good fit to a lognormal (with shape 0.5). The time that corresponds to the (normalized) x-axis T_{50} of 1 is the estimated T_{50} according to the data. In this case it is close to 20,000, as expected.



For a lognormal distribution at time T = 5000 with σ = 0.5 and T_{50} = 20,000, the PDF value is 0.34175E-5, the CDF value is 0.002781, and the failure rate is 0.3427E-5.

Functions for computing lognormal distribution PDF values, CDF values, failure rates, and for producing probability plots, are found in both Dataplot code and R code.

8.1.6.5. Gamma

Formulas for the gamma model

Formulas and Plots

There are two ways of writing (parameterizing) the gamma distribution that are common in the literature. In addition, different authors use different symbols for the shape and scale parameters. Below we show two ways of writing the gamma, with "shape" parameter $a = \alpha$, and "scale" parameter $b = 1/\beta$.

PDF:	$f(t,a,b) = \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt}$
	$f(t,\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-t/\beta}$
CDF:	$F(t) = \int_0^T f(t)dt$
Reliability:	$R(t) = 1 - F(t)$
Failure Rate:	$h(t) = f(t)/R(t)$
Mean:	a/b or $\alpha\beta$
Variance:	a/b^2 or $\alpha\beta^2$

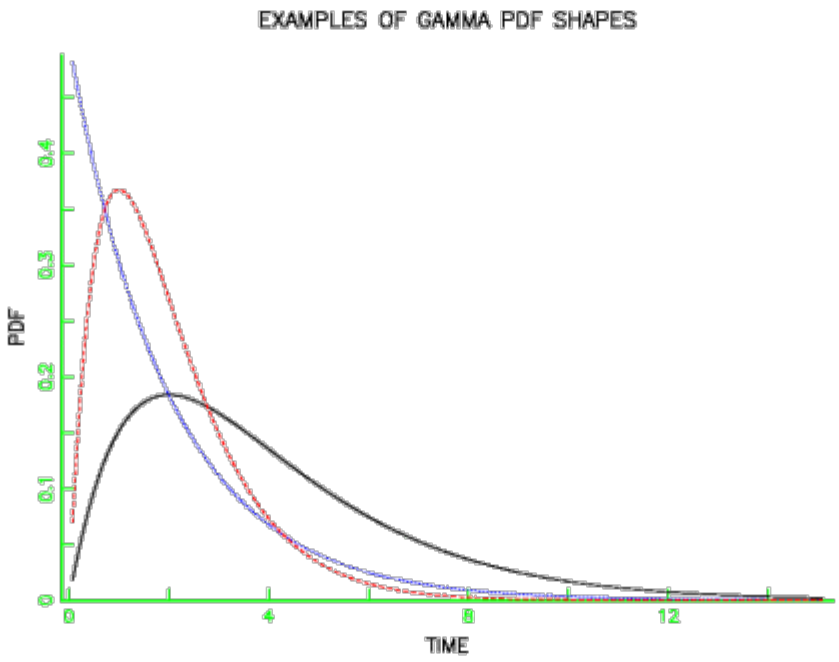
The exponential is a special case of the gamma

Note: When $a = 1$, the gamma reduces to an exponential distribution with $b = \lambda$.

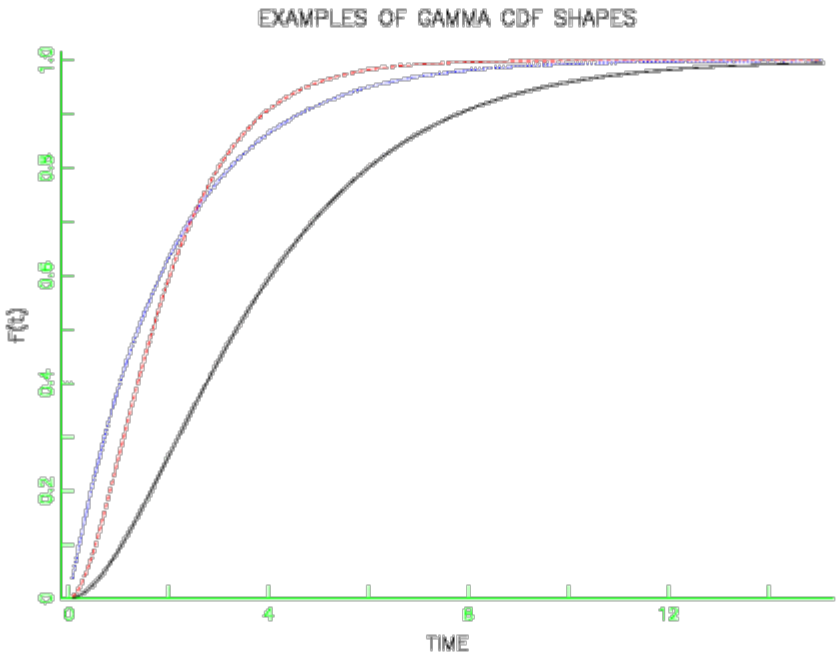
Another well-known statistical distribution, the Chi-Square, is also a special case of the gamma. A Chi-Square distribution with n degrees of freedom is the same as a gamma with $a = n/2$ and $b = 0.5$ (or $\beta = 2$).

The following plots give examples of gamma PDF, CDF and failure rate shapes.

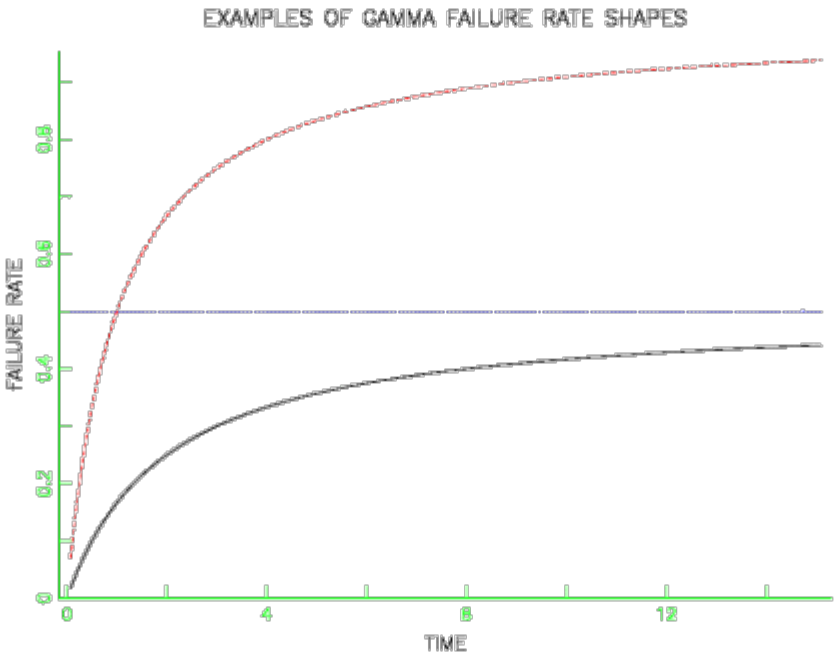
Shapes for gamma data



Gamma CDF shapes



Gamma failure rate shapes



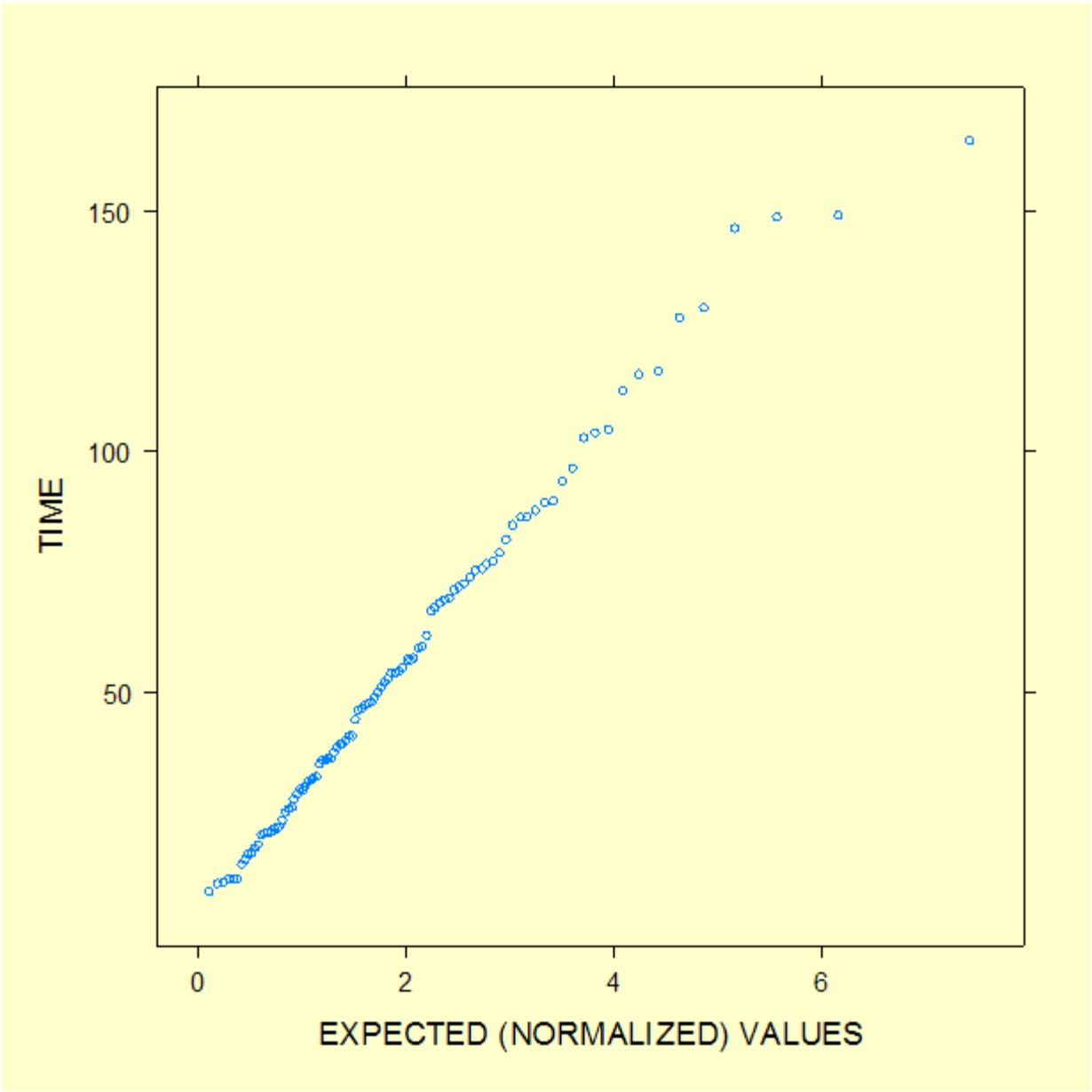
The gamma is used in "Standby" system models and also for Bayesian reliability analysis

Uses of the Gamma Distribution Model

1. The gamma is a flexible life distribution model that may offer a good fit to some sets of failure data. It is not, however, widely used as a life distribution model for common failure mechanisms.
2. The gamma does arise naturally as the time-to-first fail distribution for a system with standby exponentially distributed backups. If there are n-1 standby backup units and the system and all backups have exponential lifetimes with parameter λ, then the total lifetime has a gamma distribution with $a = n$ and $b = \lambda$. **Note**: when a is a positive integer, the gamma is sometimes called an **Erlang distribution**. The Erlang distribution is used frequently in queuing theory applications.
3. A common use of the gamma model occurs in Bayesian reliability applications. When a system follows an HPP (exponential) model with a constant repair rate λ, and it is desired to make use of prior information about possible values of λ, a gamma Bayesian prior for λ is a convenient and popular choice.

Gamma probability plot

We generated 100 random gamma data points using shape parameter $\alpha = 2$ and scale parameter $\beta = 30$. A gamma probability plot of the 100 data points is shown below.



The value of the shape parameter α can be estimated from data using

$$\hat{\alpha} = \left[\frac{\bar{t}}{s_t}\right]^2$$

the squared ratio of mean failure time to the standard deviation of the failure times.

Using an example solved in the section on standby models, where $\alpha = 2$, $\beta = 30$, and $t = 24$ months, the PDF, CDF, reliability, and failure rate are the following.

```
PDF = 0.01198
CDF = 0.19121
Reliability = 0.80879
Failure Rate = 0.01481
```

Functions for computing gamma distribution PDF values, CDF values, reliability values, failure rates, and for producing probability plots, are found in both Dataplot code and R code.

8. Assessing Product Reliability
8.1. Introduction
8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

8.1.6.6. Fatigue life (Birnbaum-Saunders)

A model based on cycles of stress causing degradation or crack growth

In 1969, Birnbaum and Saunders described a life distribution model that could be derived from a physical fatigue process where crack growth causes failure. Since one of the best ways to choose a life distribution model is to derive it from a physical/statistical argument that is consistent with the failure mechanism, the Birnbaum-Saunders fatigue life distribution is worth considering.

Formulas and shapes for the fatigue life model

Formulas and Plots for the Birnbaum-Saunders Model

The PDF, CDF, mean and variance for the Birnbaum-Saunders distribution are shown below. The parameters are: γ, a shape parameter; and μ, a scale parameter. These are the parameters we will use in our discussion, but there are other choices also common in the literature (see the parameters used for the derivation of the model).

PDF:
$$f(t) = \frac{1}{2\mu^2\gamma^2\sqrt{\pi}}\left(\frac{t^2-\mu^2}{\sqrt{\frac{t}{u}}-\sqrt{\frac{u}{t}}}\right)e^{-\frac{1}{\gamma^2}\left(\frac{t}{\mu}+\frac{\mu}{t}-2\right)}$$

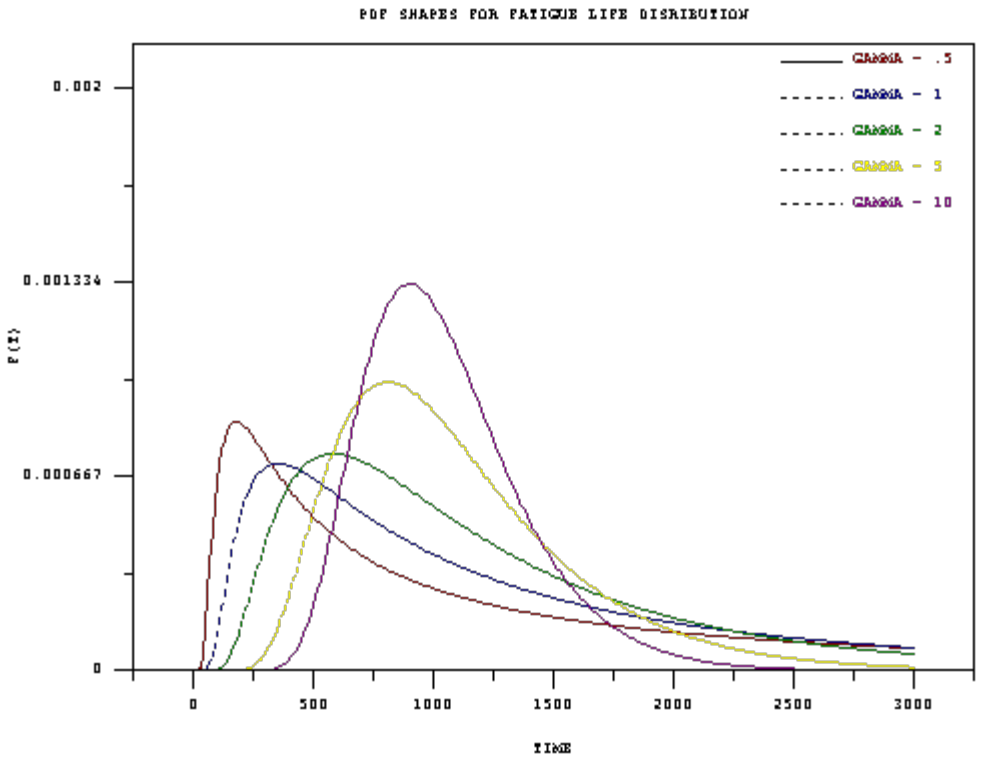
CDF:
$$F(t) = \Phi\left(\frac{1}{\gamma}\left[\sqrt{\frac{t}{\mu}}-\sqrt{\frac{\mu}{t}}\right]\right)$$

$\Phi(z)$ denotes the standard normal CDF.

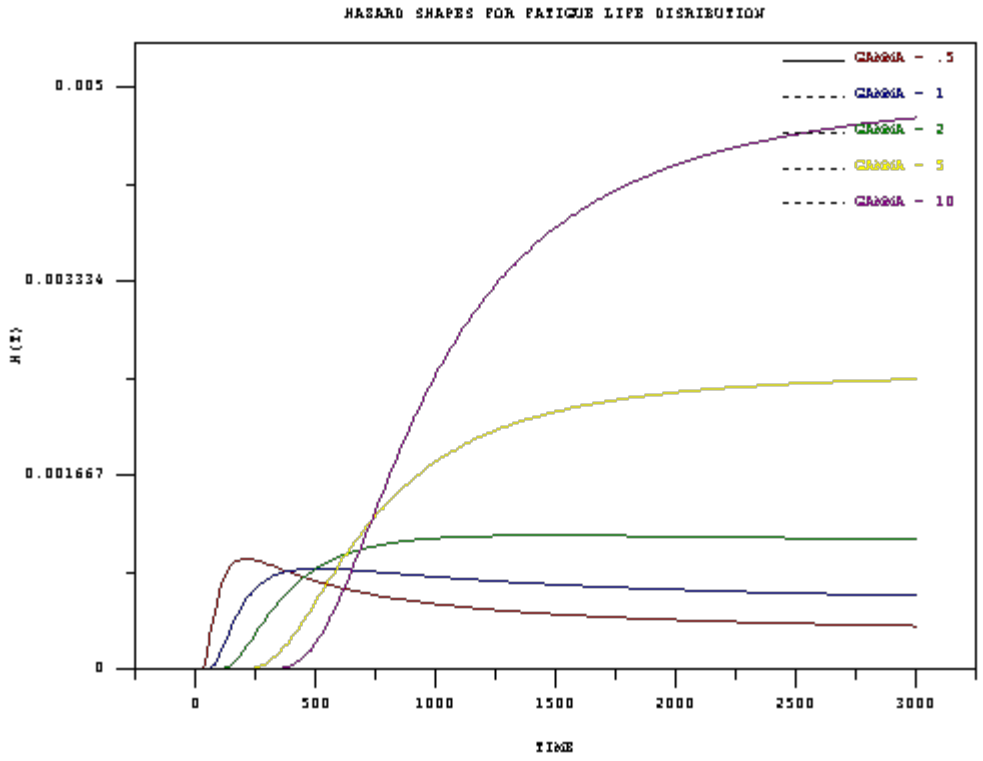
Mean:
$$\mu\left(1+\frac{\gamma^2}{2}\right)$$

Variance:
$$\mu^2\gamma^2\left(1+\frac{5\gamma^2}{4}\right)$$

PDF shapes for the model vary from highly skewed and long tailed (small gamma values) to nearly symmetric and short tailed as gamma increases. This is shown in the figure below.



Corresponding failure rate curves are shown in the next figure.



If crack growth in each stress cycle is a random

Derivation and Use of the Birnbaum-Saunders Model:

Consider a material that continually undergoes cycles of stress loads. During each cycle, a dominant crack grows towards a critical length that will cause failure. Under repeated application of *n* cycles of loads, the total extension of the dominant crack can be written as

amount independent of past cycles of growth, the Fatigue Life mode model may apply.

$$W_n = \sum_{j=1}^{n} Y_j$$

and we assume the Y_j are independent and identically distributed non-negative random variables with mean μ and variance σ^2. Suppose failure occurs at the N-th cycle, when W_n first exceeds a constant critical value w. If n is large, we can use a central limit theorem argument to conclude that

$$\Pr(N \le n) = 1 - \Pr\left(\sum_{j=1}^{n} Y_j \le w\right) = \Phi\left(\frac{\mu\sqrt{n}}{\sigma} - \frac{w}{\sigma\sqrt{n}}\right)$$

Since there are many cycles, each lasting a very short time, we can replace the discrete number of cycles N needed to reach failure by the continuous time t_f needed to reach failure. The CDF $F(t)$ of t_f is given by

$$F(t) = \Phi\left\{\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right\}$$

for appropriate choice of $\alpha = \frac{\sigma}{\sqrt{\mu w}}$ and $\beta = \frac{w}{\mu}$

Here Φ denotes the standard normal CDF. Writing the model with parameters α and β is an alternative way of writing the Birnbaum-Saunders distribution that is often used ($\alpha = \gamma$ and $\beta = \mu$, as compared to the way the formulas were parameterized earlier in this section).

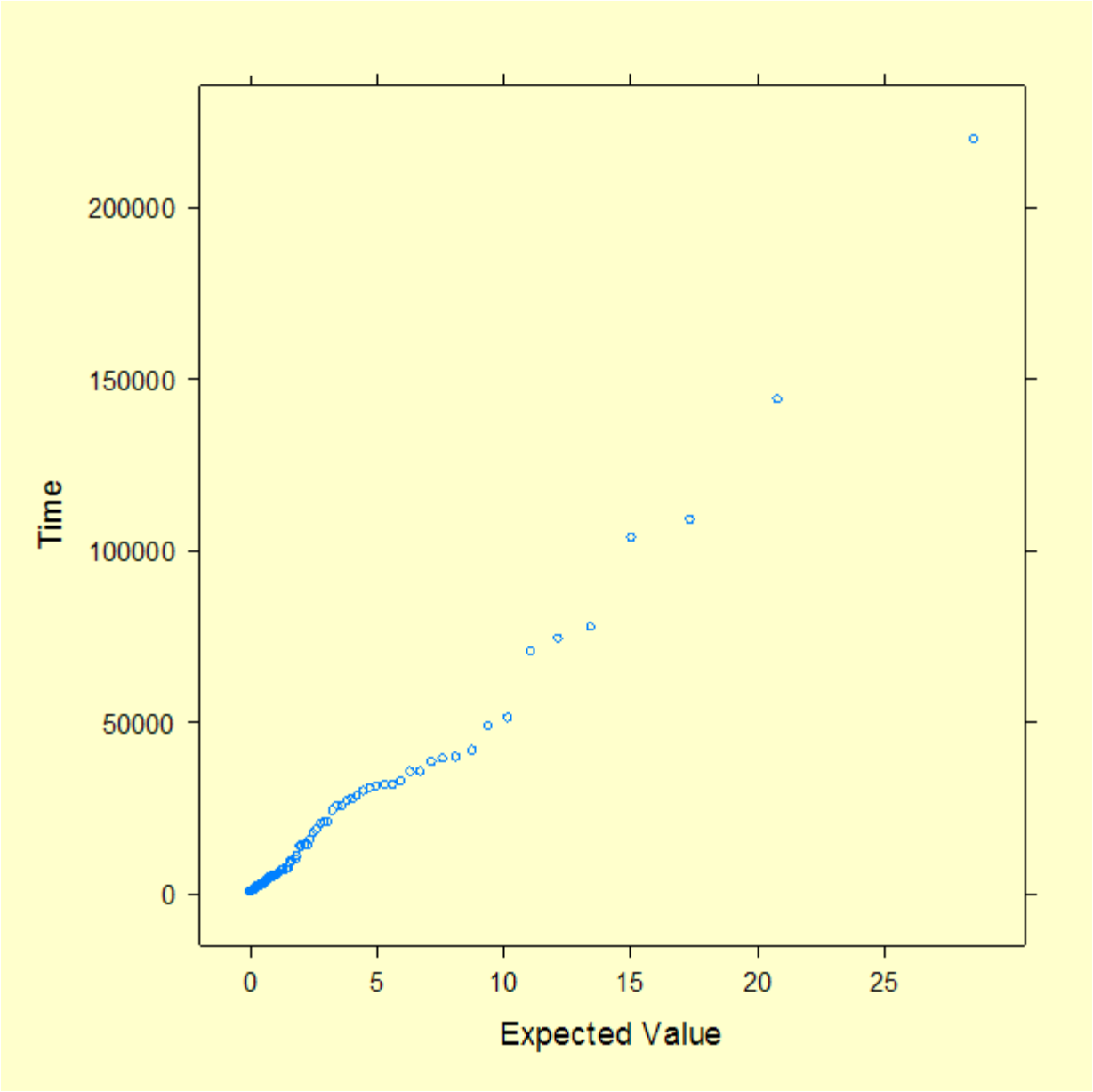
Note:
The critical assumption in the derivation, from a physical point of view, is that the crack growth during any one cycle is independent of the growth during any other cycle. Also, the growth has approximately the same random distribution, from cycle to cycle. This is a very different situation from the proportional degradation argument used to derive a log normal distribution model, with the rate of degradation at any point in time depending on the total amount of degradation that has occurred up to that time.

This kind of physical degradation is consistent with Miner's Rule.

The Birnbaum-Saunders assumption, while physically restrictive, is consistent with a deterministic model from materials physics known as Miner's Rule (Miner's Rule implies that the damage that occurs after n cycles, at a stress level that produces a fatigue life of N cycles, is proportional to n/N). So, when the physics of failure suggests Miner's Rule applies, the Birnbaum-Saunders model is a reasonable choice for a life distribution model.

Birnbaum-Saunders probability plot

We generated 100 random numbers from a Birnbaum-Saunders distribution where $\mu = 5000$ and $\gamma = 2$, and created a fatigue life probability plot of the 100 data points.



If the points on the probability plot line up roughly on a straight line, as expected for data we generated, then the data are correctly modeled by the Birnbaum-Saunders distribution.

The PDF value at time t = 4000 for a Birnbaum-Saunders (fatigue life) distribution with parameters μ = 5000 and γ = 2 is 4.987e-05 and the CDF value is 0.455.

Functions for computing Birnbaum-Saunders distribution PDF values, CDF values, and for producing probability plots, are found in both Dataplot code and R code.

8. Assessing Product Reliability
8.1. Introduction
8.1.6. What are the basic lifetime distribution models used for non-repairable populations?

8.1.6.7. Proportional hazards model

The proportional hazards model is often used in survival analysis (medical testing) studies. It is not used much with engineering data

The proportional hazards model, proposed by Cox (1972), has been used primarily in medical testing analysis, to model the effect of secondary variables on survival. It is more like an acceleration model than a specific life distribution model, and its strength lies in its ability to model and test many inferences about survival without making any specific assumptions about the form of the life distribution model.

This section will give only a brief description of the proportional hazards model, since it has limited engineering applications.

Proportional Hazards Model Assumption

Let $z = \{x, y, ...\}$ be a vector of 1 or more **explanatory** variables believed to affect lifetime. These variables may be continuous (like temperature in engineering studies, or dosage level of a particular drug in medical studies) or they may be indicator variables with the value 1 if a given factor or condition is present, and 0 otherwise.

Let the hazard rate for a nominal (or baseline) set $z_0 = (x_0,y_0, ...)$ of these variables be given by $h_0(t)$, with $h_0(t)$ denoting legitimate hazard function (failure rate) for some unspecified life distribution model.

The proportional hazard model assumes changing a stress variable (or explanatory variable) has the effect of multiplying the hazard rate by a constant.

The proportional hazards model assumes we can write the changed hazard function for a new value of z as

$$h_z(t) = g(z)h_0(t)$$

In other words, changing z, the explanatory variable vector, results in a new hazard function that is proportional to the nominal hazard function, and the proportionality constant is a function of z, $g(z)$, independent of the time variable t.

A common and useful form for $f(z)$ is the **Log Linear Model** which has the equation: $g(x) = e^{ax}$ for one variable, $g(x,y) = e^{ax+by}$ for two variables, etc.

Properties and Applications of the Proportional Hazards

Model

1. The proportional hazards model is equivalent to the acceleration factor concept if and only if the life distribution model is a Weibull (which includes the exponential model, as a special case). For a Weibull with shape parameter γ, and an acceleration factor AF between nominal use fail time t_0 and high stress fail time t_s (with $t_0 = AFt_s$) we have $g(s) = AF^{\gamma}$. In other words, $h_s(t) = AF^{\gamma} h_0(t)$.
2. Under a log-linear model assumption for $g(z)$, without any further assumptions about the life distribution model, it is possible to analyze experimental data and compute maximum likelihood estimates and use likelihood ratio tests to determine which explanatory variables are highly significant. In order to do this kind of analysis, however, special software is needed.

More details on the theory and applications of the proportional hazards model may be found in Cox and Oakes (1984).

8.1.7. What are some basic repair rate models used for repairable systems?

Models for repair rates of repairable systems

N(t), *M(t)* and *m(t)* were defined in the section on Repair Rates. Repair rate models are defined by first picking a functional form for *M(t)*, the expected number of cumulative failures by time *t*. Taking the derivative of this gives the repair rate model *m(t)*.

In the next three sections we will describe three models, of increasing complexity, for *M(t)*. They are: the Homogeneous Poisson Process, the Non-Homogeneous Poisson Process following a Power law, and the Non-Homogeneous Poisson Process following an Exponential law.

8. Assessing Product Reliability
8.1. Introduction
8.1.7. What are some basic repair rate models used for repairable systems?

8.1.7.1. Homogeneous Poisson Process (HPP)

Repair rate (ROCOF) models and formulas

The simplest useful model for $M(t)$ is $M(t) = \lambda t$ and the repair rate (or ROCOF) is the constant $m(t) = \lambda$. This model comes about when the interarrival times between failures are independent and identically distributed according to the exponential distribution, with parameter λ. This basic model is also known as a **Homogeneous Poisson Process (HPP)**. The following formulas apply:

$$F(t) = 1 - e^{-\lambda t} = \text{ CDF of the waiting time to the next failure (or "interarrival" time between failures)}$$

$$N(T) = \text{ cumulative number of failures from time 0 to time } T$$

$$P\{N(t) = k\} = \frac{(\lambda T)^k e^{-\lambda T}}{k!}$$

$$M(t) = \lambda T = \text{expected number of failures by time } T$$

$$M'(t) = m(t) = \lambda = \text{repair rate or ROCOF}$$

$$\frac{1}{\lambda} = \text{Mean Time Between Failures (MTBF)}$$

HPP model fits flat portion of "bathtub" curve

Despite the simplicity of this model, it is widely used for repairable equipment and systems throughout industry. Justification for this comes, in part, from the shape of the empirical Bathtub Curve. Most systems (or complex tools or equipment) spend most of their "lifetimes" operating in the long flat constant repair rate portion of the Bathtub Curve. The HPP is the only model that applies to that portion of the curve, so it is the most popular model for system reliability evaluation and reliability test planning.

Planning reliability assessment tests (under the HPP assumption) is covered in a later section, as is estimating the MTBF from system failure data and calculating upper and lower confidence limits.

Poisson relationship

In the HPP model, the probability of having exactly k failures by time T is given by the Poisson distribution with mean λT (see formula for $P\{N(T) = k\}$ above).

8. Assessing Product Reliability
8.1. Introduction
8.1.7. What are some basic repair rate models used for repairable systems?

8.1.7.2. Non-Homogeneous Poisson Process (NHPP) - power law

The repair rate for a NHPP following the Power law

A flexible model (that has been very successful in many applications) for the expected number of failures in the first t hours, *M(t)*, is given by the polynomial

$$M(t) = at^b, \text{ for } a, b > 0$$

The repair rate (or ROCOF) for this model is

$$m(t) = abt^{b-1} = \alpha t^{-\beta}, \text{ for } \alpha > 0, \beta < 1$$

The Power law model is very flexible and contains the HPP (exponential) model as a special case

The HPP model has a the constant repair rate *m(t)* = λ. If we substitute an arbitrary function λ(t) for λ, we have a **Non Homogeneous Poisson Process (NHPP)** with Intensity Function λ. If

$$\lambda(t) = m(t) = \alpha t^{-\beta}$$

then we have an NHPP with a **Power Law intensity function** (the "intensity function" is another name for the repair rate m(t)).

Because of the polynomial nature of the ROCOF, this model is very flexible and can model both increasing (*b*>1 or β < 0) and decreasing (0 < *b* < 1 or 0 < β < 1)) failure rates. When *b* = 1 or β = 0, the model reduces to the HPP constant repair rate model.

Probabilities of failure for all NHPP processes can easily be calculated based on the Poisson formula

Probabilities of a given number of failures for the NHPP model are calculated by a straightforward generalization of the formulas for the HPP. Thus, for any NHPP

$$P(N(T) = k) = \frac{M(T)^k}{k!} e^{-M(T)}$$

and for the Power Law model:

$$P(N(T) = k) = \frac{\left[aT^b\right]^k e^{-aT^b}}{k!} = \frac{a^k T^{bk} e^{-aT^b}}{k!}$$

The Power Law model is also called the Duane Model and the AMSAA model

Other names for the Power Law model are: the **Duane Model** and the **AMSAA model**. AMSAA stands for the United States **Army Materials System Analysis Activity**, where much theoretical work describing the Power Law model was performed in the 1970's.

It is also called a Weibull Process - but this name is misleading and should be avoided

The time to the first fail for a Power Law process has a Weibull distribution with shape parameter *b* and characteristic life *a*. For this reason, the Power Law model is sometimes called a ***Weibull Process***. *This name is confusing and should be avoided, however, since it mixes a life distribution model applicable to the lifetimes of a non-repairable population with a model for the inter-arrival times of failures of a repairable population.*

For any NHPP process with intensity function *m(t)*, the distribution function (CDF) for the inter-arrival time τ to the next failure, given a failure just occurred at time T, is given by

$$F_T(t) = 1 - e^{-\int_0^t m(T+\tau)d\tau}$$

Once a failure occurs, the waiting time to the next failure for an NHPP has a simple CDF formula

In particular, for the Power Law the waiting time to the next failure, given a failure at time T, has distribution function

$$F_T(t) = 1 - e^{-a[(T+t)^b - T^b]}$$

This inter arrival time CDF can be used to derive a simple algorithm for simulating NHPP Power Law Data.

8.1.7.3. Exponential law

The Exponential Law is another flexible NHPP model

An NHPP with ROCOF or intensity function given by

$$m(t) = e^{\alpha + \beta t}$$

is said to follow an **Exponential Law**. This is also called the **log-linear model** or the **Cox-Lewis model**.

A system whose repair rate follows this flexible model is improving if β < 0 and deteriorating if β >0. When β = 0, the Exponential Law reduces to the HPP constant repair rate model

8.1.8. How can you evaluate reliability from the "bottom-up" (component failure mode to system failure rate)?

Several simple models can be used to calculate system failure rates, starting with failure rates for failure modes within individual system components

This section deals with models and methods that apply to non-repairable components and systems. Models for failure rates (and not repair rates) are described. The next section covers models for (repairable) system reliability growth.

We use the Competing Risk Model to go from component failure modes to component failure rates. Next we use the Series Model to go from components to assemblies and systems. These models assume independence and "first failure mode to reach failure causes both the component and the system to fail".

If some components are "in parallel", so that the system can survive one (or possibly more) component failures, we have the parallel or redundant model. If an assembly has *n* identical components, at least *r* of which must be working for the system to work, we have what is known as the *r* out of *n* model.

The standby model uses redundancy like the parallel model, except that the redundant unit is in an off-state (not exercised) until called upon to replace a failed unit.

This section describes these various models. The last subsection shows how complex systems can be evaluated using the various models as building blocks.

8. Assessing Product Reliability
8.1. Introduction
8.1.8. How can you evaluate reliability from the "bottom-up" (component failure mode to system failure rate)?

8.1.8.1. Competing risk model

Use the competing risk model when the failure mechanisms are independent and the first mechanism failure causes the component to fail

Assume a (replaceable) component or unit has k different ways it can fail. These are called **failure modes** and underlying each failure mode is a **failure mechanism**.

The **Competing Risk Model** evaluates component reliability by "building up" from the reliability models for each failure mode.

The following 3 assumptions are needed:

1. Each failure mechanism leading to a particular type of failure (i.e., failure mode) proceeds independently of every other one, at least until a failure occurs.
2. The component fails when the **first** of all the competing failure mechanisms reaches a failure state.
3. Each of the k failure modes has a known life distribution model $F_i(t)$.

The competing risk model can be used when all three assumptions hold. If $R_c(t)$, $F_c(t)$, and $h_c(t)$ denote the reliability, CDF and failure rate for the component, respectively, and $R_i(t)$, $F_i(t)$ and $h_i(t)$ are the reliability, CDF and failure rate for the i-th failure mode, respectively, then the competing risk model formulas are:

Multiply reliabilities and add failure rates

$$R_c(t) = \prod_{i=1}^{k} R_i(t)$$

$$F_c(t) = 1 - \prod_{i=1}^{k} \{1 - F_i(t)\}$$

$$h_c(t) = \sum_{i=1}^{k} h_i(t)$$

Think of the competing risk model in the following way:

> All the failure mechanisms are having a race to see which can reach failure first. They are not allowed to "look over their shoulder or sideways" at the progress the other ones are

making. They just go their own way as fast as they can and the first to reach "failure" causes the component to fail.

Under these conditions the component reliability is the product of the failure mode reliabilities and the component failure rate is just the sum of the failure mode failure rates.

Note that the above holds for any arbitrary life distribution model, as long as "independence" and "first mechanism failure causes the component to fail" holds.

When we learn how to plot and analyze reliability data in later sections, the methods will be applied separately to each failure mode within the data set (considering failures due to all other modes as "censored run times"). With this approach, the competing risk model provides the glue to put the pieces back together again.

8.1.8.2. Series model

The series model is used to go from individual components to the entire system, assuming the system fails when the first component fails and all components fail or survive independently of one another

The **Series Model** is used to build up from components to sub-assemblies and systems. It only applies to non replaceable populations (or first failures of populations of systems). The assumptions and formulas for the Series Model are identical to those for the Competing Risk Model, with the *k* failure modes within a component replaced by the *n* components within a system.

The following 3 assumptions are needed:

1. Each component operates or fails independently of every other one, at least until the first component failure occurs.
2. The system fails when the first component failure occurs.
3. Each of the *n* (possibly different) components in the system has a known life distribution model $F_i(t)$.

Add failure rates and multiply reliabilities in the Series Model

When the Series Model assumptions hold we have:

$$R_S(t) = \prod_{i=1}^{n} R_i(t)$$

$$F_S(t) = 1 - \prod_{i=1}^{n} \{1 - F_i(t)\}$$

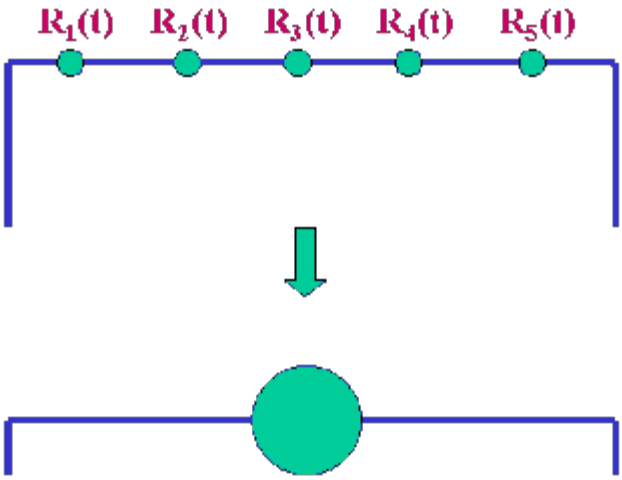
$$h_S(t) = \sum_{i=1}^{n} h_i(t)$$

with the subscript *S* referring to the entire system and the subscript *i* referring to the *i*-th component.

Note that the above holds for any arbitrary component life distribution models, as long as "independence" and "first component failure causes the system to fail" both hold.

The analogy to a series circuit is useful. The entire system has *n* components in series. The system fails when current no longer flows and each component operates or fails independently of all the others. The schematic below shows a system with 5 components in series "replaced" by an "equivalent" (as far as reliability is concerned) system with only one component.

Series System Reduced to Equivalent One Component System



$$R_s(t) = R_1(t) \times R_2(t) \times R_3(t) \times R_4(t) \times R_5(t)$$

8. Assessing Product Reliability
8.1. Introduction
8.1.8. How can you evaluate reliability from the "bottom-up" (component failure mode to system failure rate)?

8.1.8.3. Parallel or redundant model

The parallel model assumes all n components that make up a system operate independently and the system works as long as at least one component still works

The opposite of a series model, for which the first component failure causes the system to fail, is a parallel model for which all the components have to fail before the system fails. If there are *n* components, any (*n-1*) of them may be considered redundant to the remaining one (even if the components are all different). When the system is turned on, all the components operate until they fail. The system reaches failure at the time of the last component failure.

The assumptions for a parallel model are:

1. All components operate independently of one another, as far as reliability is concerned.
2. The system operates as long as at least one component is still operating. System failure occurs at the time of the last component failure.
3. The CDF for each component is known.

Multiply component CDF's to get the system CDF for a parallel model

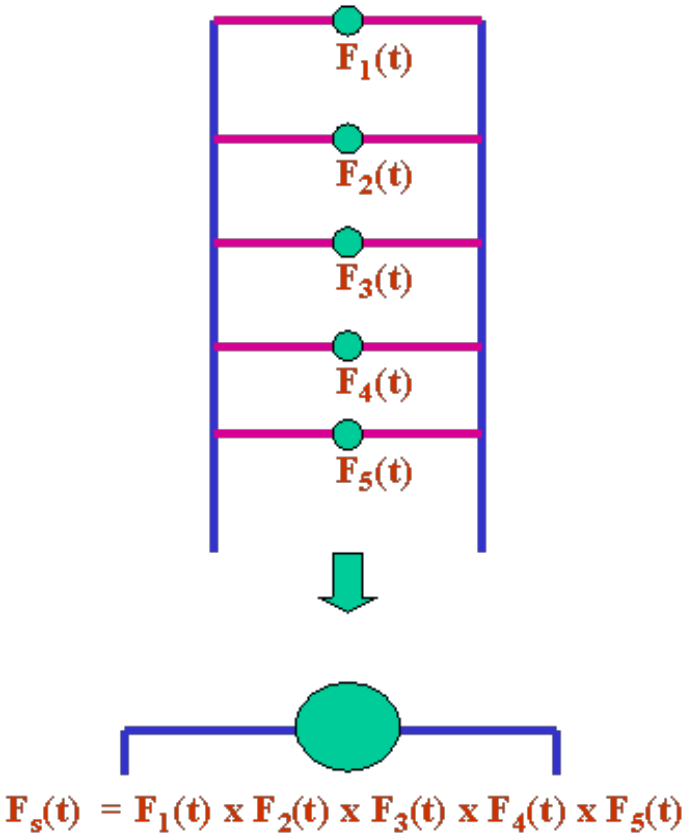
For a parallel model, the CDF $F_s(t)$ for the system is just the product of the CDF's $F_i(t)$ for the components or

$$F_s(t) = \prod_{i=1}^{n} F_i(t)$$

$R_s(t)$ and $h_s(t)$ can be evaluated using basic definitions, once we have $F_s(t)$.

The schematic below represents a parallel system with 5 components and the (reliability) equivalent 1 component system with a CDF F_s equal to the product of the 5 component CDF's.

Parallel System and Equivalent Single Component



$$F_s(t) = F_1(t) \times F_2(t) \times F_3(t) \times F_4(t) \times F_5(t)$$

8.1.8.4. R out of N model

An r out of n model is a system that survives when at least r of its components are working (any r)

An "*r* out of *n*" system contains both the series system model and the parallel system model as special cases. The system has *n* components that operate or fail independently of one another and as long as at least *r* of these components (any *r*) survive, the system survives. System failure occurs when the (*n-r+1*)th component failure occurs.

When *r* = *n*, the *r* out of *n* model reduces to the series model. When *r* = 1, the *r* out of *n* model becomes the parallel model.

We treat here the simple case where all the components are identical.

Formulas and assumptions for *r* out of *n* model (identical components):

1. All components have the identical reliability function *R(t)*.
2. All components operate independently of one another (as far as failure is concerned).
3. The system can survive any (*n-r*) of the components failing. The system fails at the instant of the (*n-r+1*)th component failure.

Formula for an r out of n system where the components are identical

System reliability is given by adding the probability of exactly *r* components surviving to time *t* to the probability of exactly (*r+1*) components surviving, and so on up to the probability of all components surviving to time *t*. These are binomial probabilities (with $p = R(t)$), so the system reliability is given by:

$$R_s(t) = \sum_{i=r}^{n} \binom{n}{i} [R(t)]^i [1 - R(t)]^{n-i}$$

Note: If we relax the assumption that all the components are identical, then $R_S(t)$ would be the sum of probabilities evaluated for all possible terms that could be formed by picking at least *r* survivors and the corresponding failures. The probability for each term is evaluated as a product of *R(t)*'s and *F(t)*'s. For example, for *n* = 4 and *r* = 2, the system

reliability would be (abbreviating the notation for *R(t)* and *F(t)* by using only *R* and *F)*

$$R_S = R_1R_2F_3F_4 + R_1R_3F_2F_4 + R_1R_4F_2F_3 + R_2R_3F_1F_4 + R_2R_4F_1F_3 + R_3R_4F_1F_2 + R_1R_2R_3F_4 + R_1R_3R_4F_2 + R_1R_2R_4F_3 + R_2R_3R_4F_1 + R_1R_2R_3R_4$$

8.1.8.5. Standby model

The Standby Model evaluates improved reliability when backup replacements are switched on when failures occur.

A **Standby Model** refers to the case in which a key component (or assembly) has an identical backup component in an "off" state until needed. When the original component fails, a switch turns on the "standby" backup component and the system continues to operate.

In the simple case, assume the non-standby part of the system has CDF *F(t)* and there are (*n-1*) identical backup units that will operate in sequence until the last one fails. At that point, the system finally fails.

The total system lifetime is the sum of *n* identically distributed random lifetimes, each having CDF *F(t)*.

Identical backup Standby model leads to convolution formulas

In other words, $T_n = t_1 + t_2 + ... + t_n$, where each t_i has CDF *F(t)* and T_n has a CDF we denote by $F_n(t)$. This can be evaluated using **convolution** formulas:

$$F_2(t) = \int_0^t F(u) f(t-u) du$$

$$F_n(t) = \int_0^t F_{n-1}(u) f(t-u) du$$

$$\text{where } f(t) \text{ is the PDF } F'(t)$$

In general, convolutions are solved numerically. However, for the special case when *F(t)* is the exponential model, the above integrations can be solved in closed form.

Exponential standby systems lead to a gamma lifetime model

Special Case: The Exponential (or Gamma) Standby Model

If *F(t)* has the exponential CDF (i.e., $F(t) = 1 - e^{-lt}$), then

$$F_2(t) = 1 - \lambda t e^{-\lambda t} - e^{-\lambda t}$$
$$f_2(t) = \lambda^2 t e^{-\lambda t}, \quad and$$
$$f_n(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}$$

and the PDF $f_n(t)$ is the well-known gamma distribution.

Example: An unmanned space probe sent out to explore the solar system has an onboard computer with reliability characterized by the exponential distribution with a **Mean Time To Failure** (MTTF) of $1/\lambda$ = 30 months (a constant failure rate of 1/30 = .033 fails per month). The probability of surviving a two year mission is only $e^{-24/30}$ = .45. If, however, a second computer is included in the probe in a standby mode, the reliability at 24 months (using the above formula for F_2) becomes .8 × .449 + .449 = .81. The failure rate at 24 months ($f_2/[1-F_2]$) reduces to [(24/900) ×.449]/.81 = .015 fails per month. At 12 months the failure rate is only .0095 fails per month, which is less than 1/3 of the failure rate calculated for the non-standby case.

Standby units (as the example shows) are an effective way of increasing reliability and reducing failure rates, especially during the early stages of product life. Their improvement effect is similar to, but greater than, that of parallel redundancy . The drawback, from a practical standpoint, is the expense of extra components that are not needed for functionality.

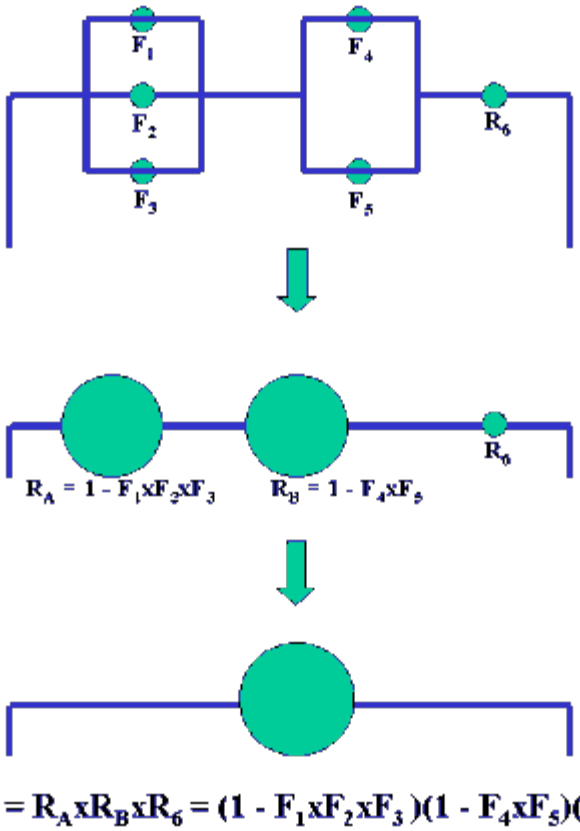
8.1.8.6. Complex systems

Often the reliability of complex systems can be evaluated by successive applications of Series and/or Parallel model formulas

Many complex systems can be diagrammed as combinations of Series components, Parallel components, *R* out of *N* components and Standby components. By using the formulas for these models, subsystems or sections of the original system can be replaced by an "equivalent" single component with a known CDF or Reliability function. Proceeding like this, it may be possible to eventually reduce the entire system to one component with a known CDF.

Below is an example of a complex system composed of both components in parallel and components in series is reduced first to a series system and finally to a one-component system.

Complex System Reduced to Equivalent One Component System



$$R_s = R_A x R_B x R_6 = (1 - F_1 x F_2 x F_3)(1 - F_4 x F_5)(R_6)$$

Note: The reduction methods described above will work for many, but not

all, systems. Some systems with a complicated operational logic structure will need a more formal structural analysis methodology. This methodology deals with subjects such as event trees, Boolean representations, coherent structures, cut sets and decompositions, and is beyond the present scope of this Handbook.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

8.1.9. How can you model reliability growth?

A reliability improvement test is a formal procedure aimed at discovering and fixing system reliability flaws

During the early stages of developing and prototyping complex systems, reliability often does not meet customer requirements. A formal test procedure aimed at discovering and fixing causes of unreliability is known as a **Reliability Improvement Test**. This test focuses on system design, system assembly and component selection weaknesses that cause failures.

A typical reliability improvement test procedure would be to run a prototype system, as the customer might for a period of several weeks, while a multidisciplined team of engineers and technicians (design, quality, reliability, manufacturing, etc.) analyze every failure that occurs. This team comes up with root causes for the failures and develops design and/or assembly improvements to hopefully eliminate or reduce the future occurrence of that type of failure. As the testing continues, the improvements the team comes up with are incorporated into the prototype, so it is expected that reliability will improve during the course of the test.

Repair rates should show an improvement trend during the course of a reliability improvement test and this can be modeled using a NHPP model

Another name for reliability improvement testing is **TAAF** testing, standing for **Test, Analyze And Fix.**

While only one model applies when a repairable system has no improvement or degradation trends (the constant repair rate HPP model), there are infinitely many models that could be used to describe a system with a decreasing repair rate (reliability growth models).

Fortunately, one or two relatively simple models have been very successful in a wide range of industrial applications. Two models that have previously been described will be used in this section. These models are the NHPP Power Law Model and the NHPP Exponential Law Model. The Power Law Model underlies the frequently used graphical technique known as Duane Plotting.

8.1.9.1. NHPP power law

If the Power Law applies, Repair Rates improve over time according to the formula $\alpha t^{-\beta}$. *The exponent* β *lies between 0 and 1 and is called the reliability growth slope*

This repairable system model was described in Section 8.1.7.2. The expected number of failures by time *t* has the form $M(t) = at^b$ and the repair rate has the form $m(t) = abt^{b-1}$. This will model improvement when $0 < b < 1$, with larger improvements coming when *b* is smaller. As we will see in the next section on Duane Plotting, it is convenient to define $\beta = 1 - b$ and $\alpha = ab$, and write the repair rate as

$$m(t) = \alpha t^{-\beta}$$

Again we have improvement when $0 < \beta < 1$, with larger improvement coming from larger values of β. β is known as the Duane Plot slope or the reliability improvement **Growth Slope**.

In terms of the original parameters for $M(t)$, we have

$$a = \frac{\alpha}{1-\beta} \text{ and } b = 1-\beta$$

Use of the Power Law model for reliability growth test data generally assumes the following:

1. While the test is ongoing, system improvements are introduced that produce continual improvements in the rate of system repair.

2. Over a long enough period of time the effect of these improvements can be modeled adequately by the continuous polynomial repair rate improvement model $\alpha t^{-\beta}$.

When an improvement test ends, the MTBF stays constant at its last achieved value

3. When the improvement test ends at test time T and no further improvement actions are ongoing, the repair rate has been reduced to $\alpha T^{-\beta}$. The repair rate remains constant from then on at this new (improved) level.

Assumption 3 means that when the test ends, the HPP constant repair rate model takes over and the MTBF for the system from then on is the reciprocal of the final repair rate or $(T^{\beta})/\alpha$. If we estimate the expected number of failures up to time T by the

actual number observed, the estimated MTBF at the end of a reliability test (following the Power Law) is:

$$\text{ESTIMATED MTBF AT END OF TEST} = \frac{T}{r(1-\beta)}$$

with T denoting the test time, r is the total number of test failures and β is the reliability growth slope. A formula for estimating β from system failure times is given in the Analysis Section for the Power Law model.

Simulated Data Example

Simulating NHPP Power Law Data

Step 1: User inputs the positive constants a and b.

Step 2: Simulate a vector of n uniform (0,1) random numbers. Call these $U_1, U_2, U_3, \ldots U_n$.

Step 3: Calculate $Y_1 = \{-1/a * \ln U_1\} ** 1/b$

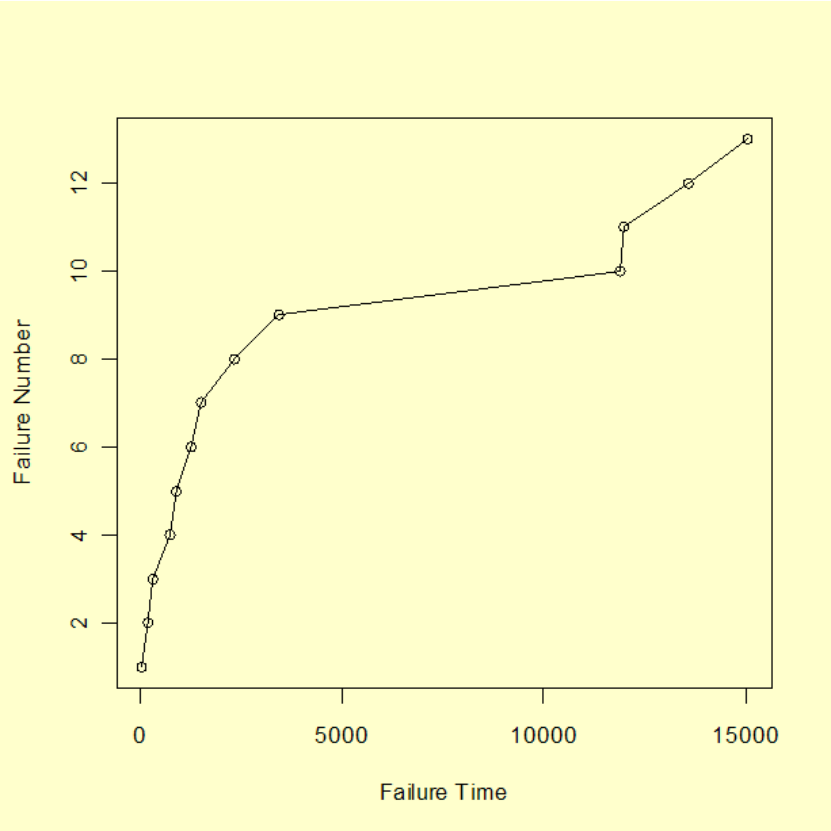
Step i: Calculate $Y_i = \{(Y_{i-1} ** b) - 1/a * \ln U_i\} ** 1/b$ for $i = 2, \ldots, n$

The n numbers $Y_1, Y_2, \ldots, Y_n$ are the desired repair times simulated from an NHPP Power Law process with parameters a, b (or $\beta = 1 - b$ and $\alpha = ab$).

Example

We generated $n = 13$ random repair times using the NHPP Power Law process with $a = 0.2$ and $b = 0.4$. The resulting data and a plot of failure number versus repair times are shown below.

Failure Number	Failure Time
1	26
2	182
3	321
4	728
5	896
6	1268
7	1507
8	2325
9	3427
10	11871
11	11978
12	13562
13	15053



The NHPP power law process can be implemented using both Dataplot code and R code.

8.1.9.2. Duane plots

A plot on log-log paper of successive MTBF estimates versus system time of fail for reliability improvement test data is called a Duane Plot

The standard estimate of the MTBF for a system with a constant repair rate (an HPP system) is T/r, with T denoting the total time the system was observed and r is the total number of failures that occurred.

If we calculate successive MTBF estimates (called Cum MTBF Estimates), every time a failure occurs for a system undergoing reliability improvement testing, we typically see a sequence of mostly increasing numbers.

In 1964, J. T. Duane observed that when he plotted these cum MTBF estimates versus the times of failure on log-log paper, the points tended to line up following a straight line. This was true for many different sets of reliability improvement data and many other engineers have seen similar results over the last three decades. This type of plot is called a **Duane Plot** and the slope β of the best line through the points is called the **reliability growth slope** or **Duane plot slope**.

Points on a Duane plot line up approximately on a straight line if the Power Law model applies

Plotting a Duane Plot is simple. If the ith failure occurs at time t_i, then plot t_i divided by i (the "y"- axis value) versus the time t_i (the "x"-axis value) on log-log graph paper. Do this for all the test failures and draw the best straight line you can following all these points.

Why does this "work"? Following the notation for repairable system models, we are plotting estimates of $\{t/M(t)\}$ versus the time of failure t. If $M(t)$ follows the Power Law (also described in the last section), then we are plotting estimates of t/at^b versus the time of fail t. This is the same as plotting $(1/a)t^{\beta}$ versus t, with $\beta = 1-b$. For a log-log scale plot, this will be a straight line with slope β and intercept (when $t = 1$) of $-\log_{10}a$.

In other words, a straight line on a Duane plot is equivalent to the NHPP Power Law Model with a reliability growth slope of $\beta = 1 - b$ and an "a" parameter equal to $10^{-\text{intercept}}$.

Note: A useful empirical rule of thumb based on Duane plots made from many reliability improvement tests across many industries is the following:

Duane plot reliability growth slopes should lie between 0.3

The reliability improvement slope for virtually all reliability improvement tests will be between 0.3 and 0.6. The lower end (0.3) describes a minimally effective test - perhaps the cross-functional team is inexperienced or the system has many failure mechanisms that are not well understood. The higher end (0.6)

and 0.6 approaches the empirical state of the art for reliability improvement activities.

Examples of Duane Plots

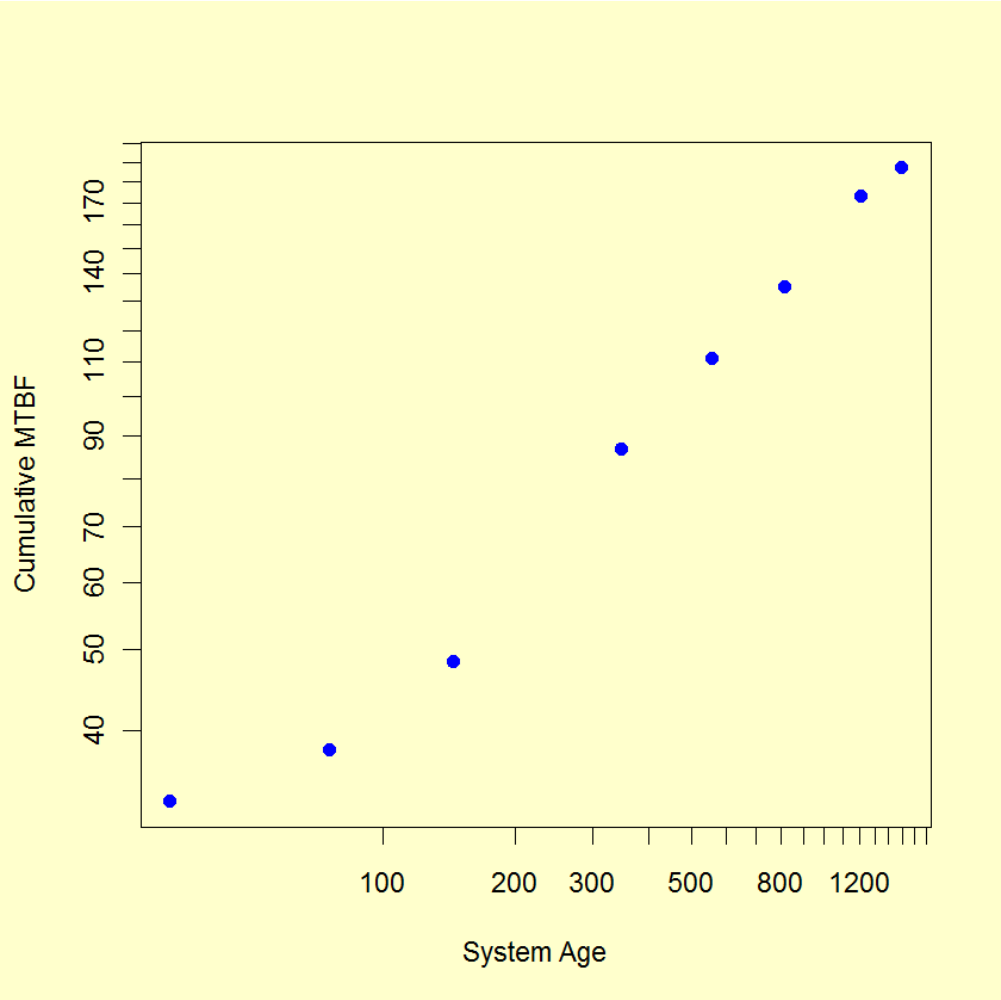
Duane Plot Example 1:

A reliability growth test lasted 1500 hours (approximately 10 weeks) and recorded 8 failures at the following system hours: 33, 76, 145, 347, 555, 811, 1212, 1499. After calculating successive cum MTBF estimates, a Duane plot shows these estimates versus system age on log vs log paper. The "best" straight line through the data points corresponds to a NHPP Power Law model with reliability growth slope β equal to the slope of the line. This line is an estimate of the theoretical model line (assuming the Power Law holds during the course of the test) and the achieved MTBF at the end of the test is given by

$$T / [r\ (1-\beta)]$$

with T denoting the total test time and *r* the number of failures. Results for this particular reliability growth test follow.

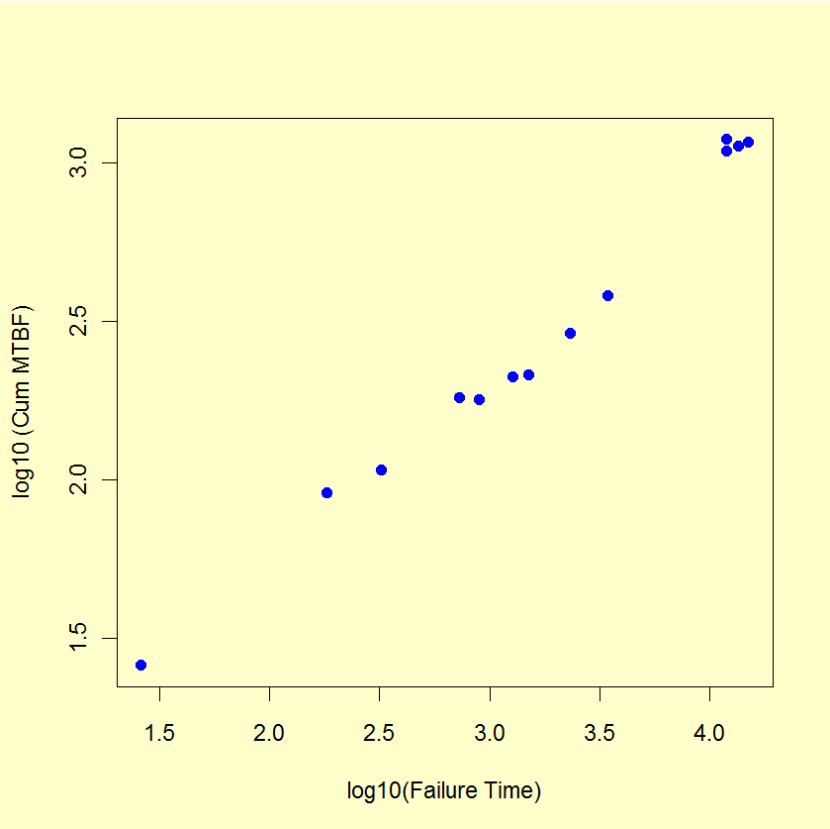
Failure #	System Age of Failure	Cum MTBF
1	33	33
2	76	38
3	145	48.3
4	347	86.8
5	555	111.0
6	811	135.2
7	1212	173.1
8	1499	187.3



The Duane plot indicates a reasonable fit to a Power Law NHPP model. The reliability improvement slope (slope of line on Duane plot) is β = 0.437 (using the formula given in the section on reliability data analysis for the Power Law model) and the estimated MTBF achieved by the end of the 1500 hour test is 1500/(8 × [1-0.437]) or 333 hours.

Duane Plot Example 2:

A Duane plot for the simulated Power Law data used in the Example in the preceding section is shown below.



Duane plots can be produced using both Dataplot code and R code.

8.1.9.3. NHPP exponential law

The Exponential Law is another useful reliability growth model to try when the Power law is not fitting well

When the data points in a Duane plot show obvious curvature, a model that might fit better is the NHPP Exponential Law.

For this model, if β < 0, the repair rate improves over time according to

$$m(t) = e^{\alpha+\beta t}$$

The corresponding cumulative expected failures model is

$$M(t) = A(1-e^{\beta t})$$

This approaches the maximum value of *A* expected failures as *t* goes to infinity, so the cumulative failures plot should clearly be bending over and asymptotically approaching a value $A = -e^{\alpha} / \beta$.

Rule of thumb: First try a Duane plot and the Power law model. If that shows obvious lack of fit, try the Exponential Law model, estimating parameters using the formulas in the Analysis Section for the Exponential law. A plot of cum fails versus time, along with the estimated *M(t)* curve, can be used to judge goodness of fit.

8.1.10. How can Bayesian methodology be used for reliability evaluation?

Several Bayesian Methods overview topics are covered in this section

This section gives an overview of the application of Bayesian techniques in reliability investigations. The following topics are covered:

- What is Bayesian Methodology ?
- Bayes Formula, Prior and Posterior Distribution Models, and Conjugate Priors
- How Bayesian Methodology is used in System Reliability Evaluation
- Advantages and Disadvantages of using Bayes Methodology

What is Bayesian Methodology?

Bayesian analysis considers population parameters to be random, not fixed

It makes a great deal of practical sense to use all the information available, old and/or new, objective or subjective, when making decisions under uncertainty. This is especially true when the consequences of the decisions can have a significant impact, financial or otherwise. Most of us make everyday personal decisions this way, using an intuitive process based on our experience and subjective judgments.

Mainstream statistical analysis, however, seeks objectivity by generally restricting the information used in an analysis to that obtained from a current set of clearly relevant data. Prior knowledge is not used except to suggest the choice of a particular population model to "fit" to the data, and this choice is later checked against the data for reasonableness.

Old information, or subjective judgment, is used to determine a prior distribution for these population parameters

Lifetime or repair models, as we saw earlier when we looked at repairable and non repairable reliability population models, have one or more unknown parameters. The **classical** statistical approach considers these parameters as fixed but unknown constants to be estimated (i.e., "guessed at") using sample data taken randomly from the population of interest. A confidence interval for an unknown parameter is really a frequency statement about the likelihood that numbers calculated from a sample capture the true parameter. Strictly speaking, one cannot make probability statements about the true parameter since it is fixed, not random.

The **Bayesian** approach, on the other hand, treats these population model parameters as random, not fixed, quantities. Before looking at the current data, we use old information, or even subjective judgments, to construct a **prior distribution model** for these parameters. This model expresses our starting assessment about how likely various values of the unknown parameters are. We then make use of the current data (via **Baye's formula**) to revise this starting assessment, deriving what is called the **posterior distribution model** for the

population model parameters. Parameter estimates, along with confidence intervals (known as **credibility intervals**), are calculated directly from the posterior distribution. **Credibility** intervals are legitimate probability statements about the unknown parameters, since these parameters now are considered random, not fixed.

It is unlikely in most applications that data will ever exist to validate a chosen prior distribution model. Parametric Bayesian prior models are chosen because of their flexibility and mathematical convenience. In particular, **conjugate priors** (defined below) are a natural and popular choice of Bayesian prior distribution models.

Bayes Formula, Prior and Posterior Distribution Models, and Conjugate Priors

Bayes formula provides the mathematical tool that combines prior knowledge with current data to produce a posterior distribution

Bayes formula is a useful equation from probability theory that expresses the conditional probability of an event A occurring, given that the event *B* has occurred (written P(A|B)), in terms of unconditional probabilities and the probability the event *B* has occurred, given that *A* has occurred. In other words, Bayes formula inverts which of the events is the conditioning event. The formula is

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{P(A) \bullet P(B \mid A)}{P(B)}$$

and *P(B)* in the denominator is further expanded by using the so-called "Law of Total Probability" to write

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)$$

with the events A_i being mutually exclusive and exhausting all possibilities and including the event *A* as one of the A_i.

The same formula, written in terms of probability density function models, takes the form:

$$g(\lambda \mid x) = \frac{f(x \mid \lambda) g(\lambda)}{\int_0^\infty f(x \mid \lambda) g(\lambda) d\lambda}$$

where $f(x|\lambda)$ is the probability model, or likelihood function, for the observed data *x* given the unknown parameter (or parameters) λ, $g(\lambda)$ is the **prior distribution** model for λ and $g(\lambda|x)$ is the **posterior distribution** model for λ given that the data *x* have been observed.

When $g(\lambda|x)$ and $g(\lambda)$ both belong to the same distribution family, $g(\lambda)$ and $f(x|\lambda)$ are called **conjugate distributions** and $g(\lambda)$ is the **conjugate prior** for $f(x|\lambda)$. For example, the Beta distribution model is a conjugate prior for the proportion of successes *p* when samples have a binomial distribution. And the Gamma model is a conjugate prior for the failure rate λ when sampling failure times or repair times from an exponentially distributed population. This latter conjugate pair (gamma, exponential) is used extensively in Bayesian system reliability applications.

How Bayes Methodology is used in System Reliability Evaluation

Bayesian system reliability evaluation assumes the system MTBF is a random quantity "chosen" according to a prior distribution model

Models and assumptions for using Bayes methodology will be described in a later section. Here we compare the classical paradigm versus the Bayesian paradigm when system reliability follows the HPP or exponential model (i.e., the flat portion of the Bathtub Curve).

Classical Paradigm For System Reliability Evaluation:

- The MTBF is one fixed unknown value - there is no “probability” associated with it
- Failure data from a test or observation period allows you to make inferences about the value of the true unknown MTBF
- No other data are used and no “judgment” - the procedure is objective and based solely on the test data and the assumed HPP model

Bayesian Paradigm For System Reliability Evaluation:

- The MTBF is a random quantity with a probability distribution
- The particular piece of equipment or system you are testing “chooses” an MTBF from this distribution and you observe failure data that follow an HPP model with that MTBF
- Prior to running the test, you already have some idea of what the MTBF probability distribution looks like based on prior test data or an consensus engineering judgment

Advantages and Disadvantages of using Bayes Methodology

Pro's and con's for using Bayesian methods

While the primary motivation to use Bayesian reliability methods is typically a desire to save on test time and materials cost, there are other factors that should also be taken into account. The table below summarizes some of these "good news" and "bad news" considerations.

Bayesian Paradigm: Advantages and Disadvantages

Pro's	Con's
• Uses prior information - this "makes sense" • If the prior information is encouraging, less new testing may be needed to confirm a desired MTBF at a given confidence • Confidence intervals are really intervals for the (random) MTBF - sometimes called "credibility intervals"	• Prior information may not be accurate - generating misleading conclusions • Way of inputting prior information (choice of prior) may not be correct • Customers may not accept validity of prior data or engineering judgements • There is no one "correct way" of inputting prior information and different approaches can give different results • Results aren't objective and

don't stand by themselves

8.2. Assumptions/Prerequisites

This section describes how life distribution models and acceleration models are typically chosen. Several graphical and analytical methods for evaluating model fit are also discussed.

Detailed contents of Section 2

2. Assumptions/Prerequisites

1. How do you choose an appropriate life distribution model?
    1. Based on failure mode
    2. Extreme value argument
    3. Multiplicative degradation argument
    4. Fatigue life (Birnbaum-Saunders) argument
    5. Empirical model fitting - distribution free (Kaplan-Meier) approach
2. How do you plot reliability data?
    1. Probability plotting
    2. Hazard and cum hazard plotting
    3. Trend and growth plotting (Duane plots)
3. How can you test reliability model assumptions?
    1. Visual tests
    2. Goodness of fit tests
    3. Likelihood ratio tests
    4. Trend tests
4. How do you choose an appropriate physical acceleration model?
5. What models and assumptions are typically made when Bayesian methods are used for reliability evaluation?

8.2.1. How do you choose an appropriate life distribution model?

Choose models that make sense, fit the data and, hopefully, have a plausible theoretical justification

Life distribution models are chosen for one or more of the following three reasons:

1. There is a physical/statistical argument that theoretically matches a failure mechanism to a life distribution model
2. A particular model has previously been used successfully for the same or a similar failure mechanism
3. A convenient model provides a good empirical fit to all the failure data

Whatever method is used to choose a model, the model should

- "make sense" - for example, don't use an exponential model with a constant failure rate to model a "wear out" failure mechanism
- pass visual and statistical tests for fitting the data.

Models like the lognormal and the Weibull are so flexible that it is not uncommon for both to fit a small set of failure data equally well. Yet, especially when projecting via acceleration models to a use condition far removed from the test data, these two models may predict failure rates that differ by orders of magnitude. That is why it is more than an academic exercise to try to find a theoretical justification for using a particular distribution.

There are several useful theoretical arguments to help guide the choice of a model

We will consider three well-known arguments of this type:

- Extreme value argument
- Multiplicative degradation argument
- Fatigue life (Birnbaum-Saunders) model

Note that physical/statistical arguments for choosing a life distribution model are typically based on individual failure modes.

For some questions,

The Kaplan-Meier technique can be used when it is appropriate to just "let the data points speak for themselves"

an "empirical" distribution-free approach can be used	without making any model assumptions. However, you generally need a considerable amount of data for this approach to be useful, and acceleration modeling is much more difficult.

8. Assessing Product Reliability
8.2. Assumptions/Prerequisites
8.2.1. How do you choose an appropriate life distribution model?

8.2.1.1. Based on failure mode

Life distribution models and physical acceleration models typically only apply at the individual failure mode level

Failure mode data are failure data sorted by types of failures. Root cause analysis must be done on each failure incident in order to characterize them by failure mode. While this may be difficult and costly, it is a key part of any serious effort to understand, model, project and improve component or system reliability.

The natural place to apply both life distribution models and physical acceleration models is at the failure mode level. Each component failure mode will typically have its own life distribution model. The same is true for acceleration models. For the most part, these models only make sense at the failure mode level, and not at the component or system level. Once each mode (or mechanism) is modeled, the bottom-up approach can be used to build up to the entire component or system.

In particular, the arguments for choosing a life distribution model described in the next 3 sections apply at the failure mode level only. These are the Extreme value argument, the Multiplicative degradation argument and the Fatigue life (Birnbaum-Saunders) model.

The distribution-free (Kaplan - Meier) approach can be applied at any level (mode, component, system, etc.).

8.2.1.2. Extreme value argument

If component or system failure occurs when the first of many competing failure processes reaches a critical point, then Extreme Value Theory suggests that the Weibull Distribution will be a good model

It is well known that the **Central Limit Theorem** suggests that normal distributions will successfully model most engineering data when the observed measurements arise from the sum of many small random sources (such as measurement errors). Practical experience validates this theory - the normal distribution "works" for many engineering data sets.

Less known is the fact that Extreme Value Theory suggests that the Weibull distribution will successfully model failure times for mechanisms for which many competing similar failure processes are "racing" to failure and the first to reach it (i.e., the minimum of a large collection of roughly comparable random failure times) produces the observed failure time. Analogously, when a large number of roughly equivalent runners are competing and the winning time is recorded for many similar races, these times are likely to follow a Weibull distribution.

Note that this does not mean that anytime there are several failure mechanisms competing to cause a component or system to fail, the Weibull model applies. One or a few of these mechanisms may dominate the others and cause almost all of the failures. Then the "minimum of a large number of roughly comparable" random failure times does not apply and the proper model should be derived from the distribution models for the few dominating mechanisms using the competing risk model.

On the other hand, there are many cases in which failure occurs at the weakest link of a large number of similar degradation processes or defect flaws. One example of this occurs when modeling catastrophic failures of capacitors caused by dielectric material breakdown. Typical dielectric material has many "flaws" or microscopic sites where a breakdown will eventually take place. These sites may be thought of as **competing with each other** to reach failure first. The Weibull model, as extreme value theory would suggest, has been very successful as a life distribution model for this failure mechanism.

8.2.1.3. Multiplicative degradation argument

The lognormal model can be applied when degradation is caused by random shocks that increase degradation at a rate proportional to the total amount already present

A brief verbal description of the multiplicative degradation argument (leading to a derivation of the lognormal model) was given under Uses of the Lognormal Distribution Model. Here a formal derivation will be outlined because it gives insight into why the lognormal has been a successful model for many failure mechanisms based on degradation processes.

Let $y_1, y_2, \ldots y_n$ be measurements of the amount of degradation for a particular failure process taken at successive discrete instants of time as the process moves towards failure. Assume the following relationships exist between the y's:

$$y_i = (1 + \varepsilon_i) y_{i-1}$$

where the ε_i are small, independent random perturbations or "shocks" to the system that move the failure process along. In other words, the increase in the amount of degradation from one instant to the next is a small random multiple of the total amount of degradation already present. This is what is meant by **multiplicative degradation**. The situation is analogous to a snowball rolling down a snow covered hill; the larger it becomes, the faster it grows because it is able to pick up even more snow.

We can express the total amount of degradation at the n-th instant of time by

$$x_n = \left(\prod_{i=1}^{n} (1 + \varepsilon_i) \right) x_0$$

where x_0 is a constant and the ε_i are small random shocks. Next we take natural logarithms of both sides and obtain:

$$\ln x_n = \sum_{i=1}^{n} \ln(1 + \varepsilon_i) + \ln x_0 \approx \sum_{i=1}^{n} \varepsilon_i + \ln x_0$$

Using a Central Limit Theorem argument we can conclude that $\ln x_n$ has approximately a normal distribution. But by

the properties of the lognormal distribution, this means that x_n (or the amount of degradation) will follow approximately a lognormal model for any n (or at any time t). Since failure occurs when the amount of degradation reaches a critical point, time of failure will be modeled successfully by a lognormal for this type of process.

Failure mechanisms that might be successfully modeled by the lognormal distribution based on the multiplicative degradation model

What kinds of failure mechanisms might be expected to follow a multiplicative degradation model? The processes listed below are likely candidates:

1. Chemical reactions leading to the formation of new compounds
2. Diffusion or migration of ions
3. Crack growth or propagation

Many semiconductor failure modes are caused by one of these three degradation processes. Therefore, it is no surprise that the lognormal model has been very successful for the following semiconductor wear out failure mechanisms:

1. Corrosion
2. Metal migration
3. Electromigration
4. Diffusion
5. Crack growth

8.2.1.4. Fatigue life (Birnbaum-Saunders) model

A model derived from random crack growth occurring during many independent cycles of stress

The derivation of the Fatigue Life model is based on repeated cycles of stress causing degradation leading to eventual failure. The typical example is crack growth. One key assumption is that the amount of degradation during any cycle is independent of the degradation in any other cycle, with the same random distribution.

When this assumption matches well with a hypothesized physical model describing the degradation process, one would expect the Birnbaum-Saunders model to be a reasonable distribution model candidate. (See the note in the derivation for comments about the difference between the lognormal model derivation and the Fatigue Life model assumptions. Also see the comment on Miner's Rule).

8.2.1.5. Empirical model fitting - distribution free (Kaplan-Meier) approach

The Kaplan-Meier procedure gives CDF estimates for complete or censored sample data without assuming a particular distribution model

The Kaplan-Meier (K-M) Product Limit procedure provides quick, simple estimates of the Reliability function or the CDF based on failure data that may even be multicensored. No underlying model (such as Weibull or lognormal) is assumed; K-M estimation is an empirical (non-parametric) procedure. Exact times of failure are required, however.

Calculating Kaplan - Meier Estimates

The steps for calculating K-M estimates are the following:

1. Order the actual failure times from t_1 through t_r, where there are r failures
2. Corresponding to each t_i, associate the number n_i, with n_i = the number of operating units just before the the i-th failure occurred at time t_i
3. Estimate $R(t_1)$ by $(n_1 - 1)/n_1$
4. Estimate $R(t_i)$ by $R(t_{i-1}) \times (n_i - 1)/n_i$
5. Estimate the CDF $F(t_i)$ by $1 - R(t_i)$

Note that unfailed units taken off test (i.e., censored) only count up to the last actual failure time before they were removed. They are included in the n_i counts up to and including that failure time, but not after.

Example of K-M estimate calculations

A simple example will illustrate the K-M procedure. Assume 20 units are on life test and 6 failures occur at the following times: 10, 32, 56, 98, 122, and 181 hours. There were 4 unfailed units removed from the test for other experiments at the following times: 50 100 125 and 150 hours. The remaining 10 unfailed units were removed from the test at 200 hours. The K-M estimates for this life test are:

$R(10) = 19/20$
$R(32) = 19/20 \times 18/19$
$R(56) = 19/20 \times 18/19 \times 16/17$
$R(98) = 19/20 \times 18/19 \times 16/17 \times 15/16$

$R(122)$ = 19/20 x 18/19 x 16/17 x 15/16 x 13/14
$R(181)$ = 19/20 x 18/19 x 16/17 x 15/16 x 13/14 x 10/11

A General Expression for K-M Estimates

A general expression for the K-M estimates can be written. Assume we have n units on test and order the observed times for these n units from t_1 to t_n. Some of these are actual failure times and some are running times for units taken off test before they fail. Keep track of all the indices corresponding to actual failure times. Then the K-M estimates are given by:

$$\hat{R}(t_i) = \prod_{\substack{j \in S \\ t_j \le t_i}} \frac{n-j}{n-j+1}$$

with the "hat" over R indicating it is an estimate and S is the set of all subscripts j such that t_j is an actual failure time. The notation $j \in$ S and t_j less than or equal to t_i means we only form products for indices j that are in S and also correspond to times of failure less than or equal to t_i.

Once values for $R(t_i)$ are calculated, the CDF estimates are $F(t_i)$ = 1 - $R(t_i)$

A small modification of K-M estimates produces better results for probability plotting

Modified K-M Estimates

The K-M estimate at the time of the last failure is $R(t_r)$= 0 and $F(t_r)$ = 1. This estimate has a pessimistic bias and cannot be plotted (without modification) on a probability plot since the CDF for standard reliability models asymptotically approaches 1 as time approaches infinity. Better estimates for graphical plotting can be obtained by modifying the K-M estimates so that they reduce to the median rank estimates for plotting Type I Censored life test data (described in the next section). Modified K-M estimates are given by the formula

$$\hat{R}(t_i) = \frac{n+0.7}{n+0.4} \prod_{\substack{j \in S \\ t_j \le t_i}} \frac{n-j+0.7}{n-j+1.7}$$

Once values for $R(t_i)$ are calculated, the CDF estimates are $F(t_i)$ = 1 - $R(t_i)$

8.2.2. How do you plot reliability data?

Create a probability plot and if the points line up approximately on a straight line, the assumed model is a reasonable fit

Graphical plots of reliability data are quick, useful visual tests of whether a particular model is consistent with the observed data. The basic idea behind virtually all graphical plotting techniques is the following:

> Points calculated from the data are plotted on a log-log scale and, as long as they line up approximately on a straight line, the analyst can conclude that the data are consistent with the assumed model.

If the reliability data consist of (possibly multicensored) failure data from a non repairable population (or a repairable population for which only time to the first failure is considered) then the models are life distribution models such as the exponential, Weibull or lognormal. If the data consist of repair times for a repairable system, then the model might be the NHPP Power Law and the plot would be a Duane Plot.

The kinds of plots we will consider for failure data from non-repairable populations are:

- Probability (CDF) plots
- Hazard and Cum Hazard plots

For repairable populations we have

- Trend plots (to check whether an HPP or exponential model applies)
- Duane plots (to check whether the NHPP Power Law applies)

Later on (Section 8.4.2.1) we will also look at plots that can be used to check acceleration model assumptions.

Note: Many of the plots discussed in this section can also be used to obtain quick estimates of model parameters. This will be covered in later sections. While there may be other, more accurate ways of estimating parameters, simple graphical estimates can be very handy, especially when other techniques require software programs that are not readily available.

8.2.2.1. Probability plotting

Use probability plots to see your data and visually check model assumptions

Probability plots are simple visual ways of summarizing reliability data by plotting CDF estimates versus time using a log-log scale.

The *x* axis is labeled "Time" and the *y* axis is labeled "cumulative percent" or "percentile". There are rules, independent of the model, for calculating plotting positions (points) from the reliability data. These only depend on the type of censoring in the data and whether exact times of failure are recorded or only readout times.

Plot each failure mode separately

Remember that different failure modes can and should be separated out and individually analyzed. When analyzing failure mode A, for example, treat failure times from failure modes B, C, etc., as censored run times. Then repeat for failure mode B, and so on.

Data points line up roughly on a straight line when the model chosen is reasonable

When the points are plotted, the analyst fits a straight line to the data (either by eye, or with the aid of a least squares fitting program). Every straight line on, say, a Weibull probability plot uniquely corresponds to a particular Weibull life distribution model and the same is true for lognormal or exponential plots. If the points follow the line reasonably well, then the model is consistent with the data. If it was your previously chosen model, there is no reason to question the choice. In addition, there is a simple way to find the parameter estimates that correspond to the fitted straight line.

Plotting positions on the x axis depend on the type of data censoring

Plotting Positions: Censored Data (Type I or Type II)

At the time t_i of the i-th failure, we need an estimate of the CDF (or the cumulative population percent failure). The simplest and most obvious estimate is just $100 \times i/n$ (with a total of n units on test). This, however, is generally an overestimate (i.e. biased). Various texts recommend corrections such as $100 \times (i\text{-}0.5)/n$ or $100 \times i/(n+1)$. Here, we recommend what are known as (approximate) **median rank** estimates.

For each time t_i of the i-th failure, calculate the CDF or percentile estimate using $100 \times (i - 0.3)/(n + 0.4)$.

Plotting Positions: Readout Data

Let the readout times be $T_1, T_2, ..., T_k$ and let the corresponding new failures recorded at each readout be $r_1, r_2, ..., r_k$. Again, there are n units on test.

For each readout time T_j, calculate the CDF or percentile estimate using

$$\frac{100 \times \sum_{i=1}^{j} r_i}{n}$$

Plotting Positions: Multicensored Data

The calculations are more complicated for multicensored data. K-M estimates (described in a preceding section) can be used to obtain plotting positions at every failure time. The more precise Modified K-M Estimates are recommended. They reduce to the Censored Type I or the Censored Type II median rank estimates when the data consist of only failures, without any removals except possibly at the end of the test.

Reliability Models

Plotting positions on the y axis depend on the reliability model

The general idea is to take the model CDF equation and write it in such a way that a function of $F(t)$ is a linear equation of a function of t. This will be clear after a few examples. In the formulas that follow, "ln" always means "natural logarithm", while "log" always means "base 10 logarithm".

a) **Exponential Model:** Rewrite the exponential CDF as

$$\ln\left(\frac{1}{1-F(t)}\right) = \lambda t \quad or,\ equivalently,$$

$$\log\left(\frac{1}{1-F(t)}\right) = \frac{\lambda}{\ln 10} t$$

If we let $y = 1/\{1 - F(t)\}$ and $x = t$, then $\log(y)$ is linear in x with slope λ / ln10. Thus, we can make an exponential probability plot by using a logarithmic y axis. Use the plotting position estimates for $F(t_i)$ described above (without the 100 × multiplier) to calculate pairs of (x_i, y_i) points.

If the data are consistent with an exponential model, the resulting plot will have points that line up almost as a straight line going through the origin with slope λ / ln10.

b) **Weibull Model:** Rewrite the Weibull CDF as

$$\ln\ln\left(\frac{1}{1-F(t)}\right) = \gamma \ln t - \gamma \ln \alpha$$

$$or, \quad \log\ln\left(\frac{1}{1-F(t)}\right) = \gamma \log t - \gamma \log \alpha$$

If we let $y = \ln [1/\{1-F(t)\}]$ and $x = t$, then $\log(y)$ is linear in $\log(x)$ with slope γ. Thus, we can make a Weibull probability plot using a log-log scale. Use the plotting

position estimates for $F(t_i)$ (without the 100 × multiplier) to calculate pairs of (x_i, y_i) points.

If the data are consistent with a Weibull model, the resulting plot will have points that line up roughly on a straight line with slope γ. This line will cross the log x axis at time $t = \alpha$ and the log y axis (i.e., the intercept) at $-\gamma \log \alpha$.

c) **Lognormal Model**: Rewrite the lognormal cdf as

$$\ln t = \sigma \Phi^{-1}\{F(t)\} + \ln T_{50}$$

$$or,\ \log t = \frac{\sigma}{\ln 10}\Phi^{-1}\{F(t)\} + \log T_{50}$$

with Φ^{-1} denoting the inverse function for the standard normal distribution (taking a probability as an argument and returning the corresponding "z" value).

If we let $y = t$ and $x = \Phi^{-1}\{F(t)\}$, then log y is linear in x with slope σ / ln10 and intercept (when $F(t) = 0.5$) of log T_{50}. We generate a lognormal probability plot using a logarithmic y axis. Use the plotting position estimates for $F(t_i)$ (without the 100 × multiplier) to calculate pairs of (x_i, y_i) points.

If the data are consistent with a lognormal model, the resulting plot will have points that line up roughly on a straight line with slope σ / ln10 and intercept T_{50} on the log y axis.

d) **Extreme Value Distribution (Type I - for minimum):** Rewrite the extreme value distribution CDF as

$$\ln\ \{-\ln(1 - F(x))\} = (x - \mu)/\beta$$

If we let $y = -\ln(1 - F(x))$, then ln y is linear in x with slope $1/\beta$ and intercept $-\mu/\beta$. We plot y versus x where the y axis is base 10 logarithmic. The points should follow a straight line with a slope of $(1/\beta)\cdot\ln 10$ and an intercept of $(-\mu/\beta)\cdot\ln 10$. The ln10 factors in the slope and intercept are needed because the plot uses a base 10 logarithmic axis.

Example

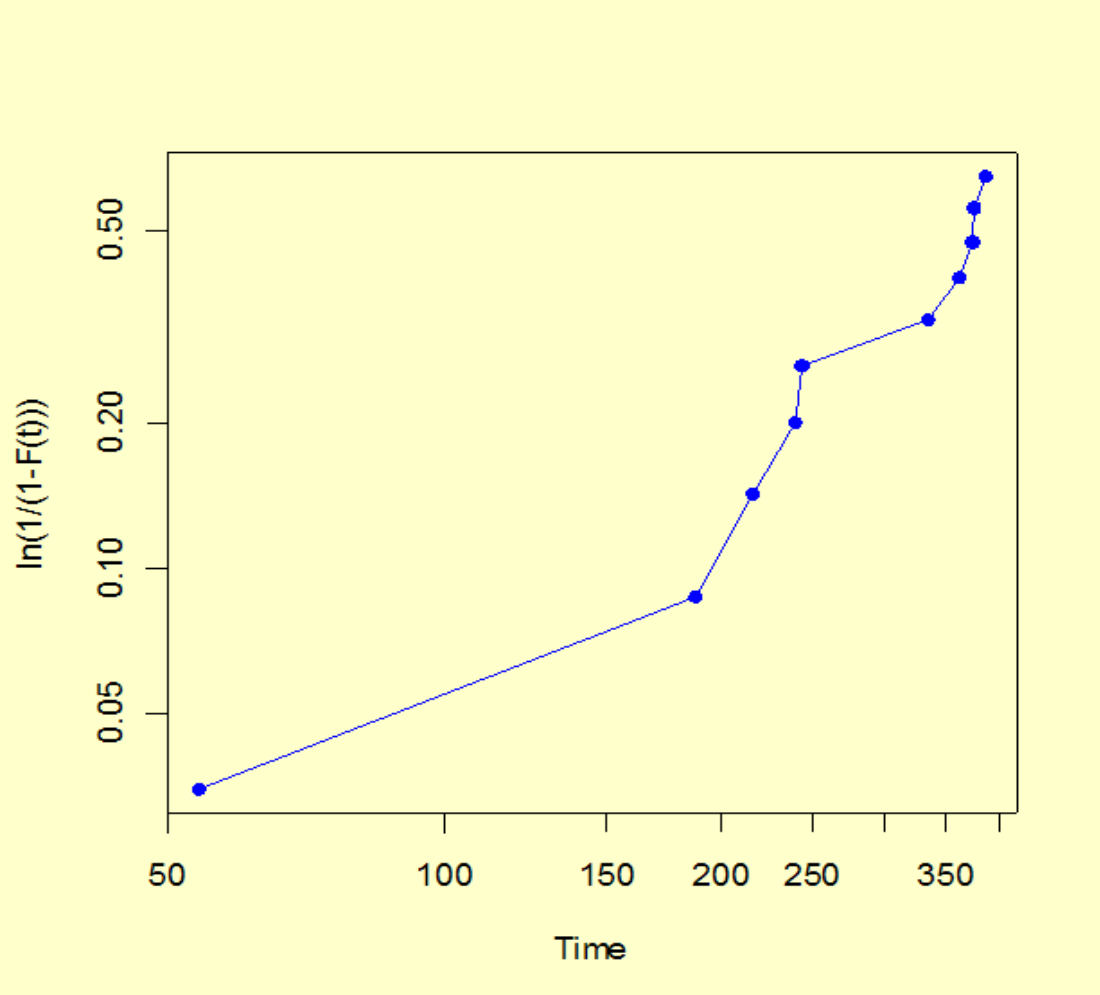
A Weibull example of probability plotting

We generated 20 random Weibull failure times with a shape parameter of $\gamma = 1.5$ and $\alpha = 500$. Assuming a test time of T = 500 hours, only 10 of these failure times would have been observed. They are, to the nearest hour: 54, 187, 216, 240, 244, 335, 361, 373, 375, and 386. We will compute plotting position CDF estimates based on these failure times, and then generate a probability plot.

(1) Failure (i)	(2) Time of Failure (x)	(3) $F(t_i)$ estimate (i-0.3)/(20+0.4)	(4) ln{1/(1-$F(t_i)$} (y)
1	54	.034	.035

2	187	.083	.087
3	216	.132	.142
4	240	.181	.200
5	244	.230	.262
6	335	.279	.328
7	361	.328	.398
8	373	.377	.474
9	375	.426	.556
10	386	.475	.645

We generate a probability plot using column (4) versus column (2) and log-log scale axes.



Note that the configuration of points appears to have some curvature. This is mostly due to the very first point on the plot (the earliest time of failure). The first few points on a probability plot have more variability than points in the central range and less attention should be paid to them when visually testing for "straightness".

Use of least squares (regression) to fit a line through the points on a

Since our data are plotted on a log-log scale, we fit a straight line using log(x) as the independent variable and log(y) as the dependent variable.

The regression produces a slope estimate of 1.46, which is close to the 1.5 value used in the simulation. The intercept is -4.114 and setting this equal to $-\gamma \log \alpha$ we estimate α = 657 (the "true" value used in the simulation was 500).

probability plot

The analyses in this section can can be implemented using both Dataplot code and R code. Both packages have special functions to automatically generate probability plots for a wide variety of distributions.

8.2.2.2. Hazard and cumulative hazard plotting

Cumulative Hazard Plotting has the same purpose as probability plotting

Similar to probability plots, cumulative hazard plots are used for visually examining distributional model assumptions for reliability data and have a similar interpretation as probability plots. The cumulative hazard plot consists of a plot of the cumulative hazard $H(t_i)$ versus the time t_i of the i-th failure. As with probability plots, the plotting positions are calculated independently of the model and a reasonable straight-line fit to the points confirms that the chosen model and the data are consistent.

Advantages of Cumulative Hazard Plotting

1. It is much easier to calculate plotting positions for multicensored data using cumulative hazard plotting techniques.
2. The most common reliability distributions, the exponential and the Weibull, are easily plotted.

Disadvantages of Cumulative Hazard Plotting

1. It is less intuitively clear just what is being plotted. In a probability plot, the cumulative percent failed is meaningful and the resulting straight-line fit can be used to identify times when desired percentages of the population will have failed. The percent cumulative hazard can increase beyond 100 % and is harder to interpret.
2. Normal cumulative hazard plotting techniques require exact times of failure and running times.
3. Since computers are able to calculate K-M estimates for probability plotting, the main advantage of cumulative hazard plotting goes away.

Since probability plots are generally more useful, we will only give a brief description of hazard plotting.

How to Make Cumulative Hazard Plots

1. Order the failure times and running times for each of the n units on test in ascending order from 1 to n. The order is called the rank of the unit. Calculate the reverse rank for each unit (reverse rank = n-rank +1).
2. Calculate a hazard "value" for every failed unit (do this only for the failed units). The hazard value for the failed unit with reverse rank k is just $1/k$.
3. Calculate the cumulative hazard values for each failed unit. The

cumulative hazard value corresponding to a particular failed unit is the sum of all the hazard values for failed units with ranks up to and including that failed unit.

4. Plot the time of failure versus the cumulative hazard value. Linear x and y scales are appropriate for an exponential distribution, while a log-log scale is appropriate for a Weibull distribution.

A life test cumulative hazard plotting example

Example: Ten units were tested at high stress test for up to 250 hours. Six failures occurred at 37, 73, 132, 195, 222 and 248 hours. Four units were taken off test without failing at the following run times: 50, 100, 200 and 250 hours. Cumulative hazard values were computed in the following table.

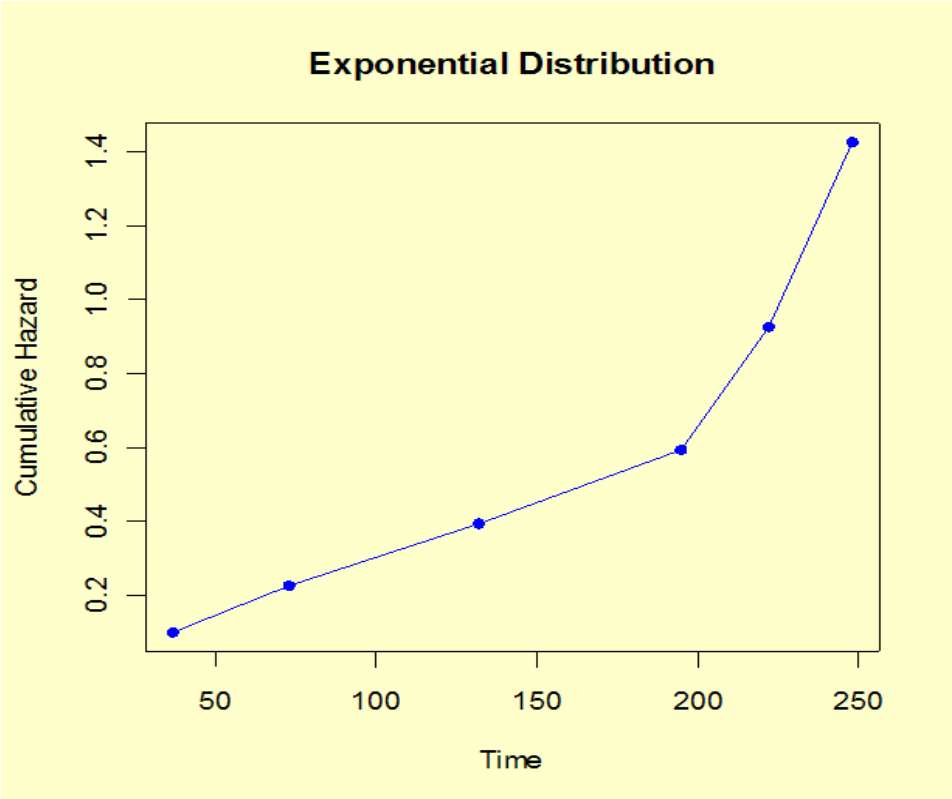
(1) Time of Event	(2) 1= failure 0=runtime	(3) Rank	(4) Reverse Rank	(5) Haz Val (2) x 1/(4)	(6) Cum Hazard Value
37	1	1	10	1/10	.10
50	0	2	9		
73	1	3	8	1/8	.225
100	0	4	7		
132	1	5	6	1/6	.391
195	1	6	5	1/5	.591
200	0	7	4		
222	1	8	3	1/3	.924
248	1	9	2	1/2	1.424
250	0	10	1		

Next ignore the rows with no cumulative hazard value and plot column (1) vs column (6).

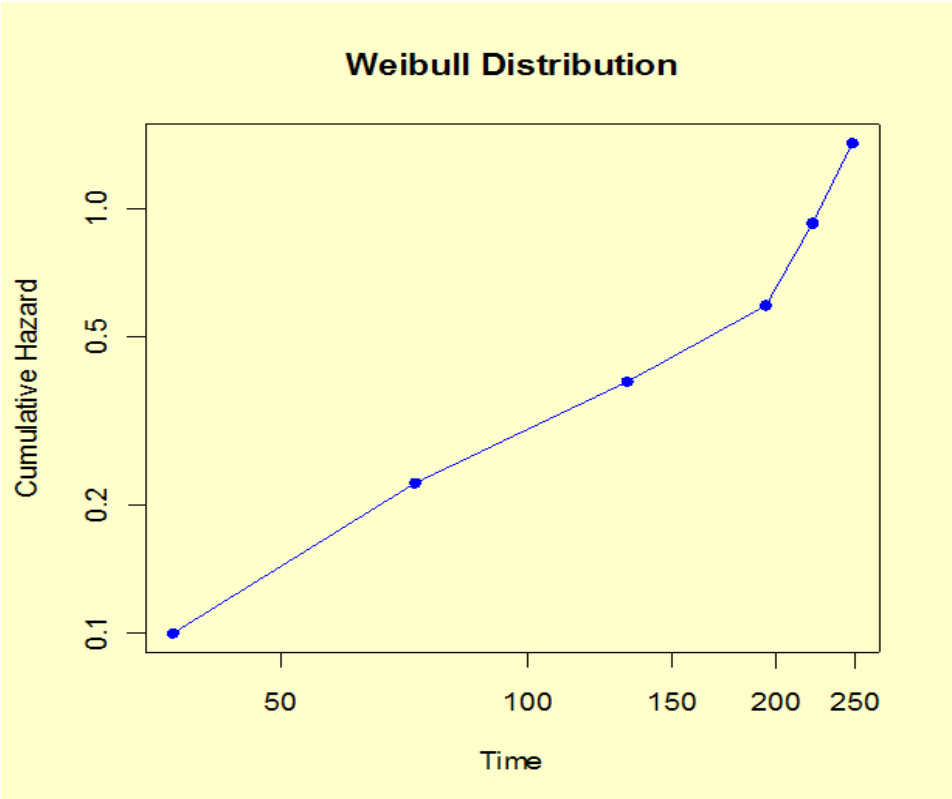
Plots of example data

Exponential and Weibull Cumulative Hazard Plots

The cumulative hazard for the exponential distribution is just $H(t) = \alpha\, t$, which is linear in t with an intercept of zero. So a simple linear graph of y = column (6) versus x = column (1) should line up as approximately a straight line going through the origin with slope λ if the exponential model is appropriate. The cumulative hazard plot for exponential distribution is shown below.



The cumulative hazard for the Weibull distribution is $H(t) = (t / \alpha)^{\gamma}$, so a plot of y versus x on a log-log scale should resemble a straight line with slope γ if the Weibull model is appropriate. The cumulative hazard plot for the Weibull distribution is shown below.



A least-squares regression fit of the data (using base 10 logarithms to transform columns (1) and (6)) indicates that the estimated slope for the Weibull distribution is 1.27, which is fairly similar to the exponential model slope of 1. The Weibull fit looks somewhat better than the exponential fit; however, with a sample of just 10, and only 6 failures, it

is difficult to pick a model from the data alone.

Software The analyses in this section can can be implemented using both Dataplot code and R code.

8.2.2.3. Trend and growth plotting (Duane plots)

Repair rates are typically either nearly constant over time or else consistently follow a good or a bad trend

Models for repairable systems were described earlier. These models are for the cumulative number of failuress (or the repair rate) over time. The two models used with most success throughout industry are the HPP (constant repair rate or "exponential" system model) and the NHPP Power Law process (the repair rate is the polynomial $m(t) = \alpha t^{-\beta}$).

Before constructing a Duane Plot, there are a few simple trend plots that often convey strong evidence of the presence or absence of a trend in the repair rate over time. If there is no trend, an HPP model is reasonable. If there is an apparent improvement or degradation trend, a Duane Plot will provide a visual check for whether the NHPP Power law model is consistent with the data.

A few simple plots can help us decide whether trends are present

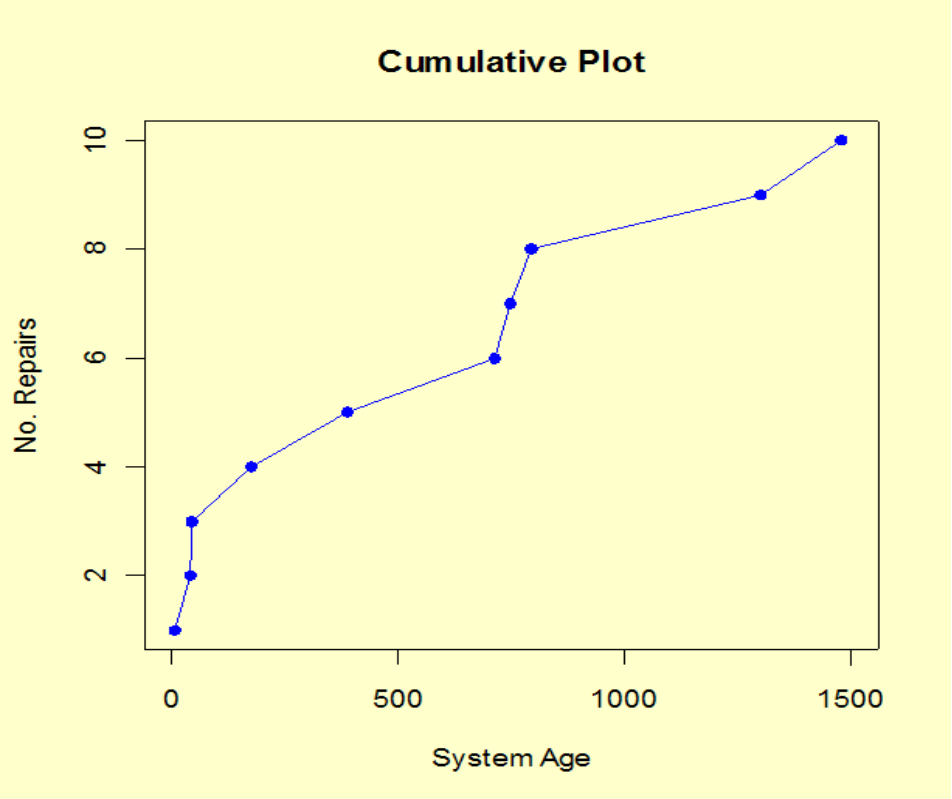
These simple visual graphical tests for trends are

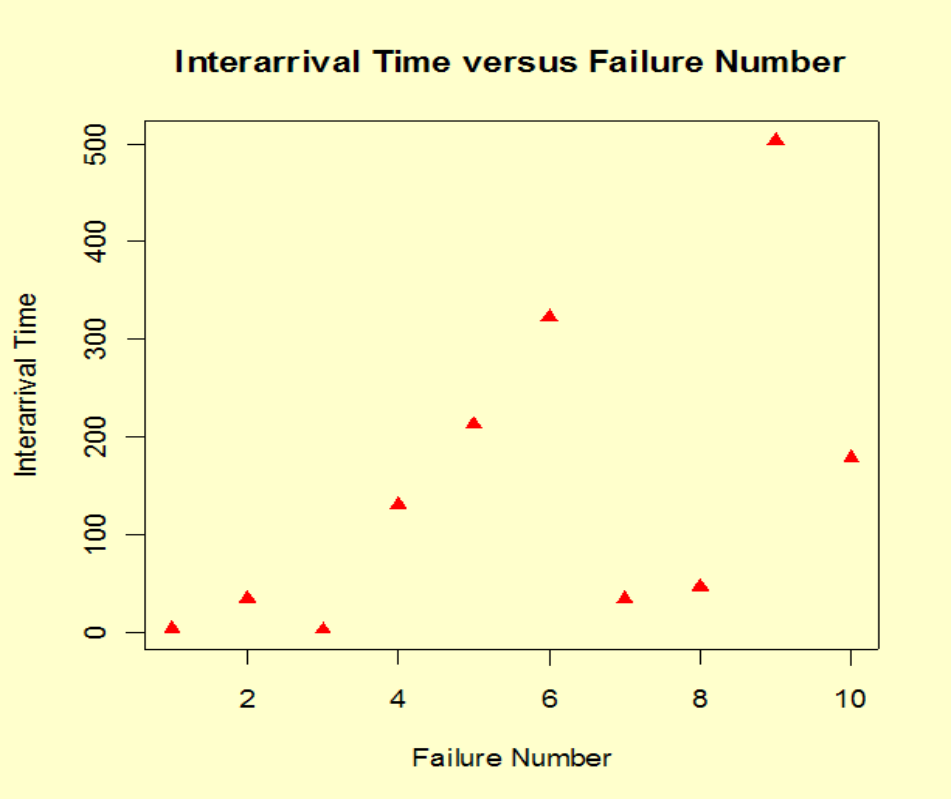
1. Plot cumulative failures versus system age (a step function that goes up every time there is a new failure). If this plot looks linear, there is no obvious improvement (or degradation) trend. A bending downward indicates improvement; bending upward indicates degradation.
2. Plot the inter arrival times between new failures (in other words, the waiting times between failures, with the time to the first failure used as the first "inter-arrival" time). If these trend up, there is improvement; if they trend down, there is degradation.
3. Plot the reciprocals of the inter-arrival times. Each reciprocal is a new failure rate estimate based only on the waiting time since the last failure. If these trend down, there is improvement; an upward trend indicates degradation.

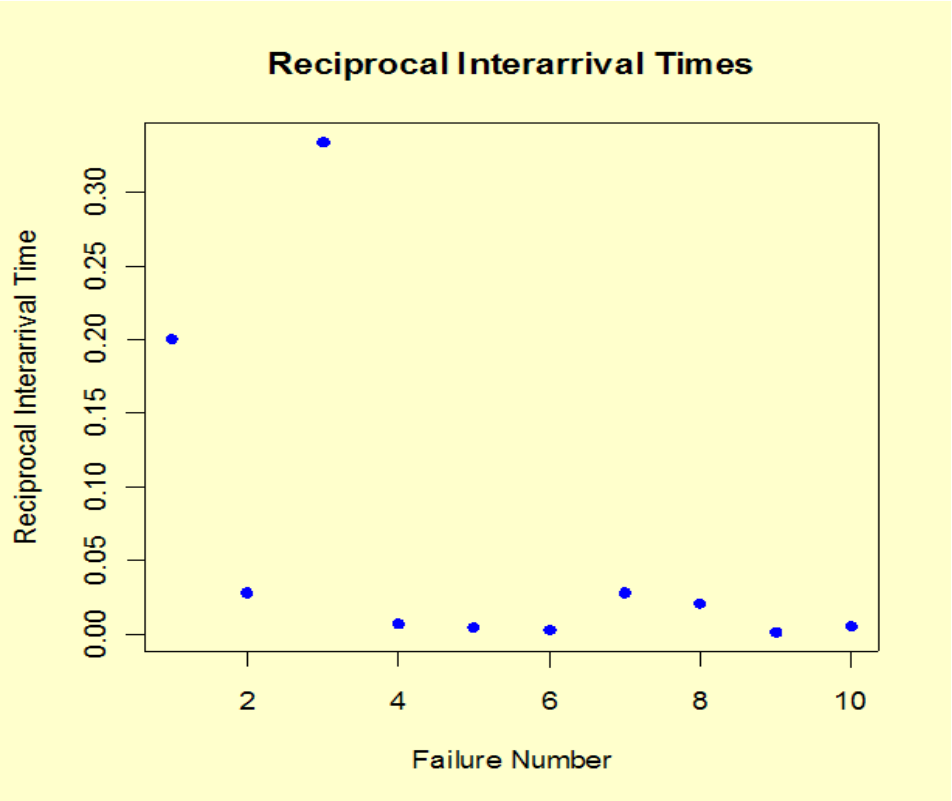
Trend plots and a Duane Plot for actual Reliability Improvement Test data

Case Study 1: Use of Trend Plots and Duane Plots with Reliability Improvement Test Data

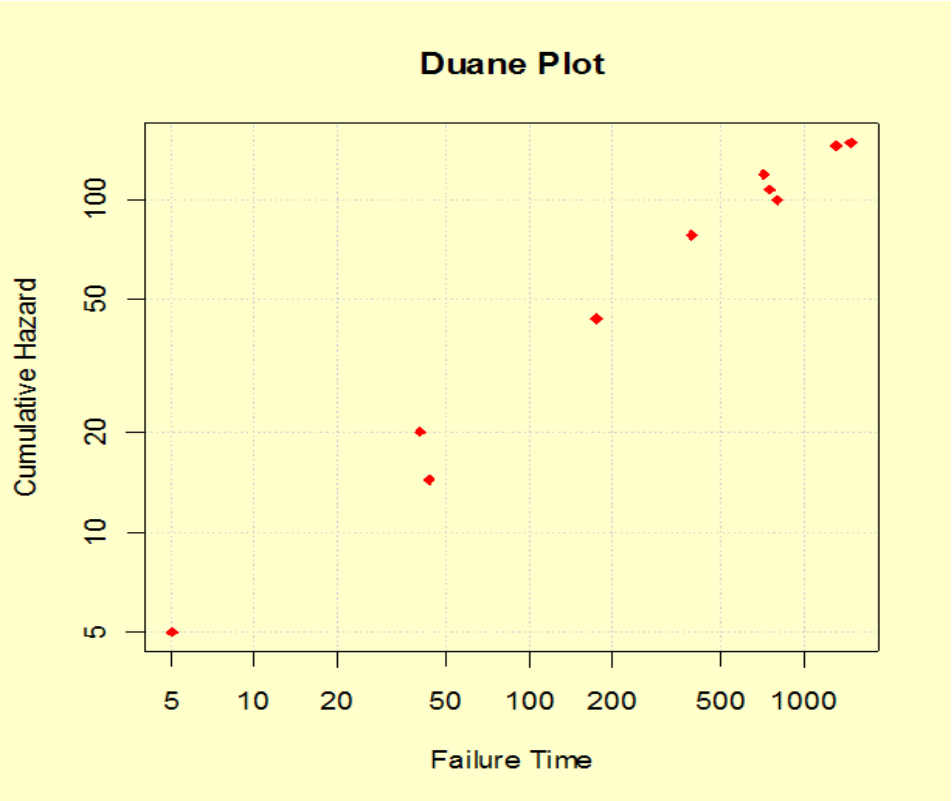
A prototype of a new, complex piece of equipment went through a 1500 operational hours Reliability Improvement Test. During the test there were 10 failures. As part of the improvement process, a cross functional Failure Review Board made sure every failure was analyzed down to the root cause and design and parts selection fixes were implemented on the prototype. The observed failure times were: 5, 40, 43, 175, 389, 712, 747, 795, 1299 and 1478 hours, with the test ending at 1500 hours. The reliability engineer on the Failure Review Board first made trend plots as described above, then made a Duane plot. These plots follow.


Cumulative Plot
No. Repairs
2
4
6
8
10
0
500
1000
1500
System Age


Interarrival Time versus Failure Number
Interarrival Time
0
100
200
300
400
500
2
4
6
8
10
Failure Number



Time	Cum MTBF
5	5
40	20
43	14.3
175	43.75
389	77.8
712	118.67
747	106.7
795	99.4
1299	144.3
1478	147.8



Comments: The three trend plots all show an improvement trend. The reason it might be useful to try all three trend plots is that a trend might show up more clearly on one plot than the others. Formal statistical tests on the significance of this visual evidence of a trend will be shown in the section on Trend Tests.

The points on the Duane Plot line up roughly as a straight line, indicating the NHPP Power Law model is consistent with the data.

Estimates for the reliability growth slope and the MTBF at the end of this test for this case study will be given in a later section.

8.2.3. How can you test reliability model assumptions?

Models are frequently necessary - but should always be checked

Since reliability models are often used to project (extrapolate) failure rates or MTBF's that are well beyond the range of the reliability data used to fit these models, it is very important to "test" whether the models chosen are consistent with whatever data are available. This section describes several ways of deciding whether a model under examination is acceptable. These are:

1. Visual Tests
2. Goodness of Fit Tests
3. Likelihood Ratio Tests
4. Trend Tests

8.2.3.1. Visual tests

A visual test of a model is a simple plot that tells us at a glance whether the model is consistent with the data

We have already seen many examples of visual tests of models. These were: Probability Plots, Cum hazard Plots, Duane Plots and Trend Plots. In all but the Trend Plots, the model was "tested' by how well the data points followed a straight line. In the case of the Trend Plots, we looked for curvature away from a straight line (cum repair plots) or increasing or decreasing size trends (inter arrival times and reciprocal inter-arrival times).

These simple plots are a powerful diagnostic tool since the human eye can often detect patterns or anomalies in the data by studying graphs. That kind of invaluable information would be lost if the analyst only used quantitative statistical tests to check model fit. Every analysis should include as many visual tests as are applicable.

Advantages of Visual Tests

1. Easy to understand and explain.
2. Can occasionally reveal patterns or anomalies in the data.
3. When a model "passes" a visual test, it is somewhat unlikely any quantitative statistical test will "reject" it (the human eye is less forgiving and more likely to detect spurious trends)

Combine visual tests with formal quantitative tests for the "best of both worlds" approach

Disadvantages of Visual Tests

1. Visual tests are subjective.
2. They do not quantify how well or how poorly a model fits the data.
3. They are of little help in choosing between two or more competing models that both appear to fit the data.
4. Simulation studies have shown that correct models may often appear to not fit well by sheer chance - it is hard to know when visual evidence is strong enough to reject what was previously believed to be a correct model.

You can retain the advantages of visual tests and remove their disadvantages by combining data plots with formal

statistical tests of goodness of fit or trend.

8.2.3.2. Goodness of fit tests

A Goodness of Fit test checks on whether your data are reasonable or highly unlikely, given an assumed distribution model

General tests for checking the hypothesis that your data are consistent with a particular model are discussed in Chapter 7. Details and examples of the Chi-Square Goodness of Fit test and the Kolmolgorov-Smirnov (K-S) test are given in Chapter 1. The Chi-Square test can be used with Type I or Type II censored data and readout data if there are enough failures and readout times. The K-S test generally requires complete samples, which limits its usefulness in reliability analysis.

These tests control the probability of rejecting a valid model as follows:

- the analyst chooses a confidence level designated by $100 \times (1 - \alpha)$.
- a test statistic is calculated from the data and compared to likely values for this statistic, assuming the model is correct.
- if the test statistic has a very unlikely value, or less than or equal to an α probability of occurring, where α is a small value like .1 or .05 or even .01, then the model is rejected.

So the risk of rejecting the right model is kept to α or less, and the choice of α usually takes into account the potential loss or difficulties incurred if the model is rejected.

8.2.3.3. Likelihood ratio tests

Likelihood Ratio Tests are a powerful, very general method of testing model assumptions. However, they require special software, not always readily available.

Likelihood functions for reliability data are described in Section 4. Two ways we use likelihood functions to choose models or verify/validate assumptions are:

1. Calculate the maximum likelihood of the sample data based on an assumed distribution model (the maximum occurs when unknown parameters are replaced by their maximum likelihood estimates). Repeat this calculation for other candidate distribution models that also appear to fit the data (based on probability plots). If all the models have the same number of unknown parameters, and there is no convincing reason to choose one particular model over another based on the failure mechanism or previous successful analyses, then pick the model with the largest likelihood value.

2. Many model assumptions can be viewed as putting restrictions on the parameters in a likelihood expression that effectively reduce the total number of unknown parameters. Some common examples are:

Examples where assumptions can be tested by the Likelihood Ratio Test

i) It is suspected that a type of data, typically modeled by a Weibull distribution, can be fit adequately by an exponential model. The exponential distribution is a special case of the Weibull, with the shape parameter γ set to 1. If we write the Weibull likelihood function for the data, the exponential model likelihood function is obtained by setting γ to 1, and the number of unknown parameters has been reduced from two to one.

ii) Assume we have n cells of data from an acceleration test, with each cell having a different operating temperature. We assume a lognormal population model applies in every cell. Without an acceleration model assumption, the likelihood of the experimental data would be the product of the likelihoods from each cell and there would be $2n$ unknown parameters (a different T_{50} and σ for each cell). If we assume an Arrhenius model applies, the total number of

parameters drops from 2*n* to just 3, the single common σ and the Arrhenius *A* and ΔH parameters. This acceleration assumption "saves" (2*n*-3) parameters.

iii) We life test samples of product from two vendors. The product is known to have a failure mechanism modeled by the Weibull distribution, and we want to know whether there is a difference in reliability between the vendors. The unrestricted likelihood of the data is the product of the two likelihoods, with 4 unknown parameters (the shape and characteristic life for each vendor population). If, however, we assume no difference between vendors, the likelihood reduces to having only two unknown parameters (the common shape and the common characteristic life). Two parameters are "lost" by the assumption of "no difference".

Clearly, we could come up with many more examples like these three, for which an important assumption can be restated as a reduction or restriction on the number of parameters used to formulate the likelihood function of the data. In all these cases, there is a simple and very useful way to test whether the assumption is consistent with the data.

The Likelihood Ratio Test Procedure

Details of the Likelihood Ratio Test procedure

Let L_1 be the maximum value of the likelihood of the data without the additional assumption. In other words, L_1 is the likelihood of the data with all the parameters unrestricted and maximum likelihood estimates substituted for these parameters.

In general, calculations are difficult and need to be built into the software you use

Let L_0 be the maximum value of the likelihood when the parameters are restricted (and reduced in number) based on the assumption. Assume *k* parameters were lost (i.e., L_0 has *k* less parameters than L_1).

Form the ratio $\lambda = L_0/L_1$. This ratio is always between 0 and 1 and the less likely the assumption is, the smaller λ will be. This can be quantified at a given confidence level as follows:

1. Calculate $\chi^2 = -2 \ln \lambda$. The smaller λ is, the larger χ^2 will be.
2. We can tell when χ^2 is significantly large by comparing it to the $100 \times (1-\alpha)$ percentile point of a Chi Square distribution with *k* degrees of freedom. χ^2 has an approximate Chi-Square distribution with *k* degrees of freedom and the approximation is usually

good, even for small sample sizes.

3. The likelihood ratio test computes χ^2 and rejects the assumption if χ^2 is larger than a Chi-Square percentile with *k* degrees of freedom, where the percentile corresponds to the confidence level chosen by the analyst.

Note: While Likelihood Ratio test procedures are very useful and widely applicable, the computations are difficult to perform by hand, especially for censored data, and appropriate software is necessary.

8.2.3.4. Trend tests

Formal Trend Tests should accompany Trend Plots and Duane Plots. Three are given in this section

In this section we look at formal statistical tests that can allow us to quantitatively determine whether or not the repair times of a system show a significant trend (which may be an improvement or a degradation trend). The section on trend and growth plotting contained a discussion of visual tests for trends - this section complements those visual tests as several numerical tests are presented.

Three statistical test procedures will be described:

1. The Reverse Arrangement Test (a simple and useful test that has the advantage of making no assumptions about a model for the possible trend)
2. The Military Handbook Test (optimal for distinguishing between "no trend' and a trend following the NHPP Power Law or Duane model)
3. The Laplace Test (optimal for distinguishing between "no trend' and a trend following the NHPP Exponential Law model)

The Reverse Arrangement Test (RAT test) is simple and makes no assumptions about what model a trend might follow

The Reverse Arrangement Test

Assume there are r repairs during the observation period and they occurred at system ages T_1, T_2, T_3, ...T_r (we set the start of the observation period to $T = 0$). Let $I_1 = T_1$, $I_2 = T_2 - T_1$, $I_3 = T_3 - T_2$, ..., $I_r = T_r - T_{r-1}$ be the inter-arrival times for repairs (i.e., the sequence of waiting times between failures). Assume the observation period ends at time $T_{end} > T_r$.

Previously, we plotted this sequence of inter-arrival times to look for evidence of trends. Now, we calculate how many instances we have of a later inter-arrival time being strictly greater than an earlier inter-arrival time. Each time that happens, we call it a *reversal*. If there are a lot of reversals (more than are likely from pure chance with no trend), we have significant evidence of an improvement trend. If there are too few reversals we have significant evidence of degradation.

A formal definition of the reversal count and some properties of this count are:

- count a reversal every time $I_j < I_k$ for some j and k with $j < k$
- this reversal count is the total number of reversals R
- for r repair times, the maximum possible number of reversals is $r(r\text{-}1)/2$
- if there are no trends, on the average one would expect to have $r(r\text{-}1)/4$ reversals.

As a simple example, assume we have 5 repair times at system ages 22, 58, 71, 156 and 225, and the observation period ended at system age 300 . First calculate the inter arrival times and obtain: 22, 36, 13, 85, 69. Next, count reversals by "putting your finger" on the first inter-arrival time, 22, and counting how many later inter arrival times are greater than that. In this case, there are 3. Continue by "moving your finger" to the second time, 36, and counting how many later times are greater. There are exactly 2. Repeating this for the third and fourth inter-arrival times (with many repairs, your finger gets very tired!) we obtain 2 and 0 reversals, respectively. Adding 3 + 2 + 2 + 0 = 7, we see that $R = 7$. The total possible number of reversals is 5x4/2 = 10 and an "average" number is half this, or 5.

In the example, we saw 7 reversals (2 more than average). Is this strong evidence for an improvement trend? The following table allows us to answer that at a 90% or 95% or 99% confidence level - the higher the confidence, the stronger the evidence of improvement (or the less likely that pure chance alone produced the result).

A useful table to check whether a reliability test has demonstrated significant improvement

Value of R Indicating Significant Improvement (One-Sided Test)

Number of Repairs	Minimum R for 90% Evidence of Improvement	Minimum R for 95% Evidence of Improvement	Minimum R for 99% Evidence of Improvement
4	6	6	-
5	9	9	10
6	12	13	14
7	16	17	19
8	20	22	24
9	25	27	30
10	31	33	36
11	37	39	43
12	43	46	50

One-sided test means before looking at the data we

expected improvement trends, or, at worst, a constant repair rate. This would be the case if we know of actions taken to improve reliability (such as occur during reliability improvement tests).

For the $r = 5$ repair times example above where we had $R = 7$, the table shows we do not (yet) have enough evidence to demonstrate a significant improvement trend. That does not mean that an improvement model is incorrect - it just means it is not yet "proved" statistically. With small numbers of repairs, it is not easy to obtain significant results.

For numbers of repairs beyond 12, there is a good approximation formula that can be used to determine whether R is large enough to be significant. Calculate

Use this formula when there are more than 12 repairs in the data set

$$z = \frac{R - \frac{r(r-1)}{4} + .5}{\sqrt{\frac{(2r+5)(r-1)r}{72}}}$$

and if $z > 1.282$, we have at least 90% significance. If $z > 1.645$, we have 95% significance, and a $z > 2.33$ indicates 99% significance since z has an approximate standard normal distribution.

That covers the (one-sided) test for significant improvement trends. If, on the other hand, we believe there may be a degradation trend (the system is wearing out or being over stressed, for example) and we want to know if the data confirms this, then we expect a low value for R and we need a table to determine when the value is low enough to be significant. The table below gives these critical values for R.

Value of R Indicating Significant Degradation Trend (One-Sided Test)

Number of Repairs	Maximum R for 90% Evidence of Degradation	Maximum R for 95% Evidence of Degradation	Maximum R for 99% Evidence of Degradation
4	0	0	-
5	1	1	0
6	3	2	1
7	5	4	2
8	8	6	4
9	11	9	6
10	14	12	9
11	18	16	12

12	23	20	16

For numbers of repairs r >12, use the approximation formula above, with R replaced by [$r(r\text{-}1)/2$ - R].

Because of the success of the Duane model with industrial improvement test data, this Trend Test is recommended

The Military Handbook Test

This test is better at finding significance when the choice is between no trend and a NHPP Power Law (Duane) model. In other words, if the data come from a system following the Power Law, this test will generally do better than any other test in terms of finding significance.

As before, we have r times of repair T_1, T_2, T_3, ...T_r with the observation period ending at time T_{end}>T_r. Calculate

$$\chi^2_{2r} = 2\sum_{i=1}^{r} \ln \frac{T_{end}}{T_i}$$

and compare this to percentiles of the chi-square distribution with 2r degrees of freedom. For a one-sided improvement test, reject no trend (or HPP) in favor of an improvement trend if the chi square value is beyond the 90 (or 95, or 99) percentile. For a one-sided degradation test, reject no trend if the chi-square value is less than the 10 (or 5, or 1) percentile.

Applying this test to the 5 repair times example, the test statistic has value 13.28 with 10 degrees of freedom, and the chi-square percentile is 79%.

The Laplace Test

This test is better at finding significance when the choice is between no trend and a NHPP Exponential model. In other words, if the data come from a system following the Exponential Law, this test will generally do better than any test in terms of finding significance.

As before, we have r times of repair T_1, T_2, T_3, ...T_r with the observation period ending at time T_{end}>T_r. Calculate

$$z = \frac{\sqrt{12r}\sum_{i=1}^{r}\left(T_i - \frac{T_{end}}{2}\right)}{rT_{end}}$$

and compare this to high (for improvement) or low (for degradation) percentiles of the standard normal distribution.

Formal tests

Case Study 1: Reliability Test Improvement Data

generally confirm the subjective information conveyed by trend plots

(Continued from earlier work)

The failure data and Trend plots and Duane plot were shown earlier. The observed failure times were: 5, 40, 43, 175, 389, 712, 747, 795, 1299 and 1478 hours, with the test ending at 1500 hours.

Reverse Arrangement Test: The inter-arrival times are: 5, 35, 3, 132, 214, 323, 35, 48, 504 and 179. The number of reversals is 33, which, according to the table above, is just significant at the 95% level.

The Military Handbook Test: The Chi-Square test statistic, using the formula given above, is 37.23 with 20 degrees of freedom and has significance level 98.9%. Since the Duane Plot looked very reasonable, this test probably gives the most precise significance assessment of how unlikely it is that sheer chance produced such an apparent improvement trend (only about 1.1% probability).

8.2.4. How do you choose an appropriate physical acceleration model?

Choosing a good acceleration model is part science and part art - but start with a good literature search

Choosing a physical acceleration model is a lot like choosing a life distribution model. First identify the failure mode and what stresses are relevant (i.e., will accelerate the failure mechanism). Then check to see if the literature contains examples of successful applications of a particular model for this mechanism.

If the literature offers little help, try the models described in earlier sections :

- Arrhenius
- The (inverse) power rule for voltage
- The exponential voltage model
- Two temperature/voltage models
- The electromigration model
- Three stress models (temperature, voltage and humidity)
- Eyring (for more than three stresses or when the above models are not satisfactory)
- The Coffin-Manson mechanical crack growth model

All but the last model (the Coffin-Manson) apply to chemical or electronic failure mechanisms, and since temperature is almost always a relevant stress for these mechanisms, the Arrhenius model is nearly always a part of any more general model. The Coffin-Manson model works well for many mechanical fatigue-related mechanisms.

Sometimes models have to be adjusted to include a **threshold level** for some stresses. In other words, failure might never occur due to a particular mechanism unless a particular stress (temperature, for example) is beyond a threshold value. A model for a temperature-dependent mechanism with a threshold at $T = T_0$ might look like

$$\text{time to fail} = f(T)/(T-T_0)$$

for which $f(T)$ could be Arrhenius. As the temperature decreases towards T_0, time to fail increases toward infinity in this (deterministic) acceleration model.

Models derived theoretically have been very successful and are convincing

In some cases, a mathematical/physical description of the failure mechanism can lead to an acceleration model. Some of the models above were originally derived that way.

Simple models are often the best

In general, use the simplest model (fewest parameters) you can. When you have chosen a model, use visual tests and formal statistical fit tests to confirm the model is consistent with your data. Continue to use the model as long as it gives results that "work," but be quick to look for a new model when it is clear the old one is no longer adequate.

There are some good quotes that apply here:

Quotes from experts on models

"All models are wrong, but some are useful." - George Box, and the principle of *Occam's Razor* (attributed to the 14th century logician William of Occam who said “Entities should not be multiplied unnecessarily” - or something equivalent to that in Latin).

A modern version of Occam's Razor is: If you have two theories that both explain the observed facts then you should use the simplest one until more evidence comes along - also called the **Law of Parsimony**.

Finally, for those who feel the above quotes place too much emphasis on simplicity, there are several appropriate quotes from Albert Einstein:

> "Make your theory as simple as possible, but no simpler"
>
> "For every complex question there is a simple and wrong solution."

8.2.5. What models and assumptions are typically made when Bayesian methods are used for reliability evaluation?

The basics of Bayesian methodology were explained earlier, along with some of the advantages and disadvantages of using this approach. Here we only consider the models and assumptions that are commonplace when applying Bayesian methodology to evaluate system reliability.

Bayesian assumptions for the gamma exponential system model

Assumptions:

1. Failure times for the system under investigation can be adequately modeled by the exponential distribution. For repairable systems, this means the HPP model applies and the system is operating in the flat portion of the bathtub curve. While Bayesian methodology can also be applied to non-repairable component populations, we will restrict ourselves to the system application in this Handbook.

2. The MTBF for the system can be regarded as chosen from a prior distribution model that is an analytic representation of our previous information or judgments about the system's reliability. The form of this prior model is the gamma distribution (the conjugate prior for the exponential model). The prior model is actually defined for λ = 1/MTBF since it is easier to do the calculations this way.

3. Our prior knowledge is used to choose the gamma parameters *a* and *b* for the prior distribution model for λ. There are many possible ways to convert "knowledge" to gamma parameters, depending on the form of the "knowledge" - we will describe three approaches.

Several ways to choose the prior gamma parameter values

i) If you have actual data from previous testing done on the system (or a system believed to have the same reliability as the one under investigation), this is the most credible prior knowledge, and the easiest to use. Simply set the gamma parameter *a* equal to the total number of failures from all the previous data, and set the parameter *b* equal to the total of all the previous test hours.

ii) A consensus method for determining *a* and *b* that

works well is the following: Assemble a group of engineers who know the system and its sub-components well from a reliability viewpoint.

- Have the group reach agreement on a reasonable MTBF they expect the system to have. They could each pick a number they would be willing to bet even money that the system would either meet or miss, and the average or median of these numbers would be their 50% best guess for the MTBF. Or they could just discuss even-money MTBF candidates until a consensus is reached.

- Repeat the process again, this time reaching agreement on a low MTBF they expect the system to exceed. A "5%" value that they are "95% confident" the system will exceed (i.e., they would give 19 to 1 odds) is a good choice. Or a "10%" value might be chosen (i.e., they would give 9 to 1 odds the actual MTBF exceeds the low MTBF). Use whichever percentile choice the group prefers.

- Call the reasonable MTBF $MTBF_{50}$ and the low MTBF you are 95% confident the system will exceed $MTBF_{05}$. These two numbers uniquely determine gamma parameters *a* and *b* that have λ percentile values at the right locations
$$\lambda_{50} = 1/MTBF_{50} \text{ and } \lambda_{95} = 1/MTBF_{05}$$
We call this method of specifying gamma prior parameters the **50/95 method** (or the 50/90 method if we use $MTBF_{10}$, etc.). A simple way to calculate *a* and *b* for this method is described below.

iii) A third way of choosing prior parameters starts the same way as the second method. Consensus is reached on an reasonable MTBF, $MTBF_{50}$. Next, however, the group decides they want a somewhat**weak prior** that will change rapidly, based on new test information. If the prior parameter "*a*" is set to 1, the gamma has a standard deviation equal to its mean, which makes it spread out, or "weak". To insure the 50th percentile is set at $\lambda_{50} = 1/ MTBF_{50}$, we have to choose $b = \ln 2 \times MTBF_{50}$, which is approximately $.6931 \times MTBF_{50}$.

Note: As we will see when we plan Bayesian tests, this *weak* prior is actually a very *friendly* prior in terms of saving test time

Many variations are possible, based on the above three methods. For example, you might have prior data from sources that you don't completely trust. Or you might question whether the data really apply to the system under investigation. You might decide to "weight" the prior data by .5, to "weaken" it. This can be implemented by setting a = .5 x the number of fails in the prior data and b = .5 times the number of test hours. That spreads out the prior distribution more, and lets it react quicker to new test data.

Consequences

After a new test is run, the posterior gamma parameters are easily obtained from the prior parameters by adding the new number of fails to "a" and the new test time to "b"

No matter how you arrive at values for the gamma prior parameters a and b, the method for incorporating new test information is the same. The new information is combined with the prior model to produce an updated or posterior distribution model for λ.

Under assumptions 1 and 2, when a new test is run with T system operating hours and r failures, the posterior distribution for λ is still a gamma, with new parameters:

$$a' = a + r,\ b' = b + T$$

In other words, add to a the number of new failures and add to b the number of new test hours to obtain the new parameters for the posterior distribution.

Use of the posterior distribution to estimate the system MTBF (with confidence, or prediction, intervals) is described in the section on estimating reliability using the Bayesian gamma model.

Obtaining Gamma Parameters

An example using the "50/95" consensus method

A group of engineers, discussing the reliability of a new piece of equipment, decide to use the 50/95 method to convert their knowledge into a Bayesian gamma prior. Consensus is reached on a likely $MTBF_{50}$ value of 600 hours and a low $MTBF_{05}$ value of 250. RT is 600/250 = 2.4. (**Note**: if the group felt that 250 was a $MTBF_{10}$ value, instead of a $MTBF_{05}$ value, then the only change needed would be to replace 0.95 in the B1 equation by 0.90. This would be the "50/90" method.)

Using software to find the root of a univariate function, the gamma prior parameters were found to be a = 2.863 and b = 1522.46. The parameters will have (approximately) a probability of 50% of l being below 1/600 = 0.001667 and a probability of 95% of λ being below 1/250 = 0.004. (The probabilities are based on the 0.001667 and 0.004 quantiles of a gamma distribution with shape parameter a = 2.863 and scale parameter b = 1522.46)

The gamma parameter estimates in this example can be produced using R code.

This example will be continued in Section 3, in which the Bayesian test time needed to confirm a 500 hour MTBF at 80% confidence will be derived.

8. Assessing Product Reliability

8.3. Reliability Data Collection

In order to assess or improve reliability, it is usually necessary to have failure data. Failure data can be obtained from field studies of system performance or from planned reliability tests, sometimes called *Life Tests*. This section focuses on how to plan reliability tests. The aim is to answer questions such as: how long should you test, what sample size do you need and what test conditions or stresses need to be run?

Detailed contents of Section 8.3

The section detailed outline follows.

3. Reliability Data Collection

1. How do you plan a reliability assessment test?
    1. Exponential life distribution (or HPP model) tests
    2. Lognormal or Weibull tests
    3. Reliability growth tests (Duane model)
    4. Accelerated life tests
    5. Bayesian gamma prior model tests

8.3.1. How do you plan a reliability assessment test?

The Plan for a reliability test ends with a detailed description of the mechanics of the test and starts with stating your assumptions and what you want to discover or prove

Planning a reliability test means:

- How long should you test?
- How many units have to be put on test?
  - For repairable systems, this is often limited to 1.
- If acceleration modeling is part of the experimental plan
  - What combination of stresses and how many experimental cells?
  - How many units go in each cell?

The answers to these questions depend on:

- What models are you assuming?
- What decisions or conclusions do you want to make after running the test and analyzing the data?
- What risks are you willing to take of making wrong decisions or conclusions?

It is not always possible, or practical, to completely answer all of these questions for every model we might want to use. This section looks at answers, or guidelines, for the following models:

- exponential or HPP Model
- Weibull or lognormal model
- Duane or NHPP Power Law model
- acceleration models
- Bayesian gamma prior model

8.3.1.1. Exponential life distribution (or HPP model) tests

Using an exponential (or HPP) model to test whether a system meets its MTBF requirement is common in industry

Exponential tests are common in industry for verifying that tools, systems or equipment are meeting their reliability requirements for Mean Time Between Failure (MTBF). The assumption is that the system has a constant failure (or repair) rate, which is the reciprocal of the MTBF. The waiting time between failures follows the exponential distribution model.

A typical test situation might be: a new complex piece of equipment or tool is installed in a factory and monitored closely for a period of several weeks to several months. If it has no more than a pre-specified number of failures during that period, the equipment "passes" its **reliability acceptance test**.

This kind of reliability test is often called a **Qualification Test** or a **Product Reliability Acceptance Test (PRAT)**. Contractual penalties may be invoked if the equipment fails the test. Everything is pegged to meeting a customer MTBF requirement at a specified confidence level.

How Long Must You Test A Piece of Equipment or a System In order to Assure a Specified MTBF at a Given Confidence?

You start with a given MTBF objective, say M, and a confidence level, say 100 × (1-α). You need one more piece of information to determine the test length: how many fails do you want to allow and still "pass" the equipment? The more fails allowed, the longer the test required. However, a longer test allowing more failures has the desirable feature of making it less likely a good piece of equipment will be rejected because of random "bad luck" during the test period.

The recommended procedure is to iterate on r = the number of allowable fails until a larger r would require an unacceptable test length. For any choice of r, the corresponding test length is quickly calculated by multiplying M (the objective) by the factor in the table below corresponding to the r-th row and the desired confidence level column.

For example, to confirm a 200-hour MTBF objective at 90% confidence, allowing up to 4 failures on the test, the test length must be 200 × 7.99 = 1598 hours. If this is unacceptably long, try allowing only 3 fails for a test length of 200 × 6.68 = 1336 hours. The shortest test would allow no fails and last 200 × 2.3 = 460 hours. All these tests guarantee a 200-hour MTBF at 90% confidence, when the equipment passes. However, the shorter test are much less "fair" to the supplier in that they have a large chance of failing a marginally acceptable piece of equipment.

Use the Test length Table to determine how long to test

Test Length Guide Table

NUMBER OF FAILURES ALLOWED	FACTOR FOR GIVEN CONFIDENCE LEVELS					
r	**50%**	**60%**	**75%**	**80%**	**90%**	**95%**
0	.693	.916	1.39	1.61	2.30	3.00
1	1.68	2.02	2.69	2.99	3.89	4.74
2	2.67	3.11	3.92	4.28	5.32	6.30
3	3.67	4.18	5.11	5.52	6.68	7.75
4	4.67	5.24	6.27	6.72	7.99	9.15
5	5.67	6.29	7.42	7.90	9.28	10.51
6	6.67	7.35	8.56	9.07	10.53	11.84
7	7.67	8.38	9.68	10.23	11.77	13.15
8	8.67	9.43	10.80	11.38	13.00	14.43
9	9.67	10.48	11.91	12.52	14.21	15.70
10	10.67	11.52	13.02	13.65	15.40	16.96
15	15.67	16.69	18.48	19.23	21.29	23.10
20	20.68	21.84	23.88	24.73	27.05	29.06

The formula to calculate the factors in the table is the following.

$FAC = .5\chi^2_{\alpha;2(r+1)}$ with $\chi^2_{\alpha;2(r+1)}$ denoting the upper 100x(1-α) percentile of the chi-square distribution with 2(r+1) degrees of freedom

Example: A new factory tool must meet a 400-hour MTBF requirement at 80% confidence. You have up to two months of 3-shift operation to decide whether the tool is acceptable. What is a good test plan?

Two months of around-the-clock operation, with some time off for maintenance and repairs, amounts to a maximum of about 1300 hours. The 80% confidence factor for r = 1 is

2.99, so a test of 400 × 2.99 = about 1200 hours (with up to 1 fail allowed) is the best that can be done.

Shorten required test times by testing more than one system

NOTE: Exponential test times can be shortened significantly if several similar tools or systems can be put on test at the same time. Test time means the same as "tool hours" and one tool operating for 1000 hours is equivalent (as far as the exponential model is concerned) to 2 tools operating for 500 hours each, or 10 tools operating for 100 hours each. Just count all the fails from all the tools and the sum of the test hours from all the tools.

8. Assessing Product Reliability
8.3. Reliability Data Collection
8.3.1. How do you plan a reliability assessment test?

8.3.1.2. Lognormal or Weibull tests

Planning reliability tests for distributions other than the exponential is difficult and involves a lot of guesswork

Planning a reliability test is not simple and straightforward when the assumed model is lognormal or Weibull. Since these models have two parameters, no estimates are possible without at least two test failures, and good estimates require considerably more than that. Because of censoring, without a good guess ahead of time at what the unknown parameters are, any test plan may fail.

However, it is often possible to make a good guess ahead of time about at least one of the unknown parameters - typically the "shape" parameter (σ for the lognormal or γ for the Weibull). With one parameter assumed known, test plans can be derived that assure the reliability or failure rate of the product tested will be acceptable.

Lognormal Case (shape parameter known): The lognormal model is used for many microelectronic wear-out failure mechanisms, such as electromigration. As a production monitor, samples of microelectronic chips taken randomly from production lots might be tested at levels of voltage and temperature that are high enough to significantly accelerate the occurrence of electromigration failures. Acceleration factors are known from previous testing and range from several hundred to several thousand.

Lognormal test plans, assuming sigma and the acceleration factor are known

The goal is to construct a test plan (put *n* units on stress test for T hours and accept the lot if no more than *r* failures occur). The following assumptions are made:

- The life distribution model is lognormal
- Sigma = σ_0 is known from past testing and does not vary appreciably from lot to lot
- Lot reliability varies because T_{50}'s (the lognormal median or 50th percentile) differ from lot to lot
- The acceleration factor from high stress to use stress is a known quantity "*A*"
- A stress time of T hours is practical as a line monitor
- A nominal use T_{50} of T_u (combined with σ_0) produces an acceptable use CDF (or use reliability function). This is equivalent to specifying an acceptable use CDF at, say, 100,000 hours to be a given value *p* and

calculating T_u via:

$$T_u = 100{,}000e^{-\sigma\Phi^{-1}(p_0)}$$

where Φ^{-1} is the inverse of the standard normal distribution

- An unacceptable use CDF of p_1 leads to a "bad" use T_{50} of T_b, using the same equation as above with p_o replaced by p_1

The acceleration factor A is used to calculate a "good" or acceptable proportion of failures p_a at stress and a "bad" or unacceptable proportion of fails p_b:

$$p_a = \Phi\left(\frac{\ln(AT/T_u)}{\sigma_0}\right), \quad p_b = \Phi\left(\frac{\ln(AT/T_b)}{\sigma_0}\right)$$

where Φ is the standard normal CDF. This reduces the reliability problem to a well-known Lot Acceptance Sampling Plan (LASP) problem, which was covered in Chapter 6.

If the sample size required to distinguish between p_a and p_b turns out to be too large, it may be necessary to increase T or test at a higher stress. The important point is that the above assumptions and equations give a methodology for planning ongoing reliability tests under a lognormal model assumption.

Weibull test plans, assuming gamma and the acceleration. factor are known

Weibull Case (shape parameter known): The assumptions and calculations are similar to those made for the lognormal:

- The life distribution model is Weibull
- Gamma = γ_0 is known from past testing and does not vary appreciably from lot to lot
- Lot reliability varies because α's (the Weibull characteristic life or 62.3 percentile) differ from lot to lot
- The acceleration factor from high stress to use stress is a known quantity "A"
- A stress time of T hours is practical as a line monitor
- A nominal use α of α_u (combined with γ_0) produces an acceptable use CDF (or use reliability function). This is equivalent to specifying an acceptable use CDF at, say, 100,000 hours to be a given value p_0 and calculating α_u

$$\alpha_u = \frac{AT}{\left[-\ln(1-p_0)\right]^{\frac{1}{\gamma_0}}}$$

- An unacceptable use CDF of p_1 leads to a "bad" use of α_b, using the same equation as above with p_o replaced by p_1

The acceleration factor A is used next to calculate a "good" or acceptable proportion of failures p_a at stress and a "bad" or unacceptable proportion of failures p_b:

$$p_a = 1 - e^{-\left(\frac{AT}{\alpha_u}\right)^{\gamma_0}} \quad , \quad p_b = 1 - e^{-\left(\frac{AT}{\alpha_b}\right)^{\gamma_0}}$$

This reduces the reliability problem to a Lot Acceptance Sampling Plan (LASP) problem, which was covered in Chapter 6.

If the sample size required to distinguish between p_a and p_b turns out to be too large, it may be necessary to increase T or test at a higher stress. The important point is that the above assumptions and equations give a methodology for planning ongoing reliability tests under a Weibull model assumption.

Planning Tests to Estimate Both Weibull or Both Lognormal Parameters

Rules-of-thumb for general lognormal or Weibull life test planning

All that can be said here are some general rules-of-thumb:

1. If you can observe at least 10 exact times of failure, estimates are usually reasonable - below 10 failures the critical shape parameter may be hard to estimate accurately. Below 5 failures, estimates are often very inaccurate.
2. With readout data, even with more than 10 total failures, you need failures in three or more readout intervals for accurate estimates.
3. When guessing how many units to put on test and for how long, try various reasonable combinations of distribution parameters to see if the corresponding calculated proportion of failures expected during the test, multiplied by the sample size, gives a reasonable number of failures.
4. As an alternative to the last rule, simulate test data from reasonable combinations of distribution parameters and see if your estimates from the simulated data are close to the parameters used in the simulation. If a test plan doesn't work well with simulated data, it is not likely to work well with real data.

8. Assessing Product Reliability
8.3. Reliability Data Collection
8.3.1. How do you plan a reliability assessment test?

8.3.1.3. Reliability growth (Duane model)

Guidelines for planning how long to run a reliability growth test

A reliability improvement test usually takes a large resource commitment, so it is important to have a way of estimating how long a test will be required. The following procedure gives a starting point for determining a test time:

1. Guess a starting value for β, the growth slope. Some guidelines were previously discussed. Pick something close to 0.3 for a conservative estimate (perhaps a new cross-functional team will be working on the improvement test or the system to be improved has many new parts with possibly unknown failure mechanisms), or close to 0.5 for a more optimistic estimate.
2. Use current data and engineering estimates to arrive at a consensus for what the starting MTBF for the system is. Call this M_1.
3. Let M_T be the target MTBF (the customer requirement). Then the improvement needed on the test is given by

   $IM = M_T/M_1$
4. A first pass estimate of the test time needed is

$$T = IM^{1/\beta}$$

This estimate comes from using the starting MTBF of M_1 as the MTBF after 1 hour on test and using the fact that the improvement from 1 hour to T hours is just T^{β}.

Make sure test time makes engineering sense

The reason the above is just a first pass estimate is it will give unrealistic (too short) test times when a high β is assumed. A very short reliability improvement test makes little sense because a minimal number of failures must be observed before the improvement team can determine design and parts changes that will "grow" reliability. And it takes time to implement these changes and observe an improved repair rate.

Iterative

Simulation methods can also be used to see if a planned test

simulation is an aid for test planning	is likely to generate data that will demonstrate an assumed growth rate.

8.3.1.4. Accelerated life tests

Accelerated testing is needed when testing even large sample sizes at use stress would yield few or no failures within a reasonable time

Accelerated life tests are component life tests with components operated at high stresses and failure data observed. While high stress testing can be performed for the sole purpose of seeing where and how failures occur and using that information to improve component designs or make better component selections, we will focus in this section on accelerated life testing for the following two purposes:

1. To study how failure is accelerated by stress and fit an acceleration model to data from multiple stress cells
2. To obtain enough failure data at high stress to accurately project (extrapolate) what the CDF at use will be.

If we already know the acceleration model (or the acceleration factor to typical use conditions from high stress test conditions), then the methods described two pages ago can be used. We assume, therefore, that the acceleration model is not known in advance.

Test planning means picking stress levels and sample sizes and test times to produce enough data to fit models and make projections

Test planning and operation for a (multiple) stress cell life test experiment consists of the following:

- Pick several combinations of the relevant stresses (the stresses that accelerate *the failure mechanism under investigation*). Each combination is a "stress cell". Note that you are planning for only one mechanism of failure at a time. Failures on test due to any other mechanism will be considered censored run times.
- Make sure stress levels used are not too high - to the point where new failure mechanisms that would never occur at use stress are introduced. Picking a maximum allowable stress level requires experience and/or good engineering judgment.
- Put random samples of components in each stress cell and run the components in each cell for fixed (but possibly different) lengths of time.
- Gather the failure data from each cell and use the data to fit an acceleration model and a life distribution model and use these models to project reliability at use stress conditions.

Test planning would be similar to topics already covered in the chapters that discussed modeling and experimental design except for one important point. When you test components in a stress cell for a fixed length test, it is typical that some (or possibly many) of the components end the test without failing. This is the censoring problem, and it greatly complicates experimental design to the point at which it becomes almost as much of an art (based on engineering judgment) as a statistical science.

An example will help illustrate the design issues. Assume a metal migration failure mode is believed to follow the 2-stress temperature voltage model given by

$$t_f = Ae^{\frac{\Delta H}{kT}}V^{\beta}$$

Normal use conditions are 4 volts and 25 degrees Celsius, and the high stress levels under consideration are 6, 8,12 volts and 85^o, 105^o and 125^o. It probably would be a waste of resources to test at (6v, 85^o), or even possibly (8v, 85^o) or (6v,105^o) since these cells are not likely to have enough stress acceleration to yield a reasonable number of failures within typical test times.

If you write all the 9 possible stress cell combinations in a 3x3 matrix with voltage increasing by rows and temperature increasing by columns, the result would look like the matrix below:

Matrix Leading to "Backward L Design"

6v, 85^o	6v, 105^o	**6v, 125^o**
8v, 85^o	8v,105^o	**8v,125^o**
12v,85^o	**12v,105^o**	**12v,125^o**

"Backwards L" designs are common in accelerated life testing. Put more experimental units in lower stress cells.

The combinations in bold are the most likely design choices covering the full range of both stresses, but still hopefully having enough acceleration to produce failures. This is the so-called **"backwards L" design** commonly used for acceleration modeling experiments.

Note: It is good design practice to put more of your test units in the lower stress cells, to make up for the fact that these cells will have a smaller proportion of units failing.

Sometimes simulation is the best way to learn whether a test plan has a chance of working

Design by Simulation:

A lengthy, but better way to choose a test matrix is the following:

- Pick an acceleration model and a life distribution model (as usual).
- Guess at the shape parameter value of the life distribution model based on literature studies or earlier experiments. The shape parameter should remain the same for all stress cells. Choose a scale parameter value at use so that the use stress CDF exactly meets requirements (i.e., for the lognormal, pick a use T_{50} that gives the desired use reliability - for a Weibull model choice, do the same for the characteristic life parameter).
- Guess at the acceleration model parameters values (ΔH and β, for the 2-stress model shown above). Again, use whatever is in the literature for similar failure mechanisms or data from earlier experiments).
- Calculate acceleration factors from any proposed test cells to use stress and divide the use scale parameter by these acceleration factors to obtain "trial" cell scale parameters.
- Simulate cell data for each proposed stress cell using the derived cell scale parameters and the guessed shape parameter.
- Check that every proposed cell has sufficient failures to give good estimates.
- Adjust the choice of stress cells and the sample size allocations until you are satisfied that, if everything goes as expected, the experiment will yield enough data to provide good estimates of the model parameters.

After you make advance estimates, it is sometimes possible to construct an optimal experimental design - but software for this is scarce

Optimal Designs:

Recent work on designing accelerated life tests has shown it is possible, for a given choice of models and assumed values of the unknown parameters, to construct an optimal design (one which will have the best chance of providing good sample estimates of the model parameters). These optimal designs typically select stress levels as far apart as possible and heavily weight the allocation of sample units to the lower stress cells. However, unless the experimenter can find software that incorporates these optimal methods for his or her particular choice of models, the methods described above are the most practical way of designing acceleration experiments.

8.3.1.5. Bayesian gamma prior model

How to plan a Bayesian test to confirm a system meets its MTBF objective

Review Bayesian Basics and assumptions, if needed. We start at the point when gamma prior parameters a and b have already been determined. Assume we have a given MTBF objective, M, and a desired confidence level of $100\times(1-\alpha)$. We want to confirm the system will have an MTBF of at least M at the $100\times(1-\alpha)$ confidence level. As in the section on classical (HPP) test plans, we pick a number of failures, r, that we can allow on the test. We need a test time T such that we can observe up to r failures and still "pass" the test. If the test time is too long (or too short), we can iterate with a different choice of r.

When the test ends, the posterior gamma distribution will have (worst case - assuming exactly r failures) new parameters of

$$a' = a + r,\ b' = b + T$$

and passing the test means that the failure rate $\lambda_{1-\alpha}$, the upper $100\times(1-\alpha)$ percentile for the posterior gamma, has to equal the target failure rate $1/M$. But this percentile is, by definition, $G^{-1}(1-\alpha; a', b')$, with G^{-1} denoting the inverse of the gamma distribution with parameters a', b'. We can find the value of T that satisfies $G^{-1}(1-\alpha; a', b') = 1/M$ by trial and error. However, based on the properties of the gamma distribution, it turns out that we can calculate T directly by using

$$T = M\times(G^{-1}(1-\alpha; a', 1)) - b$$

Special Case: The Prior Has $a = 1$ (The "Weak" Prior)

When the prior is a weak prior with a = 1, the Bayesian test is always shorter than the classical

There is a very simple way to calculate the required Bayesian test time when the prior is a weak prior with $a = 1$. Just use the Test Length Guide Table to calculate the classical test time. Call this T_c. The Bayesian test time T is just T_c minus the prior parameter b (i.e., $T = T_c - b$). If the b parameter was set equal to $(\ln 2) \times MTBF_{50}$ (where $MTBF_{50}$ is the consensus choice for an "even money" MTBF), then

$$T = T_c - (\ln 2) \times MTBF_{50}$$

This shows that when a weak prior is used, the Bayesian test

test time is always less than the corresponding classical test time. That is why this prior is also known as a **friendly prior**.

Note: In general, Bayesian test times can be shorter, or longer, than the corresponding classical test times, depending on the choice of prior parameters. However, the Bayesian time will always be shorter when the prior parameter a is less than, or equal to, 1.

Example: Calculating a Bayesian Test Time

Example A new piece of equipment has to meet a MTBF requirement of 500 hours at 80 % confidence. A group of engineers decide to use their collective experience to determine a Bayesian gamma prior using the 50/95 method described in Section 2. They think 600 hours is a likely MTBF value and they are very confident that the MTBF will exceed 250. Following the example in Section 2, they determine that the gamma prior parameters are a = 2.863 and b = 1522.46.

Now they want to determine an appropriate test time so that they can confirm a MTBF of 500 with at least 80 % confidence, provided they have no more than two failures.

We obtain a test time of 1756.117 hours using

$$500\times(G^{-1}(1\text{-}0.2;\ 2.863\text{+}2,\ 1))\ \text{-}\ 1522.46$$

To compare this result to the classical test time required, use the Test Length Guide Table. The table factor is 4.28, so the test time needed is 500 × 4.28 = 2140 hours for a non-Bayesian test. The Bayesian test saves about 384 hours, or an 18 % savings. If the test is run for 1756 hours, with no more than two failures, then an MTBF of at least 500 hours has been confirmed at 80 % confidence.

If, instead, the engineers had decided to use a weak prior with an $MTBF_{50}$ of 600, the required test time would have been

$$2140 - 600 \times \ln 2 = 1724 \text{ hours}$$

8. Assessing Product Reliability

8.4. Reliability Data Analysis

After you have obtained component or system reliability data, how do you fit life distribution models, reliability growth models, or acceleration models? How do you estimate failure rates or MTBF's and project component or system reliability at use conditions? This section answers these kinds of questions.

Detailed outline for Section 4

The detailed outline for section 4 follows.

4. Reliability Data Analysis

1. How do you estimate life distribution parameters from censored data?
    1. Graphical estimation
    2. Maximum Likelihood Estimation (MLE)
    3. A Weibull MLE example
2. How do you fit an acceleration model?
    1. Graphical estimation
    2. Maximum likelihood
    3. Fitting models using degradation data instead of failures
3. How do you project reliability at use conditions?
4. How do you compare reliability between two or more populations?
5. How do you fit system repair rate models?
    1. Constant repair rate (HPP/Exponential) model
    2. Power law (Duane) model
    3. Exponential law model
6. How do you estimate reliability using the Bayesian gamma prior model?

8.4.1. How do you estimate life distribution parameters from censored data?

Graphical estimation methods (aided by computer line fits) are easy and quick

Maximum likelihood methods are usually more precise - but require special software

Two widely used general methods will be described in this section:

- Graphical estimation
- Maximum Likelihood Estimation (MLE)

Recommendation On Which Method to Use

Maximum likelihood estimation (except when the failure data are very sparse - i.e., only a few failures) is a more precise and flexible method. However, with censored data, the method of maximum likelihood estimation requires special computer programs for distributions other than the exponential. This is no longer an obstacle since, in recent years, many statistical software packages have added reliability platforms that will calculate MLE's and most of these packages will estimate acceleration model parameters and give confidence bounds as well.

If important business decisions are based on reliability projections made from life test data and acceleration modeling, then it pays to obtain state-of-the art MLE reliability software. Otherwise, for monitoring and tracking reliability, estimation methods based on computer-augmented graphical procedures will often suffice.

8.4.1.1. Graphical estimation

The line on a probability plot uniquely identifies distributional parameters

Once you have calculated plotting positions from your failure data, and have generated the probability plot for your chosen model, parameter estimation follows easily. But along with the mechanics of graphical estimation, be aware of both the advantages and the disadvantages of graphical estimation methods.

Most probability plots have simple procedures to calculate underlying distribution parameter estimates

Graphical Estimation Mechanics:

If you draw a line through points on a probability plot, there are usually simple rules to find estimates of the slope (or shape parameter) and the scale parameter. On lognormal probability plot with time on the x-axis and cumulative percent on the y-axis, draw horizontal lines from the 34th and the 50th percentiles across to the fitted line, and drop vertical lines to the time axis from these intersection points. The time corresponding to the 50th percentile is the T_{50} estimate. Divide T_{50} by the time corresponding to the 34th percentile (this is called T_{34}). The natural logarithm of that ratio is the estimate of sigma, or the slope of the line ($\sigma = \ln(T_{50} / T_{34})$).

For a Weibull probability plot draw a horizontal line from the y-axis to the fitted line at the 62.3 percentile point. That estimation line intersects the line through the points at a time that is the estimate of the characteristic life parameter α. In order to estimate the slope of the fitted line (or the shape parameter γ), choose any two points on the fitted line and divide the change in the y variable by the change in x variable.

Using a computer generated line fitting routine removes subjectivity and can lead directly to computer

To remove the subjectivity of drawing a line through the points, a least-squares (regression) fit can be performed using the equations described in the section on probability plotting. An example of this for the Weibull was also shown in that section. Another example of a Weibull plot for the same data appears later in this section.

Finally, if you have exact times and complete samples (no censoring), many software packages have built-in Probability Plotting functions. Examples were shown in the

parameter estimates based on the plotting positions

sections describing various life distribution models.

Do probability plots even if you use some other method for the final estimates

Advantages of Graphical Methods of Estimation:

- Graphical methods are quick and easy to use and make visual sense.
- Calculations can be done with little or no special software needed.
- Visual test of model (i.e., how well the points line up) is an additional benefit.

Disadvantages of Graphical Methods of Estimation

Perhaps the worst drawback of graphical estimation is you cannot get legitimate confidence intervals for the estimates

The statistical properties of graphical estimates (i.e., how precise are they on average) are not good:

- they are biased,
- even with large samples, they are not minimum variance (i.e., most precise) estimates,
- graphical methods do not give confidence intervals for the parameters (intervals generated by a regression program for this kind of data are incorrect), and
- formal statistical tests about model fit or parameter values cannot be performed with graphical methods.

As we will see in the next section, Maximum Likelihood Estimates overcome all these disadvantages - at least for reliability data sets with a reasonably large number of failures - at a cost of losing all the advantages listed above for graphical estimation.

8.4.1.2. Maximum likelihood estimation

There is nothing visual about the maximum likelihood method - but it is a powerful method and, at least for large samples, very precise

Maximum likelihood estimation begins with writing a mathematical expression known as the ***Likelihood Function*** of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the ***Maximum Likelihood Estimates*** or **MLE's**.

Maximum likelihood estimation is a totally analytic maximization procedure. It applies to every form of censored or multicensored data, and it is even possible to use the technique across several stress cells and estimate acceleration model parameters at the same time as life distribution parameters. Moreover, MLE's and Likelihood Functions generally have very desirable large sample properties:

- they become unbiased minimum variance estimators as the sample size increases
- they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds
- likelihood functions can be used to test hypotheses about models and parameters

With small samples, MLE's may not be very precise and may even generate a line that lies above or below the data points

There are only two drawbacks to MLE's, but they are important ones:

- With small numbers of failures (less than 5, and sometimes less than 10 is small), MLE's can be heavily biased and the large sample optimality properties do not apply
- Calculating MLE's often requires specialized software for solving complex non-linear equations. This is less of a problem as time goes by, as more statistical packages are upgrading to contain MLE analysis capability every year.

Additional information about maximum likelihood estimatation can be found in Chapter 1.

Likelihood equation for censored data

Likelihood Function Examples for Reliability Data:

Let $f(t)$ be the PDF and $F(t)$ the CDF for the chosen life distribution model. Note that these are functions of t and the unknown parameters of the model. The likelihood function for Type I Censored data is:

$$L = C\left(\prod_{i=1}^{r} f(t_i)\right)\left(1-F(T)\right)^{n-r}$$

with C denoting a constant that plays no role when solving for the MLE's. Note that with no censoring, the likelihood reduces to just the product of the densities, each evaluated at a failure time. For Type II Censored Data, just replace T above by the random end of test time t_r.

The likelihood function for readout data is:

$$L = C\left(\prod_{i=1}^{k}\left(F(T_i - F(T_{i-1})\right)^{r_i}\right)\left(1-F(T)\right)^{n-\sum_{i=1}^{k} r_i}$$

with $F(T_0)$ defined to be 0.

In general, any multicensored data set likelihood will be a constant times a product of terms, one for each unit in the sample, that look like either $f(t_i)$, $[F(T_i)-F(T_{i-1})]$, or $[1-F(t_i)]$, depending on whether the unit was an exact time failure at time t_i, failed between two readouts T_{i-1} and T_i, or survived to time t_i and was not observed any longer.

The general mathematical technique for solving for MLE's involves setting partial derivatives of ln L (the derivatives are taken with respect to the unknown parameters) equal to zero and solving the resulting (usually non-linear) equations. The equation for the exponential model can easily be solved, however.

MLE for the exponential model parameter λ turns out to be just (total # of failures) divided by (total unit test time)

MLE's for the Exponential Model (Type I Censoring):

$$L = C\lambda^r e^{-\lambda\sum_{i=1}^{r} t_i}\left(e^{-\lambda(n-r)T}\right)$$

$$\ln L = \ln C + r\ln\lambda - \lambda\sum_{i=1}^{r} t_i - \lambda(n-r)T$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{r}{\lambda} - \sum_{i=1}^{r} t_i - (n-r)T = 0$$

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{r} t_i + (n-r)T}$$

Note: The MLE of the failure rate (or repair rate) in the exponential case turns out to be the total number of failures observed divided by the total unit test time. For the MLE of the MTBF, take the reciprocal of this or use the total unit test hours divided by the total observed failures.

There are examples of Weibull and lognormal MLE analysis later in this section.

8.4.1.3. A Weibull maximum likelihood estimation example

Reliability analysis using Weibull data

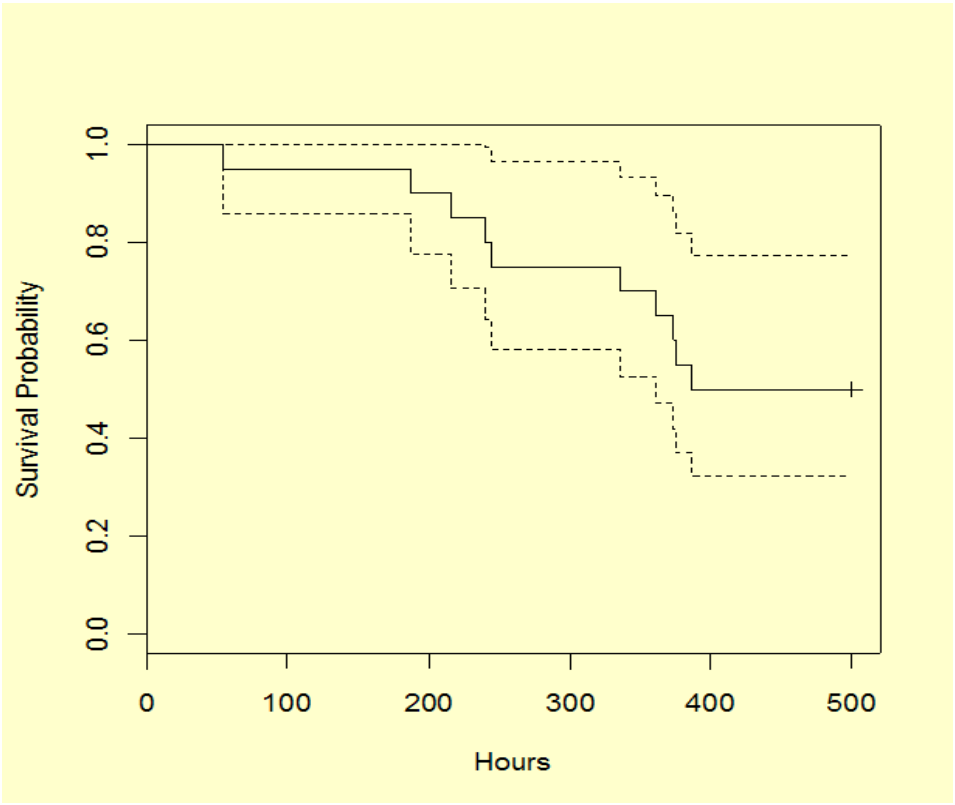
We will plot Weibull censored data and estimate parameters using data from a previous example (8.2.2.1).

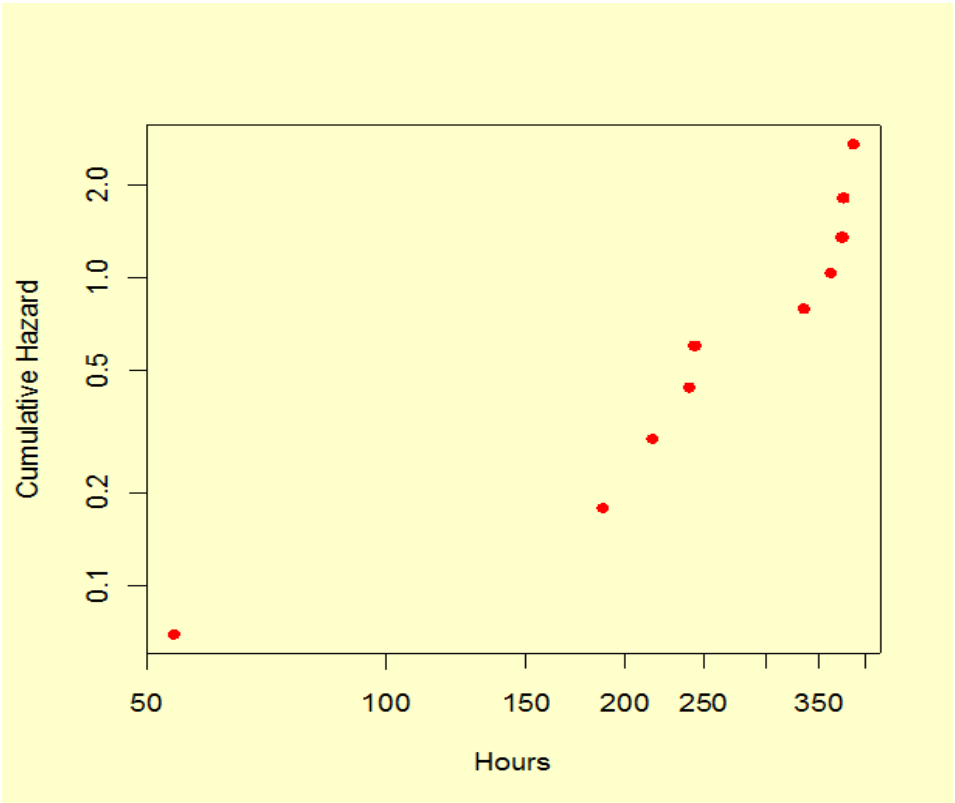
The recorded failure times were 54, 187, 216, 240, 244, 335, 361, 373, 375, and 386 hours, and 10 units that did not fail were removed from the test at 500 hours. The data are summarized in the following table.

Time	Censored	Frequency
54	0	1
187	0	1
216	0	1
240	0	1
244	0	1
335	0	1
361	0	1
373	0	1
375	0	1
386	0	1
500	1	10

The column labeled "Time" contains failure and censoring times, the "Censored" column contains a variable to indicate whether the time in column one is a failure time or a censoring time, and the "Frequency" column shows how many units failed or were censored at that time.

First, we generate a survival curve using the Kaplan-Meier method and a Weibull probability plot. **Note**: Some software packages might use the name "Product Limit Method" or "Product Limit Survival Estimates" instead of the equivalent name "Kaplan-Meier".





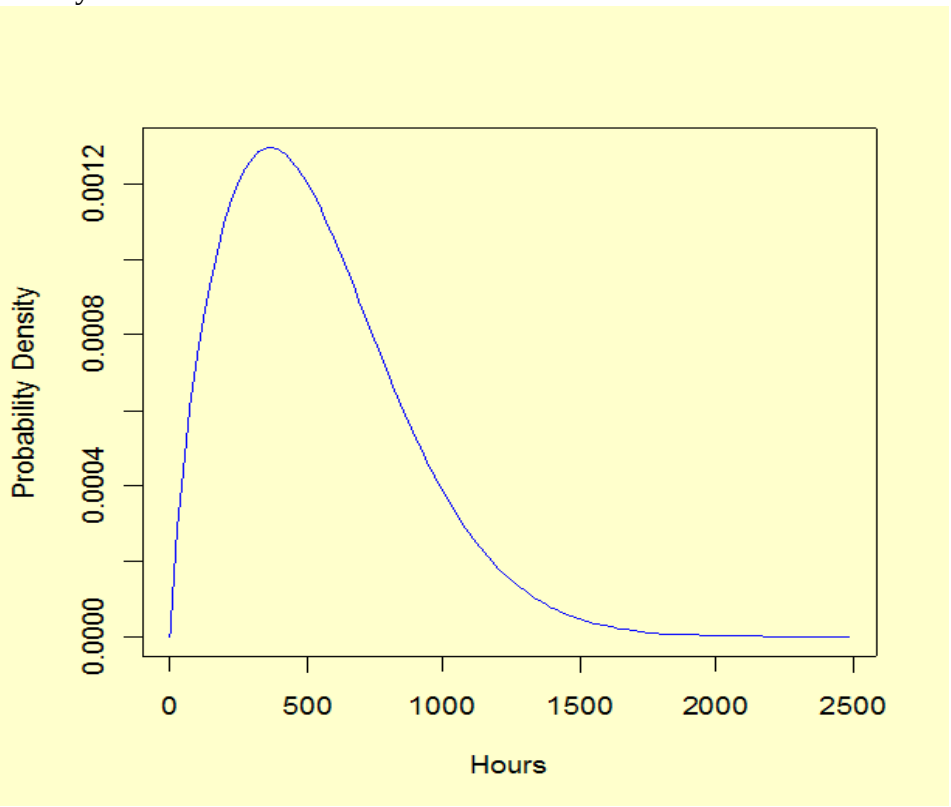
Next, we perform a regression analysis for a survival model assuming that failure times have a Weibull distribution. The Weibull characteristic life parameter (η) estimate is 606.5280 and the shape parameter (β) estimate is 1.7208.

The log-likelihood and Akaike's Information Criterion (AIC) from the model fit are -75.135 and 154.27. For comparison, we computed the AIC

for the lognormal distribution and found that it was only slightly larger than the Weibull AIC.

```
Lognormal AIC   Weibull AIC
       154.39         154.27
```

When comparing values of AIC, smaller is better. The probability density of the fitted Weibull distribution is shown below.



Based on the estimates of η and β, the lifetime expected value and standard deviation are the following.

$$\hat{\eta} = 606.5280$$
$$\hat{\beta} = 1.7208$$

$$\hat{\mu} = \hat{\eta}\,\Gamma(1 + 1/\hat{\beta}) = 540.737 \text{ hours}$$

$$\hat{\sigma} = \hat{\eta}\sqrt{\Gamma(1 + 2/\hat{\beta}) - (\Gamma(1 + 1/\hat{\beta}))^2} = 323.806 \text{ hours}$$

The greek letter, Γ, represents the gamma function.

Discussion

Maximum likelihood estimation (MLE) is an accurate and easy way to estimate life distribution parameters, provided that a good software analysis package is available. The package should also calculate confidence bounds and log-likelihood values.

The analyses in this section can can be implemented using R code.

8.4.2. How do you fit an acceleration model?

Acceleration models can be fit by either graphical procedures or maximum likelihood methods

As with estimating life distribution model parameters, there are two general approaches for estimating acceleration model parameters:

- Graphical estimation (or computer procedures based on a graphical approach)
- Maximum Likelihood Estimation (an analytic approach based on writing the likelihood of all the data across all the cells, incorporating the acceleration model).

The same comments and recommendations concerning these methods still apply. Note that it is even harder, however, to find useful software programs that will do maximum likelihood estimation across stress cells and fit and test acceleration models.

Sometimes it is possible to fit a model using degradation data

Another promising method of fitting acceleration models is sometimes possible when studying failure mechanisms characterized by a stress-induced gradual degradation process that causes the eventual failure. This approach fits models based on degradation data and has the advantage of not actually needing failures. This overcomes censoring limitations by providing measurement data at consecutive time intervals for every unit in every stress cell.

8.4.2.1. Graphical estimation

This section will discuss the following:

1. How to fit an Arrhenius model with graphical estimation
2. Graphical estimation: an Arrhenius model example
3. Fitting more complicated models

Estimate acceleration model parameters by estimating cell T_{50} values (or α values) and then using regression to fit the model across the cells

How to fit an Arrhenius Model with Graphical Estimation

Graphical methods work best (and are easiest to describe) for a simple one-stress model like the widely used Arrhenius model

$$t_f = A\exp\left\{\frac{\Delta H}{kT}\right\}$$

with T denoting temperature measured in degrees Kelvin (273.16 + degrees Celsius) and k is Boltzmann's constant (8.617×10^{-5} in eV/K).

When applying an acceleration model to a distribution of failure times, we interpret the deterministic model equation to apply at any distribution percentile we want. This is equivalent to setting the life distribution scale parameter equal to the model equation (T_{50} for the lognormal, α for the Weibull and the MTBF or $1/\lambda$ for the exponential). For the lognormal, for example, we have

$$T_{50} = Ae^{\frac{\Delta H}{kT}}$$

$$\ln T_{50} = y = \ln A + \Delta H\left(\frac{1}{kT}\right)$$

This can be written as

$$y = a + bx \quad \text{with } b = \Delta H \text{ and } x = \frac{1}{kT}$$

So, if we run several stress cells and compute T_{50} values for each cell, a plot of the natural log of these T_{50} values versus the corresponding $1/kT$ values should be roughly linear with a slope of ΔH and an intercept of $\ln A$. In practice, a computer fit of a line through these points is typically used to obtain the Arrhenius model estimates. Remember that T is in Kelvin in the above equations. For temperature in Celsius, use the

following for 1/*kT*:
11605/(*t* °C + 273.16).

An example will illustrate the procedure.

Graphical Estimation: An Arrhenius Model Example:

Arrhenius model example

Component life tests were run at three temperatures: 85 °C, 105 °C and 125 °C. The lowest temperature cell was populated with 100 components; the 105 °C cell had 50 components and the highest stress cell had 25 components. All tests were run until either all the units in the cell had failed or 1000 hours was reached. Acceleration was assumed to follow an Arrhenius model and the life distribution model for the failure mode was believed to be lognormal. The normal operating temperature for the components is 25 °C and it is desired to project the use CDF at 100,000 hours.

Test results:

Cell 1 (85 °C): 5 failures at 401, 428, 695, 725 and 738 hours. Ninety-five units were censored at 1000 hours running time.

Cell 2 (105 °C): 35 failures at 171, 187, 189, 266, 275, 285, 301, 302, 305, 316, 317, 324, 349, 350, 386, 405, 480, 493, 530, 534, 536, 567, 589, 598, 599, 614, 620, 650, 668, 685, 718, 795, 854, 917, and 926 hours. Fifteen units were censored at 1000 hours running time.

Cell 3 (125 °C): 24 failures at 24, 42, 92, 93, 141, 142, 143, 159, 181, 188, 194, 199, 207, 213, 243, 256, 259, 290, 294, 305, 392, 454, 502 and 696. One unit was censored at 1000 hours running time.

Failure analysis confirmed that all failures were due to the same failure mechanism (if any failures due to another mechanism had occurred, they would have been considered censored run times in the Arrhenius analysis).

Steps to Fitting the Distribution Model and the Arrhenius Model:

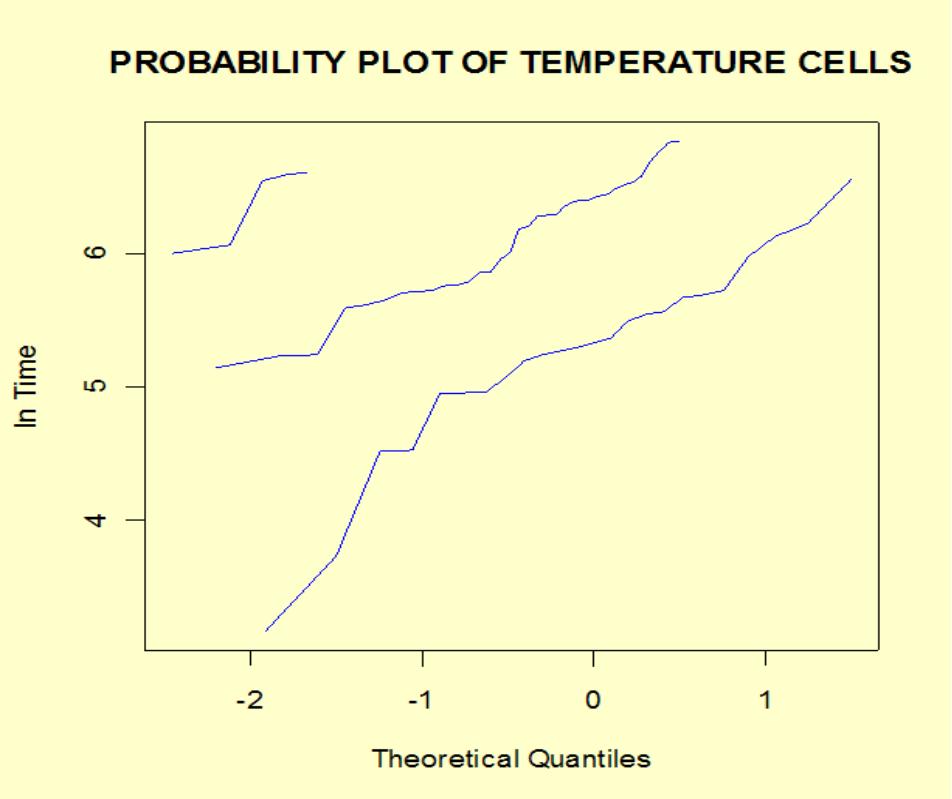
- Do plots for each cell and estimate T_{50} and sigma as previously discussed.
- Plot all the cells on the same graph and check whether the lines are roughly parallel (a necessary consequence of true acceleration).
- If probability plots indicate that the lognormal model is appropriate and that sigma is consistant among cells, plot ln T_{50} versus 11605/(*t* °C + 273.16) for each cell, check for linearity and fit a straight line through the points. Since the points have different values of precision, due to different numbers of failures in each cell, it is recommended that the number of failures in each cell be used as weights in a regression when fitting a line through the points.
- Use the slope of the line as the ΔH estimate and calculate the Arrhenius *A* constant from the intercept using $A = e^{intercept}$.
- Estimate the common sigma across all the cells by the weighted average of the individual cell sigma estimates. Use the number of

failures in a cell divided by the total number of failures in all cells as that cell's weight. This will allow cells with more failures to play a bigger role in the estimation process.

Solution for Arrhenius model example

Analysis of Multicell Arrhenius Model Data:

The following lognormal probability plot was generated for our data so that all three stress cells are plotted on the same graph.



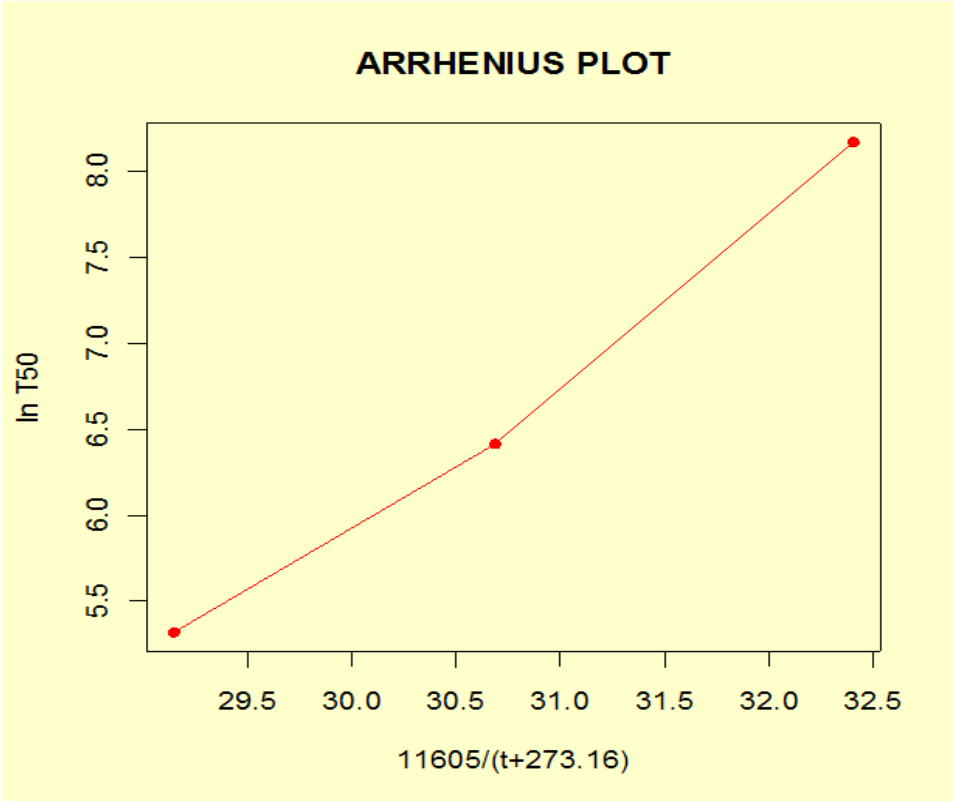
Note that the lines are somewhat straight (a check on the lognormal model) and the slopes are approximately parallel (a check on the acceleration assumption).

The cell ln T_{50} and sigma estimates are obtained from linear regression fits for each cell using the data from the probability plot. Each fit will yield a cell A_o, the ln T_{50} estimate, and A_1, the cell sigma estimate. These are summarized in the table below.

Summary of Least Squares Estimation of Cell Lognormal Parameters

Cell Number	ln T_{50}	Sigma
1 (t °C = 85)	8.168	.908
2 (t °C = 105)	6.415	.663
3 (t °C = 125)	5.319	.805

The three cells have 11605/(t °C + 273.16) values of 32.40, 30.69 and 29.15 respectively, in cell number order. The Arrhenius plot is



With only three cells, it is unlikely a straight line through the points will present obvious visual lack of fit. However, in this case, the points appear to line up very well.

Finally, the model coefficients are computed from a weighted linear fit of ln T_{50} versus 11605/(t °C + 273.16), using weights of 5, 35, and 24 for each cell. This will yield a ln A estimate of -18.312 ($A = e^{-18.312} = 0.1115 \times 10^{-7}$) and a ΔH estimate of 0.808. With this value of ΔH, the acceleration between the lowest stress cell of 85 °C and the highest of 125 °C is

$$\exp\left\{.808 \times 11605 \times \left(\frac{1}{358.16} - \frac{1}{398.16}\right)\right\} = 13.9$$

which is almost 14× acceleration. Acceleration from 125 °C to the use condition of 25 °C is 3708×. The use T_{50} is $e^{-18.312}$ x $e^{0.808 \times 11605 \times 1/298.16} = e^{13.137} = 507383$.

A single sigma estimate for all stress conditions can be calculated as a weighted average of the three sigma estimates obtained from the experimental cells. The weighted average is (5/64) × 0.908 + (35/64) × 0.663 + (24/64) × 0.805 = 0.74.

The analyses in this section can can be implemented using both Dataplot code and R code.

Fitting More Complicated models

Models Two stress models, such as the temperature/voltage model given by

involving several stresses can be fit using multiple regression

$$t_f = Ae^{\frac{\Delta H}{kT}} V^{\beta}$$

need at least four or five carefully chosen stress cells to estimate all the parameters. The Backwards L design previously described is an example of a design for this model. The bottom row of the "backward L" could be used for a plot testing the Arrhenius temperature dependence, similar to the above Arrhenius example. The right hand column could be plotted using $y = \ln T_{50}$ and $x = \ln V$, to check the voltage term in the model.

The overall model estimates should be obtained from fitting the multiple regression model

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

with

$$Y = \ln T_{50},\ b_0 = \ln A$$

$$b_1 = \Delta H,\ X_1 = 1/kT$$

$$b_2 = \beta,\ \text{and}\ x_2 = \ln V$$

Fitting this model, after setting up the Y, X1 = X_1, X2 = X_2 data vectors, provides estimates for b_0, b_1 and b_2.

Three stress models, and even Eyring models with interaction terms, can be fit by a direct extension of these methods. Graphical plots to test the model, however, are less likely to be meaningful as the model becomes more complex.

8.4.2.2. Maximum likelihood

The maximum likelihood method can be used to estimate distribution and acceleration model parameters at the same time

The likelihood equation for a multi-cell acceleration model utilizes the likelihood function for each cell, as described in section 8.4.1.2. Each cell will have unknown life distribution parameters that, in general, are different. For example, if a lognormal model is used, each cell might have its own T_{50} and sigma.

Under an acceleration assumption, however, all the cells contain samples from populations that have the same value of sigma (the slope does not change for different stress cells). Also, the T_{50} values are related to one another by the acceleration model; they all can be written using the acceleration model equation that includes the proper cell stresses.

To form the likelihood equation under the acceleration model assumption, simply rewrite each cell likelihood by replacing each cell T_{50} with its acceleration model equation equivalent and replacing each cell sigma with the same overall sigma. Then, multiply all these modified cell likelihoods together to obtain the overall likelihood equation.

Once the overall likelihood equation has been created, the maximum likelihood estimates (MLE) of sigma and the acceleration model parameters are the values that maximize this likelihood. In most cases, these values are obtained by setting partial derivatives of the log likelihood to zero and solving the resulting (non-linear) set of equations.

The method is complicated and requires specialized software

As you can see, the procedure is complicated, computationally intensive, and is only practical if appropriate software is available. MLE does have many desirable features.

- The method can, in theory at least, be used for any distribution model and acceleration model and type of censored data.
- Estimates have "optimal" statistical properties as sample sizes (i.e., numbers of failures) become large.
- Approximate confidence bounds can be calculated.
- Statistical tests of key assumptions can be made using the likelihood ratio test. Some common tests are:
    - the life distribution model versus another simpler model with fewer parameters (i.e., a 3-parameter Weibull versus a 2-parameter Weibull, or a 2-parameter Weibull versus an exponential),
    - the constant slope from cell to cell requirement of typical acceleration models, and
    - the fit of a particular acceleration model.

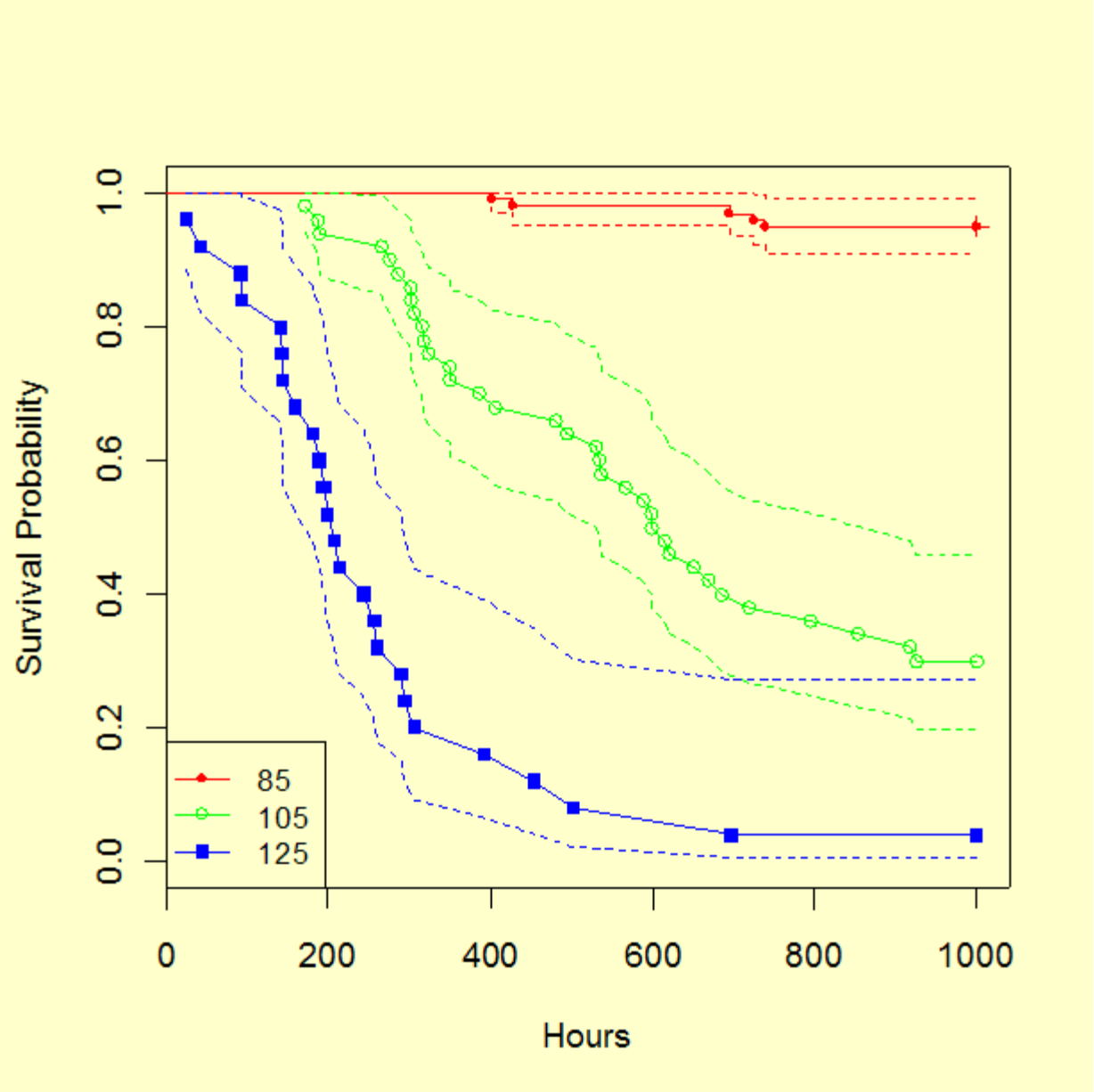
In general, the recommendations made when comparing methods of estimating life distribution model parameters also apply here. Software incorporating acceleration model analysis capability, while rare just a few years ago, is now readily available and many companies and universities have developed their own proprietary versions.

Steps For Fitting The Arrhenius Model

Use MLE to fit an Arrhenius model to example data

Data from the Arrhenius example given in section 8.4.2.1 were analyzed using MLE. The analyses in this section can can be implemented using R code.

1. We generate survival curves for each cell. All plots and estimates are based on individual cell data, without the Arrhenius model assumption.



2. The results of lognormal survival regression modeling for the three data cells are shown below.

Cell 1 - 85 °C

Parameter	Estimate	Stan. Dev	*z* Value
Intercept	8.891	0.890	9.991
ln(scale)	0.192	0.406	0.473

sigma = exp(ln(scale)) = 1.21
ln likelihood = -53.4

Cell 2 - 105 °C

Parameter	Estimate	Stan. Dev	*z* Value

```
---------   --------   ---------     -------
Intercept     6.470      0.108        60.14
ln(scale)    -0.336      0.129        -2.60

sigma = exp(ln(scale)) = 0.715
ln likelihood = -265.2
```

```
Cell 3 - 125 °C
  Parameter   Estimate   Stan. Dev     z Value
  ---------   --------   ---------     -------
  Intercept      5.33      0.163        32.82
  ln(scale)     -0.21      0.146        -1.44

  sigma = exp(ln(scale)) = 0.81
  ln likelihood = -156.5
```

The cell ln likelihood values are -53.4, -265.2 and -156.5, respectively. Adding them together yields a total ln likelihood of -475.1 for all the data fit with separate lognormal parameters for each cell (no Arrhenius model assumption).

3. Fit the Arrhenius model to all data using MLE.

```
Parameter   Estimate   Stan. Dev     z Value
---------   --------   ---------     -------
Intercept   -19.906      2.3204       -8.58
l/kT          0.863      0.0761       11.34
ln(scale)    -0.259      0.0928       -2.79

sigma = exp(ln(scale))Scale = 0.772
ln likelihood = -476.7
```

4. The likelihood ratio test statistic for the Arrhenius model fit (which also incorporates the single sigma acceleration assumption) is -2ln λ, where λ denotes the ratio of the likelihood values with (L_0), and without (L_1) the Arrhenius model assumption so that

$$-2\ln \lambda = -2\ln (L_0/L_1) = -2(\ln L_0 - \ln L_1).$$

Using the results from steps 2 and 3, we have -2ln λ = -2(-476.7 - (-475.1)) = 3.2. The degrees of freedom for the Chi-Square test statistic is 6 - 3 = 3, since six parameters were reduced to three under the acceleration model assumption. The chance of obtaining a value 3.2 or higher is 36.3% for a Chi-Square distribution with 3 degrees of freedom, which indicates an acceptable model (no significant lack of fit).

This completes the Arrhenius model analysis of the three cells of data. If different cells of data have different voltages, then a new variable "ln *V*" could be added as an effect to fit the Inverse Power Law voltage model. In fact, several effects can be included at once if more than one stress varies across cells. Cross product stress terms could also be included by adding these columns to the spreadsheet and adding them in the model as additional "effects".

Example Comparing Graphical Estimates and MLE

Arrhenius example comparing graphical and MLE method results

The results from the three-stress-cell Arrhenius example using graphical and MLE methods for estimating parameters are shown in the table below.

	Graphical Estimates		**MLE**	
	ln T_{50}	Sigma	ln T_{50}	Sigma
Cell 1	8.17	0.91	8.89	1.21
Cell 2	6.42	0.66	6.47	0.71

Cell 3	5.32	0.81	5.33	0.81

Acceleration Model Overall Estimates

	ΔH	Sigma	ln *A*
Graphical	0.808	0.74	-18.312
MLE	0.863	0.77	-19.91

Note that when there are a lot of failures and little censoring, the two methods are in fairly close agreement. Both methods are also in close agreement on the Arrhenius model results. However, even small differences can be important when projecting reliability numbers at use conditions. In this example, the CDF at 25 °C and 100,000 hours projects to 0.014 using the graphical estimates and only 0.003 using the MLE.

MLE method tests models and gives confidence intervals

The maximum likelihood method allows us to test whether parallel lines (a single sigma) are reasonable and whether the Arrhenius model is acceptable. The likelihood ratio tests for the three example data cells indicated that a single sigma and the Arrhenius model are appropriate. In addition, we can compute confidence intervals for all estimated parameters based on the MLE results.

8. Assessing Product Reliability
8.4. Reliability Data Analysis
8.4.2. How do you fit an acceleration model?

8.4.2.3. Fitting models using degradation data instead of failures

If you can fit models using degradation data, you don't need actual test failures

When failure can be related directly to a change over time in a measurable product parameter, it opens up the possibility of measuring degradation over time and using that data to extrapolate when failure will occur. That allows us to fit acceleration models and life distribution models without actually waiting for failures to occur.

This overview of degradation modeling assumes you have chosen a life distribution model and an acceleration model and offers an alternative to the accelerated testing methodology based on failure data, previously described. The following topics are covered.

- Common assumptions
- Advantages
- Drawbacks
- A simple method
- A more accurate approach for a special case
- Example

More details can be found in Nelson (1990, pages 521-544) or Tobias and Trindade (1995, pages 197-203).

Common Assumptions When Modeling Degradation Data

You need a measurable parameter that drifts (degrades) linearly to a critical failure value

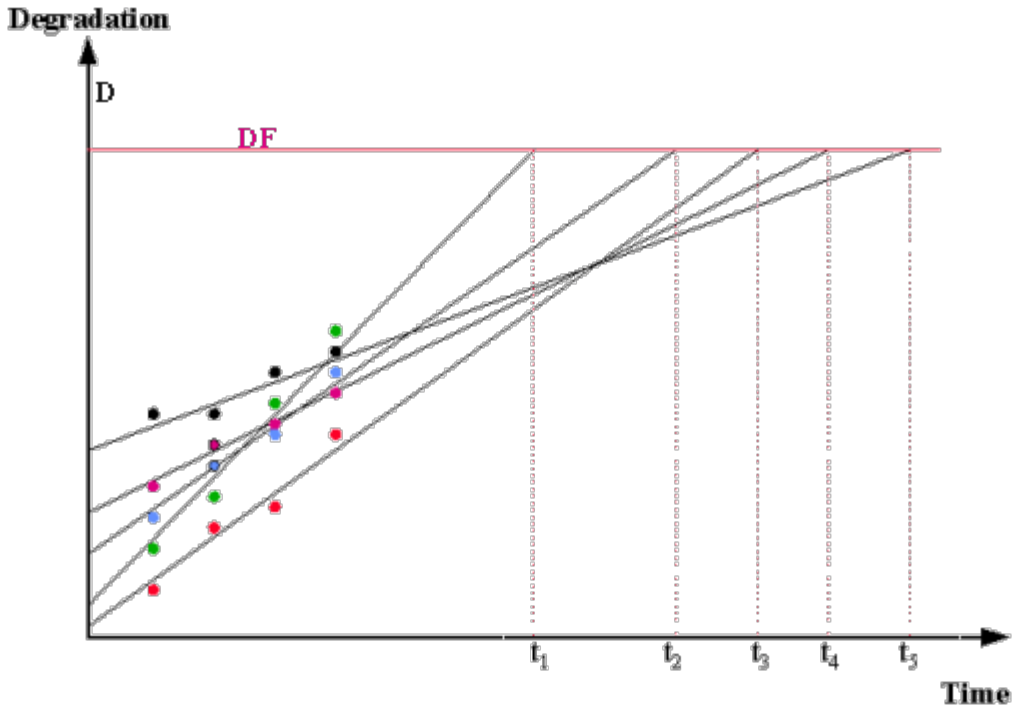
Two common assumptions typically made when degradation data are modeled are the following:

1. A parameter D, that can be measured over time, drifts monotonically (upwards, or downwards) towards a specified critical value DF. When it reaches DF, failure occurs.
2. The drift, measured in terms of D, is linear over time with a slope (or **rate of degradation**) R, that depends on the relevant stress the unit is operating under and also the (random) characteristics of the unit being measured. Note: It may be necessary to define D as a transformation of some standard parameter in order to obtain linearity - logarithms or powers are sometimes needed.

The figure below illustrates these assumptions by showing degradation plots of five units on test. Degradation readings for each unit are taken at the same four time points and straight lines fit through these readings on a unit-by-unit basis. These lines are then extended up to a critical (failure) degradation value. The projected times of failure for these units are then read off the plot. The are: t , t ,

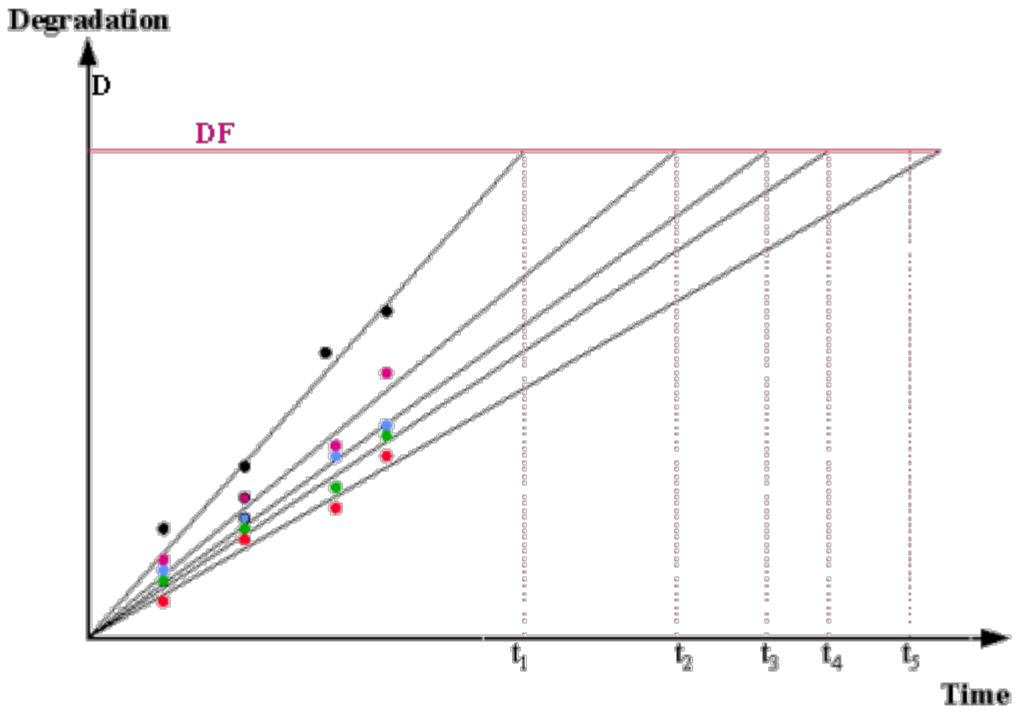
$...,t_5$.

Plot of linear degradation trends for five units read out at four time points



In many practical situations, D starts at 0 at time zero, and all the linear theoretical degradation lines start at the origin. This is the case when D is a "% change" parameter, or failure is defined as a change of a specified magnitude in a parameter, regardless of its starting value. Lines all starting at the origin simplify the analysis since we don't have to characterize the population starting value for D, and the "distance" any unit "travels" to reach failure is always the constant DF. For these situations, the degradation lines would look as follows.

Often, the degradation lines go through the origin - as when % change is the measurable parameter increasing to a failure level



It is also common to assume the effect of measurement error, when reading values of D, has relatively little impact on the accuracy of model estimates.

Advantages of Modeling Based on Degradation Data

Modeling based on complete samples of measurement data, even with low stress cells, offers many advantages

1. Every degradation readout for every test unit contributes a data point. This leads to large amounts of useful data, even if there are very few failures.
2. You don't have to run tests long enough to obtain significant numbers of failures.
3. You can run low stress cells that are much closer to use conditions and obtain meaningful degradation data. The same cells would be a waste of time to run if failures were needed for modeling. Since these cells are more typical of use conditions, it makes sense to have them influence model parameters.
4. Simple plots of degradation vs time can be used to visually test the linear degradation assumption.

Drawbacks to Modeling Based on Degradation Data

Degradation may not proceed in a smooth, linear fashion towards what the customer calls "failure"

1. For many failure mechanisms, it is difficult or impossible to find a measurable parameter that degrades to a critical value in such a way that reaching that critical value is equivalent to what the customer calls a failure.
2. Degradation trends may vary erratically from unit to unit, with no apparent way to transform them into linear trends.
3. Sometimes degradation trends are reversible and a few units appear to "heal themselves" or get better. This kind of behavior does not follow typical assumptions and is difficult to model.
4. Measurement error may be significant and overwhelm small degradation trends, especially at low stresses.
5. Even when degradation trends behave according to assumptions and the chosen models fit well, the final results may not be consistent with an analysis based on actual failure data. This probably means that the failure mechanism depends on more than a simple continuous degradation process.

Because of the last listed drawback, it is a good idea to have at least one high-stress cell where enough real failures occur to do a standard life distribution model analysis. The parameter estimates obtained can be compared to the predictions from the degradation data analysis, as a "reality" check.

A Simple Method For Modeling Degradation Data

A simple approach is to extend each unit's degradation line until a projected "failure time" is obtained

1. As shown in the figures above, fit a line through each unit's degradation readings. This can be done by hand, but using a least squares regression program is better.
2. Take the equation of the fitted line, substitute DF for Y and solve for X. This value of X is the "projected time of fail" for that unit.
3. Repeat for every unit in a stress cell until a complete sample of (projected) times of failure is obtained for the cell.
4. Use the failure times to compute life distribution parameter estimates for a cell. Under the fairly typical assumption of a lognormal model, this is very simple. Take natural logarithms of all failure times and treat the resulting data as a sample from a normal distribution. Compute the sample mean

and the sample standard deviation. These are estimates of ln T_{50} and σ, respectively, for the cell.

5. Assuming there are k cells with varying stress, fit an appropriate acceleration model using the cell ln T_{50} values, as described in the graphical estimation section. A single sigma estimate is obtained by taking the square root of the average of the cell σ^2 estimates (assuming the same number of units each cell). If the cells have n_j units on test, where the n_j values are not all equal, use the pooled sum-of-squares estimate across all k cells calculated by

$$\hat{\sigma}^2 = \frac{1}{\sum_{j=1}^{k}(n_j - 1)} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

A More Accurate Regression Approach For the Case When D = 0 at time 0 and the "Distance To Fail" DF is the Same for All Units

Models can be fit using all the degradation readings and linear regression

Let the degradation measurement for the i-th unit at the j-th readout time in the k-th stress cell be given by D_{ijk}, and let the corresponding readout time be denoted by t_{jk}. That readout gives a degradation rate (or slope) estimate of D_{ijk}/t_{jk}. This follows from the linear assumption or:

(Rate of degradation) × (Time on test) = (Amount of degradation)

Based on that readout alone, an estimate of the natural logarithm of the time to fail for that unit is

$$y_{ijk} = \ln \mathrm{DF} - (\ln D_{ijk} - \ln t_{jk}).$$

This follows from the basic formula connecting linear degradation with failure time

(rate of degradation) × (time of failure) = DF

by solving for (time of failure) and taking natural logarithms.

For an Arrhenius model analysis, with

$$t_f = Ae^{\Delta H/kT},$$
$$y_{ijk} = a + bx_k$$

with the x_k values equal to $1/KT$. Here T is the temperature of the k-th cell, measured in Kelvin (273.16 + degrees Celsius) and K is Boltzmann's constant (8.617×10^{-5} in eV/ unit Kelvin). Use a linear regression program to estimate $a = \ln A$ and $b = \Delta H$. If we further assume t_f has a lognormal distribution, the mean square residual error from the regression fit is an estimate of σ^2 (with σ the lognormal sigma).

One way to think about this model is as follows: each unit has a random rate R

of degradation. Since t_f = DF/R, it follows from a characterization property of the normal distribution that if t_f is lognormal, then R must also have a lognormal distribution (assuming DF and R are independent). After we take logarithms, ln R has a normal distribution with a mean determined by the acceleration model parameters. The randomness in R comes from the variability in physical characteristics from unit to unit, due to material and processing differences.

Note: The estimate of sigma based on this simple graphical approach might tend to be too large because it includes an adder due to the measurement error that occurs when making the degradation readouts. This is generally assumed to have only a small impact.

Example: Arrhenius Degradation Analysis

An example using the regression approach to fit an Arrhenius model

A component has a critical parameter that studies show degrades linearly over time at a rate that varies with operating temperature. A component failure based on this parameter occurs when the parameter value changes by 30% or more. Fifteen components were tested under 3 different temperature conditions (5 at 65 °C, 5 at 85 °C and the last 5 at 105 °C). Degradation percent values were read out at 200, 500 and 1000 hours. The readings are given by unit in the following three temperature cell tables.

65 °C

	200 hr	500 hr	1000 hr
Unit 1	0.87	1.48	2.81
Unit 2	0.33	0.96	2.13
Unit 3	0.94	2.91	5.67
Unit 4	0.72	1.98	4.28
Unit 5	0.66	0.99	2.14

85 °C

	200 hr	500 hr	1000 hr
Unit 1	1.41	2.47	5.71
Unit 2	3.61	8.99	17.69
Unit 3	2.13	5.72	11.54
Unit 4	4.36	9.82	19.55
Unit 5	6.91	17.37	34.84

105 °C

	200 hr	500 hr	1000 hr
Unit 1	24.58	62.02	124.10
Unit 2	9.73	24.07	48.06
Unit 3	4.74	11.53	23.72
Unit 4	23.61	58.21	117.20
Unit 5	10.90	27.85	54.97

Note that one unit failed in the 85 °C cell and four units failed in the 105 °C cell. Because there were so few failures, it would be impossible to fit a life distribution model in any cell but the 105 °C cell, and therefore no acceleration model can be fit using failure data. We will fit an Arrhenius/lognormal model, using the degradation data.

Solution:

Fit the model to the degradation data

From the above tables, first create a variable (DEG) with 45 degradation values starting with the first row in the first table and proceeding to the last row in the last table. Next, create a temperature variable (TEMP) that has 15 repetitions of 65, followed by 15 repetitions of 85 and then 15 repetitions of 105. Finally, create a time variable (TIME) that corresponds to readout times.

Fit the Arrhenius/lognormal equation, $y_{ijk} = a + b\,x_{ijk}$, where

$$y_{ijk} = \ln(30) - (\ln(\text{DEG}) - \ln(\text{TIME}))$$

and

$$x_{ijk} = 100000 / [8.617*(\text{TEMP} + 273.16)].$$

The linear regression results are the following.

Parameter	Estimate	Stan. Dev	*t* Value
a	-18.94337	1.83343	-10.33
b	0.81877	0.05641	14.52

Residual standard deviation = 0.5611
Residual degrees of freedom = 45

The Arrhenius model parameter estimates are: $\ln A = -18.94$; $\Delta H = 0.82$. An estimate of the lognormal sigma is $\sigma = 0.56$.

The analyses in this section can can be implemented using both Dataplot code and R code.

8.4.3. How do you project reliability at use conditions?

When projecting from high stress to use conditions, having a correct acceleration model and life distribution model is critical

General Considerations

Reliability projections based on failure data from high stress tests are based on assuming we know the correct acceleration model for the failure mechanism under investigation and we are also using the correct life distribution model. This is because we are extrapolating "backwards" - trying to describe failure behavior in the early tail of the life distribution, where we have little or no actual data.

For example, with an acceleration factor of 5000 (and some are much larger than this), the first 100,000 hours of use life is "over" by 20 hours into the test. Most, or all, of the test failures typically come later in time and are used to fit a life distribution model with only the first 20 hours or less being of practical use. Many distributions may be flexible enough to adequately fit the data at the percentiles where the points are, and yet differ from the data by orders of magnitude in the very early percentiles (sometimes referred to as the early "tail" of the distribution).

However, it is frequently necessary to test at high stress (to obtain any failures at all!) and project backwards to use. When doing this bear in mind two important points:

Project for each failure mechanism separately

- Distribution models, and especially acceleration models, should be applied only to a single failure mechanism at a time. Separate out failure mechanisms when doing the data analysis and use the competing risk model to build up to a total component failure rate
- Try to find theoretical justification for the chosen models, or at least a successful history of their use for the same or very similar mechanisms. (Choosing models solely based on empirical fit is like extrapolating from quicksand to a mirage.)

How to Project from High Stress to Use Stress

Two types of use-condition reliability projections are common:

1. Projection to use conditions after completing a multiple

stress cell experiment and successfully fitting both a life distribution model and an acceleration model
2. Projection to use conditions after a single cell at high stress is run as a line reliability monitor.

Arrhenius model projection example

The Arrhenius example from the graphical estimation and the MLE estimation sections ended by comparing use projections of the CDF at 100,000 hours. This is a projection of the first type. We know from the Arrhenius model assumption that the T_{50} at 25 °C is just

$$Ae^{\Delta H / k(25+273.16)}$$

Using the graphical model estimates for ln A and we have

$$T_{50} \text{ at use} = e^{-18.312} \times e^{0.808 \times 11605/298.16}$$
$$= e^{13.137} = 507383$$

and combining this T_{50} with the estimate of the common sigma of 0.74 allows us to easily estimate the CDF or failure rate after any number of hours of operation at use conditions.

In particular, the CDF value of a lognormal at T/T_{50} (where time T = 100,000, T_{50} = 507383, and sigma = 0.74) is 0.014, which matches the answer given in the MLE estimation section as the graphical projection of the CDF at 100,000 hours at a use temperature of 25 °C.

If the life distribution model had been Weibull, the same type of analysis would be performed by letting the characteristic life parameter α vary with stress according to the acceleration model, while the shape parameter γ is constant for all stress conditions.

The second type of use projection was used in the section on lognormal and Weibull tests, in which we judged new lots of product by looking at the proportion of failures in a sample tested at high stress. The assumptions we made were:

- we knew the acceleration factor between use and high stress
- the shape parameter (sigma for the lognormal, gamma for the Weibull) is also known and does not change significantly from lot to lot.

With these assumptions, we can take any proportion of failures we see from a high stress test and project a use CDF or failure rate. For a T-hour high stress test and an acceleration factor of A from high stress to use stress, an observed proportion p is converted to a use CDF at 100,000 hours for a lognormal model using:

$$T_{50Stress} = \mathrm{T}\times G^{-1}(p, 0, \sigma)$$
$$\mathrm{CDF} = G((100000/(A\times T_{50Stress})), 0, \sigma).$$

where $G(q, \mu, \sigma)$ is the lognormal distribution function with mean μ and standard deviation σ.

If the model is Weibull, we can find the use CDF or failure rate with:

$$A_{Stress} = \mathrm{T}\times W^{-1}(p, \gamma, 1)$$
$$\mathrm{CDF} = W((100000/(A\times A_{Stress})), \gamma, 1).$$

where $W(q, \gamma, \alpha)$ is the Weibull distribution function with shape parameter γ and scale parameter α.

The analyses in this section can can be implemented using both Dataplot code and R code.

8.4.4. How do you compare reliability between two or more populations?

Several methods for comparing reliability between populations are described

Comparing reliability among populations based on samples of failure data usually means asking whether the samples came from populations with the same reliability function (or CDF). Three techniques already described can be used to answer this question for censored reliability data. These are:

- Comparing sample proportion failures
- Likelihood ratio test comparisons
- Lifetime regression comparisons

Comparing Sample Proportion Failures

Assume each sample is a random sample from possibly a different lot, vendor or production plant. All the samples are tested under the same conditions. Each has an observed proportion of failures on test. Call these sample proportions of failures p_1, p_2, p_3, ...p_n. Could these all have come from equivalent populations?

This is a question covered in Chapter 7 for two populations, and for more than two populations, and the techniques described there apply equally well here.

Likelihood Ratio Test Comparisons

The Likelihood Ratio test was described earlier. In this application, the Likelihood ratio λ has as a denominator the product of all the Likelihoods of all the samples assuming each population has its own unique set of parameters. The numerator is the product of the Likelihoods assuming the parameters are exactly the same for each population. The test looks at whether -2lnλ is unusually large, in which case it is unlikely the populations have the same parameters (or reliability functions).

This procedure is very effective if, and only if, it is built into the analysis software package being used and this software covers the models and situations of interest to the analyst.

Lifetime Regression Comparisons

Lifetime regression is similar to maximum likelihood and

likelihood ratio test methods. Each sample is assumed to have come from a population with the same shape parameter and a wide range of questions about the scale parameter (which is often assumed to be a "measure" of lot-to-lot or vendor-to-vendor quality) can be formulated and tested for significance.

For a complicated, but realistic example, assume a company manufactures memory chips and can use chips with some known defects ("partial goods") in many applications. However, there is a question of whether the reliability of "partial good" chips is equivalent to "all good" chips. There exists lots of customer reliability data to answer this question. However the data are difficult to analyze because they contain several different vintages with known reliability differences as well as chips manufactured at many different locations. How can the partial good vs all good question be resolved?

A lifetime regression model can be constructed with variables included that change the scale parameter based on vintage, location, partial versus all good, and any other relevant variables. Then, a good lifetime regression program will sort out which, if any, of these factors are significant and, in particular, whether there is a significant difference between "partial good" and "all good".

8. Assessing Product Reliability
8.4. Reliability Data Analysis

8.4.5. How do you fit system repair rate models?

Fitting models discussed earlier

This subsection describes how to fit system repair rate models when you have actual failure data. The data could come from from observing a system in normal operation or from running tests such as Reliability Improvement tests.

The three models covered are the constant repair rate (HPP/exponential) model, the power law (Duane) model and the exponential law model.

8.4.5.1. Constant repair rate (HPP/exponential) model

This section covers estimating MTBF's and calculating upper and lower confidence bounds

The HPP or exponential model is widely used for two reasons:

- Most systems spend most of their useful lifetimes operating in the flat constant repair rate portion of the bathtub curve
- It is easy to plan tests, estimate the MTBF and calculate confidence intervals when assuming the exponential model.

This section covers the following:

1. **Estimating the MTBF (or repair rate/failure rate)**
2. **How to use the MTBF confidence interval factors**
3. **Tables of MTBF confidence interval factors**
4. **Confidence interval equation and "zero fails" case**
5. **Calculation of confidence intervals**
6. **Example**

Estimating the MTBF (or repair rate/failure rate)

For the HPP system model, as well as for the non repairable exponential population model, there is only one unknown parameter λ (or equivalently, the MTBF = $1/\lambda$). The method used for estimation is the same for the HPP model and for the exponential population model.

The best estimate of the MTBF is just "Total Time" divided by "Total Failures"

The estimate of the MTBF is

$$\hat{MTBF} = \frac{Total\ System(s)\ operation\ time}{Total\ number\ of\ failures}$$

$$\hat{\lambda} = \frac{1}{\hat{MTBF}} = \frac{Total\ number\ of\ failures}{Total\ System(s)\ (or\ units)\ operation\ time}$$

This estimate is the maximum likelihood estimate whether the data are censored or complete, or from a repairable system or a non-repairable population.

Confidence Interval Factors multiply the estimated

How To Use the MTBF Confidence Interval Factors

1. Estimate the MTBF by the standard estimate (total unit test hours divided by total failures)
2. Pick a confidence level (i.e., pick 100x(1-α)). For 95%, α = .05; for 90%, α = .1; for 80%, α = .2 and for 60%, α = .4

MTBF to obtain lower and upper bounds on the true MTBF

3. Read off a lower and an upper factor from the confidence interval tables for the given confidence level and number of failures *r*
4. Multiply the MTBF estimate by the lower and upper factors to obtain $MTBF_{lower}$ and $MTBF_{upper}$
5. When *r* (the number of failures) = 0, multiply the total unit test hours by the "0 row" lower factor to obtain a 100 × (1-α/2)% one-sided lower bound for the MTBF. There is no upper bound when *r* = 0.
6. Use ($MTBF_{lower}$, $MTBF_{upper}$) as a 100×(1-α)% confidence interval for the MTBF λ (*r* > 0)
7. Use $MTBF_{lower}$ as a (one-sided) lower 100×(1-α/2)% limit for the MTBF
8. Use $MTBF_{upper}$ as a (one-sided) upper 100×(1-α/2)% limit for the MTBF
9. Use ($1/MTBF_{upper}$, $1/MTBF_{lower}$) as a 100×(1-α)% confidence interval for λ
10. Use $1/MTBF_{upper}$ as a (one-sided) lower 100×(1-α/2)% limit for λ
11. Use $1/MTBF_{lower}$ as a (one-sided) upper 100×(1-α/2)% limit for λ

Tables of MTBF Confidence Interval Factors

Confidence bound factor tables for 60, 80, 90 and 95% confidence

Confidence Interval Factors to Multiply MTBF Estimate

	60%		80%	
Num Fails r	**Lower for MTBF**	**Upper for MTBF**	**Lower for MTBF**	**Upper for MTBF**
0	0.6213	-	0.4343	-
1	0.3340	4.4814	0.2571	9.4912
2	0.4674	2.4260	0.3758	3.7607
3	0.5440	1.9543	0.4490	2.7222
4	0.5952	1.7416	0.5004	2.2926
5	0.6324	1.6184	0.5391	2.0554
6	0.6611	1.5370	0.5697	1.9036
7	0.6841	1.4788	0.5947	1.7974
8	0.7030	1.4347	0.6156	1.7182
9	0.7189	1.4000	0.6335	1.6567
10	0.7326	1.3719	0.6491	1.6074
11	0.7444	1.3485	0.6627	1.5668
12	0.7548	1.3288	0.6749	1.5327
13	0.7641	1.3118	0.6857	1.5036
14	0.7724	1.2970	0.6955	1.4784
15	0.7799	1.2840	0.7045	1.4564
20	0.8088	1.2367	0.7395	1.3769
25	0.8288	1.2063	0.7643	1.3267
30	0.8436	1.1848	0.7830	1.2915
35	0.8552	1.1687	0.7978	1.2652

40	0.8645	1.1560	0.8099	1.2446
45	0.8722	1.1456	0.8200	1.2280
50	0.8788	1.1371	0.8286	1.2142
75	0.9012	1.1090	0.8585	1.1694
100	0.9145	1.0929	0.8766	1.1439
500	0.9614	1.0401	0.9436	1.0603

Confidence Interval Factors to Multiply MTBF Estimate

	90%		95%	
Num Fails	**Lower for MTBF**	**Upper for MTBF**	**Lower for MTBF**	**Upper for MTBF**
0	0.3338	-	0.2711	-
1	0.2108	19.4958	0.1795	39.4978
2	0.3177	5.6281	0.2768	8.2573
3	0.3869	3.6689	0.3422	4.8491
4	0.4370	2.9276	0.3906	3.6702
5	0.4756	2.5379	0.4285	3.0798
6	0.5067	2.2962	0.4594	2.7249
7	0.5324	2.1307	0.4853	2.4872
8	0.5542	2.0096	0.5075	2.3163
9	0.5731	1.9168	0.5268	2.1869
10	0.5895	1.8432	0.5438	2.0853
11	0.6041	1.7831	0.5589	2.0032
12	0.6172	1.7330	0.5725	1.9353
13	0.6290	1.6906	0.5848	1.8781
14	0.6397	1.6541	0.5960	1.8291
15	0.6494	1.6223	0.6063	1.7867
20	0.6882	1.5089	0.6475	1.6371
25	0.7160	1.4383	0.6774	1.5452
30	0.7373	1.3893	0.7005	1.4822
35	0.7542	1.3529	0.7190	1.4357
40	0.7682	1.3247	0.7344	1.3997
45	0.7800	1.3020	0.7473	1.3710
50	0.7901	1.2832	0.7585	1.3473
75	0.8252	1.2226	0.7978	1.2714
100	0.8469	1.1885	0.8222	1.2290
500	0.9287	1.0781	0.9161	1.0938

Confidence Interval Equation and "Zero Fails" Case

Formulas for confidence bound factors -

Confidence bounds for the typical Type I censoring situation are obtained from chi-square distribution tables or programs. The formula for calculating confidence intervals is:

even for "zero fails" case

$$P\left[\frac{MTBF \cdot 2r}{\chi^2_{1-\alpha/2,2(r+1)}} \leq True\ MTBF \leq \frac{MTBF \cdot 2r}{\chi^2_{\alpha/2,2r}}\right] \geq 1-\alpha$$

In this formula, $X^2_{\alpha/2,2r}$ is a value that the chi-square statistic with $2r$ degrees of freedom is less than with probability $\alpha/2$. In other words, the left-hand tail of the distribution has probability $\alpha/2$. An even simpler version of this formula can be written using T = the total unit test time:

$$P\left[\frac{2T}{\chi^2_{1-\alpha/2,2(r+1)}} \leq True\ MTBF \leq \frac{2T}{\chi^2_{\alpha/2,2r}}\right] \geq 1-\alpha$$

These bounds are exact for the case of one or more repairable systems on test for a fixed time. They are also exact when non repairable units are on test for a fixed time and failures are replaced with new units during the course of the test. For other situations, they are approximate.

When there are zero failures during the test or operation time, only a (one-sided) MTBF lower bound exists, and this is given by

$$MTBF_{lower} = T/(-\ln \alpha)$$

The interpretation of this bound is the following: if the true MTBF were any lower than $MTBF_{lower}$, we would have seen at least one failure during T hours of test with probability at least $1-\alpha$. Therefore, we are $100(1-\alpha)$ % confident that the true MTBF is not lower than $MTBF_{lower}$.

Calculation of confidence limits

A one-sided, lower $100(1-\alpha/2)$ % confidence bound for the MTBF is given by

$$LOWER = 2T/G^{-1}(1-\alpha/2, [2(r+1)])$$

where T is the total unit or system test time, r is the total number of failures, and $G(q,v)$ is the X^2 distribution function with shape parameter v.

A one-sided, upper $100(1-\alpha/2)$ % confidence bound for the MTBF is given by

$$UPPER = 2T/G^{-1}(\alpha/2, [2r])$$

The two intervals together, (LOWER, UPPER), are a $100(1-\alpha)$ % two-sided confidence interval for the true MTBF.

Please use caution when using CDF and inverse CDF functions in commercial software because some functions require left-tail probabilities and others require right-tail probabilities. In the left-tail case, $\alpha/2$ is used for the upper bound because 2T is being divided by the smaller percentile, and $1-\alpha/2$ is used for the lower bound because 2T is divided by the larger percentile. For the right-tail case, $1-\alpha/2$ is used to compute the upper bound

and $\alpha/2$ is used to compute the lower bound. Our formulas for $G^{-1}(q,v)$ assume the inverse CDF function requires left-tail probabilities.

Example

Example showing how to calculate confidence limits

A system was observed for two calendar months of operation, during which time it was in operation for 800 hours and had 2 failures.

The MTBF estimate is 800/2 = 400 hours. A 90 %, two-sided confidence interval is given by (400×0.3177, 400×5.6281) = (127, 2251). The same interval could have been obtained using

$$\text{LOWER} = 1600/G^{-1}(0.95,6)$$
$$\text{UPPER} = 1600/G^{-1}(0.05,4)$$

Note that 127 is a 95 % lower limit for the true MTBF. The customer is usually only concerned with the lower limit and one-sided lower limits are often used for statements of contractual requirements.

Zero fails confidence limit calculation

What could we have said if the system had no failures? For a 95 % lower confidence limit on the true MTBF, we either use the 0 failures factor from the 90 % confidence interval table and calculate 800 × 0.3338 = 267, or we use $T/(\ln \alpha)$ = 800/(ln 0.05) = 267.

The analyses in this section can can be implemented using both Dataplot code and R code.

8.4.5.2. Power law (Duane) model

The Power Law (Duane) model has been very successful in modeling industrial reliability improvement data

Brief Review of Power Law Model and Duane Plots

Recall that the Power Law is a NHPP with the expected number of fails, M(t), and the repair rate, M'(t) = m(t), given by:

$$M(t) = at^b, M'(t) = abt^{b-1} = \alpha t^{-\beta}$$

The parameter β = 1-b is called the Reliability Growth Slope and typical industry values for growth slopes during reliability improvement tests are in the .3 to .6 range.

If a system is observed for a fixed time of T hours and failures occur at times t_1, t_2, t_3, ..., t_r (with the start of the test or observation period being time 0), a Duane plot is a plot of (t_i / i) versus t_i on log-log graph paper. If the data are consistent with a Power Law model, the points in a Duane Plot will roughly follow a straight line with slope β and intercept (where t = 1 on the log-log paper) of $-\log_{10}a$.

MLE's for the Power Law model are given

Estimates for the Power Law Model

Computer aided graphical estimates can easily be obtained by doing a regression fit of Y = ln (t_i / i) vs X = ln t_i. The slope is the β estimate and $e^{-\text{intercept}}$ is the a estimate. The estimate of b is 1-β.

However, better estimates can easily be calculated. These are modified maximum likelihood estimates (corrected to eliminate bias). The formulas are given below for a fixed time of T hours, and r failures occurring at times t_1, t_2, t_3, ..., t_r.

$$\hat{\beta} = 1 - \frac{r-1}{\sum_{i=1}^{r} \ln\left(\frac{T}{t_i}\right)}, \quad \hat{a} = \frac{r}{T^{1-\hat{\beta}}}$$

$$\hat{b} = 1 - \hat{\beta} = \frac{r-1}{\sum_{i=1}^{r} \ln\left(\frac{T}{t_i}\right)}$$

The estimated MTBF at the end of the test (or observation) period is

$$\hat{\text{MTBF (AT END OF TEST)}} = \frac{T}{r(1-\hat{\beta})} = \frac{T}{r\hat{b}}$$

Approximate confidence bounds for the MTBF at end of test are given

Approximate Confidence Bounds for the MTBF at End of Test

We give an approximate 100(1-α) % confidence interval (M_L, M_U) for the MTBF at the end of the test. Note that M_L is a 100(1-α/2) % one-sided lower confidence bound and M_U is a 100(1-α/2) % one-sided upper confidence bound. The formulas are:

$$M_L = MTBF \frac{r(r-1)}{\left[r + \frac{z^2_{1-\alpha/2}}{4} + \sqrt{r\frac{z^2_{1-\alpha/2}}{2} + \frac{z^4_{1-\alpha/2}}{16}}\right]^2}$$

$$M_U = MTBF \frac{r(r-1)}{\left(r - z_{1-\alpha/2}\sqrt{\frac{r}{2}}\right)^2}$$

where $z_{1-\alpha/2}$ is the 100(1-α/2) percentile point of the standard normal distribution.

Case Study 1: Reliability Improvement Test Data Continued

Fitting the power law model to case study 1 failure data

This case study was introduced in section 2, where we did various plots of the data, including a Duane Plot. The case study was continued when we discussed trend tests and verified that significant improvement had taken place. Now we will complete the case study data analysis.

The observed failure times were: 5, 40, 43, 175, 389, 712, 747, 795, 1299 and 1478 hours, with the test ending at 1500 hours. We estimate β, a, and the MTBF at the end of test, along with a 100(1-α) % confidence interval for the true MTBF at the end of test (assuming, of course, that the Power Law model holds). The parameters and confidence intervals for the power law model were

estimated to be the following.

Estimate of β = 0.5165
Estimate of a = 0.2913
Estimate of MTBF at the end of the test = 310.234
80 % two-sided confidence interval:
(157.7139 , 548.5565)
90 % one-sided lower confidence limit = 157.7139

The analyses in this section can can be implemented using R code.

8.4.5.3. Exponential law model

Estimates of the parameters of the Exponential Law model can be obtained from either a graphical procedure or maximum likelihood estimation

Recall from section 1 that the Exponential Law refers to a NHPP process with repair rate M'(t) = m(t) = $e^{\alpha+\beta t}$. This model has not been used nearly as much in industrial applications as the Power Law model, and it is more difficult to analyze. Only a brief description will be given here.

Since the expected number of failures is given by M(t) = $(1/\beta)e^{\alpha+\beta t}$ and ln M(t) = $-\alpha\ln\beta+\beta t$, a plot of the cum fails versus time of failure on a log-linear scale should roughly follow a straight line with slope β. Doing a regression fit of y = ln cum fails versus x = time of failure will provide estimates of the slope β and the intercept $-\alpha\ln\beta$.

Alternatively, maximum likelihood estimates can be obtained from the following pair of equations:

$$\sum_{i=1}^{r} t_i + \frac{r}{\beta} - \frac{rT}{1-e^{-\beta T}} = 0$$

$$\alpha = \ln\left(\frac{r\beta}{e^{-\beta T}-1}\right)$$

The first equation is non-linear and must be solved iteratively to obtain the maximum likelihood estimate for β. Then, this estimate is substituted into the second equation to solve for the maximum likelihood estimate for α.

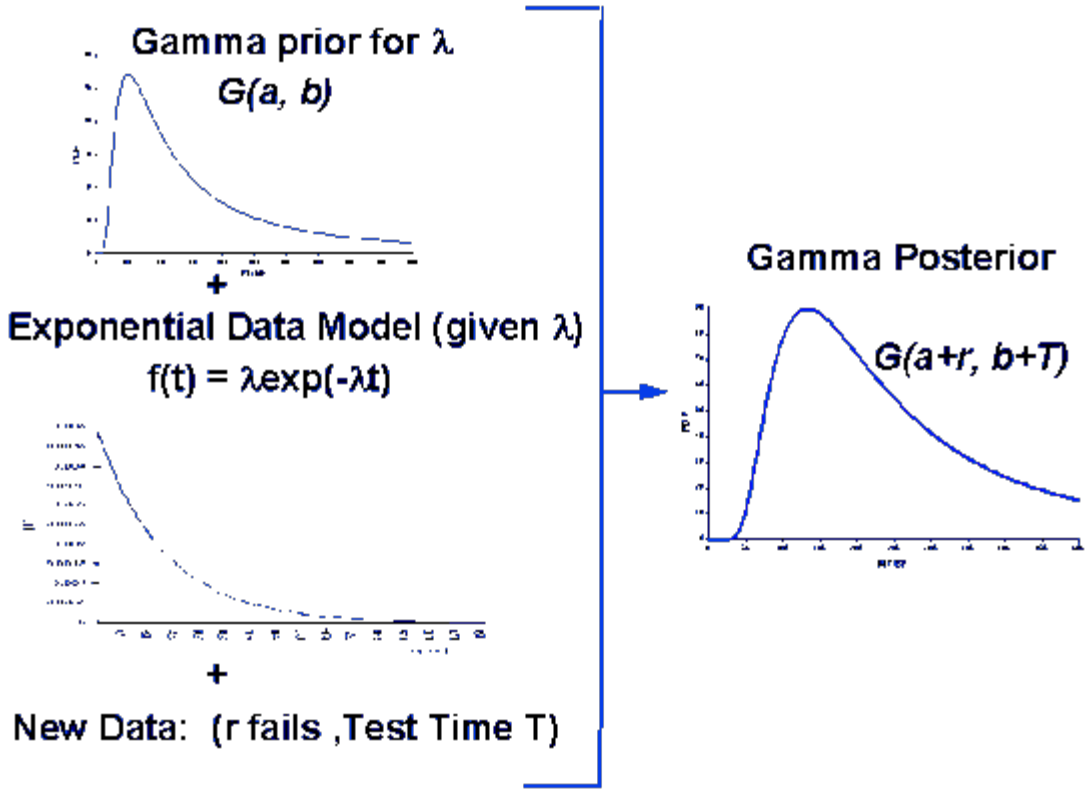
8. Assessing Product Reliability
8.4. Reliability Data Analysis

8.4.6. How do you estimate reliability using the Bayesian gamma prior model?

The Bayesian paradigm was introduced in Section 1 and Section 2 described the assumptions underlying the gamma/exponential system model (including several methods to transform prior data and engineering judgment into gamma prior parameters "*a*" and "*b*"). Finally, we saw in Section 3 how to use this Bayesian system model to calculate the required test time needed to confirm a system MTBF at a given confidence level.

Review of Bayesian procedure for the gamma exponential system model

The goal of Bayesian reliability procedures is to obtain as accurate a posterior distribution as possible, and then use this distribution to calculate failure rate (or MTBF) estimates with confidence intervals (called **credibility intervals** by Bayesians). The figure below summarizes the steps in this process.



How to estimate the MTBF with bounds, based on the

Once the test has been run, and *r* failures observed, the posterior gamma parameters are:

$$a' = a + r,\ b' = b + T$$

and a (median) estimate for the MTBF is calculated by

$$1 / G^{-1}(0.5, a', (1/b'))$$

posterior distribution

where $G(q, \gamma, \beta)$ represents the gamma distribution with shape parameter γ, and scale parameter β. Some people prefer to use the reciprocal of the mean of the posterior distribution as their estimate for the MTBF. The mean is the **minimum mean square error** (MSE) estimator of λ, but using the reciprocal of the mean to estimate the MTBF is always more conservative than the "even money" 50% estimator.

A lower 80% bound for the MTBF is obtained from

$$1 / G^{-1}(0.8, a', (1/b'))$$

and, in general, a lower 100(1-α) % lower bound is given by

$$1 / G^{-1}((1-\alpha), a', (1/b')).$$

A two-sided 100(1-α) % credibility interval for the MTBF is

$$[1 / G^{-1}((1-\alpha/2), a', (1/b')),$$
$$1 / G^{-1}((\alpha/2), a', (1/b'))].$$

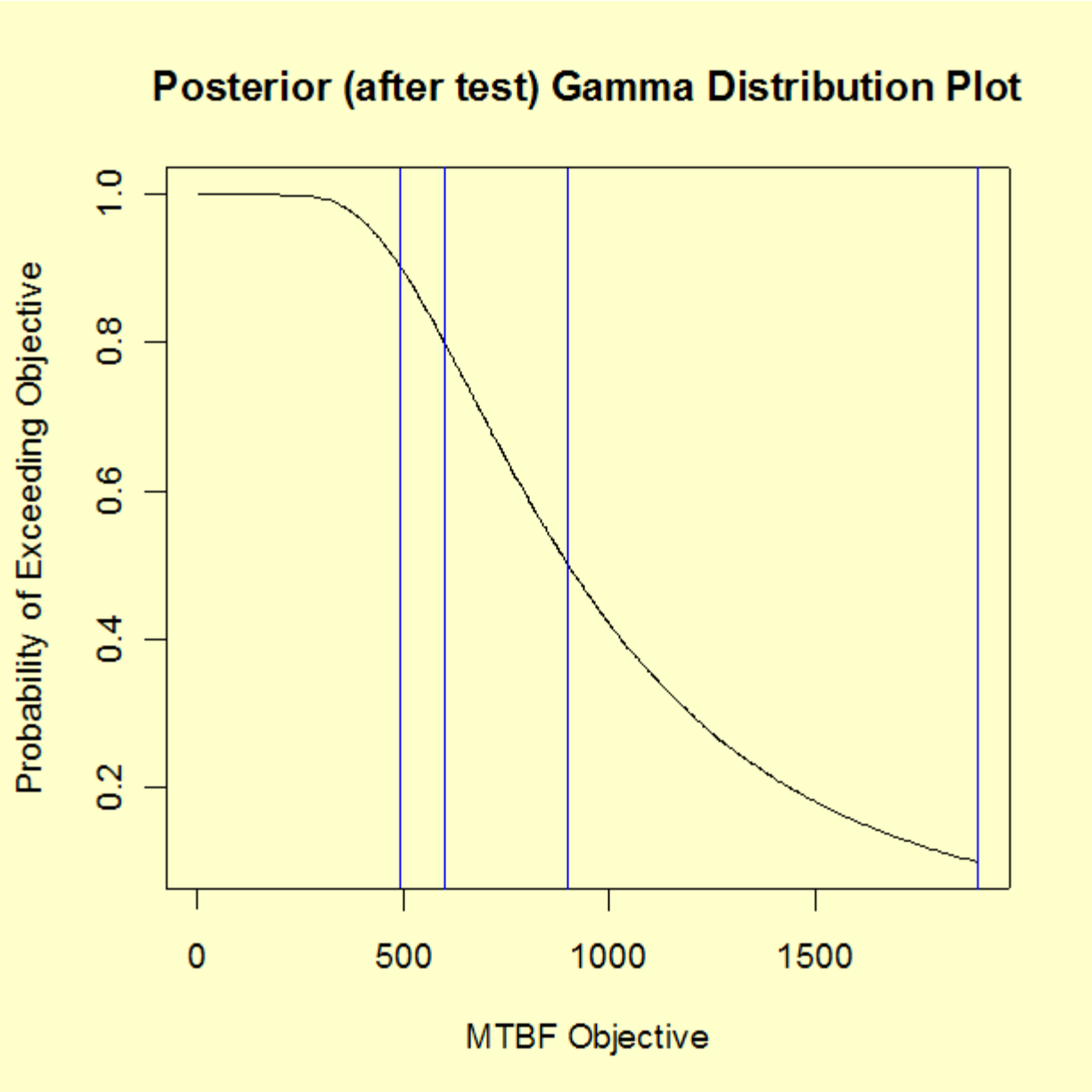
Finally, the $G((1/M), a', (1/b'))$ calculates the probability that MTBF is greater than M.

Example

A Bayesian example to estimate the MTBF and calculate upper and lower bounds

A system has completed a reliability test aimed at confirming a 600 hour MTBF at an 80% confidence level. Before the test, a gamma prior with $a = 2$, $b = 1400$ was agreed upon, based on testing at the vendor's location. Bayesian test planning calculations, allowing up to 2 new failures, called for a test of 1909 hours. When that test was run, there actually were exactly two failures. What can be said about the system?

The posterior gamma CDF has parameters $a' = 4$ and $b' = 3309$. The plot below shows CDF values on the y-axis, plotted against $1/\lambda$ = MTBF, on the x-axis. By going from probability, on the y-axis, across to the curve and down to the MTBF, we can estimate any MTBF percentile point.



The MTBF values are shown below.

$1 / G^{-1}(0.9, 4, (1/3309))$	= 495 hours
$1 / G^{-1}(0.8, 4, (1/3309))$	= 600 hours (as expected)
$1 / G^{-1}(0.5, 4, (1/3309))$	= 901 hours
$1 / G^{-1}(0.1, 4, (1/3309))$	= 1897 hours

The test has confirmed a 600 hour MTBF at 80 % confidence, a 495 hour MTBF at 90 % confidence and (495, 1897) is a 90 % credibility interval for the MTBF. A single number (point) estimate for the system MTBF would be 901 hours. Alternatively, you might want to use the reciprocal of the mean of the posterior distribution (b'/a') = 3309/4 = 827 hours as a single estimate. The reciprocal mean is more conservative, in this case it is a 57 % lower bound ($G((4/3309), 4, (1/3309))$).

The analyses in this section can can be implemented using R code.

8.4.7. References For Chapter 8: Assessing Product Reliability

Aitchison, J. and Brown, J. A. C.,(1957), *The Log-normal distribution*, Cambridge University Press, New York and London.

Ascher, H. (1981), "Weibull Distribution vs Weibull Process," *Proceedings Annual Reliability and Maintainability Symposium*, pp. 426-431.

Ascher, H. and Feingold, H. (1984), *Repairable Systems Reliability*, Marcel Dekker, Inc., New York.

Bain, L.J. and Englehardt, M. (1991), *Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods, 2nd ed*., Marcel Dekker, New York.

Barlow, R. E. and Proschan, F. (1975), *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York.

Birnbaum, Z.W., and Saunders, S.C. (1968), "A Probabilistic Interpretation of Miner's Rule," *SIAM Journal of Applied Mathematics*, Vol. 16, pp. 637-652.

Birnbaum, Z.W., and Saunders, S.C. (1969), "A New Family of Life Distributions," *Journal of Applied Probability*, Vol. 6, pp. 319-327.

Cox, D.R. and Lewis, P.A.W. (1966), *The Statistical Analysis of Series of Events*, John Wiley & Sons, Inc., New York.

Cox, D.R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, B 34, pp. 187-220.

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London, New York.

Crow, L.H. (1974), "Reliability Analysis for Complex Repairable Systems", *Reliability and Biometry*, F. Proschan and R.J. Serfling, eds., SIAM, Philadelphia, pp 379- 410.

Crow, L.H. (1975), "On Tracking Reliability Growth," *Proceedings Annual Reliability and Maintainability Symposium*, pp. 438-443.

Crow, L.H. (1982), "Confidence Interval Procedures for the Weibull Process With Applications to Reliability Growth," *Technometrics*,

24(1):67-72.

Crow, L.H. (1990), "Evaluating the Reliability of Repairable Systems," *Proceedings Annual Reliability and Maintainability Symposium*, pp. 275-279.

Crow, L.H. (1993), "Confidence Intervals on the Reliability of Repairable Systems," *Proceedings Annual Reliability and Maintainability Symposium*, pp. 126-134

Duane, J.T. (1964), "Learning Curve Approach to Reliability Monitoring," *IEEE Transactions On Aerospace*, 2, pp. 563-566.

Gumbel, E. J. (1954), *Statistical Theory of Extreme Values and Some Practical Applications*, National Bureau of Standards Applied Mathematics Series 33, U.S. Government Printing Office, Washington, D.C.

Hahn, G.J., and Shapiro, S.S. (1967), *Statistical Models in Engineering*, John Wiley & Sons, Inc., New York.

Hoyland, A., and Rausand, M. (1994), *System Reliability Theory*, John Wiley & Sons, Inc., New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994), *Continuous Univariate Distributions Volume 1*, 2nd edition, John Wiley & Sons, Inc., New York.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995), *Continuous Univariate Distributions Volume 2*, 2nd edition, John Wiley & Sons, Inc., New York.

Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53: 457-481.

Kalbfleisch, J.D., and Prentice, R.L. (1980), *The Statistical Analysis of Failure Data*, John Wiley & Sons, Inc., New York.

Kielpinski, T.J., and Nelson, W.(1975), "Optimum Accelerated Life-Tests for the Normal and Lognormal Life Distributins," *IEEE Transactions on Reliability*, Vol. R-24, 5, pp. 310-320.

Klinger, D.J., Nakada, Y., and Menendez, M.A. (1990), *AT&T Reliability Manual*, Van Nostrand Reinhold, Inc, New York.

Kolmogorov, A.N. (1941), "On A Logarithmic Normal Distribution Law Of The Dimensions Of Particles Under Pulverization," *Dokl. Akad Nauk*, USSR 31, 2, pp. 99-101.

Kovalenko, I.N., Kuznetsov, N.Y., and Pegg, P.A. (1997), *Mathematical Theory of Reliability of Time Dependent Systems with Practical Applications*, John Wiley & Sons, Inc., New York.

Landzberg, A.H., and Norris, K.C. (1969), "Reliability of Controlled

Collapse Interconnections." *IBM Journal Of Research and Development*, Vol. 13, 3.

Lawless, J.F. (1982), *Statistical Models and Methods For Lifetime Data*, John Wiley & Sons, Inc., New York.

Leon, R. (1997-1999), JMP Statistical Tutorials on the Web at http://www.nist.gov/cgi-bin/exit_nist.cgi?url=http://web.utk.edu/~leon/jmp/.

Mann, N.R., Schafer, R.E. and Singpurwalla, N.D. (1974), *Methods For Statistical Analysis Of Reliability & Life Data*, John Wiley & Sons, Inc., New York.

Martz, H.F., and Waller, R.A. (1982), *Bayesian Reliability Analysis*, Krieger Publishing Company, Malabar, Florida.

Meeker, W.Q., and Escobar, L.A. (1998), *Statistical Methods for Reliability Data*, John Wiley & Sons, Inc., New York.

Meeker, W.Q., and Hahn, G.J. (1985), "How to Plan an Accelerated Life Test - Some Practical Guidelines," *ASC Basic References In Quality Control: Statistical Techniques - Vol. 10*, ASQC , Milwaukee, Wisconsin.

Meeker, W.Q., and Nelson, W. (1975), "Optimum Accelerated Life-Tests for the Weibull and Extreme Value Distributions," *IEEE Transactions on Reliability*, Vol. R-24, 5, pp. 321-322.

Michael, J.R., and Schucany, W.R. (1986), "Analysis of Data From Censored Samples," *Goodness of Fit Techniques*, ed. by D'Agostino, R.B., and Stephens, M.A., Marcel Dekker, New York.

MIL-HDBK-189 (1981), *Reliability Growth Management*, U.S. Government Printing Office.

MIL-HDBK-217F ((1986), *Reliability Prediction of Electronic Equipment*, U.S. Government Printing Office.

MIL-STD-1635 (EC) (1978), *Reliability Growth Testing*, U.S. Government Printing Office.

Nelson, W. (1990), *Accelerated Testing*, John Wiley & Sons, Inc., New York.

Nelson, W. (1982), *Applied Life Data Analysis*, John Wiley & Sons, Inc., New York.

O'Connor, P.D.T. (1991), *Practical Reliability Engineering (Third Edition)*, John Wiley & Sons, Inc., New York.

Peck, D., and Trapp, O.D. (1980), *Accelerated Testing Handbook*, Technology Associates and Bell Telephone Laboratories, Portola, Calif.

Pore, M., and Tobias, P. (1998), "How Exact are 'Exact' Exponential

System MTBF Confidence Bounds?", *1998 Proceedings of the Section on Physical and Engineering Sciences of the American Statistical Association*.

SEMI E10-0701, (2001), *Standard For Definition and Measurement of Equipment Reliability, Availability and Maintainability (RAM)*, Semiconductor Equipment and Materials International, Mountainview, CA.

Tobias, P. A., and Trindade, D. C. (1995), *Applied Reliability*, 2nd edition, Chapman and Hall, London, New York.

医课汇 公众号 专业医疗器械资讯平台 WECHAT OF HLONGMED
hlongmed.com 医疗器械咨询服务 MEDICAL DEVICE CONSULTING SERVICES
医课培训平台 医疗器械任职培训 WEB TRAINING CENTER
医械宝 医疗器械知识平台 KNOWLEDG ECENTEROF MEDICAL DEVICE
MDCPP.COM 医械云专业平台 KNOWLEDG ECENTEROF MEDICAL DEVICE