6. Process or Product Monitoring and Control

This chapter presents techniques for monitoring and controlling processes and signaling when corrective actions are necessary.

6. Process or Product Monitoring and Control - Detailed Table of Contents [6.]

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.

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?

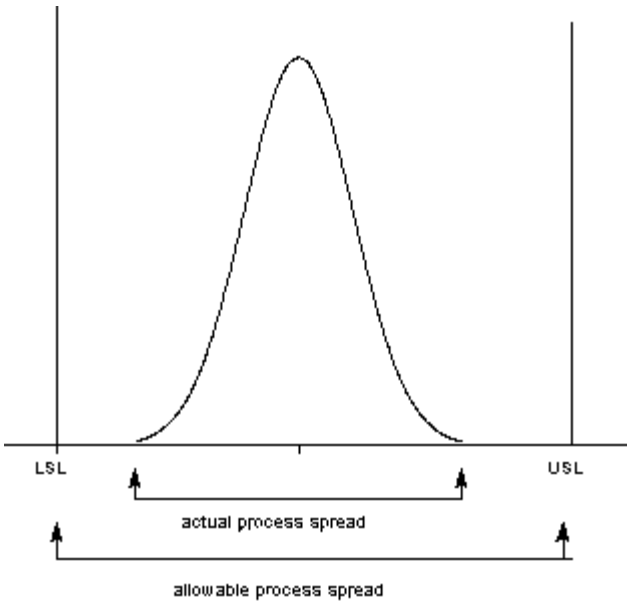
Process capability compares the output of an *in-control* process to the specification limits by using *capability indices*. The comparison is made by forming the ratio of the spread between the process specifications (the specification "width") to the spread of the process values, as measured by 6 process 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 $\bar{x} \le USL$.

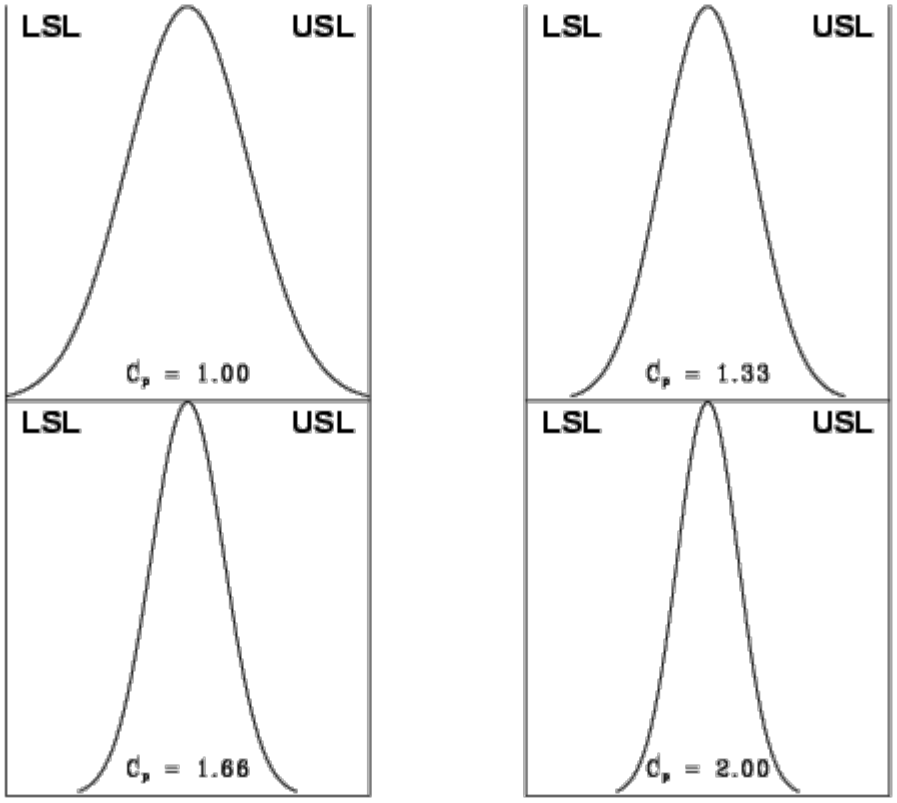
The estimator for the C_p index, adjusted by the k factor, is

$$\hat{C}_{pk} = \hat{C}_p(1 - \hat{k})$$

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

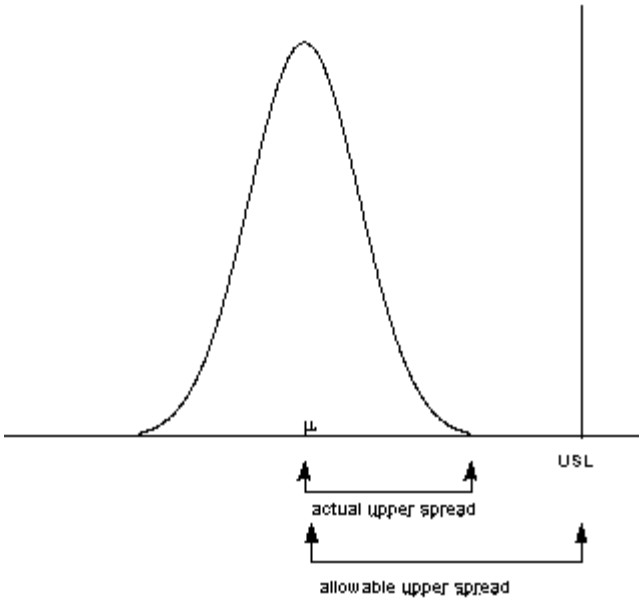
$$C_{pl} = \frac{\text{allowable lower spread}}{\text{actual lower spread}} = \frac{\mu - LSL}{3\sigma}$$

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.

6.2. 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.

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.

6.2.3.1. 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,

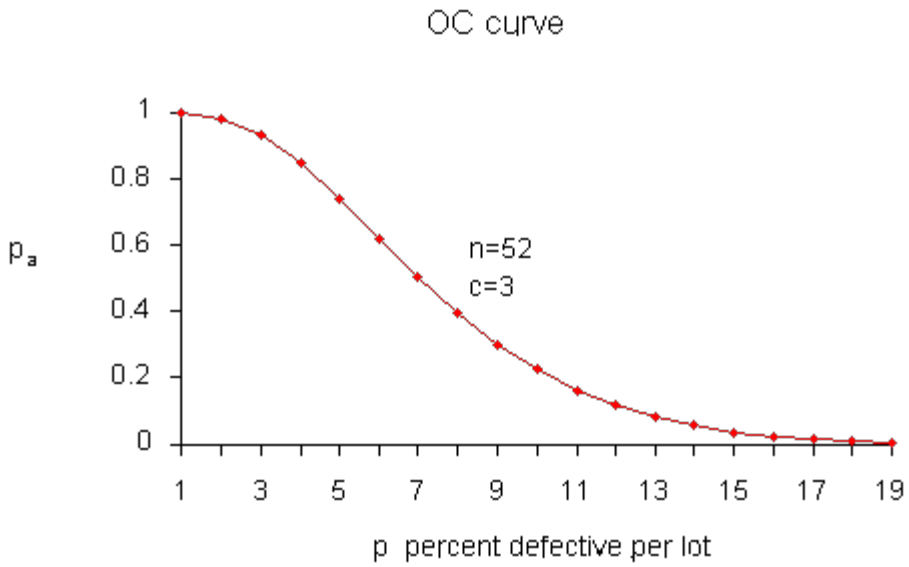
information (and 105D). The interested reader is referred to references such as (Montgomery (2000), Schilling, tables 11-2 to 11-17, and Duncan, pages 214 - 248).

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

$$AOQ = (.930)(.03)(10000\text{-}52) / 10000 = 0.02775.$$

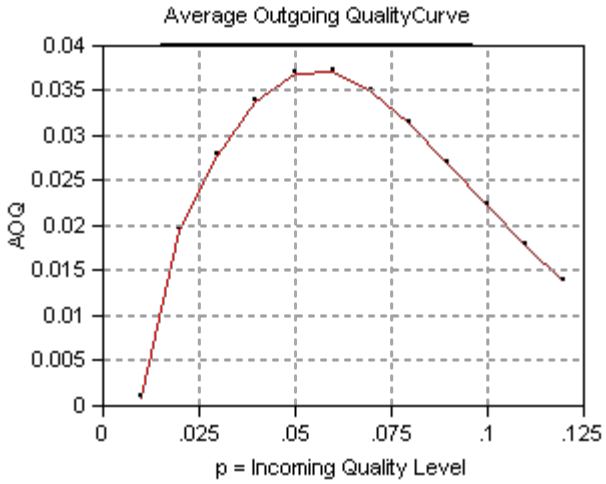
Sample table of AOQ versus p

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.)

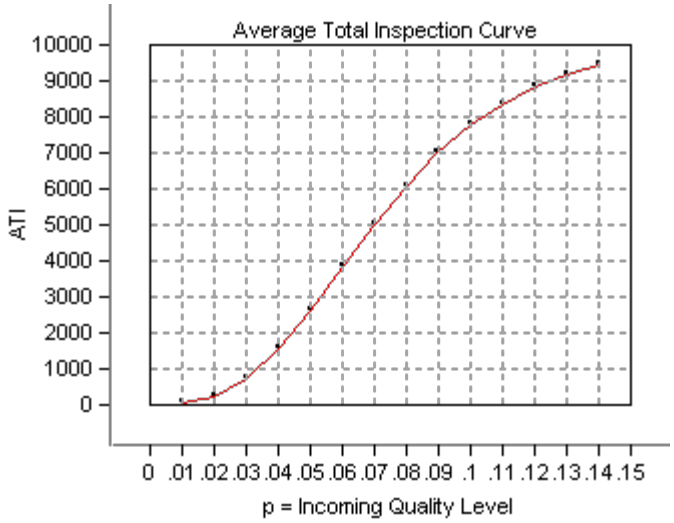
Sample table of ATI versus p

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

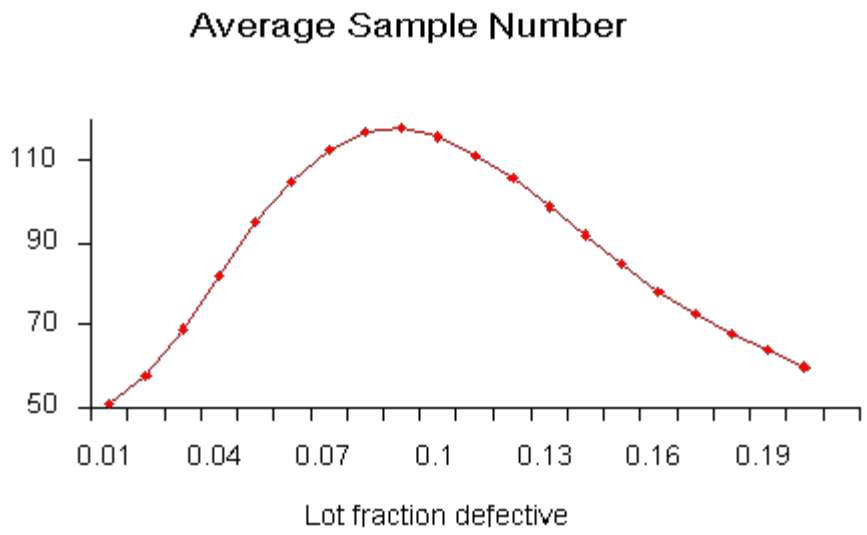
$$50(.445) + 150(.555) = 106$$

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?

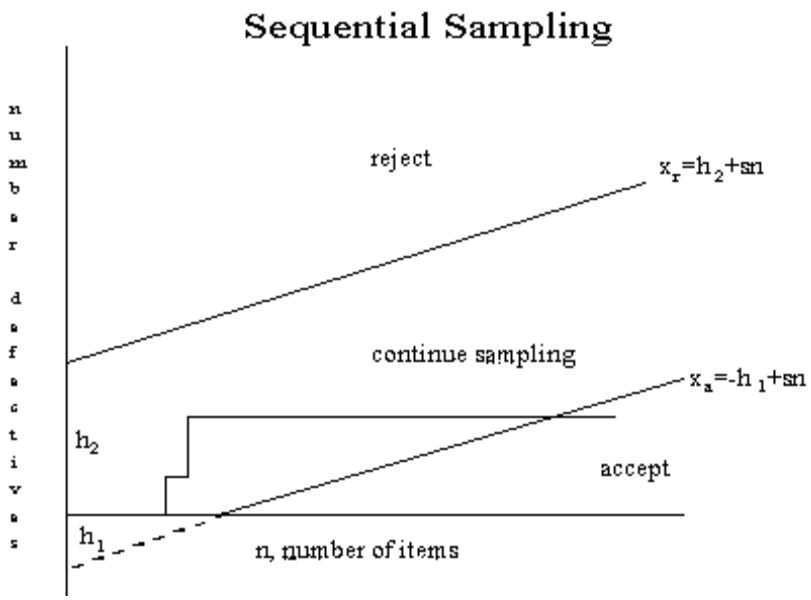
Sequential Sampling

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

$$P_a(f,i) = \frac{fP + (1-f)P^i}{f + (1-f)P^i}$$

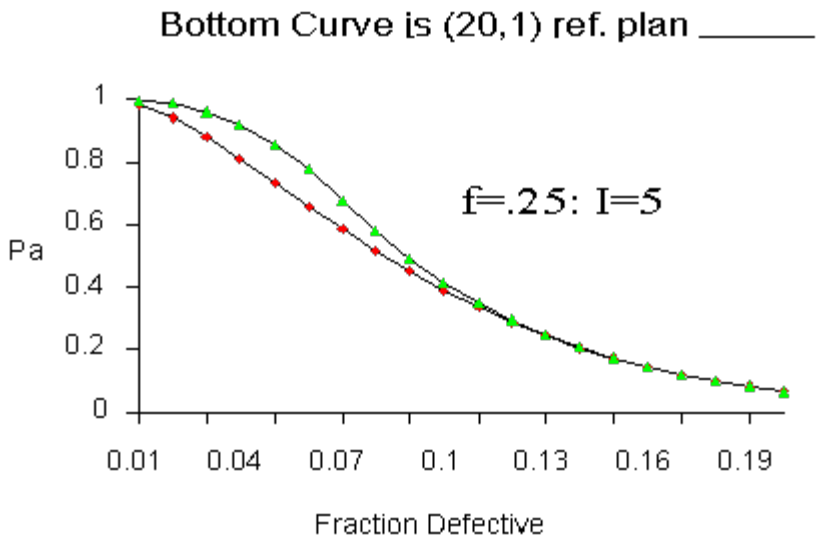
where P is the probability of accepting a lot with a given proportion of incoming defectives p, from the OC curve of the single sampling plan.

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. Process or Product Monitoring and Control

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.

6.3.1. What are Control Charts?

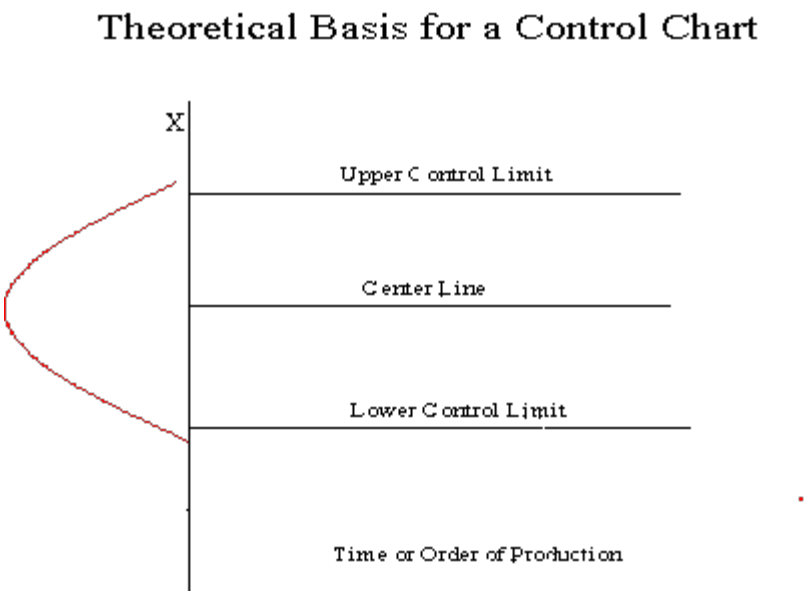
Comparison of univariate and multivariate control data

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.

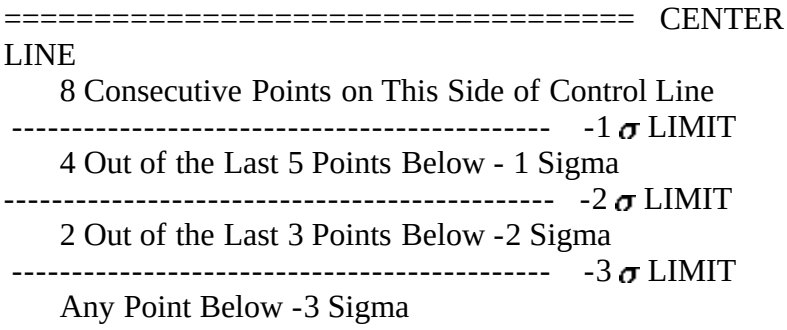
- If a known, preventable act changes the way the tool or process would behave (power goes out, consumable is corrupted or bad quality, etc.)

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



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 ± 3σ 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.

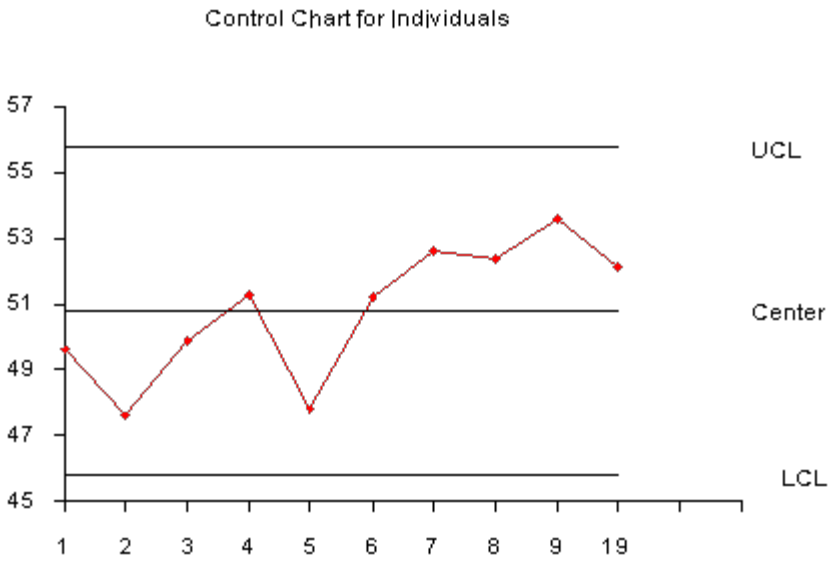
$$UCL = \bar{x} + 3\frac{\overline{MR}}{1.128} = 50.81 + 3\frac{1.8778}{1.128} = 55.8041$$

$$\text{Center Line} = \bar{x} = 50.81$$

$$LCL = \bar{x} - 3\frac{\overline{MR}}{1.128} = 50.81 - 3\frac{1.8778}{1.128} = 45.8159$$

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

$$S_m = \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0) \quad \text{or} \quad S'_m = \frac{1}{\sigma_{\bar{x}}} \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_0)$$

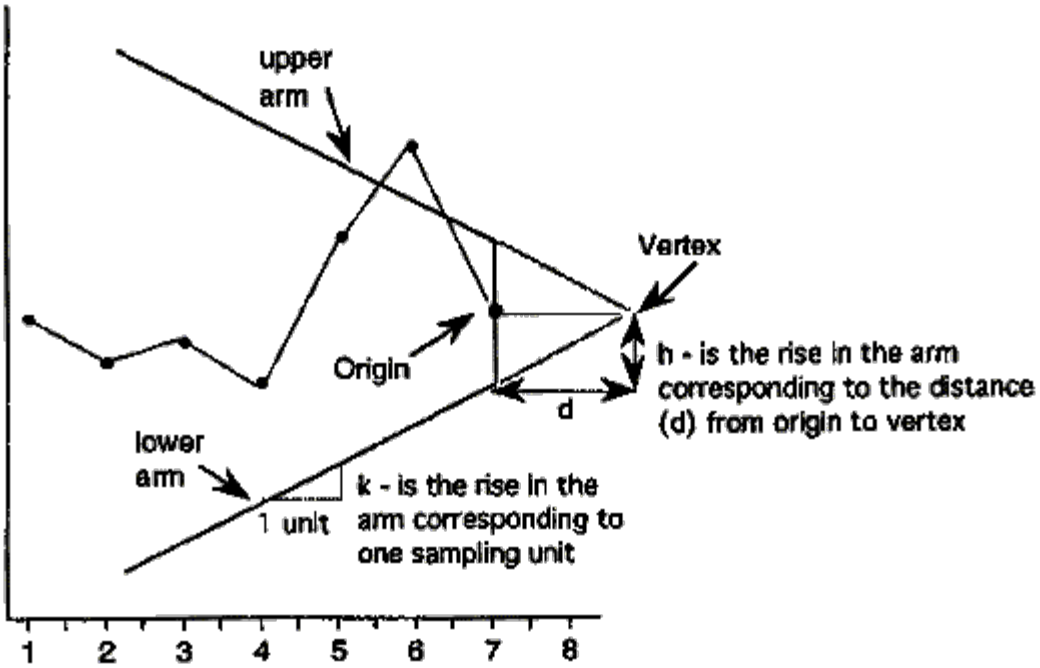
against the sample number *m*, where $\hat{\mu}_0$ is the estimate of the in-control mean and $\sigma_{\bar{x}}$ is the known (or estimated) standard deviation of the sample means. The choice of which of these two quantities is plotted is usually determined by the statistical software package. In either 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).

$$k = \frac{\delta \sigma_x}{2}$$

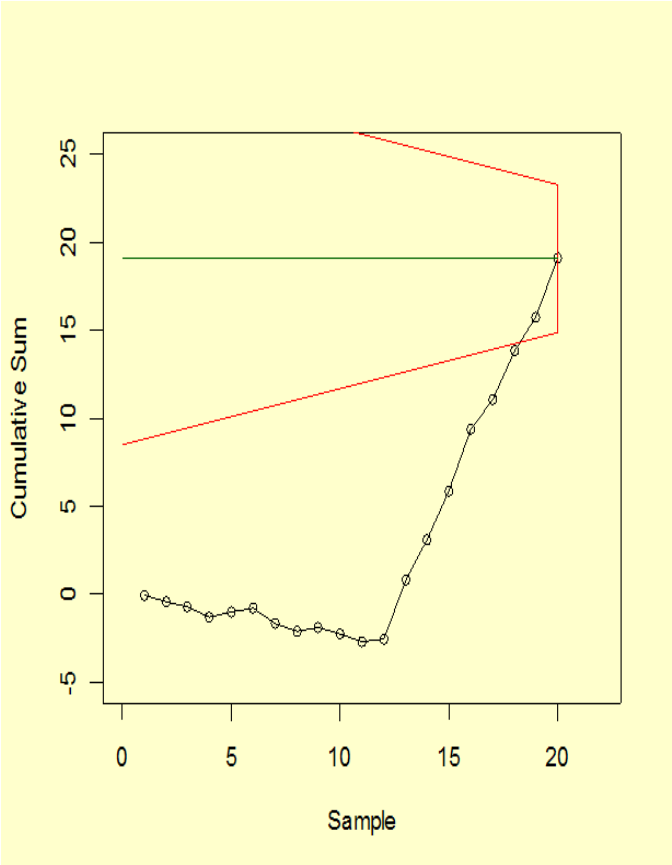
$$d = \frac{2}{\delta^2} \ln\left(\frac{1-\beta}{\alpha}\right)$$

$$h = dk$$

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 $\delta = 1$.

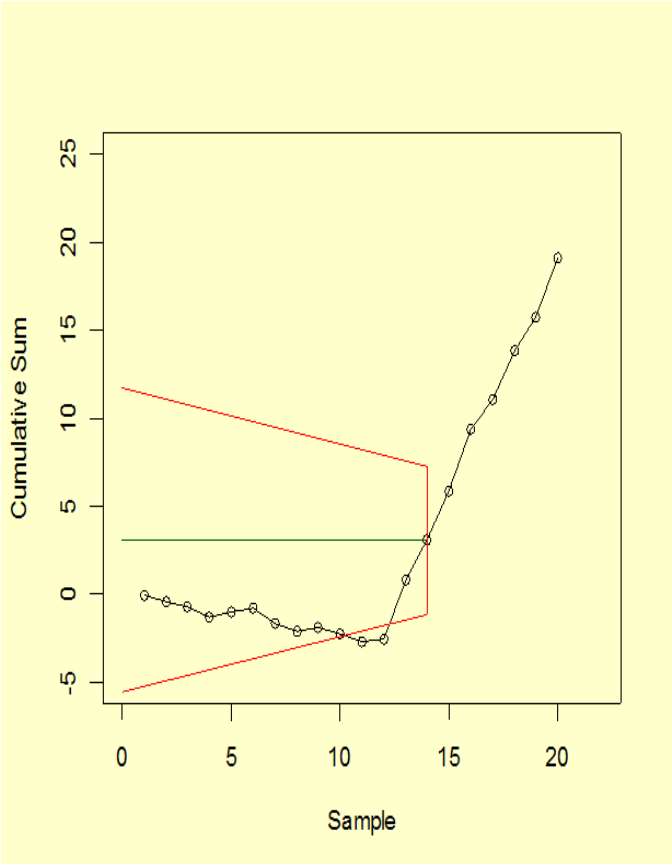
CUSUM Chart with V-Mask

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
```

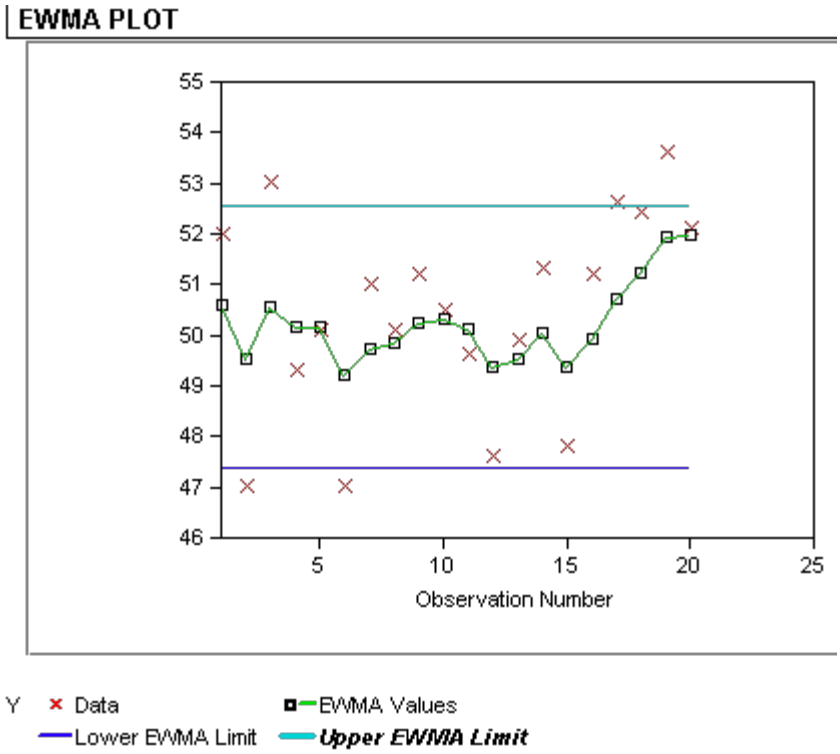
EWMA statistics for sample data

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

$$\bar{c} = \frac{\text{total number of defects}}{\text{total number of samples}} = \frac{400}{25} = 16$$

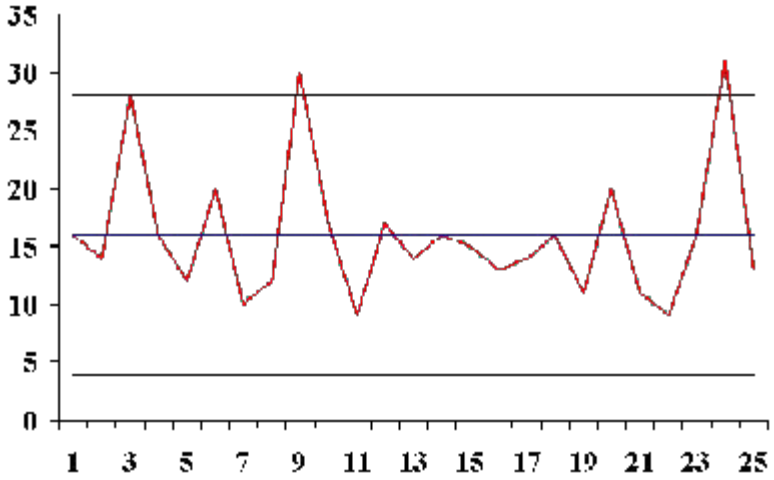
$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 16 + 2\sqrt{16} = 28$$

$$LCL = c - k\sqrt{c}$$

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.

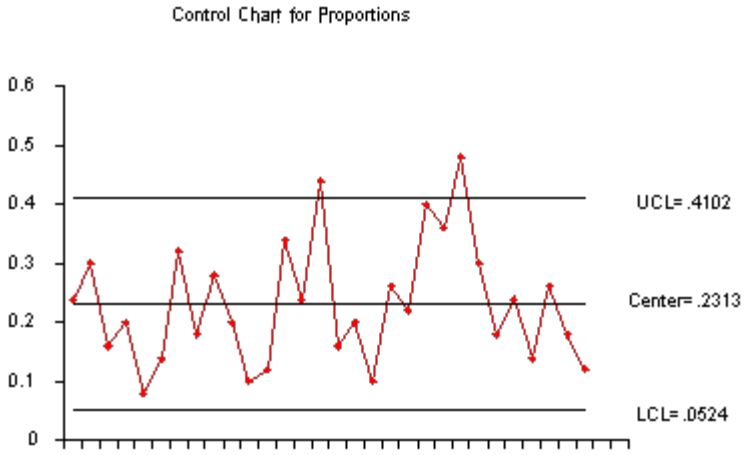
On each wafer 50 chips are measured and a defective is defined whenever a misregistration, in terms of horizontal and/or vertical distances from the center, is recorded. The results are

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

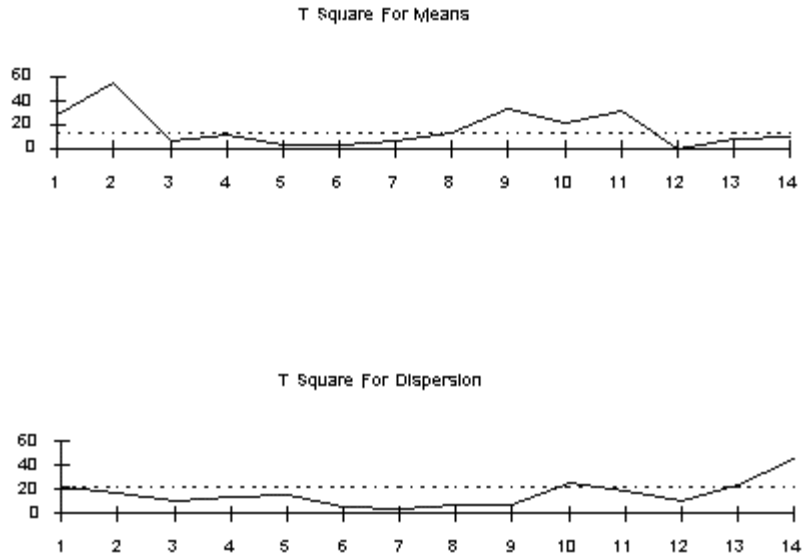
Nowadays, modern computers in general and the PC in particular have made complex calculations accessible and during the last decade, multivariate control charts were given more attention. In fact, the multivariate charts which display the Hotelling T^2 statistic became so popular that they sometimes 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 $N_p(\mathbf{m}, \mathbf{\Sigma})$ 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).

6.3.4.3. 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

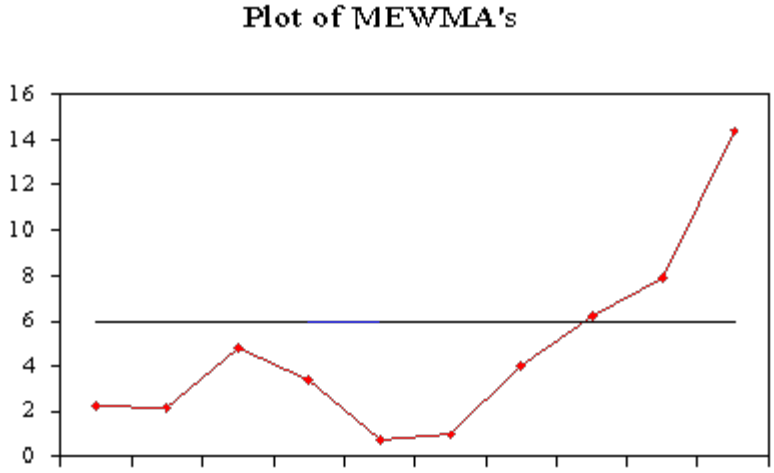
Here is an example of the application of an MEWMA control chart. To faciltate comparison with existing literature, we used data from Lowry et al. The data were simulated from a bivariate normal distribution with unit variances and a correlation coefficient of 0.5. The value for $\lambda = .10$ and the 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:

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.

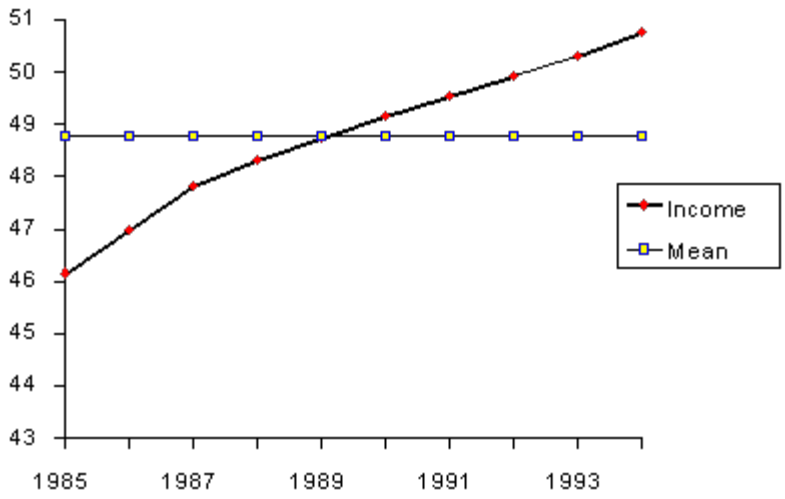
	Squared

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 $M = 4$.

		Interim Steps	
Period	**Value**	**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:

$$\alpha \sum_{i=0}^{t-1} (1-\alpha)^i = \alpha\left[\frac{1-(1-\alpha)^t}{1-(1-\alpha)}\right] = 1-(1-\alpha)^t$$

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 $\alpha - \Delta$ and $\alpha + \Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.

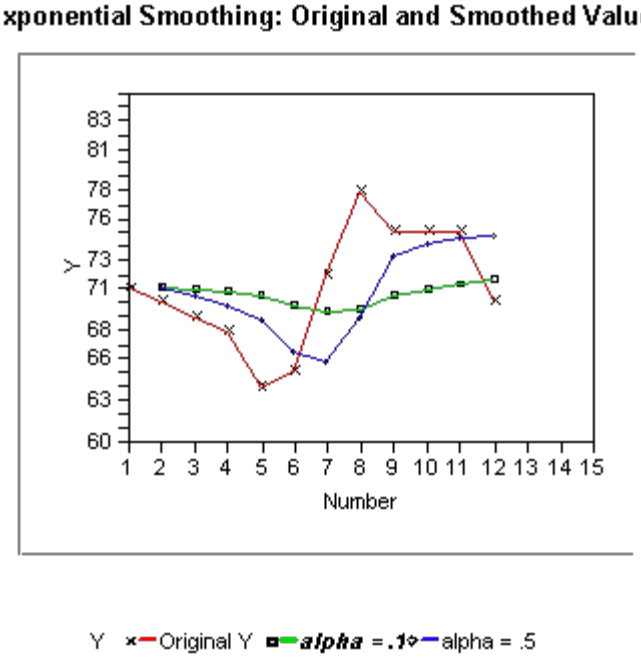
Nonlinear optimizers can be

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

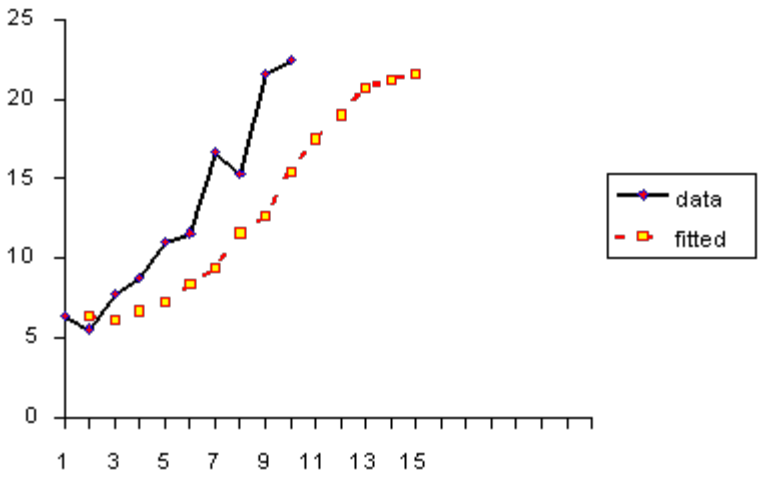
Let us demonstrate this with the following data set smoothed with an α of 0.3:

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



NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

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

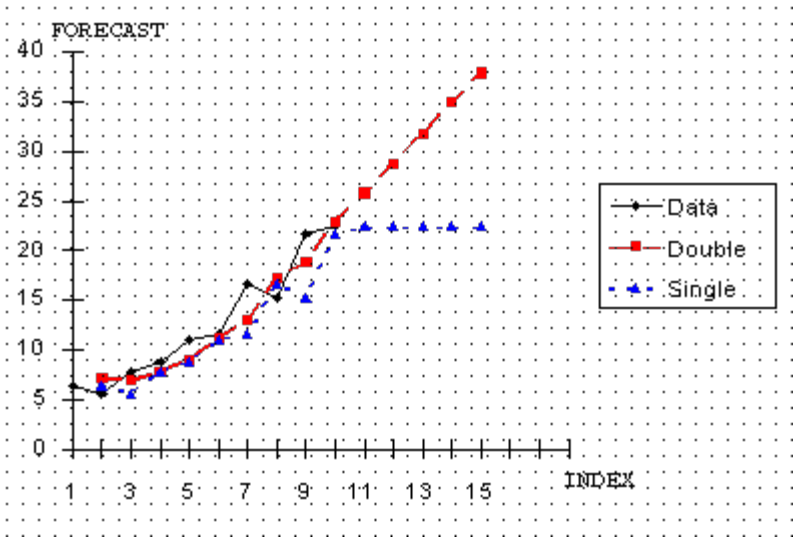
Table showing single and double exponential smoothing forecasts

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.

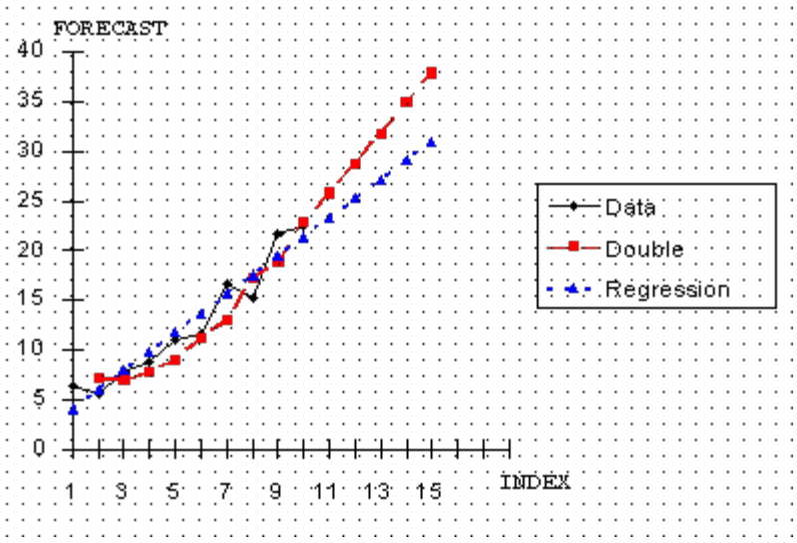


This graph indicates that double smoothing follows the data much closer than single smoothing. Furthermore, for forecasting single smoothing cannot do better than projecting a straight horizontal line, which is not very likely to occur in reality. So in this case double smoothing is preferred.

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

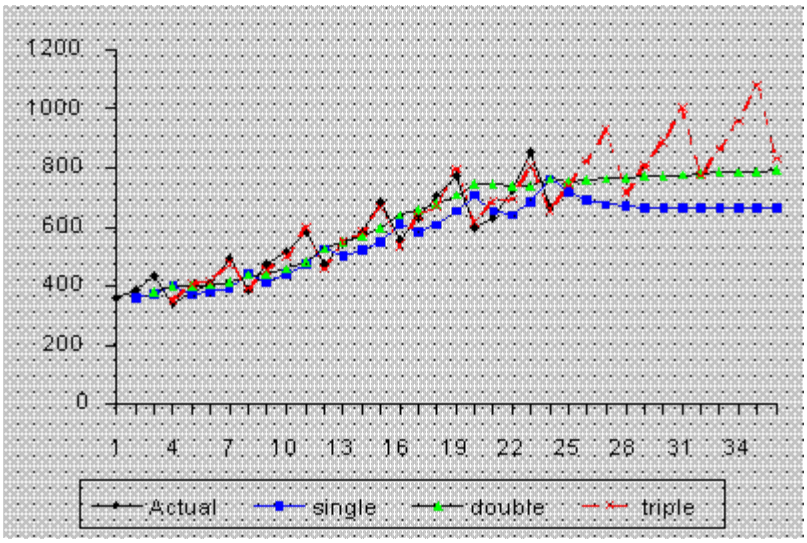
This example shows comparison of single, double and triple exponential smoothing for a data set.

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

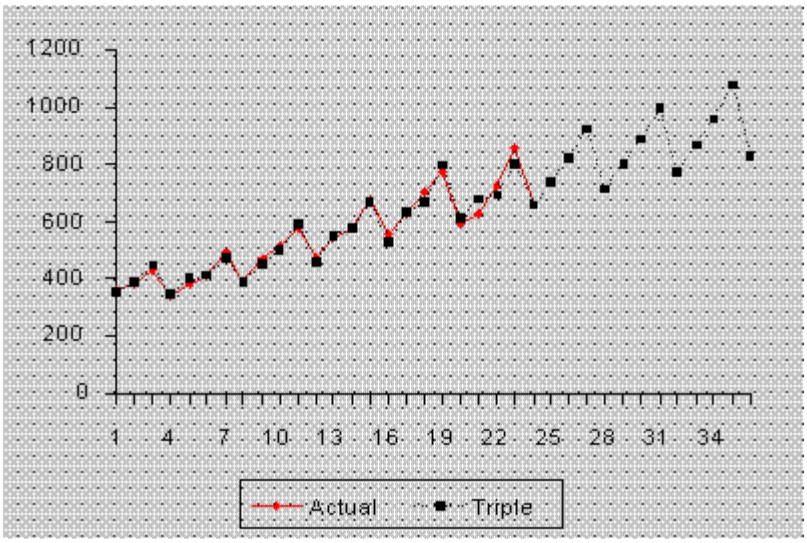
	Quarter	Period	Sales		Quarter	Period	Sales
90	1	1	362	**93**	1	13	544
	2	2	385		2	14	582
	3	3	432		3	15	681
	4	4	341		4	16	557
91	1	5	382	**94**	1	17	628
	2	6	409		2	18	707
	3	7	498		3	19	773
	4	8	387		4	20	592
92	1	9	473	**95**	1	21	627
	2	10	513		2	22	725
	3	11	582		3	23	854
	4	12	474		4	24	661

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:

$$b_1 = \frac{1}{4}\left[\frac{Y_5 - Y_1}{4}\right] + \left[\frac{Y_6 - Y_2}{4}\right] + \left[\frac{Y_7 - Y_3}{4}\right] + \left[\frac{Y_8 - Y_4}{4}\right]$$
$$= \frac{1}{4}\left[\frac{382 - 362}{4}\right] + \left[\frac{409 - 385}{4}\right] + \left[\frac{498 - 432}{4}\right] + \left[\frac{387 - 341}{4}\right]$$
$$= \frac{5 + 6 + 16.5 + 11.5}{4} = 9.75$$

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

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

6.4.4.1.2. 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.

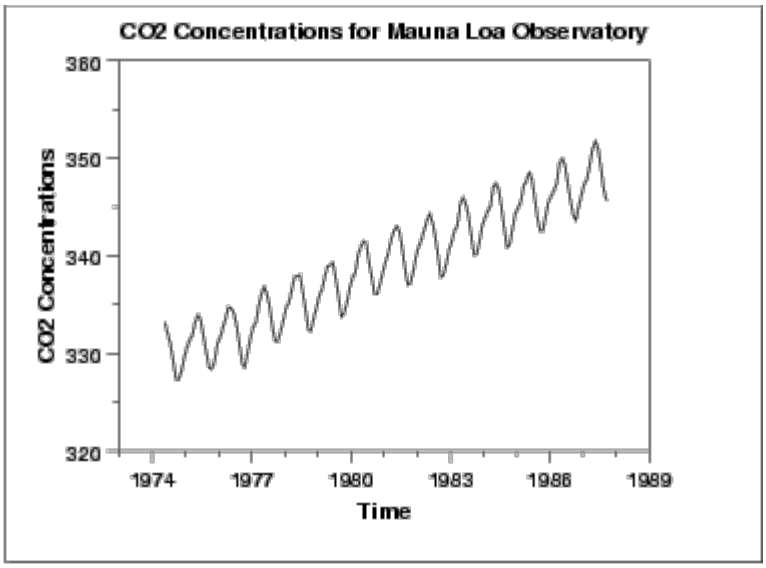
The above techniques are intended to generate series with

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.

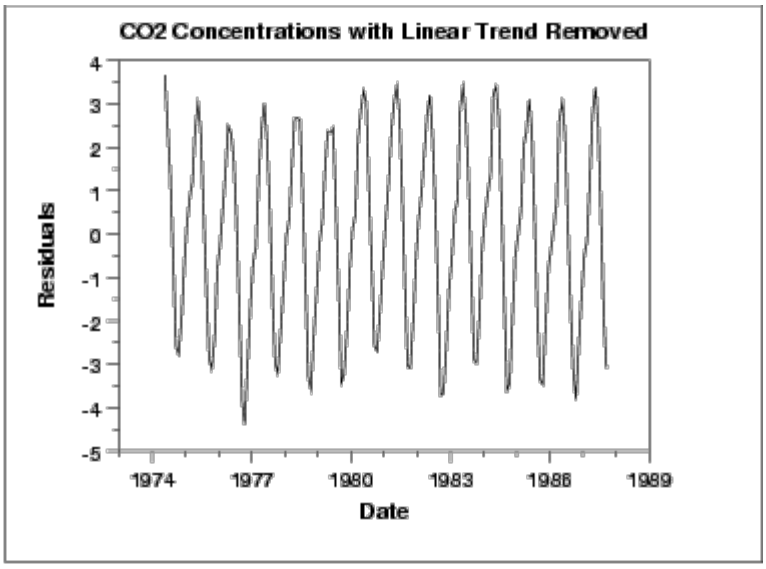
Run Sequence Plot



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.

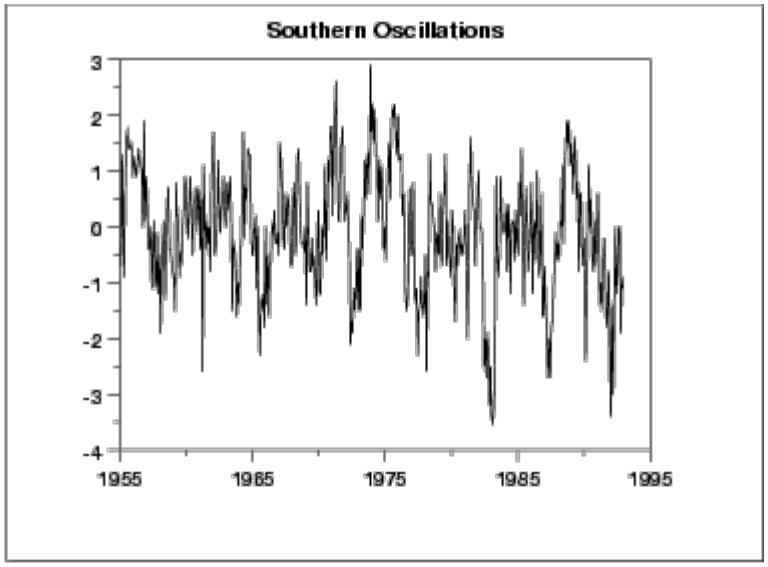
Both the seasonal subseries plot and the box plot assume that

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

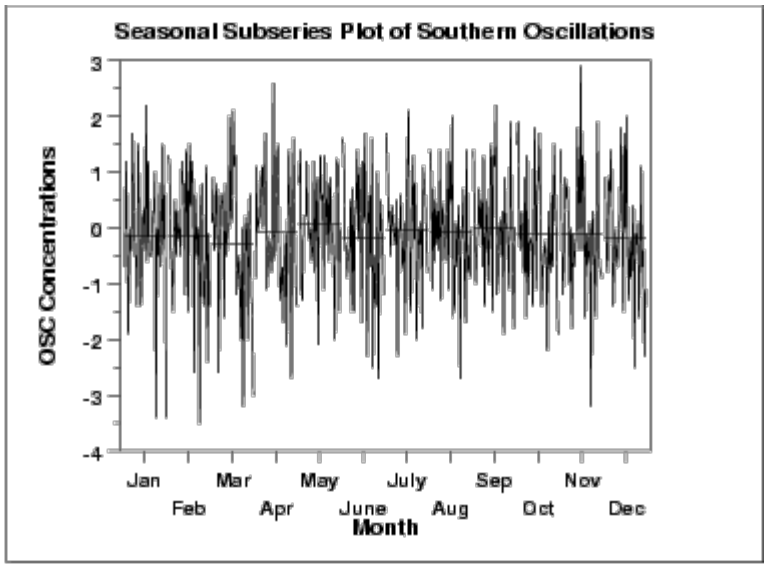
The following plots are from a data set of southern oscillations for predicting el nino.

Run Sequence Plot



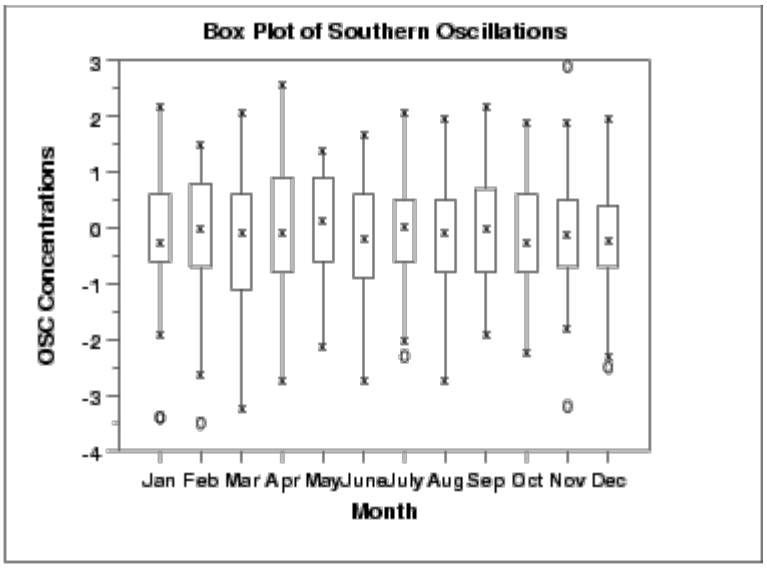
No obvious periodic patterns are apparent in the run sequence plot.

Seasonal Subseries Plot



The means for each month are relatively close and show no obvious pattern.

Box Plot



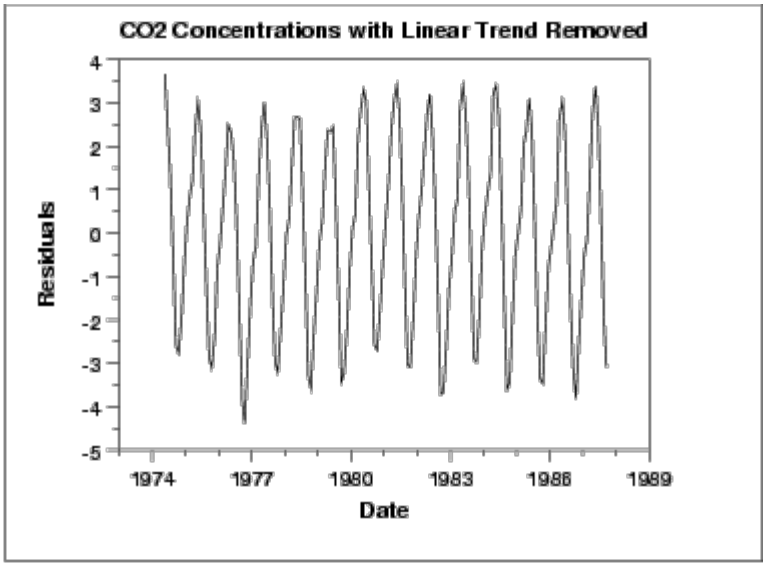
As with the seasonal subseries plot, no obvious seasonal pattern is apparent.

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

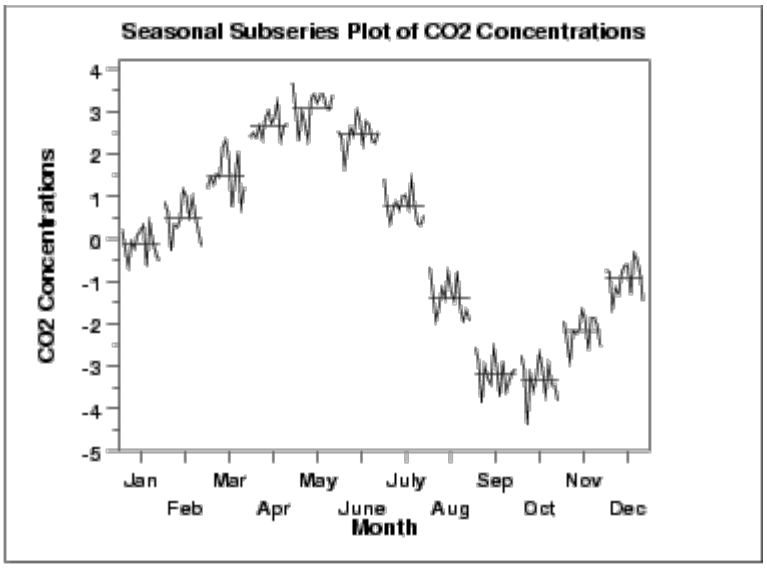
The following plots are from a data set of monthly CO2 concentrations. A linear trend has been removed from these data.

Run Sequence Plot



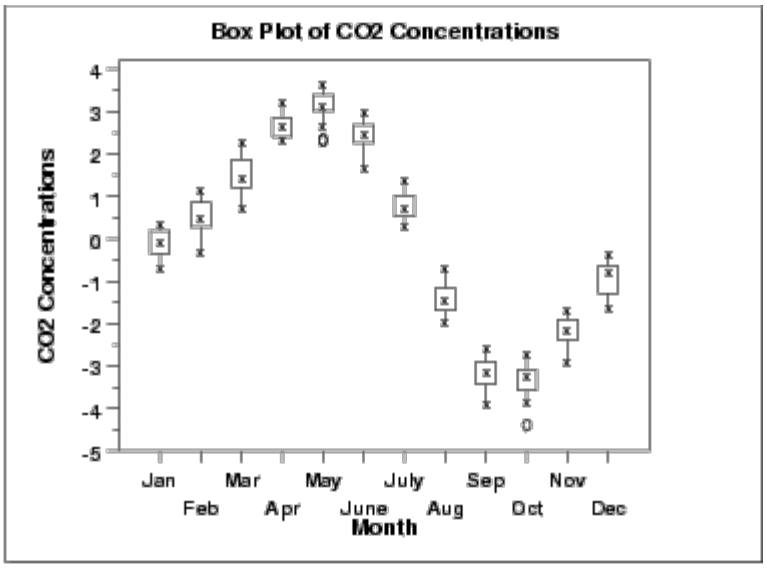
This plot shows periodic behavior. However, it is difficult to determine the nature of the seasonality from this plot.

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

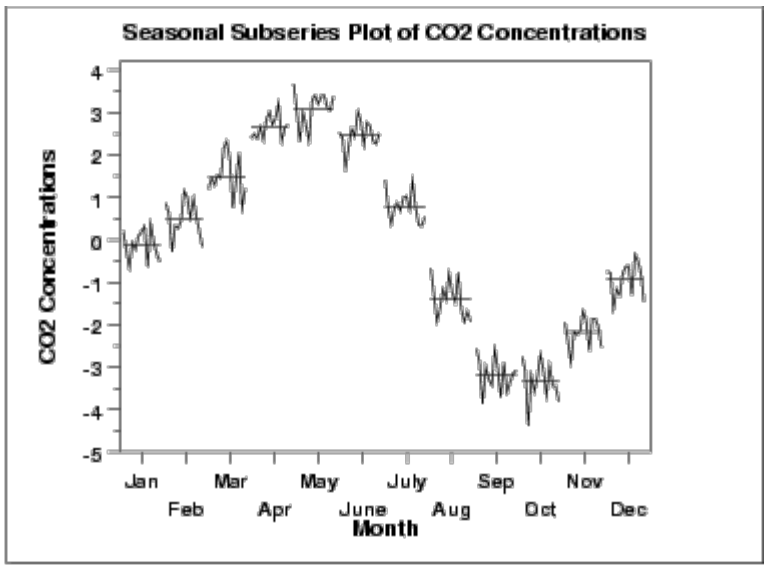
Purpose

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.

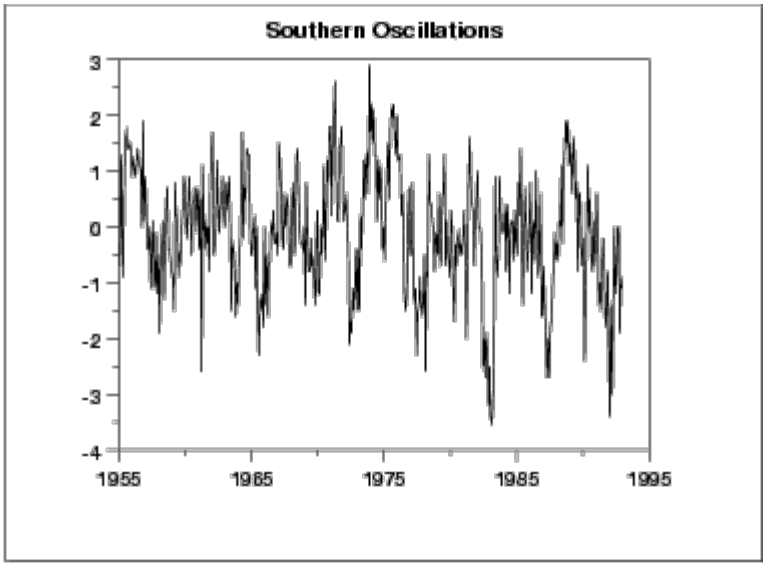
6.4.4.6.1. Model Identification for Southern Oscillations Data

Example for Southern Oscillations

We show typical series of plots for the initial model identification stages of Box-Jenkins modeling for two different examples.

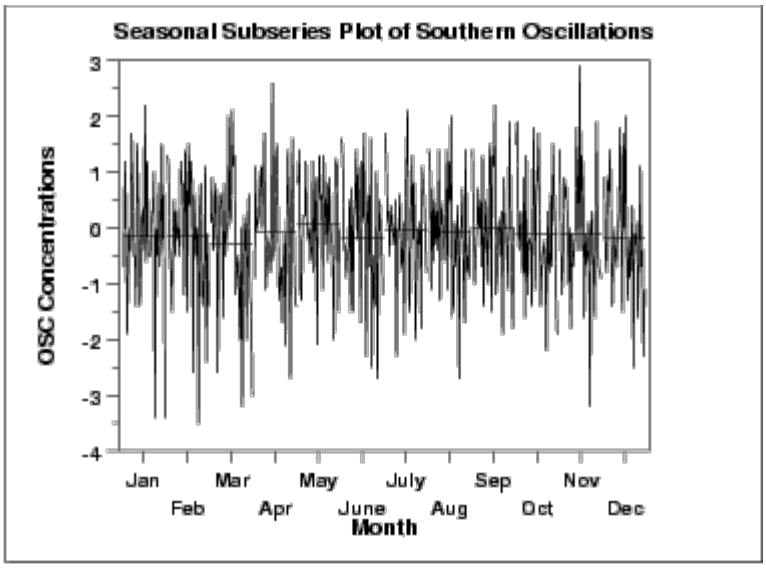
The first example is for the southern oscillations data set. We start with the run sequence plot and seasonal subseries plot to determine if we need to address stationarity and seasonality.

Run Sequence Plot



The run sequence plot indicates stationarity.

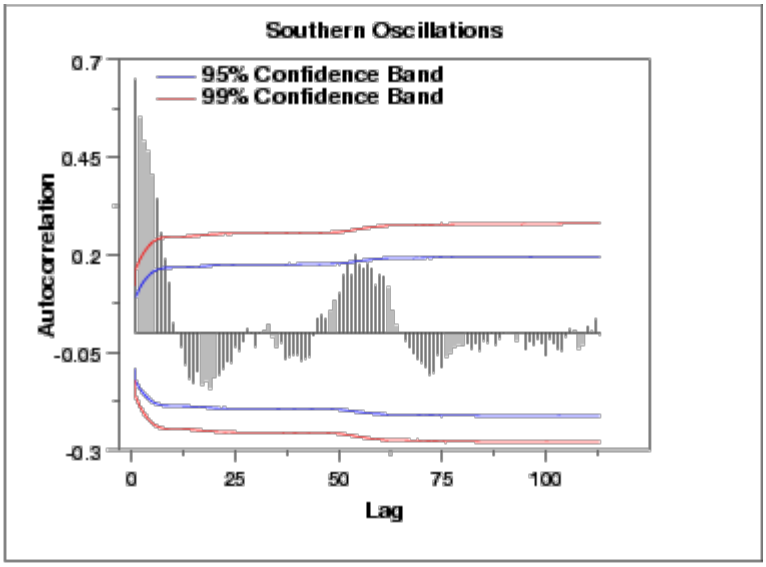
Seasonal Subseries Plot



The seasonal subseries plot indicates that there is no significant seasonality.

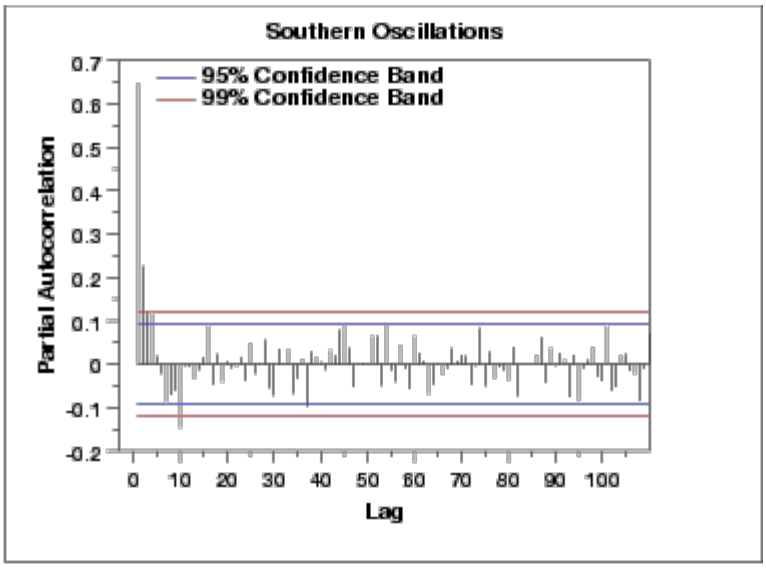
Since the above plots show that this series does not exhibit any significant non-stationarity or seasonality, we generate the autocorrelation and partial autocorrelation plots of the raw data.

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.

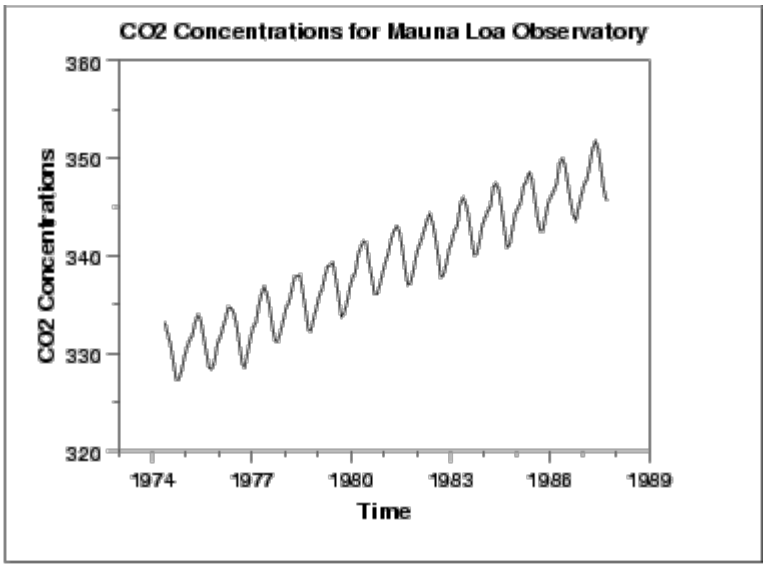
In summary, our intial attempt would be to fit an AR(2) model with no seasonal terms and no differencing or trend removal. Model validation should be performed before accepting this as a final model.

6.4.4.6.2. Model Identification for the CO_2 Concentrations Data

Example for Monthly CO_2 Concentrations

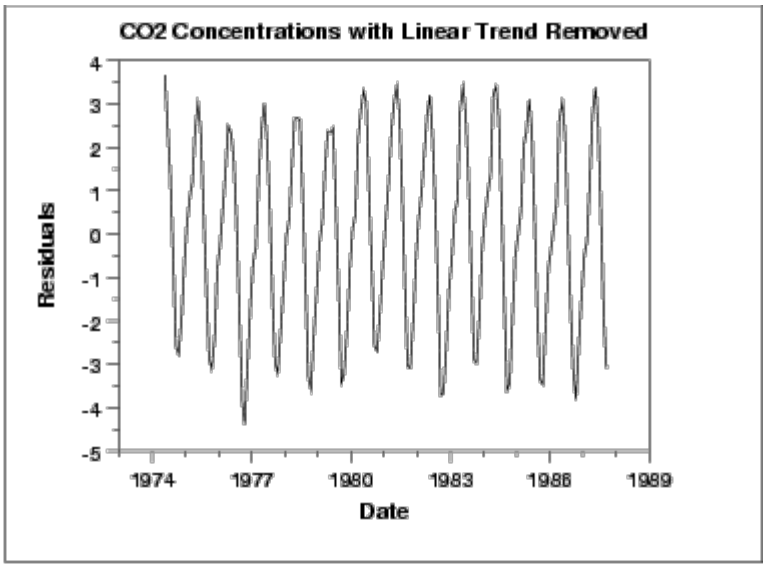
The second example is for the monthly CO_2 concentrations data set. As before, we start with the run sequence plot to check for stationarity.

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.

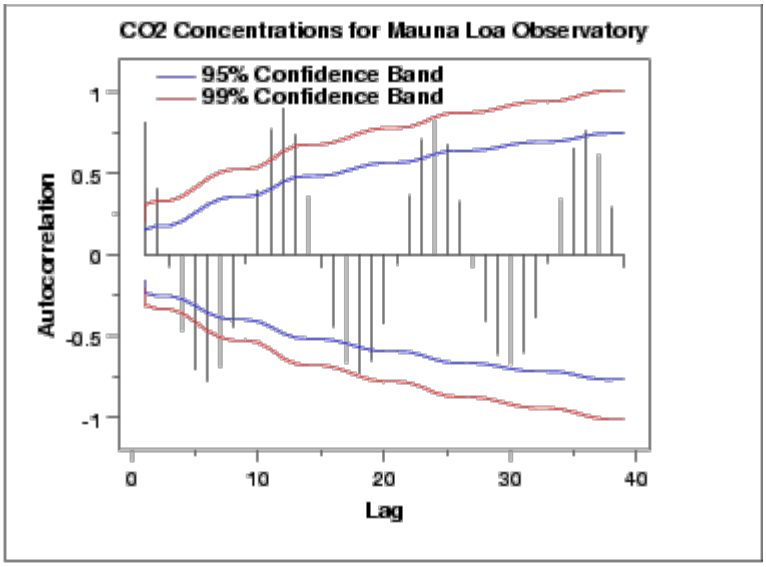
Linear Trend Removed



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.

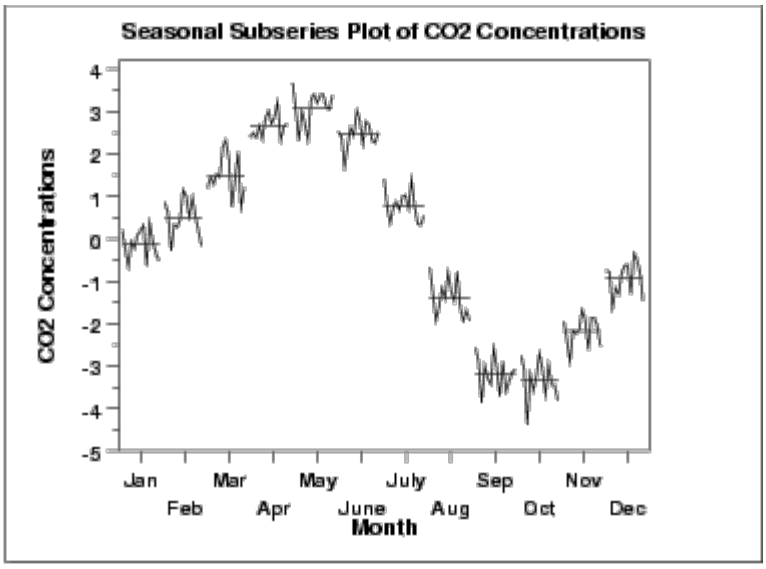
Autocorrelation Plot



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.

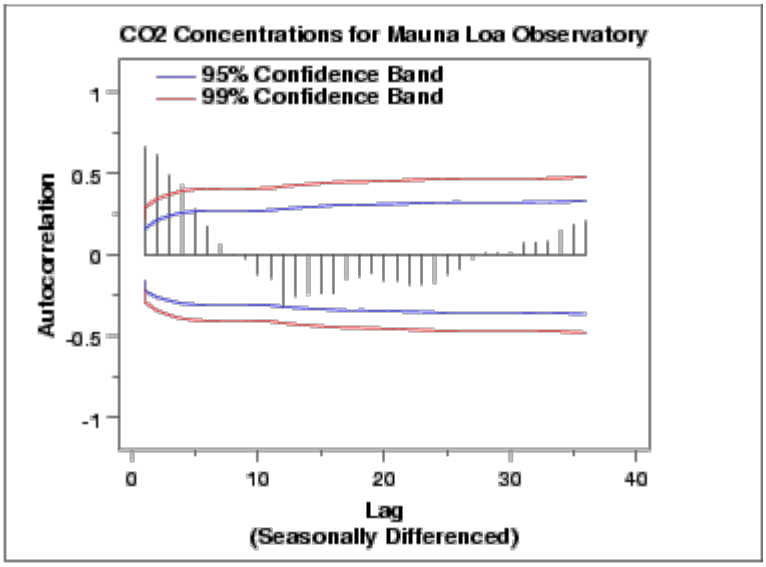
Seasonal Subseries Plot



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.

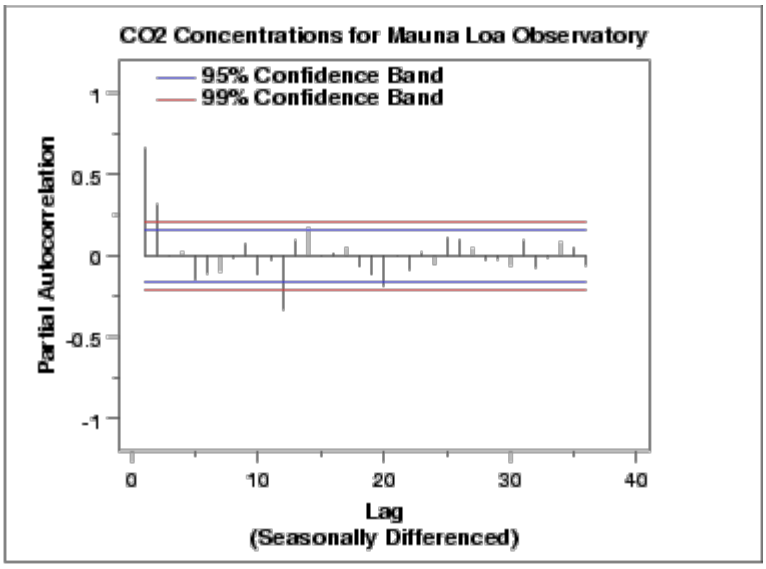
Autocorrelation Plot for Seasonally Differenced Data



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.

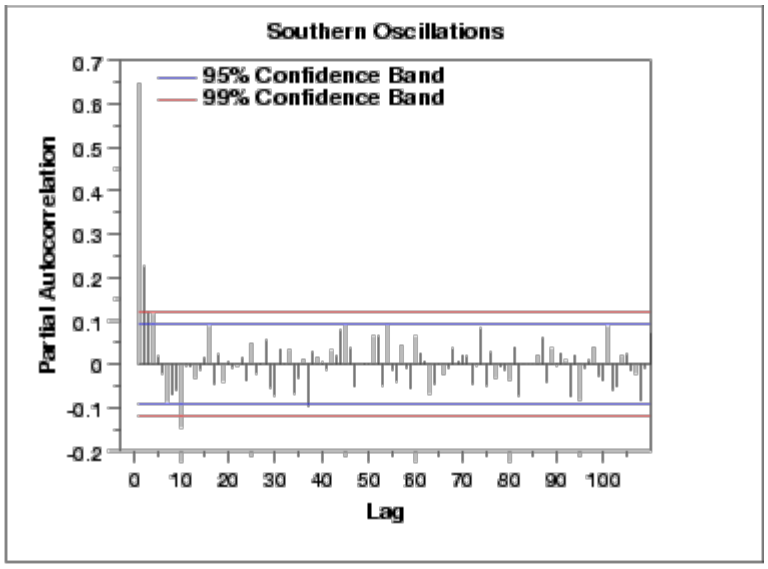
The partial autocorrelation at lag ***k*** is the autocorrelation between X_t and X_{t-k} that is not accounted for by lags 1 through ***k*-1**.

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

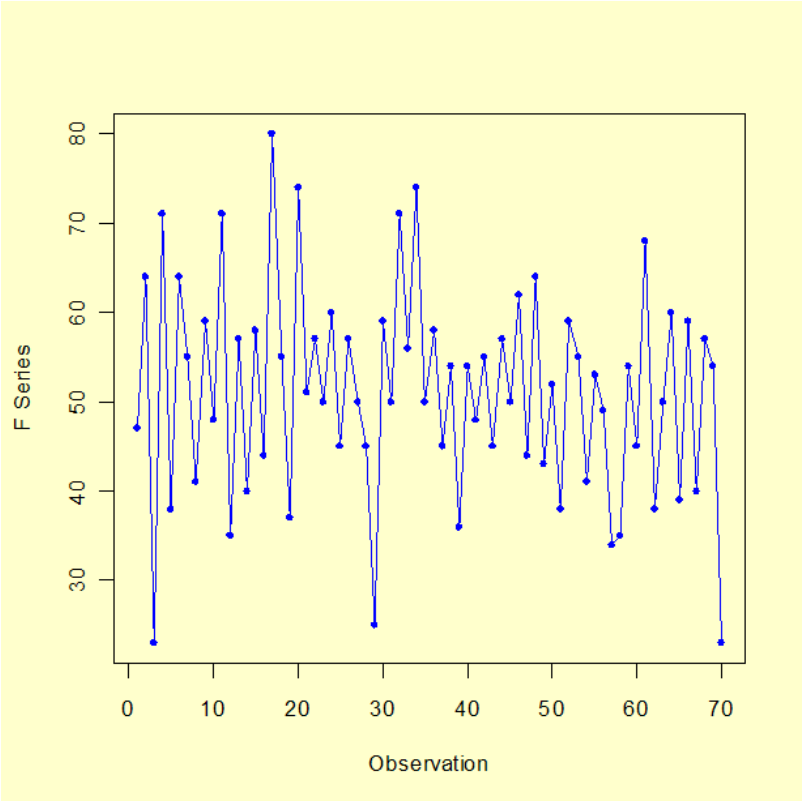
100 Bureau Drive, M/S 8900
Gaithersburg, MD 20899-8900

Sign Up for ITL Newsletters & Bulletins:
Enter email address Go

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.



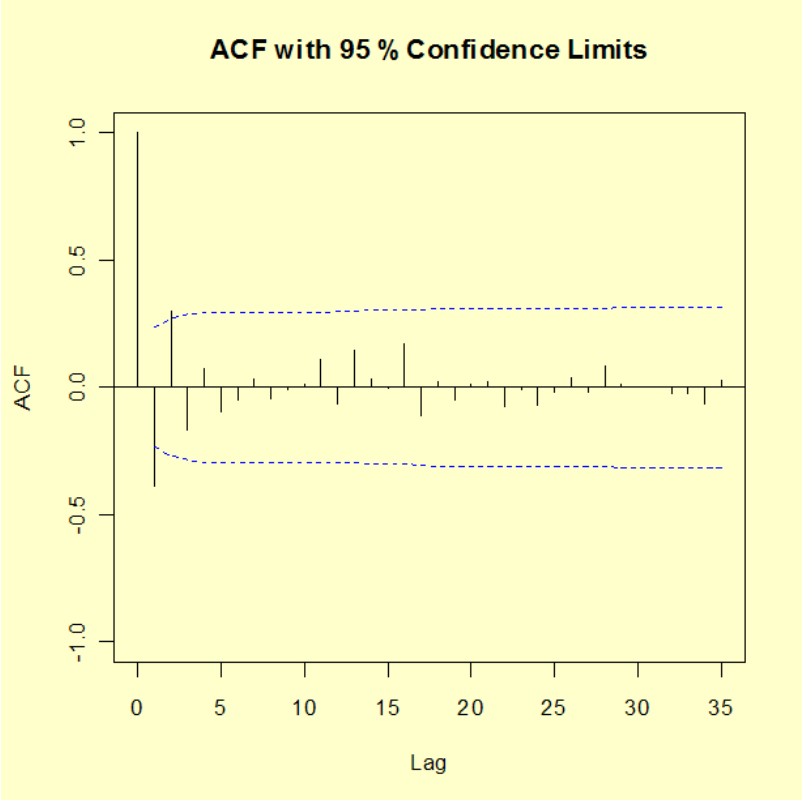
The data do not appear to have a seasonal component or a noticeable trend. (The stationarity of the series was verified by fitting a straight line to the data versus time period. The slope was not found to be significantly different from zero (p-value = 0.2).)

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

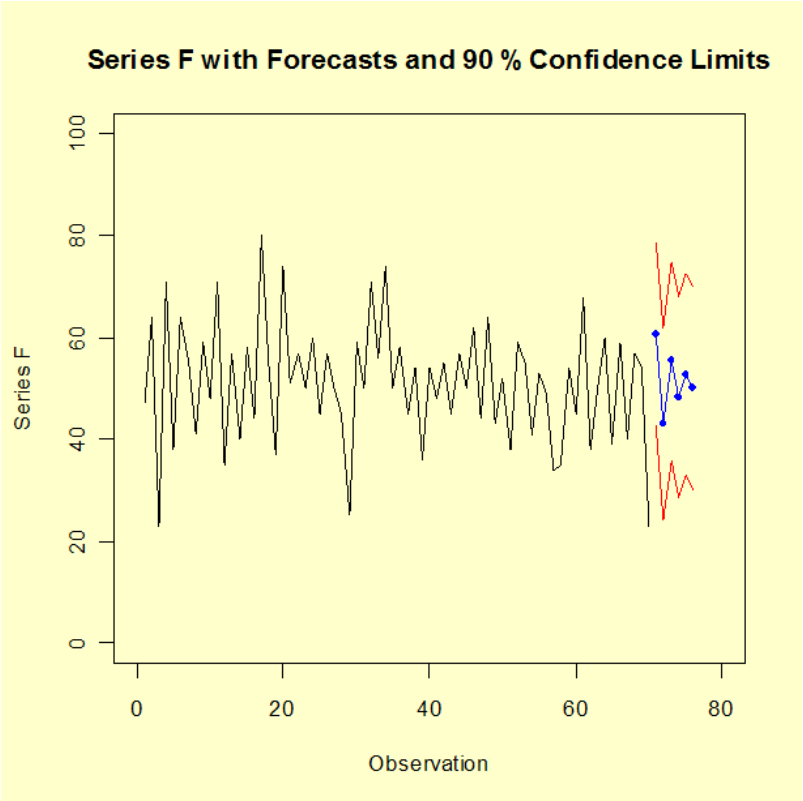
```
φ2          0.1797      0.1202

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

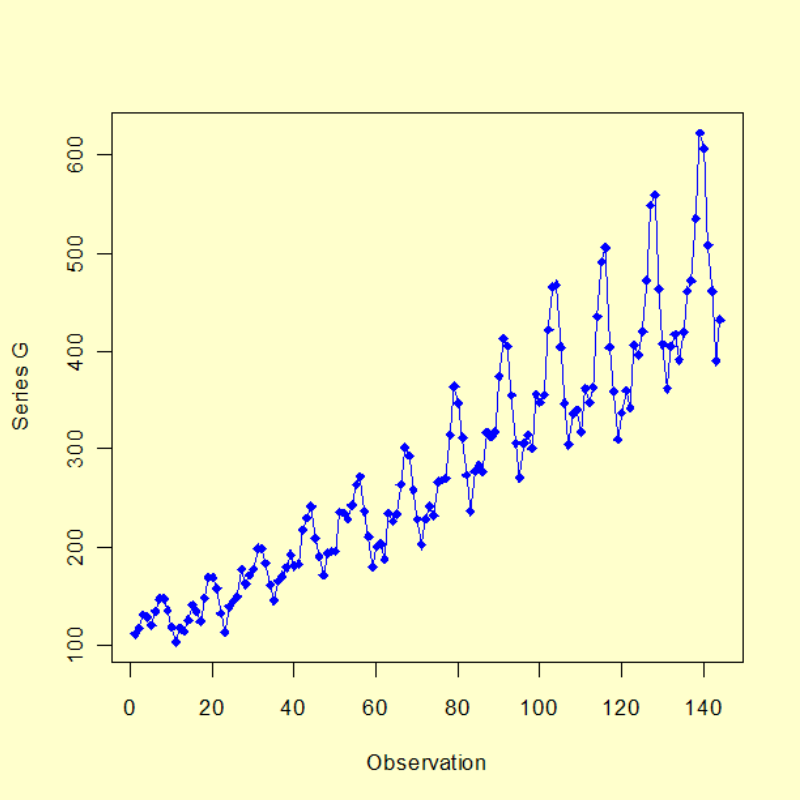
The "historical" data and forecasted values (with 90 % confidence limits) are shown in the graph below.



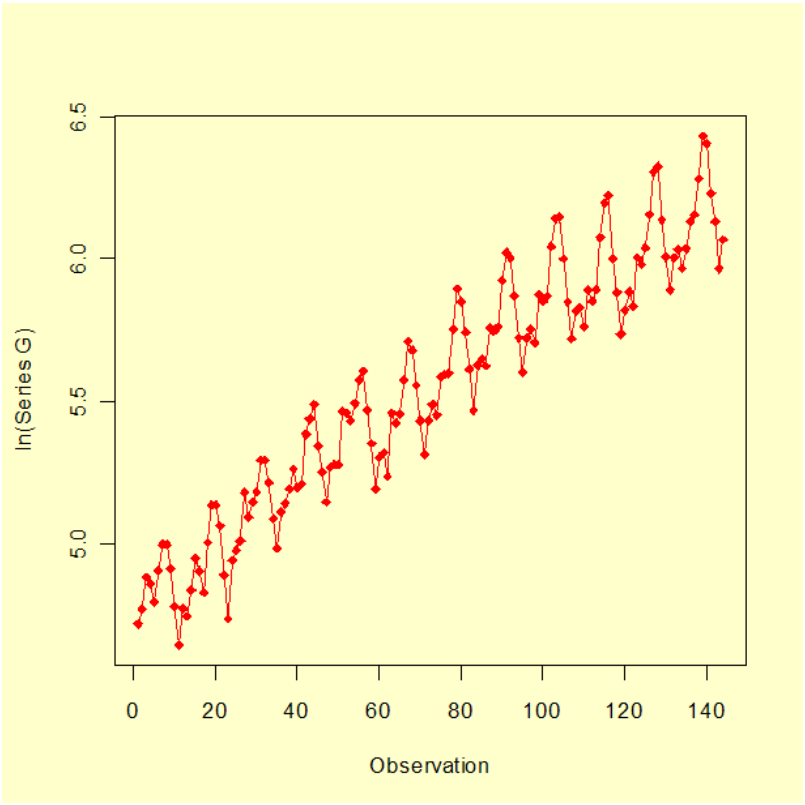
6.4.4.10. Box-Jenkins Analysis on Seasonal Data

Series G

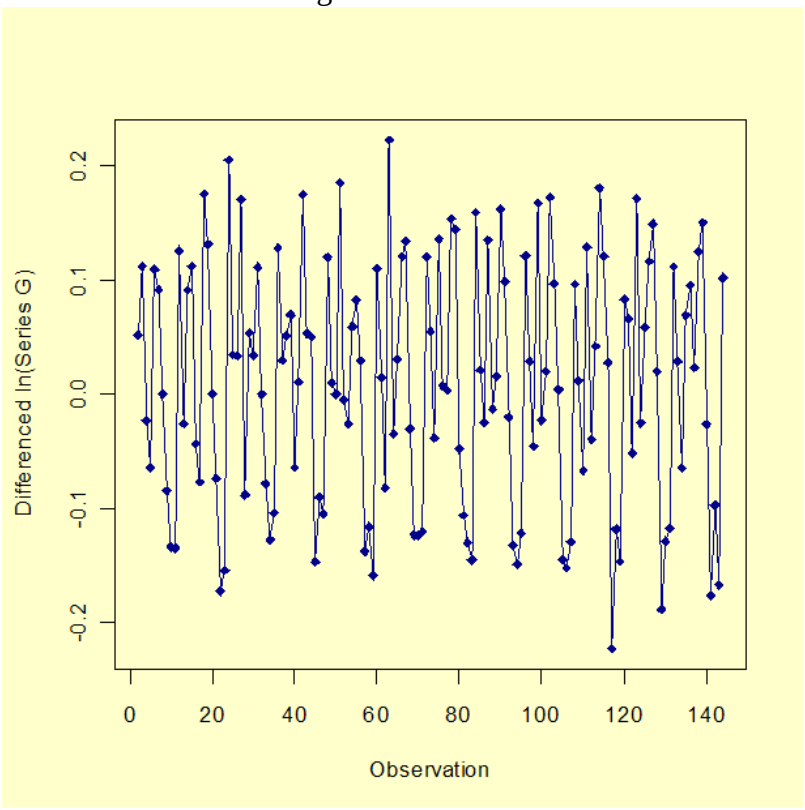
This example illustrates a Box-Jenkins time series analysis for seasonal data using the series G data set in Box, Jenkins, and Reinsel, 1994. A plot of the 144 observations is shown below.



Non-constant variance can be removed by performing a natural log transformation.

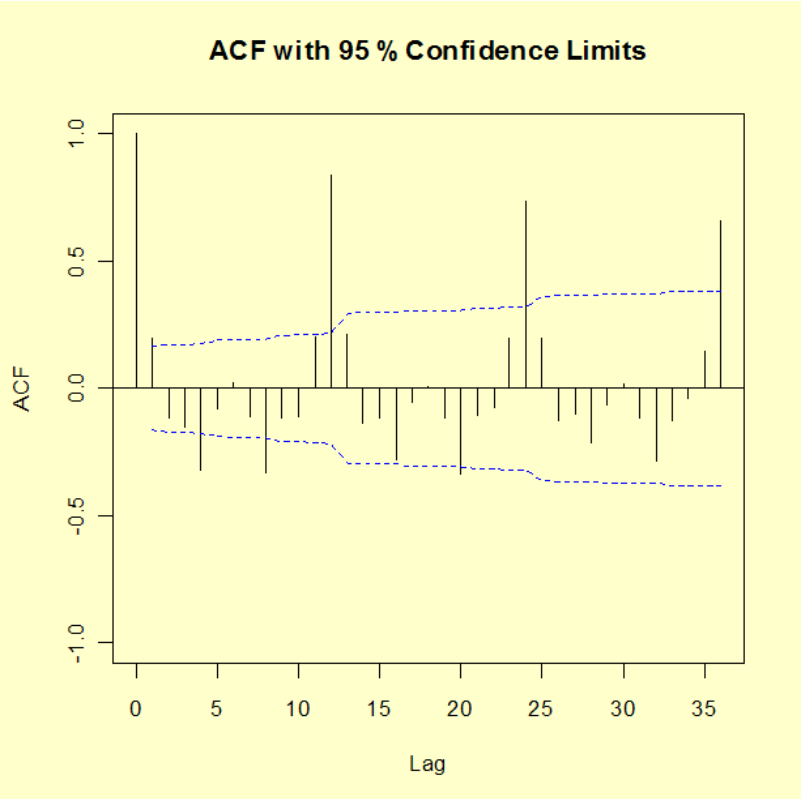


Next, we remove trend in the series by taking first differences. The resulting series is shown below.



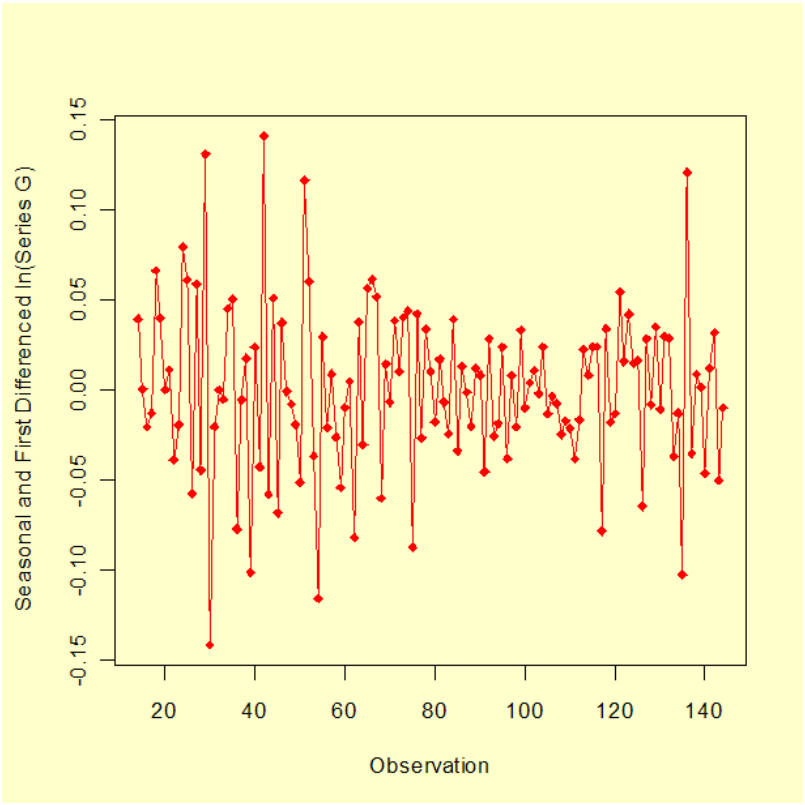
Analyzing Autocorrelation Plot for Seasonality

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.

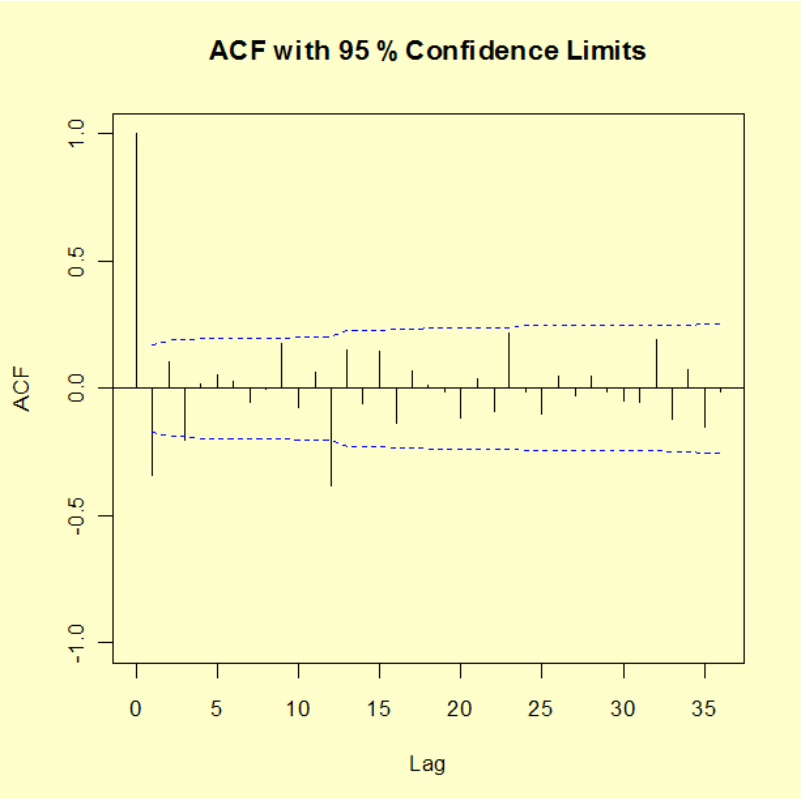
A plot of Series G after taking the natural log, first differencing, and seasonal differencing is shown below.



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 $Q > 41.3371$

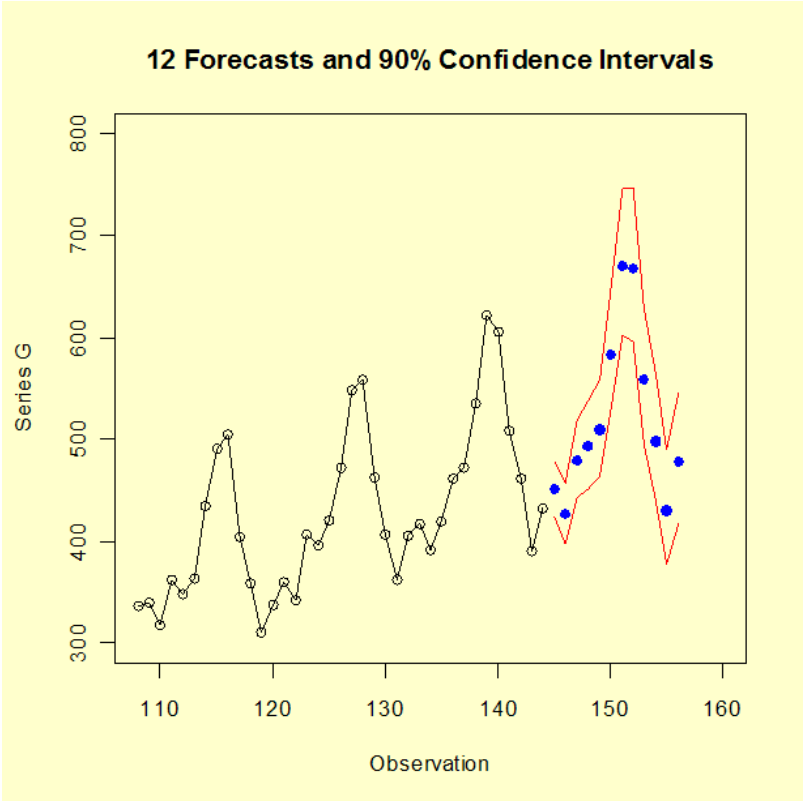
Since the null hypothesis of the Box-Ljung test is not rejected we conclude that the fitted model is adequate.

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$[\boldsymbol{a_t}]$ = $\mathbf{0}$
- $E(a_t a'_{t-k}) = 0 \qquad k \neq 0$

  $E(a_t a'_{t-k}) = \Sigma_a \qquad k = 0$

  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

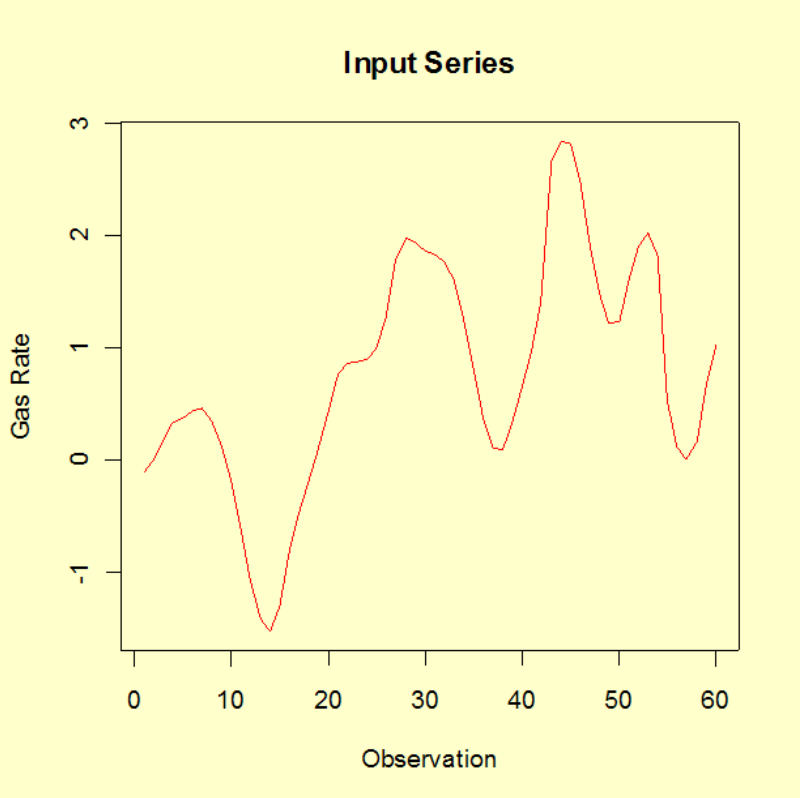
The gas furnace data from Box, Jenkins, and Reinsel, 1994 is used to illustrate the analysis of a bivariate time series. Inside the gas furnace, air and methane were combined in order to obtain a mixture of gases containing CO_2 (carbon dioxide). The input series is the methane gas feedrate described by

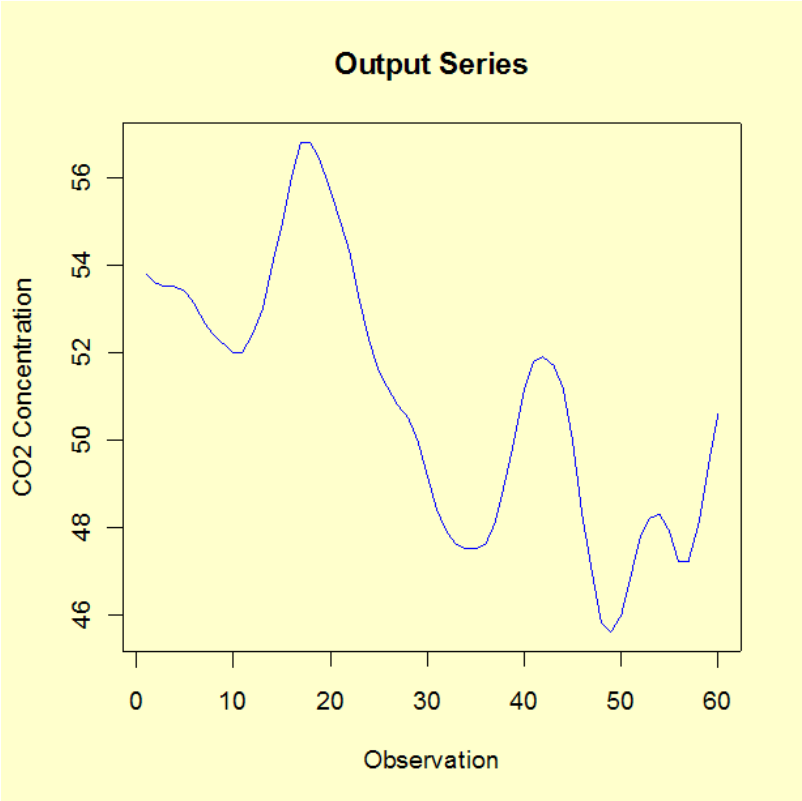
$$\text{Methane Gas Input Feed} = 0.60 - 0.04\,X(t)$$

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. Process or Product Monitoring and Control

6.5. Tutorials

Tutorial contents

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

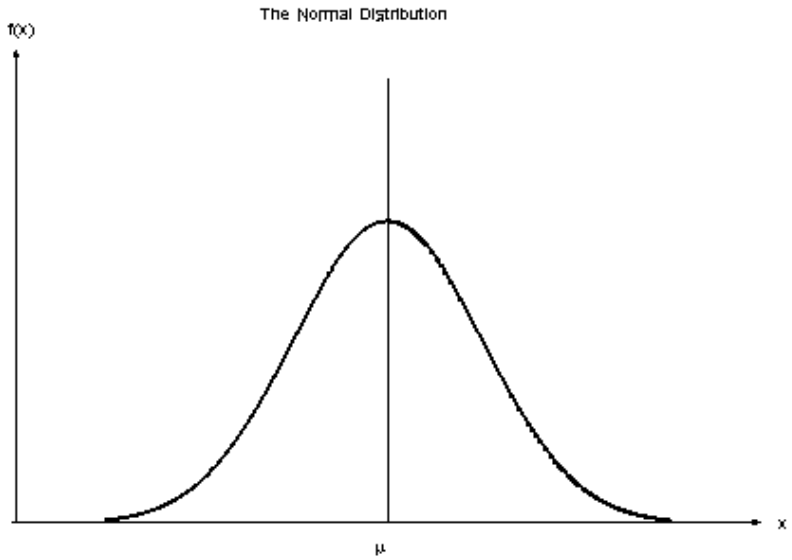
The parameters of the normal distribution are the mean μ and the standard deviation σ (or the variance σ^2). A special notation is employed to indicate that X is normally distributed with these parameters, namely

$$X \sim N(\mu, \sigma) \text{ or } X \sim N(\mu, \sigma^2).$$

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 $\mathbf{A} = (a_{ij})$ 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 $\boldsymbol{\Sigma}$ 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^* and $\bar{\bar{x}}^*$ 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

$$n(\bar{x}^{(future)} - \bar{\bar{x}}^*)'(S_p^*)^{-1}(\bar{x}^{(future)} - \bar{\bar{x}}^*)$$

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^* and $\bar{\bar{x}}^*$. 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

6.6.1. Lithography Process

Lithography Process This case study illustrates the use of control charts in analyzing a lithography process.

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
```

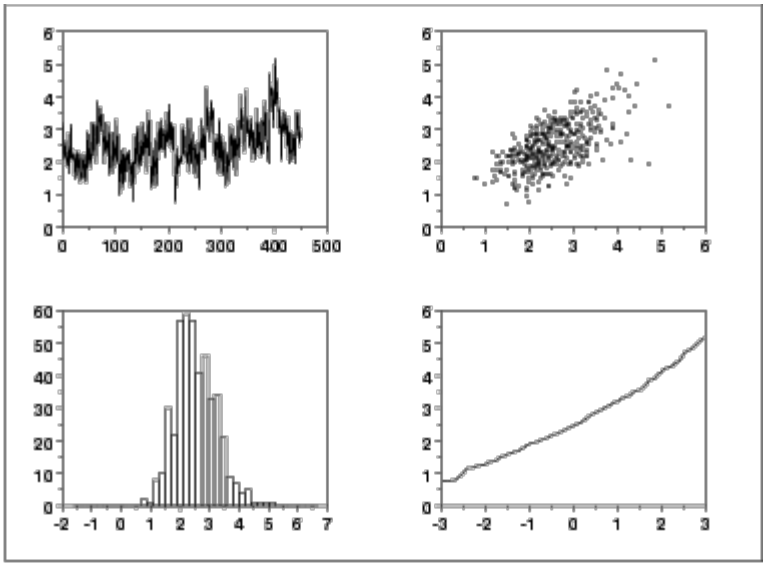
```
28          1     Bot     3.575446     410
2.755446
28          2     Top     3.024903     411
2.202903
28          2     Lef     3.099192     412
2.275192
28          2     Cen     2.048139     413
1.222139
28          2     Rgt     2.927978     414
2.099978
28          2     Bot     3.15257      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.

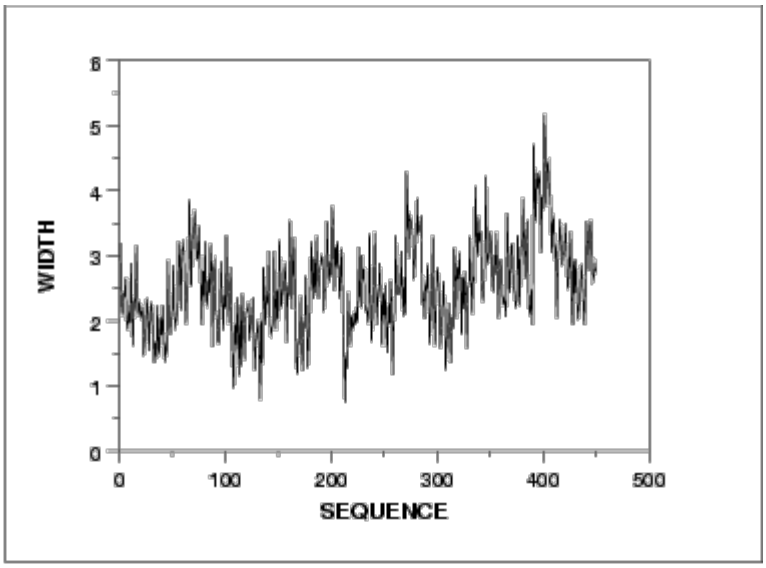
2. The lag plot (upper right) indicates that there is some mild autocorrelation in the data. This is not unexpected as the data are grouped in a logical order of the three factors (i.e., not randomly) and the run sequence plot indicates that there are factor effects.

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

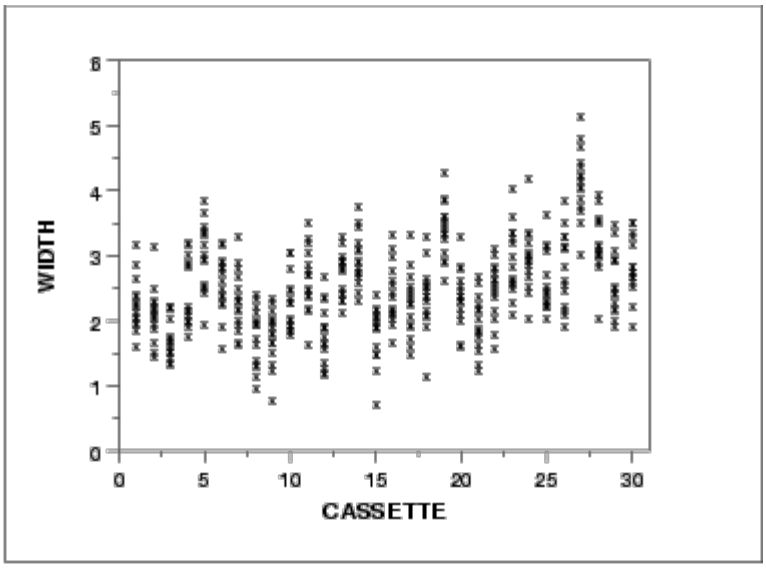
```
Sample size     = 450
Mean            =   2.53228
Median          =   2.45334
Minimum         =   0.74655
Maximum         =   5.16867
Range           =   4.42212
Stan. Dev.      =   0.69376
Autocorrelation =   0.60726
```

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.

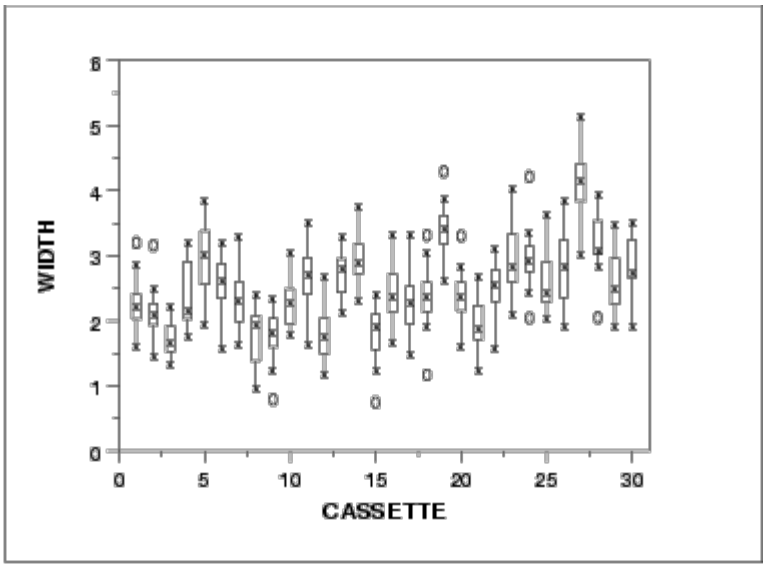
Plot response against individual factors

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

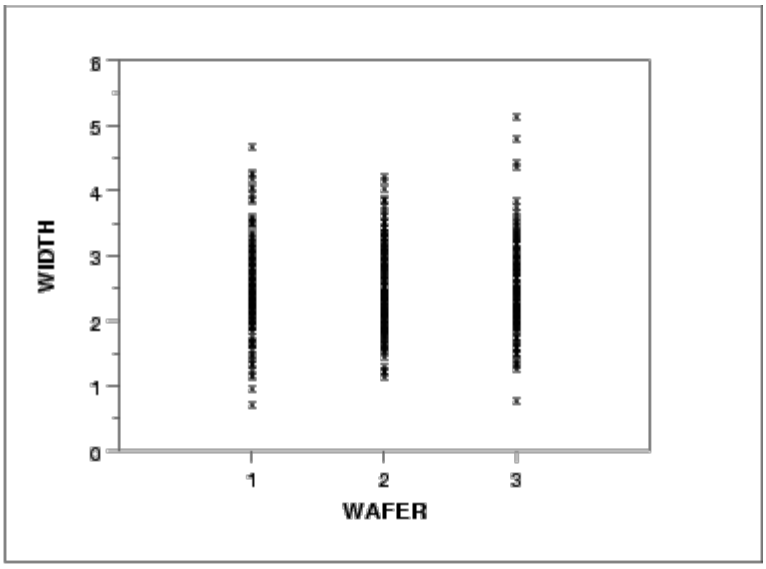


Interpretation

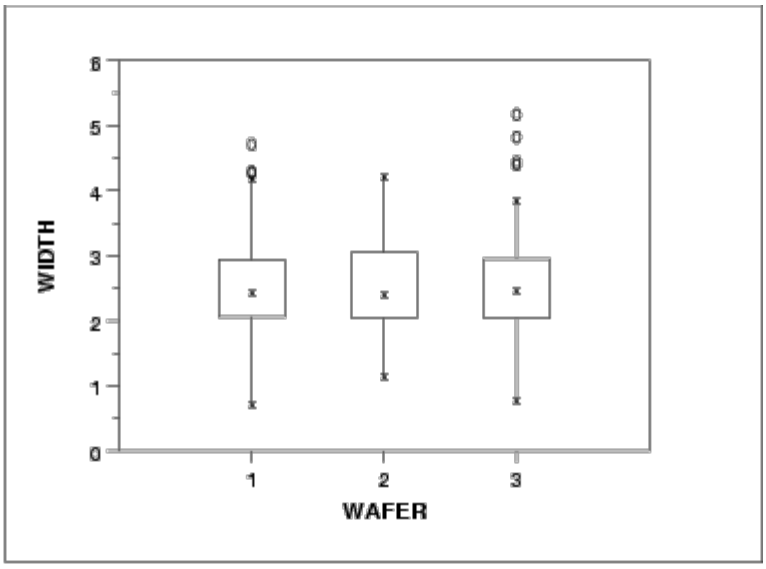
We can make the following conclusions based on the above scatter and box plots.

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

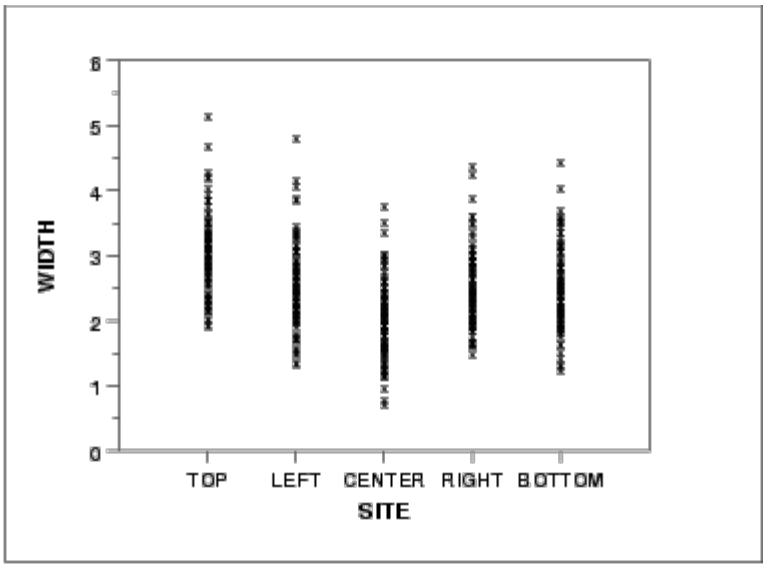


Interpretation

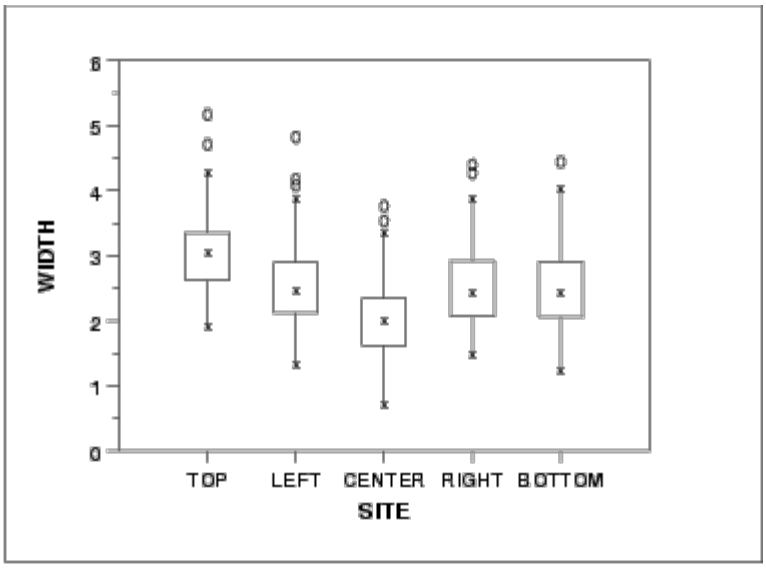
We can make the following conclusions based on the above scatter and box plots.

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

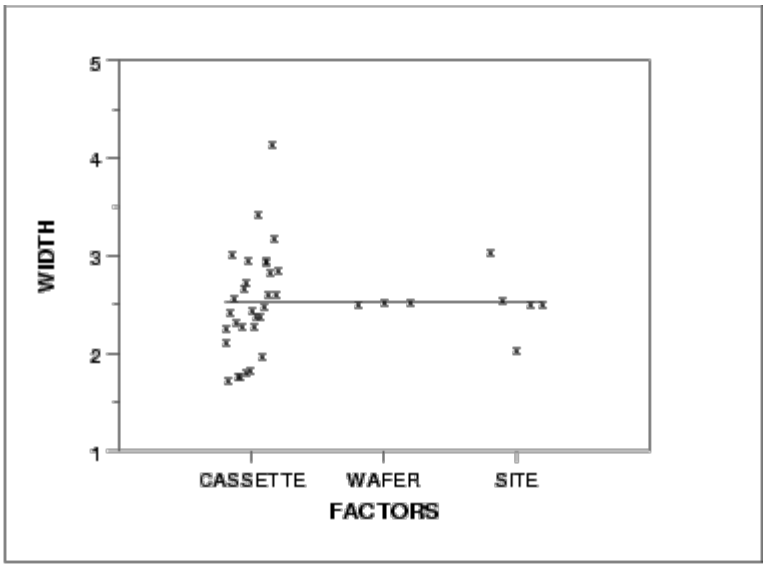
We can make the following conclusions based on the above scatter and box plots.

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.

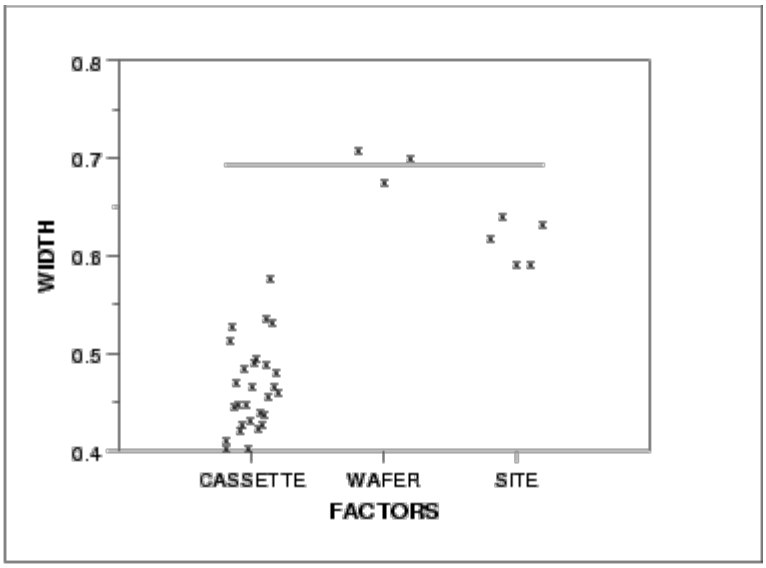
DOE mean and sd plots

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

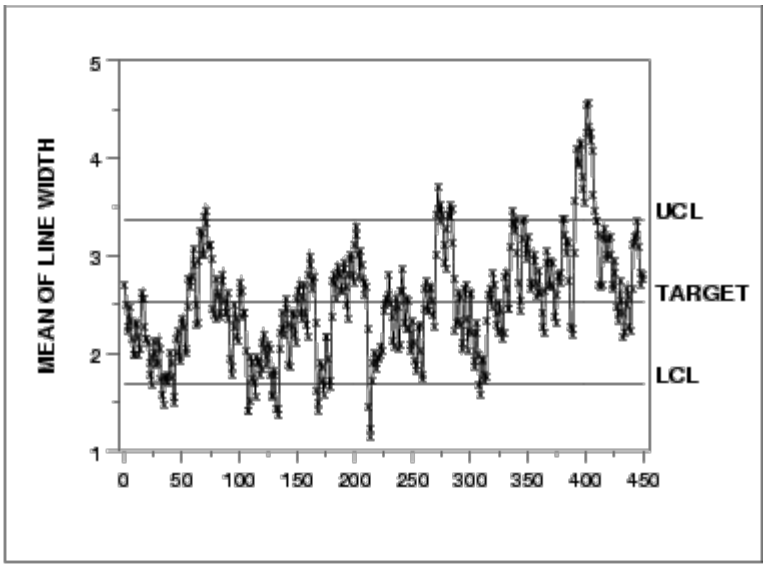
Control charts for subgroups

The resulting classical Shewhart control charts for each possible subgroup are shown below.

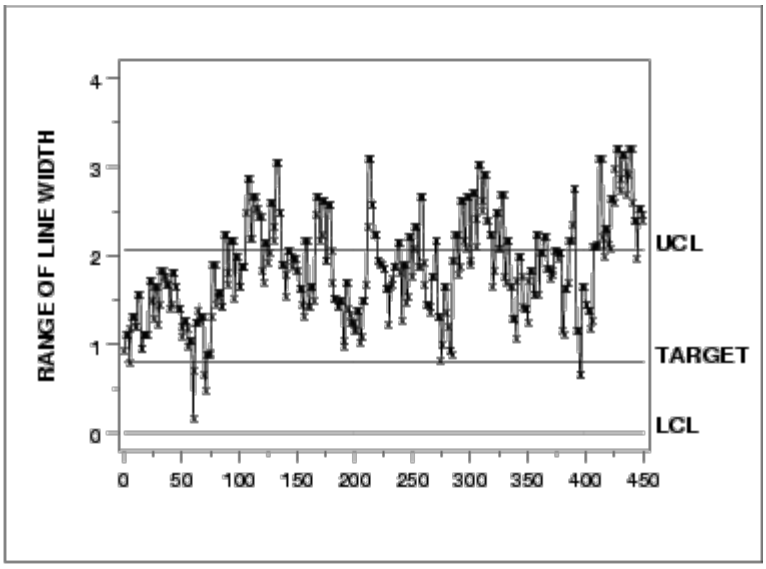
Site as subgroup

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



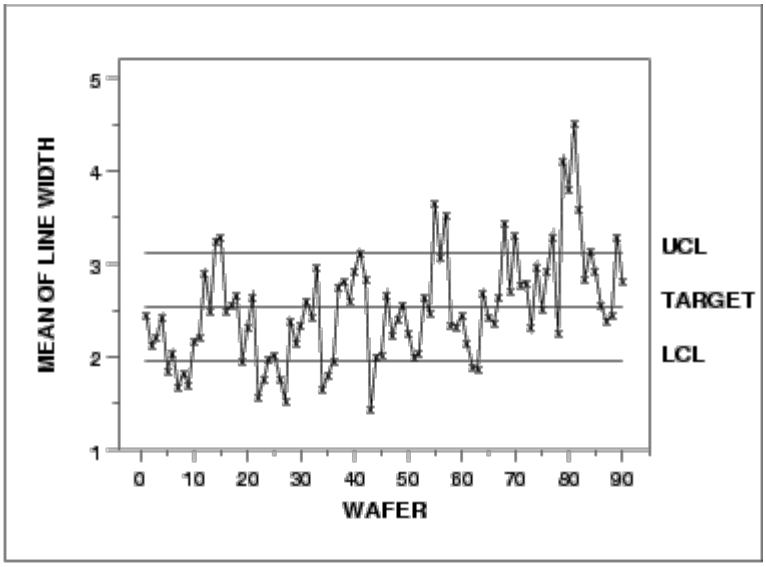
Moving range control chart



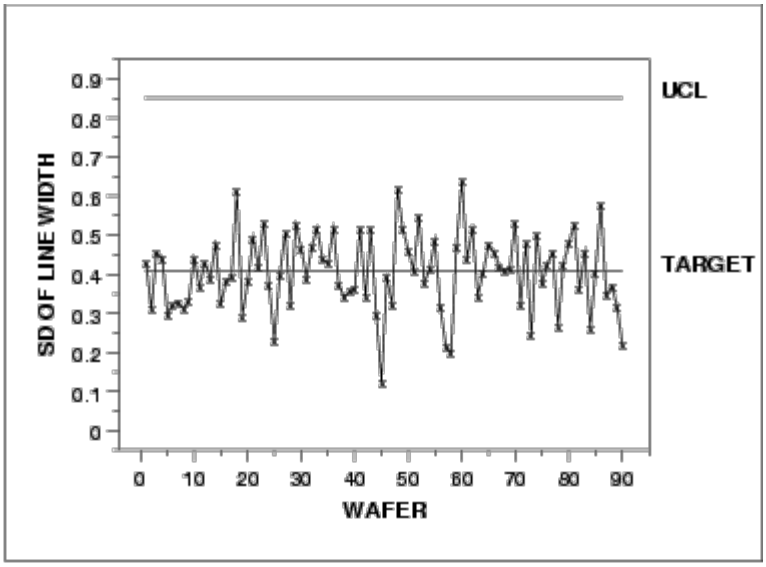
Wafer as subgroup

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

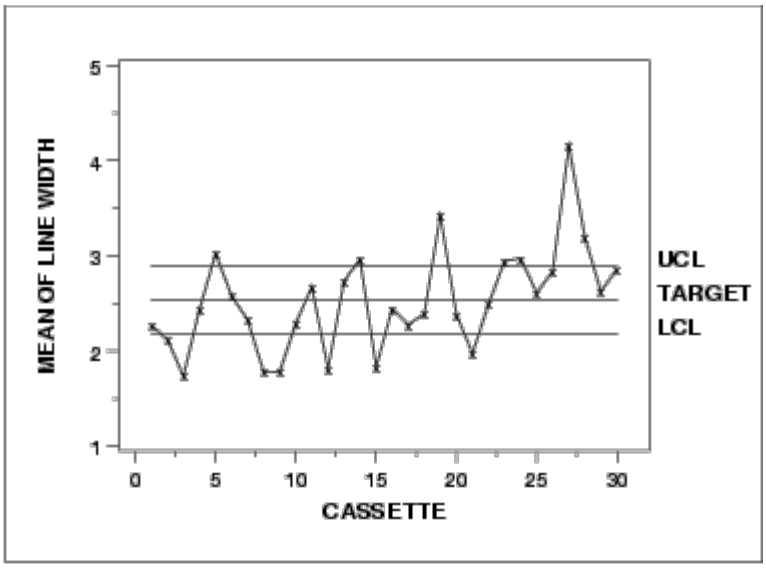


There is no LCL for the standard deviation chart because of the small subgroup size.

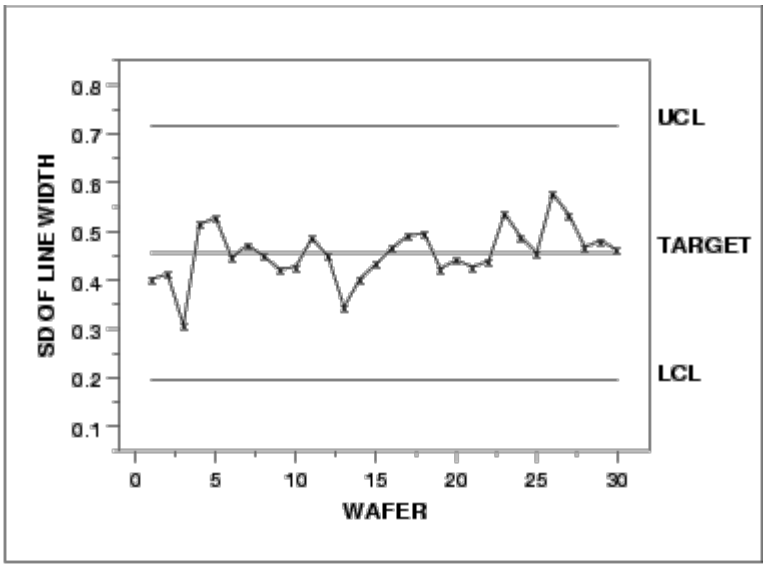
Cassette as subgroup

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

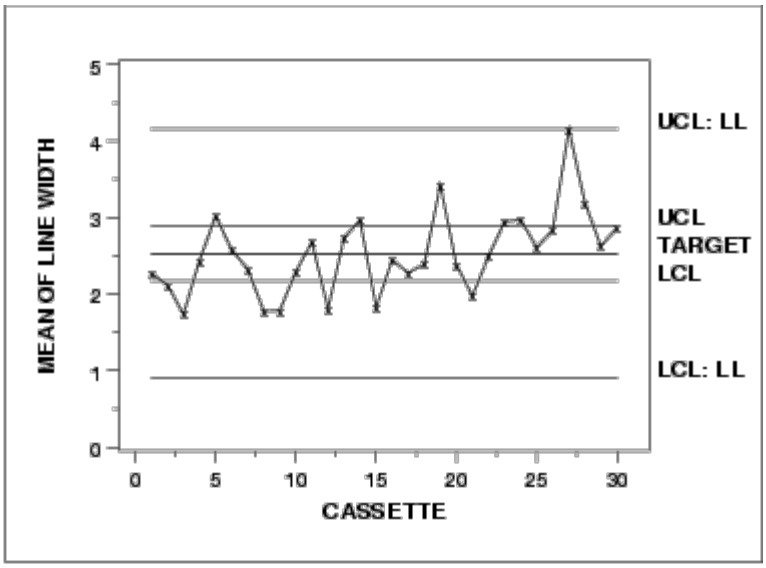
We could create a non-standard chart that would plot all the individual data values and group them together in a boxplot type format by lot. The control limits could be generated to monitor the individual data values while the lot-to-lot variation would be monitored by the patterns of the groupings. 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

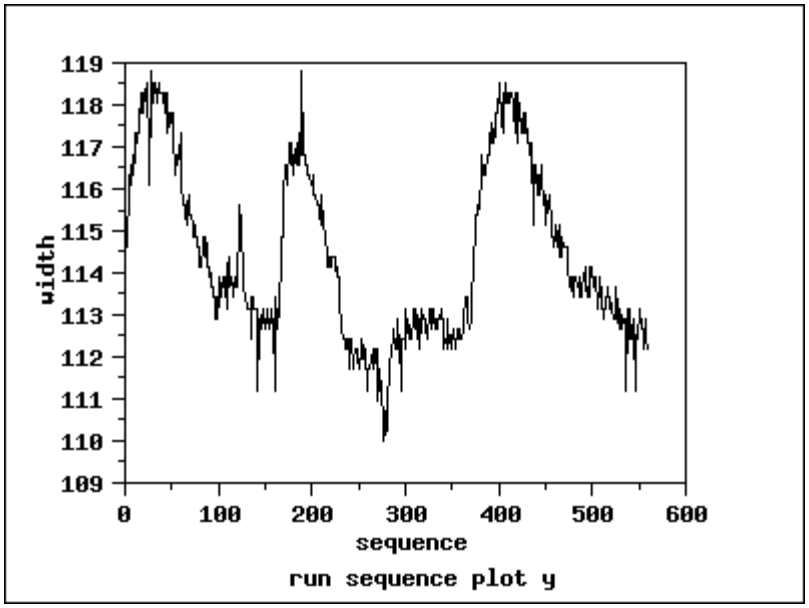
Check for Stationarity, Outliers, Seasonality

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

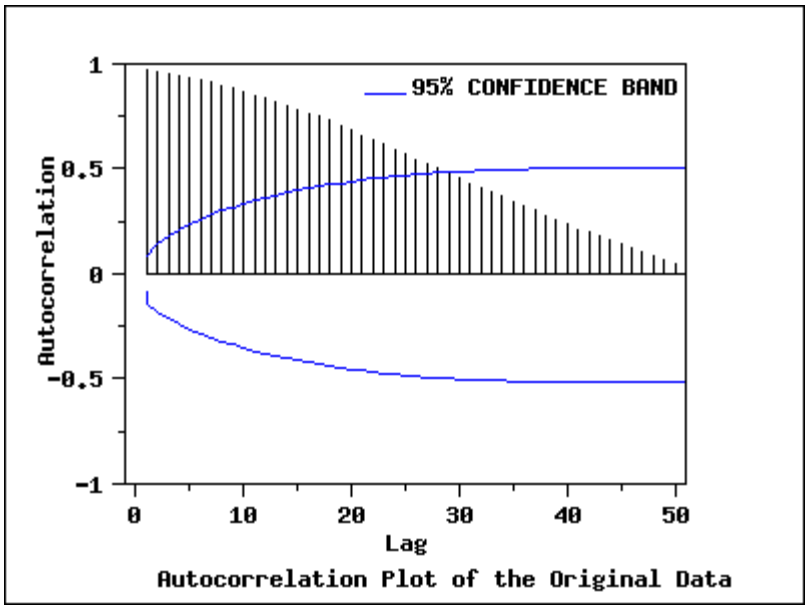
We can make the following conclusions from the run sequence plot.

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

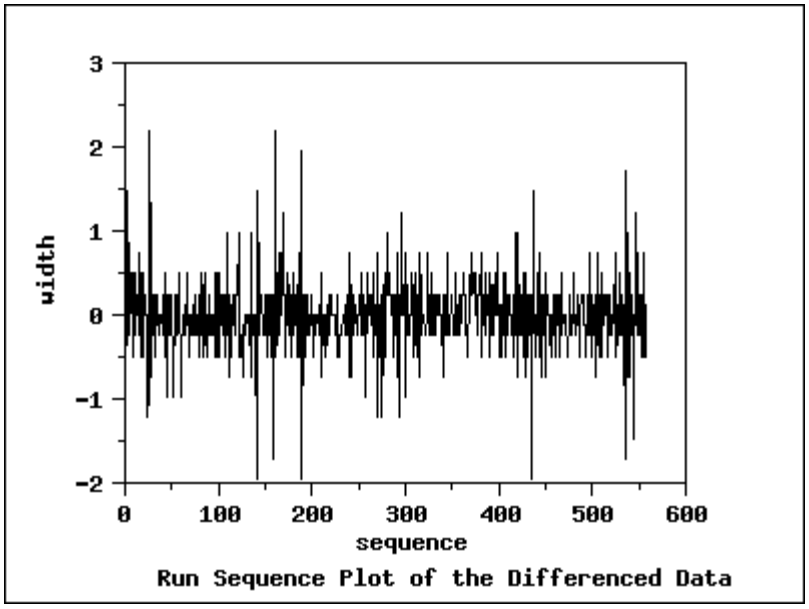


Interpretation of the Autocorrelation Plot

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



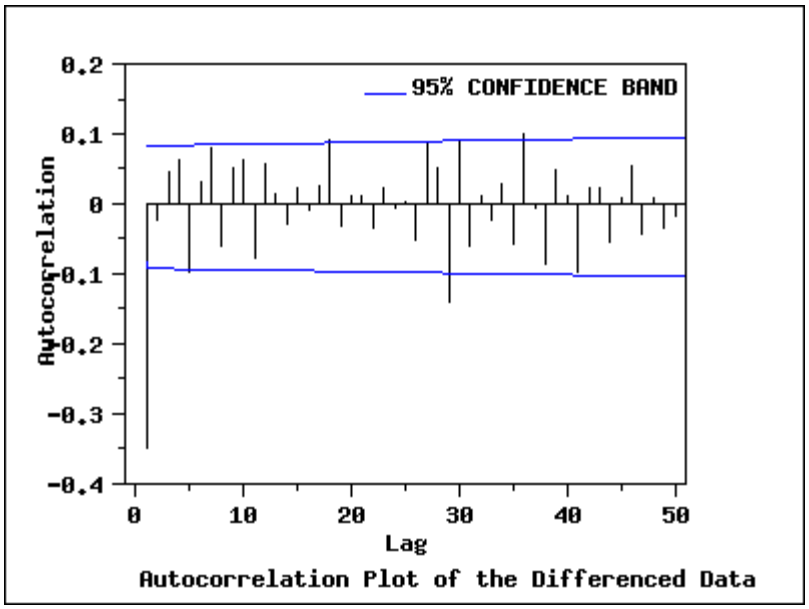
Interpretation of the Run Sequence Plot

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

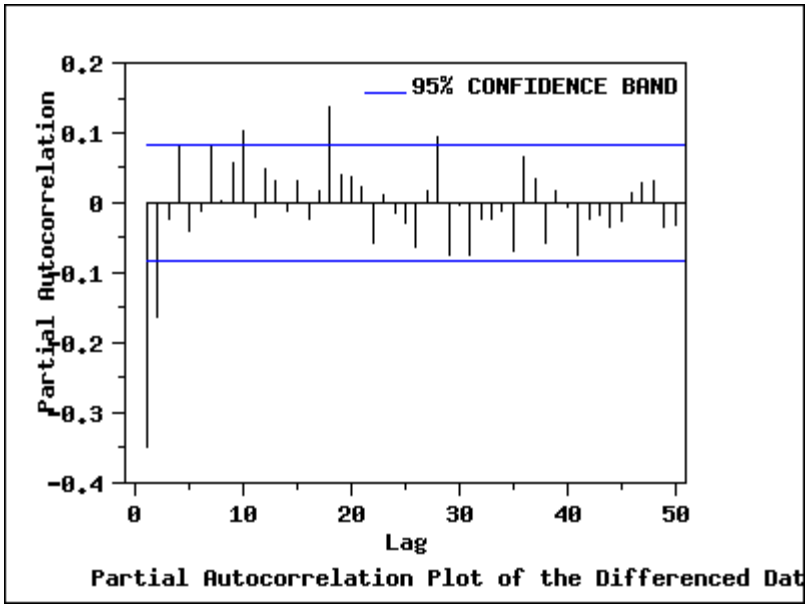


Interpretation of the Autocorrelation Plot of the Differenced Data

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.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

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

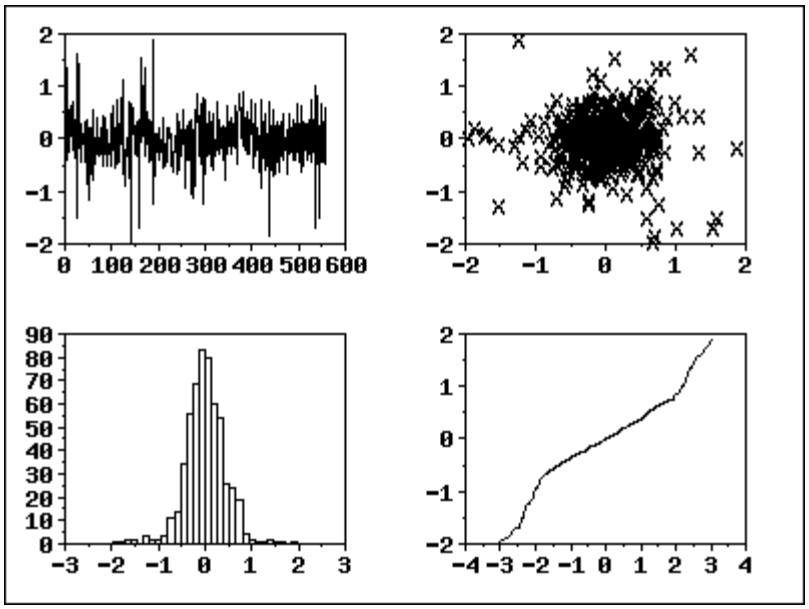
Residuals

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.

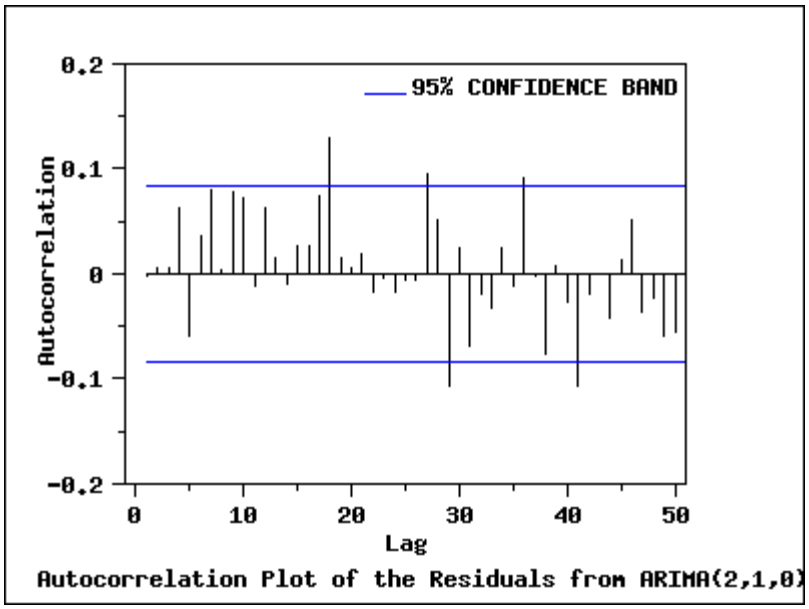
1. The run sequence plot shows that the residuals do not violate the assumption of constant location and scale. It also shows that most of the residuals are in the range (-1, 1).

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

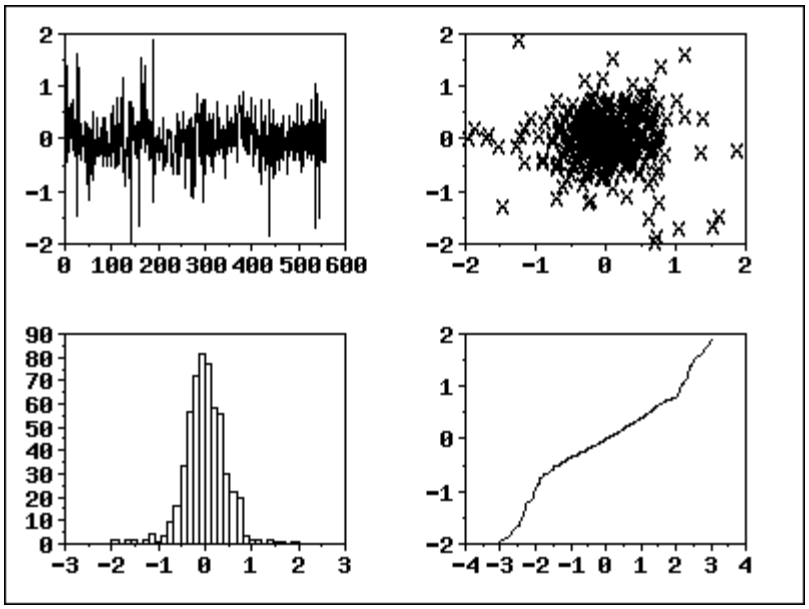
The autocorrelation plot shows that for the first 25 lags, all sample autocorrelations except those at lags 7 and 18 fall inside the 95 % confidence bounds indicating the residuals appear to be random.

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).

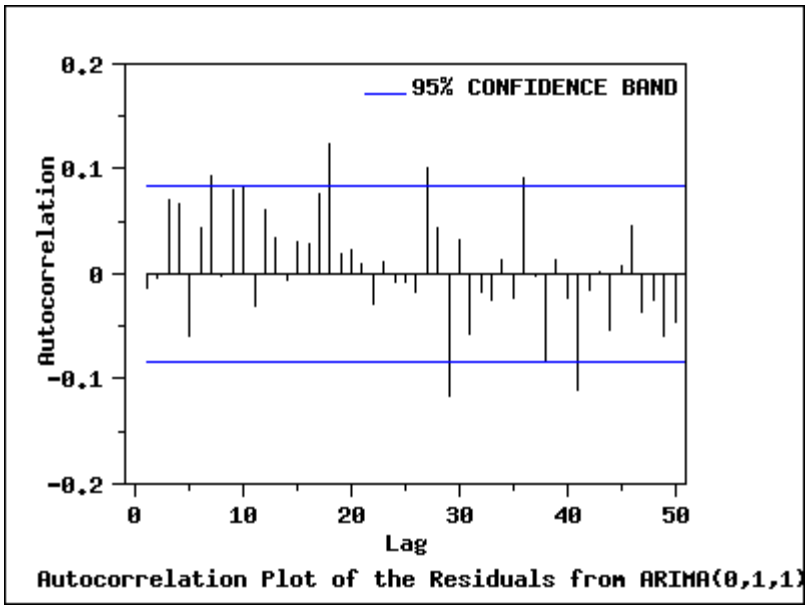
2. The lag plot indicates that the residuals are not autocorrelated at lag 1.

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.