1.3.5. Quantitative Techniques

Confirmatory Statistics

The techniques discussed in this section are classical statistical methods as opposed to EDA techniques. EDA and classical techniques are not mutually exclusive and can be used in a complamentary fashion. For example, the analysis can start with some simple graphical techniques such as the 4-plot followed by the classical confirmatory methods discussed herein to provide more rigorous statments about the conclusions. If the classical methods yield different conclusions than the graphical analysis, then some effort should be invested to explain why. Often this is an indication that some of the assumptions of the classical techniques are violated.

Many of the quantitative techniques fall into two broad categories:

1. Interval estimation
2. Hypothesis tests

Interval Estimates

It is common in statistics to estimate a parameter from a sample of data. The value of the parameter using all of the possible data, not just the sample data, is called the population parameter or true value of the parameter. An estimate of the true parameter value is made using the sample data. This is called a point estimate or a sample estimate.

For example, the most commonly used measure of location is the mean. The population, or true, mean is the sum of all the members of the given population divided by the number of members in the population. As it is typically impractical to measure every member of the population, a random sample is drawn from the population. The sample mean is calculated by summing the values in the sample and dividing by the number of values in the sample. This sample mean is then used as the point estimate of the population mean.

Interval estimates expand on point estimates by incorporating the uncertainty of the point estimate. In the example for the mean above, different samples from the same population will generate different values for the sample mean. An interval estimate quantifies this uncertainty in the sample estimate by computing lower and upper values of an interval which will, with a given level of confidence (i.e., probability), contain the population parameter.

Hypothesis Tests

Hypothesis tests also address the uncertainty of the sample estimate. However, instead of providing an interval, a hypothesis test attempts to refute a specific claim about a population parameter based on the sample data. For example, the hypothesis might be one of the following:

- the population mean is equal to 10
- the population standard deviation is equal to 5
- the means from two populations are equal
- the standard deviations from 5 populations are equal

To reject a hypothesis is to conclude that it is false. However, to accept a hypothesis does not mean that it is true, only that we do not have evidence to believe otherwise. Thus hypothesis tests are usually stated in terms of both a condition that is doubted (null hypothesis) and a condition that is believed (alternative hypothesis).

A common format for a hypothesis test is:

H_0:	A statement of the null hypothesis, e.g., two population means are equal.
H_a:	A statement of the alternative hypothesis, e.g., two population means are not equal.
Test Statistic:	The test statistic is based on the specific hypothesis test.
Significance Level:	The significance level, α, defines the sensitivity of the test. A value of α = 0.05 means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true. This is also called the type I error. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are commonly used. The probability of rejecting the null hypothesis when it is in fact false is called the power of the test and is denoted by 1 - β. Its complement, the probability of accepting the null hypothesis when the alternative hypothesis is, in fact, true (type II error), is called β and can only be computed for a specific alternative hypothesis.

Critical Region: The critical region encompasses those values of the test statistic that lead to a rejection of the null hypothesis. Based on the distribution of the test statistic and the significance level, a cut-off value for the test statistic is computed. Values either above or below or both (depending on the direction of the test) this cut-off define the critical region.

Practical Versus Statistical Significance

It is important to distinguish between statistical significance and practical significance. Statistical significance simply means that we reject the null hypothesis. The ability of the test to detect differences that lead to rejection of the null hypothesis depends on the sample size. For example, for a particularly large sample, the test may reject the null hypothesis that two process means are equivalent. However, in practice the difference between the two means may be relatively small to the point of having no real engineering significance. Similarly, if the sample size is small, a difference that is large in engineering terms may not lead to rejection of the null hypothesis. The analyst should not just blindly apply the tests, but should combine engineering judgement with statistical analysis.

Bootstrap Uncertainty Estimates

In some cases, it is possible to mathematically derive appropriate uncertainty intervals. This is particularly true for intervals based on the assumption of a normal distribution. However, there are many cases in which it is not possible to mathematically derive the uncertainty. In these cases, the bootstrap provides a method for empirically determining an appropriate interval.

Table of Contents

Some of the more common classical quantitative techniques are listed below. This list of quantitative techniques is by no means meant to be exhaustive. Additional discussions of classical statistical techniques are contained in the product comparisons chapter.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.5.1. Measures of Location

Location

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

Definition of Location

The first step is to define what we mean by a typical value. For univariate data, there are three common definitions:

1. mean - the mean is the sum of the data points divided by the number of data points. That is,

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

The mean is that value that is most commonly referred to as the average. We will use the term average as a synonym for the mean and the term typical value to refer generically to measures of location.

2. median - the median is the value of the point which has half the data smaller than that point and half the data larger than that point. That is, if X_1, X_2, ... ,X_N is a random sample sorted from smallest value to largest value, then the median is defined as:

$$\tilde{Y} = Y_{(N+1)/2} \qquad \text{if } N \text{ is odd}$$

$$\tilde{Y} = (Y_{N/2} + Y_{(N/2)+1})/2 \qquad \text{if } N \text{ is even}$$

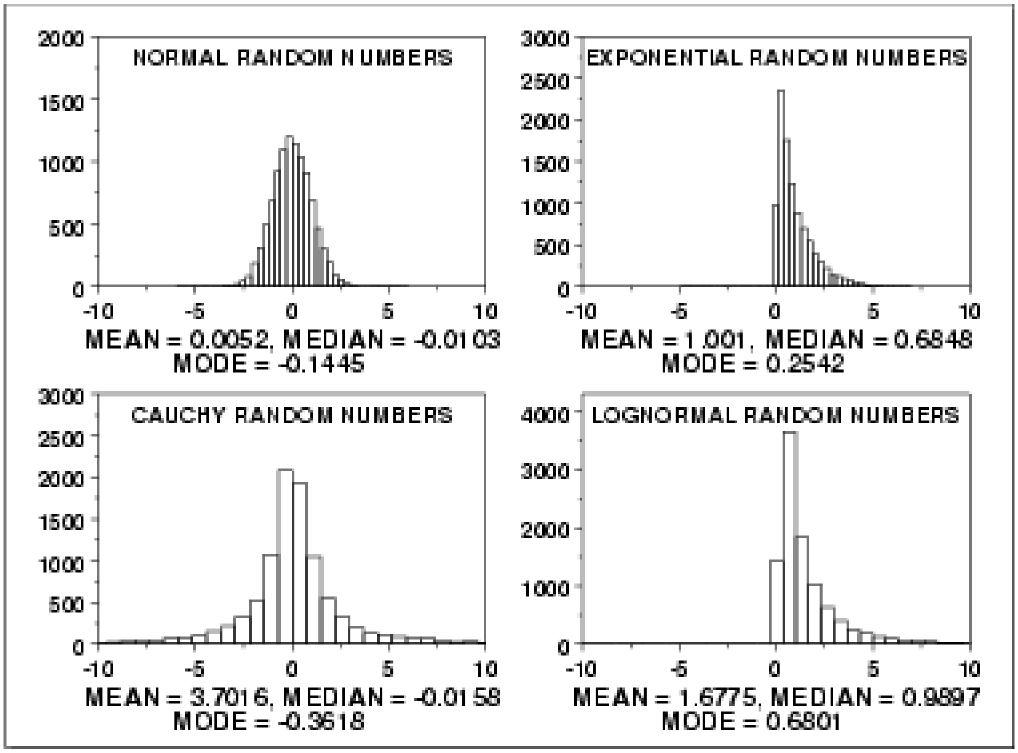
3. mode - the mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique. The mode is typically used in a qualitative fashion. For example, there may be a single dominant hump in the data perhaps two or more smaller humps in the data. This is usually evident from a histogram of the data.

When taking samples from continuous populations, we need to be somewhat careful in how we define the mode. That is, any specific value may not occur more than once if the data are continuous. What may be a more meaningful, if less exact measure, is the midpoint of the class interval of the histogram with the highest peak.

Why Different Measures

A natural question is why we have more than one measure of the typical value. The following example helps to explain why these alternative definitions are useful and necessary.

This plot shows histograms for 10,000 random numbers generated from a normal, an exponential, a Cauchy, and a lognormal distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The mean is 0.005, the median is -0.010, and the mode is -0.144 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The normal distribution is a symmetric distribution with well-behaved tails and a single peak at the center of the distribution. By symmetric, we mean that the distribution can be folded about an axis so that the 2 sides coincide. That is, it behaves the same to the left and right of some center point. For a normal distribution, the mean, median, and mode are actually equivalent. The histogram above generates similar estimates for the mean, median, and mode. Therefore, if a histogram or normal probability plot indicates that your data are approximated well by a normal distribution, then it is reasonable to use the mean as the location estimator.

Exponential Distribution

The second histogram is a sample from an exponential distribution. The mean is 1.001, the median is 0.684, and the mode is 0.254 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The exponential distribution is a skewed, i. e., not symmetric, distribution. For skewed distributions, the mean and median are not the same. The mean will be pulled in the direction of the skewness. That is, if the right tail is heavier than the left tail, the mean will be greater than the median. Likewise, if the left tail is heavier than the right tail, the mean will be less than the median.

For skewed distributions, it is not at all obvious whether the mean, the median, or the mode is the more meaningful measure of the typical value. In this case, all three measures are useful.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution. The mean is 3.70, the median is -0.016, and the mode is -0.362 (the mode is computed as the midpoint of the histogram interval with the highest peak).

For better visual comparison with the other data sets, we restricted the histogram of the Cauchy distribution to values between -10 and 10. The full Cauchy data set in fact has a minimum of approximately -29,000 and a maximum of approximately 89,000.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. The Cauchy distribution has the interesting property that collecting more data does not provide a more accurate estimate of the mean. That is, the sampling distribution of the mean is equivalent to the sampling distribution of the original data. This means that for the Cauchy distribution the mean is useless as a measure of the typical value. For this histogram, the mean of 3.7 is well above the vast majority of the data. This is caused by a few very extreme values in the tail. However, the median does provide a useful measure for the typical value.

Although the Cauchy distribution is an extreme case, it does illustrate the importance of heavy tails in measuring the mean. Extreme values in the tails distort the mean. However, these extreme values do not distort the median since the median is based on ranks. In general, for data with extreme values in the tails, the median provides a better estimate of location than does the mean.

Lognormal Distribution

The fourth histogram is a sample from a lognormal distribution. The mean is 1.677, the median is 0.989, and the mode is 0.680 (the mode is computed as the midpoint of the histogram interval with the highest peak).

The lognormal is also a skewed distribution. Therefore the mean and median do not provide similar estimates for the location. As with the exponential distribution, there is no obvious answer to the question of which is the more meaningful measure of location.

Robustness

There are various alternatives to the mean and median for measuring location. These alternatives were developed to address non-normal data since the mean is an optimal estimator if in fact your data are normal.

Tukey and Mosteller defined two types of robustness where robustness is a lack of susceptibility to the effects of nonnormality.

1. Robustness of validity means that the confidence intervals for the population location have a 95% chance of covering the population location regardless of what the underlying distribution is.
2. Robustness of efficiency refers to high effectiveness in the face of non-normal tails. That is, confidence intervals for the population location tend to be almost as narrow as the best that could be done if we knew the true shape of the distributuion.

The mean is an example of an estimator that is the best we can do if the underlying distribution is normal. However, it lacks robustness of validity. That is, confidence intervals based on the mean tend not to be precise if the underlying distribution is in fact not normal.

The median is an example of a an estimator that tends to have robustness of validity but not robustness of efficiency.

The alternative measures of location try to balance these two concepts of robustness. That is, the confidence intervals for the case when the data are normal should be almost as narrow as the confidence intervals based on the mean. However, they should maintain their validity even if the underlying data are not normal. In particular, these alternatives address the problem of heavy-tailed distributions.

Alternative Measures of Location

A few of the more common alternative location measures are:

1. Mid-Mean - computes a mean using the data between the 25th and 75th percentiles.
2. Trimmed Mean - similar to the mid-mean except different percentile values are used. A common choice is to trim 5% of the points in both the lower and upper tails, i.e., calculate the mean for data between the 5th and 95th percentiles.
3. Winsorized Mean - similar to the trimmed mean. However, instead of trimming the points, they are set to the lowest (or highest) value. For example, all data below the 5th percentile are set equal to the value of the 5th percentile and all data greater than the 95th percentile are set equal to the 95th percentile.
4. Mid-range = (smallest + largest)/2.

The first three alternative location estimators defined above have the advantage of the median in the sense that they are not unduly affected by extremes in the tails. However, they generate estimates that are closer to the mean for data that are normal (or nearly so).

The mid-range, since it is based on the two most extreme points, is not robust. Its use is typically restricted to situations in which the behavior at the extreme points is relevant.

Case Study

The uniform random numbers case study compares the performance of several different location estimators for a particular non-normal distribution.

Software

Most general purpose statistical software programs, including Dataplot, can compute at least some of the measures of location discussed above.

1.3.5.2. Confidence Limits for the Mean

Purpose: Interval Estimate for Mean

Confidence limits for the mean (Snedecor and Cochran, 1989) are an interval estimate for the mean. Interval estimates are often desirable because the estimate of the mean varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a lower and upper limit for the mean. The interval estimate gives an indication of how much uncertainty there is in our estimate of the true mean. The narrower the interval, the more precise is our estimate.

Confidence limits are expressed in terms of a confidence coefficient. Although the choice of confidence coefficient is somewhat arbitrary, in practice 90%, 95%, and 99% intervals are often used, with 95% being the most commonly used.

As a technical note, a 95% confidence interval does **not** mean that there is a 95% probability that the interval contains the true mean. The interval computed from a given sample either contains the true mean or it does not. Instead, the level of confidence is associated with the method of calculating the interval. The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean. That is, for a 95% confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95% of these intervals would contain the true mean.

Definition: Confidence Interval

Confidence limits are defined as:

$$\bar{Y} \pm t_{(\alpha/2,N-1)} s/\sqrt{N}$$

where $\bar{Y}$ is the sample mean, s is the sample standard deviation, N is the sample size, α is the desired significance level, and $t_{(\alpha/2,N-1)}$ is the upper critical value of the t distribution with N - 1 degrees of freedom. Note that the confidence coefficient is 1 - α.

From the formula, it is clear that the width of the interval is controlled by two factors:

1. As N increases, the interval gets narrower from the $\sqrt{N}$ term.

   That is, one way to obtain more precise estimates for the mean is to increase the sample size.
2. The larger the sample standard deviation, the larger the confidence interval.

This simply means that noisy data, i.e., data with a large standard deviation, are going to generate wider intervals than data with a smaller standard deviation.

Definition: Hypothesis Test

To test whether the population mean has a specific value, μ_0, against the two-sided alternative that it does not have a value μ_0, the confidence interval is converted to hypothesis-test form. The test is a one-sample *t*-test, and it is defined as:

H_0:	$\mu = \mu_0$
H_a:	$\mu \neq \mu_0$
Test Statistic:	$T = (\bar{Y} - \mu_0)/(s/\sqrt{N})$ where $\bar{Y}$, N, and s are defined as above.
Significance Level:	α. The most commonly used value for α is 0.05.
Critical Region:	Reject the null hypothesis that the mean is a specified value, μ_0, if $$T < -t_{(\alpha/2, N-1)}$$ or $$T > t_{(\alpha/2, N-1)}$$

Sample Output for Confidence Interval

Dataplot generated the following output for a confidence interval from the ZARR13.DAT data set:

```
                 CONFIDENCE LIMITS FOR MEAN
                         (2-SIDED)

        NUMBER OF OBSERVATIONS     =      195
        MEAN                       =    9.261460
        STANDARD DEVIATION         =   0.2278881E-01
        STANDARD DEVIATION OF MEAN =   0.1631940E-02

 CONFIDENCE    T     T X SD(MEAN)     LOWER        UPPER
 VALUE (%)   VALUE                    LIMIT        LIMIT
---------------------------------------------------------
    50.000   0.676  0.110279E-02   9.26036      9.26256
    75.000   1.154  0.188294E-02   9.25958      9.26334
    90.000   1.653  0.269718E-02   9.25876      9.26416
    95.000   1.972  0.321862E-02   9.25824      9.26468
    99.000   2.601  0.424534E-02   9.25721      9.26571
    99.900   3.341  0.545297E-02   9.25601      9.26691
    99.990   3.973  0.648365E-02   9.25498      9.26794
    99.999   4.536  0.740309E-02   9.25406      9.26886
```

Interpretation of the Sample Output

The first few lines print the sample statistics used in calculating the confidence interval. The table shows the confidence interval for several different significance levels. The first column lists the confidence level (which is 1 - α expressed as a percent), the second column lists the t-value (i.e., $t_{(\alpha/2,N-1)}$), the third column lists the t-value times the standard error (the standard error is $s/\sqrt{N}$), the fourth column lists the lower confidence limit, and the fifth column lists the upper confidence limit. For example, for a 95% confidence interval, we go to the row identified by 95.000 in the first column and extract an interval of (9.25824, 9.26468) from the last two columns.

Output from other statistical software may look somewhat different from the above output.

Sample Output for t Test

Dataplot generated the following output for a one-sample *t*-test from the ZARR13.DAT data set:

```
                          T TEST
                        (1-SAMPLE)
                    MU0 =     5.000000
 NULL HYPOTHESIS UNDER TEST--MEAN MU = 5.000000

 SAMPLE:
    NUMBER OF OBSERVATIONS      =      195
    MEAN                        =    9.261460
    STANDARD DEVIATION          =   0.2278881E-01
    STANDARD DEVIATION OF MEAN  =   0.1631940E-02

 TEST:
    MEAN-MU0                    =    4.261460
    T TEST STATISTIC VALUE      =    2611.284
    DEGREES OF FREEDOM          =    194.0000
    T TEST STATISTIC CDF VALUE  =    1.000000

                 ALTERNATIVE-          ALTERNATIVE-
  ALTERNATIVE-   HYPOTHESIS            HYPOTHESIS
  HYPOTHESIS     ACCEPTANCE INTERVAL   CONCLUSION
 MU <> 5.000000   (0,0.025) (0.975,1)    ACCEPT
 MU  < 5.000000   (0,0.05)               REJECT
 MU  > 5.000000   (0.95,1)               ACCEPT
```

Interpretation of Sample Output

We are testing the hypothesis that the population mean is 5. The output is divided into three sections.

1. The first section prints the sample statistics used in the computation of the *t*-test.
2. The second section prints the *t*-test statistic value, the degrees of freedom, and the cumulative distribution function (cdf) value of the *t*-test statistic. The *t*-test statistic cdf value is an alternative way of expressing the critical value. This cdf value is compared to the acceptance intervals printed in section three. For an upper one-tailed test, the alternative hypothesis acceptance interval is $(1 - \alpha, 1)$, the alternative hypothesis acceptance interval for a lower one-tailed test is $(0, \alpha)$, and the alternative hypothesis acceptance interval for a two-tailed test is $(1 - \alpha/2, 1)$ or $(0, \alpha/2)$. Note that accepting the alternative hypothesis is equivalent to rejecting the null hypothesis.
3. The third section prints the conclusions for a 95% test since this is the most common case. Results are given in terms of the alternative hypothesis for the two-tailed test and for the one-tailed test in both directions. The alternative hypothesis acceptance interval column is stated in terms of the cdf value printed in section two. The last column specifies whether the alternative hypothesis is accepted or rejected. For a different significance level, the appropriate conclusion can be drawn from the *t*-test statistic cdf value printed in section two. For example, for a significance level of 0.10, the corresponding alternative hypothesis acceptance intervals are (0,0.05) and (0.95,1), (0, 0.10), and (0.90,1).

Output from other statistical software may look somewhat different from the above output.

Questions

Confidence limits for the mean can be used to answer the following questions:

1. What is a reasonable estimate for the mean?
2. How much variability is there in the estimate of the mean?
3. Does a given target value fall within the confidence limits?

Related Techniques

Two-Sample T-Test

Confidence intervals for other location estimators such as the median or mid-mean tend to be mathematically difficult or intractable. For these cases, confidence intervals can be obtained using the bootstrap.

Case Study

Heat flow meter data.

Software

Confidence limits for the mean and one-sample t-tests are available in just about all general purpose statistical software programs, including Dataplot.

1.3.5.3. Two-Sample *t*-Test for Equal Means

Purpose: Test if two population means are equal

The two-sample *t*-test (Snedecor and Cochran, 1989) is used to determine if two population means are equal. A common application of this is to test if a new process or treatment is superior to a current process or treatment.

There are several variations on this test.

1. The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are the two samples, then X_i corresponds to Y_i. For paired samples, the difference $X_i - Y_i$ is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.
2. The variances of the two samples may be assumed to be equal or unequal. Equal variances yields somewhat simpler formulas, although with computers this is no longer a significant issue.
3. The null hypothesis might be that the two population means are not equal ($\mu_1 \neq \mu_2$). If so, this must be converted to the form that the difference between the two population means is equal to some constant ($\mu_1 - \mu_2 = d_0$). This form might be preferred if you only want to adopt a new process or treatment if it exceeds the current treatment by some threshold value.

Definition

The two sample t test for unpaired data is defined as:

H$_0$: $\mu_1 = \mu_2$

H$_a$: $\mu_1 \ne \mu_2$

Test Statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

where N_1 and N_2 are the sample sizes, $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances.

If equal variances are assumed, then the formula reduces to:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Significance Level: α.

Critical Region: Reject the null hypothesis that the two means are equal if

$$T < -t_{(\alpha/2,\nu)}$$

or

$$T > t_{(\alpha/2,\nu)}$$

where $t_{(\alpha/2,\nu)}$ is the critical value of the t distribution with ν degrees of freedom where

$$\nu = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

If equal variances are assumed, then

$$\nu = N_1 + N_2 - 2$$

Sample Output

Dataplot generated the following output for the t test from the AUTO83B.DAT data set:

```
                            T TEST
                          (2-SAMPLE)
 NULL HYPOTHESIS UNDER TEST--POPULATION MEANS MU1 = MU2

 SAMPLE 1:
    NUMBER OF OBSERVATIONS      =      249
    MEAN                        =    20.14458
    STANDARD DEVIATION          =    6.414700
    STANDARD DEVIATION OF MEAN  =   0.4065151

 SAMPLE 2:
    NUMBER OF OBSERVATIONS      =       79
    MEAN                        =    30.48101
    STANDARD DEVIATION          =    6.107710
    STANDARD DEVIATION OF MEAN  =   0.6871710

 IF     ASSUME SIGMA1 = SIGMA2:
    POOLED STANDARD DEVIATION   =    6.342600
    DIFFERENCE (DEL) IN MEANS   =   -10.33643
    STANDARD DEVIATION OF DEL   =   0.8190135
    T TEST STATISTIC VALUE      =   -12.62059
    DEGREES OF FREEDOM          =    326.0000
    T TEST STATISTIC CDF VALUE  =    0.000000

 IF NOT ASSUME SIGMA1 = SIGMA2:
    STANDARD DEVIATION SAMPLE 1 =    6.414700
    STANDARD DEVIATION SAMPLE 2 =    6.107710
    BARTLETT CDF VALUE          =    0.402799
    DIFFERENCE (DEL) IN MEANS   =   -10.33643
    STANDARD DEVIATION OF DEL   =   0.7984100
    T TEST STATISTIC VALUE      =   -12.94627
    EQUIVALENT DEG. OF FREEDOM  =    136.8750
    T TEST STATISTIC CDF VALUE  =    0.000000

                  ALTERNATIVE-                ALTERNATIVE-
 ALTERNATIVE-     HYPOTHESIS                  HYPOTHESIS
 HYPOTHESIS       ACCEPTANCE INTERVAL         CONCLUSION
 MU1 <> MU2        (0,0.025) (0.975,1)          ACCEPT
 MU1 < MU2         (0,0.05)                     ACCEPT
 MU1 > MU2         (0.95,1)                     REJECT
```

Interpretation of Sample Output

We are testing the hypothesis that the population mean is equal for the two samples. The output is divided into five sections.

1. The first section prints the sample statistics for sample one used in the computation of the *t*-test.
2. The second section prints the sample statistics for sample two used in the computation of the *t*-test.
3. The third section prints the pooled standard deviation, the difference in the means, the *t*-test statistic value, the degrees of freedom, and the cumulative distribution function (cdf) value of the *t*-test statistic under the assumption that the standard deviations are equal. The *t*-test statistic cdf value is an alternative way of expressing the critical value. This cdf value is compared to the acceptance intervals printed in section five. For an upper one-tailed test, the acceptance interval is $(0,1 - \alpha)$, the acceptance interval for a two-tailed test is $(\alpha/2, 1 - \alpha/2)$, and the acceptance interval for a lower one-tailed test is $(\alpha,1)$.
4. The fourth section prints the pooled standard deviation, the difference in the means, the *t*-test statistic value, the degrees of freedom, and the cumulative distribution function (cdf) value of the *t*-test statistic under the assumption that the standard deviations are not equal. The *t*-test statistic cdf value is an alternative way of expressing the critical value. cdf value is compared to the acceptance intervals printed in section five. For an upper one-tailed test, the alternative hypothesis acceptance interval is $(1 - \alpha,1)$, the alternative hypothesis acceptance interval for a lower one-tailed test is $(0,\alpha)$, and the alternative hypothesis acceptance interval for a two-tailed test is $(1 - \alpha/2,1)$ or $(0,\alpha/2)$. Note that accepting the alternative hypothesis is equivalent to rejecting the null hypothesis.
5. The fifth section prints the conclusions for a 95% test under the assumption that the standard deviations are not equal since a 95% test is the most common case. Results are given in terms of the alternative hypothesis for the two-tailed test and for the one-tailed test in both directions. The alternative hypothesis acceptance interval column is stated in terms of the cdf value printed in section four. The last column specifies whether the alternative hypothesis is accepted or rejected. For a different significance level, the appropriate conclusion can be drawn from the *t*-test statistic cdf value printed in section four. For example, for a significance level of 0.10, the corresponding alternative hypothesis acceptance intervals are (0,0.05) and (0.95,1), (0, 0.10), and (0.90,1).

Output from other statistical software may look somewhat different from the above output.

Questions	Two-sample *t*-tests can be used to answer the following questions: 1. Is process 1 equivalent to process 2? 2. Is the new process better than the current process? 3. Is the new process better than the current process by at least some pre-determined threshold amount?
Related Techniques	Confidence Limits for the Mean Analysis of Variance
Case Study	Ceramic strength data.
Software	Two-sample *t*-tests are available in just about all general purpose statistical software programs, including Dataplot.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.5.3.1. Data Used for Two-Sample *t*-Test

Data Used for Two-Sample t-Test Example

The following is the data used for the two-sample *t*-test example. The first column is miles per gallon for U.S. cars and the second column is miles per gallon for Japanese cars. For the *t*-test example, rows with the second column equal to -999 were deleted.

```
    18             24
    15             27
    18             27
    16             25
    17             31
    15             35
    14             24
    14             19
    14             28
    15             23
    15             27
    14             20
    15             22
    14             18
    22             20
    18             31
    21             32
    21             31
    10             32
    10             24
    11             26
     9             29
    28             24
    25             24
    19             33
    16             33
    17             32
    19             28
```

18	19
14	32
14	34
14	26
14	30
12	22
13	22
13	33
18	39
22	36
19	28
18	27
23	21
26	24
25	30
20	34
21	32
13	38
14	37
15	30
14	31
17	37
11	32
13	47
12	41
13	45
15	34
13	33
13	24
14	32
22	39
28	35
13	32
14	37
13	38
14	34
15	34
12	32
13	33
13	32
14	25
13	24
12	37
13	31
18	36
16	36

18	34
18	38
23	32
11	38
12	32
13	-999
12	-999
18	-999
21	-999
19	-999
21	-999
15	-999
16	-999
15	-999
11	-999
20	-999
21	-999
19	-999
15	-999
26	-999
25	-999
16	-999
16	-999
18	-999
16	-999
13	-999
14	-999
14	-999
14	-999
28	-999
19	-999
18	-999
15	-999
15	-999
16	-999
15	-999
16	-999
14	-999
17	-999
16	-999
15	-999
18	-999
21	-999
20	-999
13	-999
23	-999

```
20          -999
23          -999
18          -999
19          -999
25          -999
26          -999
18          -999
16          -999
16          -999
15          -999
22          -999
22          -999
24          -999
23          -999
29          -999
25          -999
20          -999
18          -999
19          -999
18          -999
27          -999
13          -999
17          -999
13          -999
13          -999
13          -999
30          -999
26          -999
18          -999
17          -999
16          -999
15          -999
18          -999
21          -999
19          -999
19          -999
16          -999
16          -999
16          -999
16          -999
25          -999
26          -999
31          -999
34          -999
36          -999
20          -999
```

```
19          -999
20          -999
19          -999
21          -999
20          -999
25          -999
21          -999
19          -999
21          -999
21          -999
19          -999
18          -999
19          -999
18          -999
18          -999
18          -999
30          -999
31          -999
23          -999
24          -999
22          -999
20          -999
22          -999
20          -999
21          -999
17          -999
18          -999
17          -999
18          -999
17          -999
16          -999
19          -999
19          -999
36          -999
27          -999
23          -999
24          -999
34          -999
35          -999
28          -999
29          -999
27          -999
34          -999
32          -999
28          -999
26          -999
```

```
24          -999
19          -999
28          -999
24          -999
27          -999
27          -999
26          -999
24          -999
30          -999
39          -999
35          -999
34          -999
30          -999
22          -999
27          -999
20          -999
18          -999
28          -999
27          -999
34          -999
31          -999
29          -999
27          -999
24          -999
23          -999
38          -999
36          -999
25          -999
38          -999
26          -999
22          -999
36          -999
27          -999
27          -999
32          -999
28          -999
31          -999
```

1.3.5.4. One-Factor ANOVA

Purpose: Test for Equal Means Across Groups

One factor analysis of variance (Snedecor and Cochran, 1989) is a special case of analysis of variance (ANOVA), for one factor of interest, and a generalization of the two-sample *t*-test. The two-sample *t*-test is used to decide whether two groups (levels) of a factor have the same mean. One-way analysis of variance generalizes this to levels where *k*, the number of levels, is greater than or equal to 2.

For example, data collected on, say, five instruments have one factor (instruments) at five levels. The ANOVA tests whether instruments have a significant effect on the results.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of 1-factor ANOVA, including the details for the mathematical computations of one-way analysis of variance.

The model for the analysis of variance can be stated in two mathematically equivalent ways. In the following discussion, each level of each factor is called a cell. For the one-way case, a cell and a level are equivalent since there is only one factor. In the following, the subscript ***i*** refers to the level and the subscript ***j*** refers to the observation within a level. For example, Y_{23} refers to the third observation in the second level.

The first model is

$$Y_{ij} = \mu_i + E_{ij}$$

This model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These estimated cell means are the predicted values of the model and the differences between the response variable and the estimated cell means are the residuals. That is

$$\hat{Y}_{ij} = \hat{\mu}_i$$

$$R_{ij} = Y_{ij} - \hat{\mu}_i$$

The second model is

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

This model decomposes the response into an overall (grand) mean, the effect of the *i*th factor level, and an error term. The analysis of variance provides estimates of the grand mean and the effect of the *i*th factor level. The predicted values and the residuals of the model are

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i$$

$$R_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i$$

The distinction between these models is that the second model divides the cell mean into an overall mean and the effect of the *i*th factor level. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, E_{ij}, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

Sample Output

Dataplot generated the following output for the one-way analysis of variance from the GEAR.DAT data set.

```
      NUMBER OF OBSERVATIONS           =      100
      NUMBER OF FACTORS                =        1
      NUMBER OF LEVELS FOR FACTOR  1  =       10
      BALANCED CASE
      RESIDUAL    STANDARD DEVIATION   =    0.59385783970E-02
      RESIDUAL    DEGREES OF FREEDOM   =       90
      REPLICATION CASE
      REPLICATION STANDARD DEVIATION   =    0.59385774657E-02
      REPLICATION DEGREES OF FREEDOM   =       90
      NUMBER OF DISTINCT CELLS         =       10

                         *****************
                         *  ANOVA TABLE  *
                         *****************

SOURCE              DF  SUM OF SQUARES    MEAN SQUARE   F STATISTIC    F CDF SIG
-------------------------------------------------------------------------------
TOTAL (CORRECTED)   99        0.003903       0.000039
-------------------------------------------------------------------------------
FACTOR  1            9        0.000729       0.000081        2.2969   97.734%   *
-------------------------------------------------------------------------------
RESIDUAL            90        0.003174       0.000035

      RESIDUAL    STANDARD DEVIATION =        0.00593857840
      RESIDUAL    DEGREES OF FREEDOM =           90
      REPLICATION STANDARD DEVIATION =        0.00593857747
      REPLICATION DEGREES OF FREEDOM =           90

                         ****************
                         *  ESTIMATION  *
                         ****************

      GRAND MEAN                       =    0.99764001369E+00
      GRAND STANDARD DEVIATION         =    0.62789078802E-02


             LEVEL-ID      NI       MEAN      EFFECT     SD(EFFECT)
--------------------------------------------------------------------
FACTOR 1--    1.00000     10.    0.99800    0.00036    0.00178
        --    2.00000     10.    0.99910    0.00146    0.00178
        --    3.00000     10.    0.99540   -0.00224    0.00178
        --    4.00000     10.    0.99820    0.00056    0.00178
        --    5.00000     10.    0.99190   -0.00574    0.00178
        --    6.00000     10.    0.99880    0.00116    0.00178
        --    7.00000     10.    1.00150    0.00386    0.00178
```

```
  --     8.00000     10.    1.00040    0.00276    0.00178
  --     9.00000     10.    0.99830    0.00066    0.00178
  --    10.00000     10.    0.99480   -0.00284    0.00178


       MODEL             RESIDUAL STANDARD DEVIATION
 -------------------------------------------------------
 CONSTANT             ONLY--        0.0062789079
 CONSTANT & FACTOR  1 ONLY--        0.0059385784
```

Interpretation of Sample Output

The output is divided into three sections.

1. The first section prints the number of observations (100), the number of factors (10), and the number of levels for each factor (10 levels for factor 1). It also prints some overall summary statistics. In particular, the residual standard deviation is 0.0059. The smaller the residual standard deviation, the more we have accounted for the variance in the data.
2. The second section prints an ANOVA table. The ANOVA table decomposes the variance into the following component sum of squares:
   - Total sum of squares. The degrees of freedom for this entry is the number of observations minus one.
   - Sum of squares for the factor. The degrees of freedom for this entry is the number of levels minus one. The mean square is the sum of squares divided by the number of degrees of freedom.
   - Residual sum of squares. The degrees of freedom is the total degrees of freedom minus the factor degrees of freedom. The mean square is the sum of squares divided by the number of degrees of freedom.

   That is, it summarizes how much of the variance in the data (total sum of squares) is accounted for by the factor effect (factor sum of squares) and how much is random error (residual sum of squares). Ideally, we would like most of the variance to be explained by the factor effect. The ANOVA table provides a formal F test for the factor effect. The F-statistic is the mean square for the factor divided by the mean square for the error. This statistic follows an F distribution with (*k*-1) and (*N*-*k*) degrees of freedom. If the F CDF column for the factor effect is greater than 95%, then the factor is significant at the 5% level.
3. The third section prints an estimation section. It prints an overall mean and overall standard deviation. Then for each level of each factor, it prints the number of observations, the mean for the observations of each cell ($\hat{\mu}_i$ in the above terminology), the factor effect ($\hat{\alpha}_j$ in the above terminology), and the standard deviation of the factor effect. Finally, it prints the residual standard deviation for the various possible models. For the one-way ANOVA, the two models are the constant model, i.e.,

   $$Y_i = A_0 + E_i$$

   and the model with a factor effect

   $$Y_{ij} = \mu + \alpha_i + E_{ij}$$

For these data, including the factor effect reduces the residual standard deviation from 0.00623 to 0.0059. That is, although the factor is statistically significant, it has minimal improvement over a simple constant model. This is because the factor is just barely significant.

Output from other statistical software may look somewhat different from the above output.

In addition to the quantitative ANOVA output, it is recommended that any analysis of variance be complemented with model validation. At a minimum, this should include

1. A run sequence plot of the residuals.
2. A normal probability plot of the residuals.
3. A scatter plot of the predicted values against the residuals.

Question

The analysis of variance can be used to answer the following question

- Are means the same across groups in the data?

Importance

The analysis of uncertainty depends on whether the factor significantly affects the outcome.

Related Techniques

Two-sample *t*-test
Multi-factor analysis of variance
Regression
Box plot

Software

Most general purpose statistical software programs, including Dataplot, can generate an analysis of variance.

1.3.5.5. Multi-factor Analysis of Variance

Purpose: Detect significant factors

The analysis of variance (ANOVA) (Neter, Wasserman, and Kunter, 1990) is used to detect significant factors in a multi-factor model. In the multi-factor model, there is a response (dependent) variable and one or more factor (independent) variables. This is a common model in designed experiments where the experimenter sets the values for each of the factor variables and then measures the response variable.

Each factor can take on a certain number of values. These are referred to as the levels of a factor. The number of levels can vary betweeen factors. For designed experiments, the number of levels for a given factor tends to be small. Each factor and level combination is a cell. Balanced designs are those in which the cells have an equal number of observations and unbalanced designs are those in which the number of observations varies among cells. It is customary to use balanced designs in designed experiments.

Definition

The Product and Process Comparisons chapter (chapter 7) contains a more extensive discussion of 2-factor ANOVA, including the details for the mathematical computations.

The model for the analysis of variance can be stated in two mathematically equivalent ways. We explain the model for a two-way ANOVA (the concepts are the same for additional factors). In the following discussion, each combination of factors and levels is called a cell. In the following, the subscript ***i*** refers to the level of factor 1, ***j*** refers to the level of factor 2, and the subscript ***k*** refers to the *k*th observation within the (*i*,*j*)th cell. For example, Y_{235} refers to the fifth observation in the second level of factor 1 and the third level of factor 2.

The first model is

$$Y_{ijk} = \mu_{ij} + E_{ijk}$$

This model decomposes the response into a mean for each cell and an error term. The analysis of variance provides estimates for each cell mean. These cell means are the predicted values of the model and the differences between the response variable and the estimated cell means are the residuals. That is

$$\hat{Y}_{ijk} = \hat{\mu}_{ij}$$

$$R_{ijk} = Y_{ijk} - \hat{\mu}_{ij}$$

The second model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$$

This model decomposes the response into an overall (grand) mean, factor effects ($\hat{\alpha}_i$ and $\hat{\beta}_j$ represent the effects of the *i*th level of the first factor and the *j*th level of the second factor, respectively), and an error term. The analysis of variance provides estimates of the grand mean and the factor effects. The predicted values and the residuals of the model are

$$\hat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$$

$$R_{ijk} = Y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$$

The distinction between these models is that the second model divides the cell mean into an overall mean and factor effects. This second model makes the factor effect more explicit, so we will emphasize this approach.

Model Validation

Note that the ANOVA model assumes that the error term, E_{ijk}, should follow the assumptions for a univariate measurement process. That is, after performing an analysis of variance, the model should be validated by analyzing the residuals.

Sample Output

Dataplot generated the following ANOVA output for the JAHANMI2.DAT data set:

```
                  **********************************
                  **********************************
                  **  4-WAY ANALYSIS OF VARIANCE  **
                  **********************************
                  **********************************

        NUMBER OF OBSERVATIONS           =      480
        NUMBER OF FACTORS                =        4
        NUMBER OF LEVELS FOR FACTOR  1  =        2
        NUMBER OF LEVELS FOR FACTOR  2  =        2
        NUMBER OF LEVELS FOR FACTOR  3  =        2
        NUMBER OF LEVELS FOR FACTOR  4  =        2
        BALANCED CASE
        RESIDUAL    STANDARD DEVIATION   =    0.63057727814E+02
        RESIDUAL    DEGREES OF FREEDOM   =      475
        REPLICATION CASE
        REPLICATION STANDARD DEVIATION   =    0.61890106201E+02
        REPLICATION DEGREES OF FREEDOM   =      464
        NUMBER OF DISTINCT CELLS         =       16

                           *****************
                           *  ANOVA TABLE  *
                           *****************

  SOURCE               DF SUM OF SQUARES    MEAN SQUARE   F STATISTIC    F CDF SIG
  -------------------------------------------------------------------------------
  TOTAL (CORRECTED)   479 2668446.000000    5570.868652
  -------------------------------------------------------------------------------
  FACTOR  1             1   26672.726562   26672.726562        6.7080  99.011%  **
  FACTOR  2             1   11524.053711   11524.053711        2.8982  91.067%
  FACTOR  3             1   14380.633789   14380.633789        3.6166  94.219%
  FACTOR  4             1  727143.125000  727143.125000      182.8703 100.000%  **
  -------------------------------------------------------------------------------
  RESIDUAL            475 1888731.500000    3976.276855

        RESIDUAL    STANDARD DEVIATION =       63.05772781
```

```
          RESIDUAL    DEGREES OF FREEDOM =           475
          REPLICATION STANDARD DEVIATION =       61.89010620
          REPLICATION DEGREES OF FREEDOM =           464
          LACK OF FIT F RATIO =       2.6447 = THE   99.7269% POINT OF THE
          F DISTRIBUTION WITH     11 AND    464 DEGREES OF FREEDOM

                                     ***************
                                     *  ESTIMATION  *
                                     ***************

          GRAND MEAN                        =    0.65007739258E+03
          GRAND STANDARD DEVIATION          =    0.74638252258E+02

                 LEVEL-ID      NI       MEAN      EFFECT     SD(EFFECT)
    --------------------------------------------------------------------
    FACTOR 1--    -1.00000    240.   657.53168    7.45428    2.87818
            --     1.00000    240.   642.62286   -7.45453    2.87818
    FACTOR 2--    -1.00000    240.   645.17755   -4.89984    2.87818
            --     1.00000    240.   654.97723    4.89984    2.87818
    FACTOR 3--    -1.00000    240.   655.55084    5.47345    2.87818
            --     1.00000    240.   644.60376   -5.47363    2.87818
    FACTOR 4--     1.00000    240.   688.99890   38.92151    2.87818
            --     2.00000    240.   611.15594  -38.92145    2.87818

            MODEL                  RESIDUAL STANDARD DEVIATION
    -------------------------------------------------------
    CONSTANT             ONLY--       74.6382522583
    CONSTANT & FACTOR  1 ONLY--       74.3419036865
    CONSTANT & FACTOR  2 ONLY--       74.5548019409
    CONSTANT & FACTOR  3 ONLY--       74.5147094727
    CONSTANT & FACTOR  4 ONLY--       63.7284545898
    CONSTANT & ALL 4 FACTORS --       63.0577278137
```

Interpretation of Sample Output

The output is divided into three sections.

1. The first section prints the number of observations (480), the number of factors (4), and the number of levels for each factor (2 levels for each factor). It also prints some overall summary statistics. In particular, the residual standard deviation is 63.058. The smaller the residual standard deviation, the more we have accounted for the variance in the data.
2. The second section prints an ANOVA table. The ANOVA table decomposes the variance into the following component sum of squares:
   - Total sum of squares. The degrees of freedom for this entry is the number of observations minus one.
   - Sum of squares for each of the factors. The degrees of freedom for these entries are the number of levels for the factor minus one. The mean square is the sum of squares divided by the number of degrees of freedom.
   - Residual sum of squares. The degrees of freedom is the total degrees of freedom minus the sum of the factor degrees of freedom. The mean square is the sum of squares divided by the number of degrees of freedom.

That is, it summarizes how much of the variance in the data (total sum of squares) is accounted for by the factor effects (factor sum of squares) and how much is random error (residual sum of squares). Ideally, we would like most of the variance to be explained by the factor effects. The ANOVA table provides a formal F test for the factor effects. The F-statistic is the mean square for the factor divided by the mean square for the error. This statistic follows an F distribution with (k-1) and (N-k) degrees of freedom where k is the number of levels for the given factor. If the F CDF column for the factor effect is greater than 95%, then the factor is significant at the 5% level. Here, we see that the size of the effect of factor 4 dominates the size of the other effects. The F test shows that factors one and four are significant at the 1% level while factors two and three are not significant at the 5% level.

3. The third section is an estimation section. It prints an overall mean and overall standard deviation. Then for each level of each factor, it prints the number of observations, the mean for the observations of each cell ($\hat{\mu}_{ij}$ in the above terminology), the factor effects ($\hat{\alpha}_i$ and $\hat{\beta}_j$ in the above terminology), and the standard deviation of the factor effect. Finally, it prints the residual standard deviation for the various possible models. For the four-way ANOVA here, it prints the constant model

$$Y_i = A_0 + E_i$$

a model with each factor individually, and the model with all four factors included.

For these data, we see that including factor 4 has a significant impact on the residual standard deviation (63.73 when only the factor 4 effect is included compared to 63.058 when all four factors are included).

Output from other statistical software may look somewhat different from the above output.

In addition to the quantitative ANOVA output, it is recommended that any analysis of variance be complemented with model validation. At a minimum, this should include

1. A run sequence plot of the residuals.
2. A normal probability plot of the residuals.
3. A scatter plot of the predicted values against the residuals.

Questions

The analysis of variance can be used to answer the following questions:

1. Do any of the factors have a significant effect?
2. Which is the most important factor?
3. Can we account for most of the variability in the data?

Related Techniques

One-factor analysis of variance
Two-sample *t*-test
Box plot
Block plot
Dex mean plot

Case Study

The quantitative ANOVA approach can be contrasted with the more graphical EDA approach in the ceramic strength case study.

Software Most general purpose statistical software programs, including Dataplot, can perform multi-factor analysis of variance.

1.3.5.6. Measures of Scale

Scale, Variability, or Spread

A fundamental task in many statistical analyses is to characterize the *spread*, or variability, of a data set. Measures of scale are simply attempts to estimate this variability.

When assessing the variability of a data set, there are two key components:

1. How spread out are the data values near the center?
2. How spread out are the tails?

Different numerical summaries will give different weight to these two elements. The choice of scale estimator is often driven by which of these components you want to emphasize.

The histogram is an effective graphical technique for showing both of these components of the spread.

Definitions of Variability

For univariate data, there are several common numerical measures of the spread:

1. variance - the variance is defined as

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

   where $\bar{Y}$ is the mean of the data.

   The variance is roughly the arithmetic average of the squared distance from the mean. Squaring the distance from the mean has the effect of giving greater weight to values that are further from the mean. For example, a point 2 units from the mean adds 4 to the above sum while a point 10 units from the mean adds 100 to the sum. Although the variance is intended to be an overall measure of spread, it can be greatly affected by the tail behavior.

2. standard deviation - the standard deviation is the square root of the variance. That is,

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

The standard deviation restores the units of the spread to the original data units (the variance squares the units).

3. range - the range is the largest value minus the smallest value in a data set. Note that this measure is based only on the lowest and highest extreme values in the sample. The spread near the center of the data is not captured at all.
4. average absolute deviation - the average absolute deviation (AAD) is defined as

$$AAD = \sum_{i=1}^{N}(|Y_i - \bar{Y}|)/N$$

where $\bar{Y}$ is the mean of the data and $|Y|$ is the absolute value of Y. This measure does not square the distance from the mean, so it is less affected by extreme observations than are the variance and standard deviation.

5. median absolute deviation - the median absolute deviation (MAD) is defined as

$$MAD = median(|Y_i - \tilde{Y}|)$$

where $\tilde{Y}$ is the median of the data and $|Y|$ is the absolute value of Y. This is a variation of the average absolute deviation that is even less affected by extremes in the tail because the data in the tails have less influence on the calculation of the median than they do on the mean.

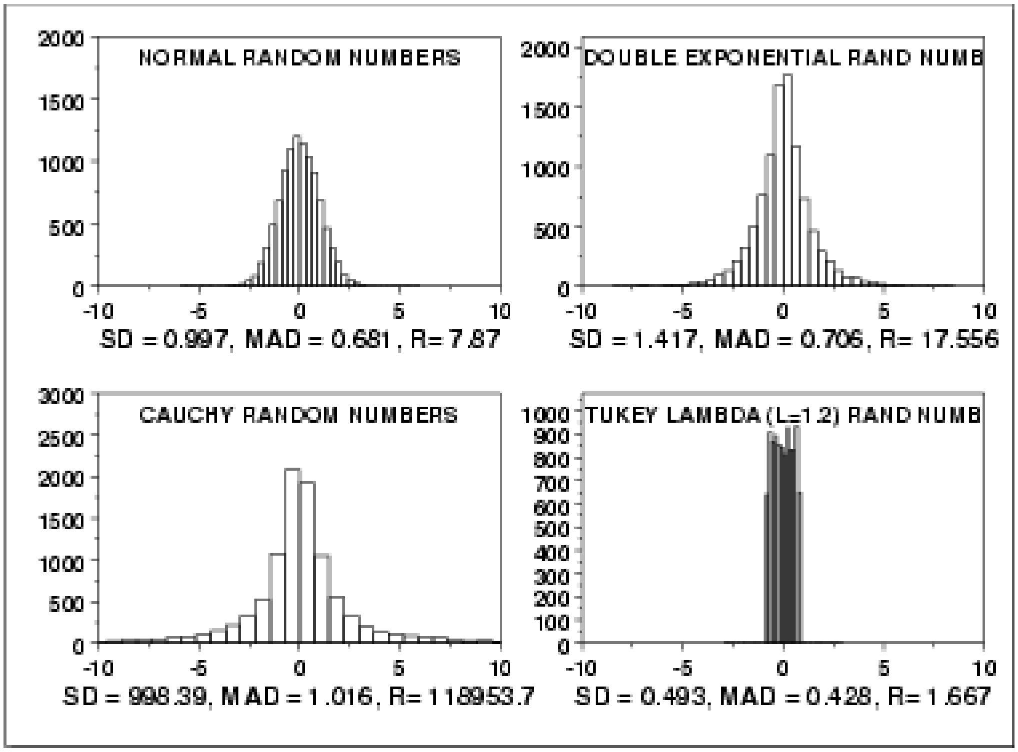
6. interquartile range - this is the value of the 75th percentile minus the value of the 25th percentile. This measure of scale attempts to measure the variability of points near the center.

In summary, the variance, standard deviation, average absolute deviation, and median absolute deviation measure both aspects of the variability; that is, the variability near the center and the variability in the tails. They differ in that the average absolute deviation and median absolute deviation do not give undue weight to the tail behavior. On the other hand, the range only uses the two most extreme points and the interquartile range only uses the middle portion of the data.

Why Different Measures?

The following example helps to clarify why these alternative defintions of spread are useful and necessary.

This plot shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Tukey-Lambda distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The standard deviation is 0.997, the median absolute deviation is 0.681, and the range is 7.87.

The normal distribution is a symmetric distribution with well-behaved tails and a single peak at the center of the distribution. By symmetric, we mean that the distribution can be folded about an axis so that the two sides coincide. That is, it behaves the same to the left and right of some center point. In this case, the median absolute deviation is a bit less than the standard deviation due to the downweighting of the tails. The range of a little less than 8 indicates the extreme values fall within about 4 standard deviations of the mean. If a histogram or normal probability plot indicates that your data are approximated well by a normal distribution, then it is reasonable to use the standard deviation as the spread estimator.

Double Exponential Distribution

The second histogram is a sample from a double exponential distribution. The standard deviation is 1.417, the median absolute deviation is 0.706, and the range is 17.556.

Comparing the double exponential and the normal histograms shows that the double exponential has a stronger peak at the center, decays more rapidly near the center, and has much longer tails. Due to the longer tails, the standard deviation tends to be inflated compared to the normal. On the other hand, the median absolute deviation is only slightly larger than it is for the normal data. The longer tails are clearly reflected in the value of the range, which shows that the extremes fall about 12 standard deviations from the mean compared to about 4 for the normal data.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution. The standard deviation is 998.389, the median absolute deviation is 1.16, and the range is 118,953.6.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. The Cauchy distribution has the interesting property that collecting more data does not provide a more accurate estimate for the mean or standard deviation. That is, the sampling distribution of the means and standard deviation are equivalent to the sampling distribution of the original data. That means that for the Cauchy distribution the standard deviation is useless as a measure of the spread. From the histogram, it is clear that just about all the data are between about -5 and 5. However, a few very extreme values cause both the standard deviation and range to be extremely large. However, the median absolute deviation is only slightly larger than it is for the normal distribution. In this case, the median absolute deviation is clearly the better measure of spread.

Although the Cauchy distribution is an extreme case, it does illustrate the importance of heavy tails in measuring the spread. Extreme values in the tails can distort the standard deviation. However, these extreme values do not distort the median absolute deviation since the median absolute deviation is based on ranks. In general, for data with extreme values in the tails, the median absolute deviation or interquartile range can provide a more stable estimate of spread than the standard deviation.

Tukey-Lambda Distribution

The fourth histogram is a sample from a Tukey lambda distribution with shape parameter α = 1.2. The standard deviation is 0.49, the median absolute deviation is 0.427, and the range is 1.666.

The Tukey lambda distribution has a range limited to $(-1/\lambda, 1/\lambda)$. That is, it has truncated tails. In this case the standard deviation and median absolute deviation have closer values than for the other three examples which have significant tails.

Robustness

Tukey and Mosteller defined two types of robustness where robustness is a lack of susceptibility to the effects of nonnormality.

1. Robustness of validity means that the confidence intervals for a measure of the population spread (e.g., the standard deviation) have a 95% chance of covering the true value (i.e., the population value) of that measure of spread regardless of the underlying distribution.
2. Robustness of efficiency refers to high effectiveness in the face of non-normal tails. That is, confidence intervals for the measure of spread tend to be almost as narrow as the best that could be done if we knew the true shape of the distribution.

The standard deviation is an example of an estimator that is the best we can do if the underlying distribution is normal. However, it lacks robustness of validity. That is, confidence intervals based on the standard deviation tend to lack precision if the underlying distribution is in fact not normal.

The median absolute deviation and the interquartile range are estimates of scale that have robustness of validity. However, they are not particularly strong for robustness of efficiency.

If histograms and probability plots indicate that your data are in fact reasonably approximated by a normal distribution, then it makes sense to use the standard deviation as the estimate of scale. However, if your data are not normal, and in particular if there are long tails, then using an alternative measure such as the median absolute deviation, average absolute deviation, or interquartile range makes sense. The range is used in some applications, such as quality control, for its simplicity. In addition, comparing the range to the standard deviation gives an indication of the spread of the data in the tails.

Since the range is determined by the two most extreme points in the data set, we should be cautious about its use for large values of N.

Tukey and Mosteller give a scale estimator that has both robustness of validity and robustness of efficiency. However, it is more complicated and we do not give the formula here.

Software

Most general purpose statistical software programs, including Dataplot, can generate at least some of the measures of scale discusssed above.

1.3.5.7. Bartlett's Test

Purpose: Test for Homogeneity of Variances

Bartlett's test (Snedecor and Cochran, 1983) is used to test if *k* samples have equal variances. Equal variances across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

Bartlett's test is sensitive to departures from normality. That is, if your samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality.

Definition

The Bartlett test is defined as:

H_0: $\sigma_1 = \sigma_2 = \ldots = \sigma_k$

H_a: $\sigma_i \neq \sigma_j$ for at least one pair (i,j).

Test Statistic: The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln s_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i - 1)) - 1/(N-k))}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{k}(N_i - 1)s_i^2/(N-k)$$

Significance Level: α

Critical Region: The variances are judged to be unequal if,

$$T > \chi^2_{(\alpha, k-1)}$$

where $\chi^2_{(\alpha, k-1)}$ is the upper critical value of the chi-square distribution with k - 1 degrees of freedom and a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that χ^2_{α} is the upper critical value from the chi-square distribution and $\chi^2_{1-\alpha}$ is the lower critical value from the chi-square distribution. Note that this is the opposite of some texts and software programs. In particular, Dataplot uses the opposite convention.

An alternate definition (Dixon and Massey, 1969) is based on an approximation to the F distribution. This definition is given in the Product and Process Comparisons chapter (chapter 7).

Sample Output Dataplot generated the following output for Bartlett's test using the GEAR.DAT data set:

```
                    BARTLETT TEST
                (STANDARD DEFINITION)
NULL HYPOTHESIS UNDER TEST--ALL SIGMA(I) ARE EQUAL

TEST:
   DEGREES OF FREEDOM                =    9.000000

   TEST STATISTIC VALUE              =    20.78580
   CUTOFF: 95% PERCENT POINT         =    16.91898
   CUTOFF: 99% PERCENT POINT         =    21.66600

   CHI-SQUARE CDF VALUE              =    0.986364

  NULL           NULL HYPOTHESIS        NULL HYPOTHESIS
  HYPOTHESIS     ACCEPTANCE INTERVAL    CONCLUSION
ALL SIGMA EQUAL     (0.000,0.950)         REJECT
```

Interpretation of Sample Output

We are testing the hypothesis that the group variances are all equal. The output is divided into two sections.

1. The first section prints the value of the Bartlett test statistic, the degrees of freedom (k-1), the upper critical value of the chi-square distribution corresponding to significance levels of 0.05 (the 95% percent point) and 0.01 (the 99% percent point). We reject the null hypothesis at that significance level if the value of the Bartlett test statistic is greater than the corresponding critical value.
2. The second section prints the conclusion for a 95% test.

Output from other statistical software may look somewhat different from the above output.

Question

Bartlett's test can be used to answer the following question:

- Is the assumption of equal variances valid?

Importance

Bartlett's test is useful whenever the assumption of equal variances is made. In particular, this assumption is made for the frequently used one-way analysis of variance. In this case, Bartlett's or Levene's test should be applied to verify the assumption.

Related Techniques

Standard Deviation Plot
Box Plot
Levene Test
Chi-Square Test
Analysis of Variance

Case Study

Heat flow meter data

Software

The Bartlett test is available in many general purpose statistical software programs, including Dataplot.

1.3.5.8. Chi-Square Test for the Standard Deviation

Purpose: Test if standard deviation is equal to a specified value

A chi-square test (Snedecor and Cochran, 1983) can be used to test if the standard deviation of a population is equal to a specified value. This test can be either a two-sided test or a one-sided test. The two-sided version tests against the alternative that the true standard deviation is either less than or greater than the specified value. The one-sided version only tests in one direction. The choice of a two-sided or one-sided test is determined by the problem. For example, if we are testing a new process, we may only be concerned if its variability is greater than the variability of the current process.

Definition

The chi-square hypothesis test is defined as:

H_0: $\sigma = \sigma_0$

H_a: $\sigma < \sigma_0$ for a lower one-tailed test

$\sigma > \sigma_0$ for an upper one-tailed test

$\sigma \neq \sigma_0$ for a two-tailed test

Test Statistic: $T = (N-1)(s/\sigma_0)^2$

where N is the sample size and s is the sample standard deviation. The key element of this formula is the ratio s/σ_0 which compares the ratio of the sample standard deviation to the target standard deviation. The more this ratio deviates from 1, the more likely we are to reject the null hypothesis.

Significance Level: α.

Critical Region: Reject the null hypothesis that the standard deviation is a specified value, σ_0, if

$T > \chi^2_{(\alpha,N-1)}$ for an upper one-tailed alternative

$T < \chi^2_{(1-\alpha,N-1)}$ for a lower one-tailed alternative

$T < \chi^2_{(1-\alpha/2,N-1)}$ for a two-tailed test

or

$T > \chi^2_{(\alpha/2,N-1)}$

where $\chi^2_{(.,N-1)}$ is the critical value of the chi-square distribution with N - 1 degrees of freedom.

In the above formulas for the critical regions, the Handbook follows the convention that χ^2_{α} is the upper critical value from the chi-square distribution and $\chi^2_{1-\alpha}$ is the lower critical value from the chi-square distribution. Note that this is the opposite of some texts and software programs. In particular, Dataplot uses the opposite convention.

The formula for the hypothesis test can easily be converted to form an interval estimate for the standard deviation:

$$\sqrt{\frac{(N-1)s^2}{\chi^2_{(\alpha/2,N-1)}}} \le \sigma \le \sqrt{\frac{(N-1)s^2}{\chi^2_{(1-\alpha/2,N-1)}}}$$

Sample Output

Dataplot generated the following output for a chi-square test from the GEAR.DAT data set:

```
                      CHI-SQUARED TEST
                    SIGMA0 =   0.1000000
 NULL HYPOTHESIS UNDER TEST--STANDARD DEVIATION SIGMA = .1000000

 SAMPLE:
    NUMBER OF OBSERVATIONS        =      100
    MEAN                          =    0.9976400
    STANDARD DEVIATION S          =    0.6278908E-02

 TEST:
    S/SIGMA0                      =    0.6278908E-01
    CHI-SQUARED STATISTIC         =    0.3903044
    DEGREES OF FREEDOM            =     99.00000
    CHI-SQUARED CDF VALUE         =     0.000000

                      ALTERNATIVE-            ALTERNATIVE-
    ALTERNATIVE-      HYPOTHESIS              HYPOTHESIS
    HYPOTHESIS        ACCEPTANCE INTERVAL     CONCLUSION
 SIGMA <> .1000000     (0,0.025), (0.975,1)    ACCEPT
 SIGMA <  .1000000     (0,0.05)                ACCEPT
 SIGMA >  .1000000     (0.95,1)                REJECT
```

Interpretation of Sample Output

We are testing the hypothesis that the population standard deviation is 0.1. The output is divided into three sections.

1. The first section prints the sample statistics used in the computation of the chi-square test.
2. The second section prints the chi-square test statistic value, the degrees of freedom, and the cumulative distribution function (cdf) value of the chi-square test statistic. The chi-square test statistic cdf value is an alternative way of expressing the critical value. This cdf value is compared to the acceptance intervals printed in section three. For an upper one-tailed test, the alternative hypothesis acceptance interval is (1 - α,1), the alternative hypothesis acceptance interval for a lower one-tailed test is (0,α), and the alternative hypothesis acceptance interval for a two-tailed test is (1 - α/2,1) or (0,α/2). Note that accepting the alternative hypothesis is equivalent to rejecting the null hypothesis.

3. The third section prints the conclusions for a 95% test since this is the most common case. Results are given in terms of the alternative hypothesis for the two-tailed test and for the one-tailed test in both directions. The alternative hypothesis acceptance interval column is stated in terms of the cdf value printed in section two. The last column specifies whether the alternative hypothesis is accepted or rejected. For a different significance level, the appropriate conclusion can be drawn from the chi-square test statistic cdf value printed in section two. For example, for a significance level of 0.10, the corresponding alternative hypothesis acceptance intervals are (0,0.05) and (0.95,1), (0, 0.10), and (0.90,1).

Output from other statistical software may look somewhat different from the above output.

Questions

The chi-square test can be used to answer the following questions:

1. Is the standard deviation equal to some pre-determined threshold value?
2. Is the standard deviation greater than some pre-determined threshold value?
3. Is the standard deviation less than some pre-determined threshold value?

Related Techniques

F Test
Bartlett Test
Levene Test

Software

The chi-square test for the standard deviation is available in many general purpose statistical software programs, including Dataplot.

1.3.5.8.1. Data Used for Chi-Square Test for the Standard Deviation

Data Used for Chi-Square Test for the Standard Deviation Example

The following are the data used for the chi-square test for the standard deviation example. The first column is gear diameter and the second column is batch number. Only the first column is used for this example.

```
   1.006       1.000
   0.996       1.000
   0.998       1.000
   1.000       1.000
   0.992       1.000
   0.993       1.000
   1.002       1.000
   0.999       1.000
   0.994       1.000
   1.000       1.000
   0.998       2.000
   1.006       2.000
   1.000       2.000
   1.002       2.000
   0.997       2.000
   0.998       2.000
   0.996       2.000
   1.000       2.000
   1.006       2.000
   0.988       2.000
   0.991       3.000
   0.987       3.000
   0.997       3.000
   0.999       3.000
   0.995       3.000
   0.994       3.000
   1.000       3.000
```

```
0.999      3.000
0.996      3.000
0.996      3.000
1.005      4.000
1.002      4.000
0.994      4.000
1.000      4.000
0.995      4.000
0.994      4.000
0.998      4.000
0.996      4.000
1.002      4.000
0.996      4.000
0.998      5.000
0.998      5.000
0.982      5.000
0.990      5.000
1.002      5.000
0.984      5.000
0.996      5.000
0.993      5.000
0.980      5.000
0.996      5.000
1.009      6.000
1.013      6.000
1.009      6.000
0.997      6.000
0.988      6.000
1.002      6.000
0.995      6.000
0.998      6.000
0.981      6.000
0.996      6.000
0.990      7.000
1.004      7.000
0.996      7.000
1.001      7.000
0.998      7.000
1.000      7.000
1.018      7.000
1.010      7.000
0.996      7.000
1.002      7.000
0.998      8.000
1.000      8.000
1.006      8.000
```

1.000	8.000
1.002	8.000
0.996	8.000
0.998	8.000
0.996	8.000
1.002	8.000
1.006	8.000
1.002	9.000
0.998	9.000
0.996	9.000
0.995	9.000
0.996	9.000
1.004	9.000
1.004	9.000
0.998	9.000
0.999	9.000
0.991	9.000
0.991	10.000
0.995	10.000
0.984	10.000
0.994	10.000
0.997	10.000
0.997	10.000
0.991	10.000
0.998	10.000
1.004	10.000
0.997	10.000

1.3.5.9. F-Test for Equality of Two Standard Deviations

Purpose: Test if standard deviations from two populations are equal

An F-test (Snedecor and Cochran, 1983) is used to test if the standard deviations of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the standard deviations are not equal. The one-tailed version only tests in one direction, that is the standard deviation from the first population is either greater than or less than (but not both) the second population standard deviation . The choice is determined by the problem. For example, if we are testing a new process, we may only be interested in knowing if the new process is less variable than the old process.

Definition

The F hypothesis test is defined as:

H_0:	$\sigma_1 = \sigma_2$	
H_a:	$\sigma_1 < \sigma_2$	for a lower one tailed test
	$\sigma_1 > \sigma_2$	for an upper one tailed test
	$\sigma_1 \neq \sigma_2$	for a two tailed test
Test Statistic:	$F = s_1^2 / s_2^2$ where s_1^2 and s_2^2 are the sample variances. The more this ratio deviates from 1, the stronger the evidence for unequal population variances.	
Significance Level:	α	

Critical Region: The hypothesis that the two standard deviations are equal is rejected if

$F > F_{(\alpha,N1-1,N2-1)}$ for an upper one-tailed test

$F < F_{(1-\alpha,N1-1,N2-1)}$ for a lower one-tailed test

$F < F_{(1-\alpha/2,N1-1,N2-1)}$ for a two-tailed test

or

$F > F_{(\alpha/2,N1-1,N2-1)}$

where $F_{(\alpha,k-1,N-k)}$ is the critical value of the F distribution with ν_1 and ν_2 degrees of freedom and a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that F_α is the upper critical value from the F distribution and $F_{1-\alpha}$ is the lower critical value from the F distribution. Note that this is the opposite of the designation used by some texts and software programs. In particular, Dataplot uses the opposite convention.

Sample Output Dataplot generated the following output for an F-test from the JAHANMI2.DAT data set:

```
                      F TEST
NULL HYPOTHESIS UNDER TEST--SIGMA1 = SIGMA2
ALTERNATIVE HYPOTHESIS UNDER TEST--SIGMA1 NOT EQUAL SIGMA2

SAMPLE 1:
   NUMBER OF OBSERVATIONS      =      240
   MEAN                        =    688.9987
   STANDARD DEVIATION          =    65.54909

SAMPLE 2:
   NUMBER OF OBSERVATIONS      =      240
   MEAN                        =    611.1559
   STANDARD DEVIATION          =    61.85425

TEST:
   STANDARD DEV. (NUMERATOR)   =    65.54909
   STANDARD DEV. (DENOMINATOR) =    61.85425
   F TEST STATISTIC VALUE      =    1.123037
   DEG. OF FREEDOM (NUMER.)    =    239.0000
   DEG. OF FREEDOM (DENOM.)    =    239.0000
   F TEST STATISTIC CDF VALUE  =    0.814808

  NULL           NULL HYPOTHESIS        NULL HYPOTHESIS
  HYPOTHESIS     ACCEPTANCE INTERVAL    CONCLUSION
SIGMA1 = SIGMA2    (0.000,0.950)          ACCEPT
```

Interpretation of Sample Output

We are testing the hypothesis that the standard deviations for sample one and sample two are equal. The output is divided into four sections.

1. The first section prints the sample statistics for sample one used in the computation of the F-test.
2. The second section prints the sample statistics for sample two used in the computation of the F-test.
3. The third section prints the numerator and denominator standard deviations, the F-test statistic value, the degrees of freedom, and the cumulative distribution function (cdf) value of the F-test statistic. The F-test statistic cdf value is an alternative way of expressing the critical value. This cdf value is compared to the acceptance interval printed in section four. The acceptance interval for a two-tailed test is (0,1 - α).
4. The fourth section prints the conclusions for a 95% test since this is the most common case. Results are printed for an upper one-tailed test. The acceptance interval column is stated in terms of the cdf value printed in section three. The last column specifies whether the null hypothesis is accepted or rejected. For a different significance level, the appropriate conclusion can be drawn from the F-test statistic cdf value printed in section four. For example, for a significance level of 0.10, the corresponding acceptance interval become (0.000,0.9000).

Output from other statistical software may look somewhat different from the above output.

Questions

The F-test can be used to answer the following questions:

1. Do two samples come from populations with equal standard deviations?
2. Does a new process, treatment, or test reduce the variability of the current process?

Related Techniques

Quantile-Quantile Plot
Bihistogram
Chi-Square Test
Bartlett's Test
Levene Test

Case Study

Ceramic strength data.

Software

The F-test for equality of two standard deviations is available in many general purpose statistical software programs, including Dataplot.

1.3.5.10. Levene Test for Equality of Variances

Purpose: Test for Homogeneity of Variances

Levene's test (Levene 1960) is used to test if *k* samples have equal variances. Equal variances across samples is called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene test can be used to verify that assumption.

Levene's test is an alternative to the Bartlett test. The Levene test is less sensitive than the Bartlett test to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal, distribution, then Bartlett's test has better performance.

Definition

The Levene test is defined as:

H_0: $\sigma_1 = \sigma_2 = \ldots = \sigma_k$

H_a: $\sigma_i \neq \sigma_j$ for at least one pair (i,j).

Test Statistic: Given a variable Y with sample of size N divided into k subgroups, where N_i is the sample size of the ith subgroup, the Levene test statistic is defined as:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_{i.})^2}$$

where Z_{ij} can have one of the following three definitions:

1. $$Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$$

   where $\bar{Y}_{i.}$ is the mean of the ith subgroup.

2. $$Z_{ij} = |Y_{ij} - \tilde{Y}_{i.}|$$

   where $\tilde{Y}_{i.}$ is the median of the ith subgroup.

3. $$Z_{ij} = |Y_{ij} - \bar{Y}'_{i.}|$$

   where $\bar{Y}'_{i.}$ is the 10% trimmed mean of the *i*th subgroup.

$\bar{Z}_{i.}$ are the group means of the Z_{ij} and $\bar{Z}_{..}$ is the overall mean of the Z_{ij}.

The three choices for defining Z_{ij} determine the robustness and power of Levene's test. By robustness, we mean the ability of the test to not falsely detect unequal variances when the underlying data are not normally distributed and the variables are in fact equal. By power, we mean the ability of the test to detect unequal variances when the variances are in fact unequal.

Levene's original paper only proposed using the mean. Brown and Forsythe (1974)) extended Levene's test to use either the median or the trimmed mean in addition to the mean. They performed Monte Carlo studies that indicated that using the trimmed mean performed best when the underlying data followed a Cauchy distribution (i.e., heavy-tailed) and the median performed best when the underlying data followed a χ^2_4 (i.e., skewed) distribution. Using the mean provided the best power for symmetric, moderate-tailed, distributions.

Although the optimal choice depends on the underlying distribution, the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. If you have knowledge of the underlying distribution of the data, this may indicate using one of the other choices.

Significance Level: α

Critical Region: The Levene test rejects the hypothesis that the variances are equal if

$$W > F_{(\alpha,k-1,N-k)}$$

where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F distribution with k - 1 and N - k degrees of freedom at a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that F_{α} is the upper critical value from the F distribution and $F_{1-\alpha}$ is the lower critical value. Note that this is the opposite of some texts and software programs. In particular, Dataplot uses the opposite convention.

Sample Output Dataplot generated the following output for Levene's test using the GEAR.DAT data set:

```
                 LEVENE F-TEST FOR SHIFT IN VARIATION
                        (ASSUMPTION: NORMALITY)

1. STATISTICS
      NUMBER OF OBSERVATIONS    =      100
      NUMBER OF GROUPS          =       10
      LEVENE F TEST STATISTIC   =    1.705910


2. FOR LEVENE TEST STATISTIC
      0          % POINT    =           0.
      50         % POINT    =    0.9339308
      75         % POINT    =     1.296365
      90         % POINT    =     1.702053
      95         % POINT    =     1.985595
      99         % POINT    =     2.610880
      99.9       % POINT    =     3.478882


         90.09152       % Point:     1.705910

3. CONCLUSION (AT THE 5% LEVEL):
      THERE IS NO SHIFT IN VARIATION.
      THUS: HOMOGENEOUS WITH RESPECT TO VARIATION.
```

Interpretation of Sample Output

We are testing the hypothesis that the group variances are equal. The output is divided into three sections.

1. The first section prints the number of observations (*N*), the number of groups (*k*), and the value of the Levene test statistic.
2. The second section prints the upper critical value of the F distribution corresponding to various significance levels. The value in the first column, the confidence level of the test, is equivalent to 100(1-α). We reject the null hypothesis at that significance level if the value of the Levene F test statistic printed in section one is greater than the critical value printed in the last column.
3. The third section prints the conclusion for a 95% test. For a different significance level, the appropriate conclusion can be drawn from the table printed in section two. For example, for α = 0.10, we look at the row for 90% confidence and compare the critical value 1.702 to the Levene test statistic 1.7059. Since the test statistic is greater than the critical value, we reject the null hypothesis at the α = 0.10 level.

Output from other statistical software may look somewhat different from the above output.

Question

Levene's test can be used to answer the following question:

- Is the assumption of equal variances valid?

Related Techniques

Standard Deviation Plot
Box Plot
Bartlett Test
Chi-Square Test
Analysis of Variance

Software

The Levene test is available in some general purpose statistical software programs, including Dataplot.

1.3.5.11. Measures of Skewness and Kurtosis

Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

Definition of Skewness

For univariate data $Y_1, Y_2, ..., Y_N$, the formula for skewness is:

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is heavier than the right tail. Similarly, skewed right means that the right tail is heavier than the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

Definition of Kurtosis

For univariate data $Y_1, Y_2, ..., Y_N$, the formula for kurtosis is:

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points.

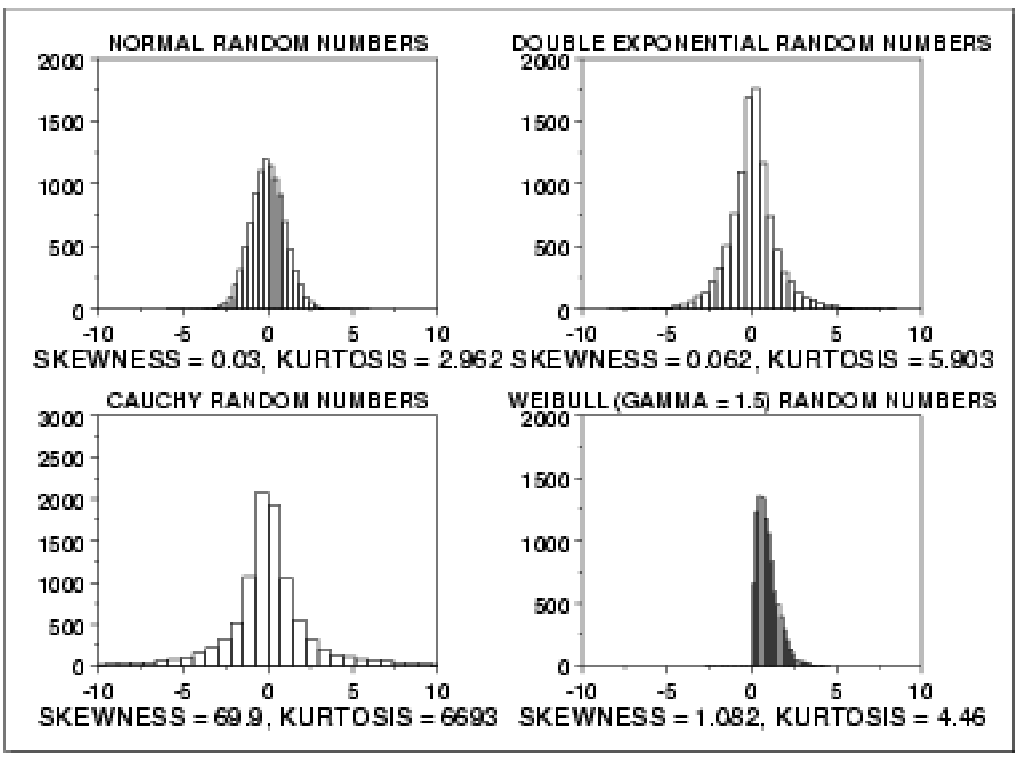
The kurtosis for a standard normal distribution is three. For this reason, excess kurtosis is defined as

$$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4} - 3$$

so that the standard normal distribution has a kurtosis of zero. Positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

Examples

The following example shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Weibull distribution.



Normal Distribution

The first histogram is a sample from a normal distribution. The normal distribution is a symmetric distribution with well-behaved tails. This is indicated by the skewness of 0.03. The kurtosis of 2.96 is near the expected value of 3. The histogram verifies the symmetry.

Double Exponential Distribution

The second histogram is a sample from a double exponential distribution. The double exponential is a symmetric distribution. Compared to the normal, it has a stronger peak, more rapid decay, and heavier tails. That is, we would expect a skewness near zero and a kurtosis higher than 3. The skewness is 0.06 and the kurtosis is 5.9.

Cauchy Distribution

The third histogram is a sample from a Cauchy distribution.

For better visual comparison with the other data sets, we restricted the histogram of the Cauchy distribution to values between -10 and 10. The full data set for the Cauchy data in fact has a minimum of approximately -29,000 and a maximum of approximately 89,000.

The Cauchy distribution is a symmetric distribution with heavy tails and a single peak at the center of the distribution. Since it is symmetric, we would expect a skewness near zero. Due to the heavier tails, we might expect the kurtosis to be larger than for a normal distribution. In fact the skewness is 69.99 and the kurtosis is 6,693. These extremely high values can be explained by the heavy tails. Just as the mean and standard deviation can be distorted by extreme values in the tails, so too can the skewness and kurtosis measures.

Weibull Distribution

The fourth histogram is a sample from a Weibull distribution with shape parameter 1.5. The Weibull distribution is a skewed distribution with the amount of skewness depending on the value of the shape parameter. The degree of decay as we move away from the center also depends on the value of the shape parameter. For this data set, the skewness is 1.08 and the kurtosis is 4.46, which indicates moderate skewness and kurtosis.

Dealing with Skewness and Kurtosis

Many classical statistical tests and intervals depend on normality assumptions. Significant skewness and kurtosis clearly indicate that data are not normal. If a data set exhibits significant skewness or kurtosis (as indicated by a histogram or the numerical measures), what can we do about it?

One approach is to apply some type of transformation to try to make the data normal, or more nearly normal. The Box-Cox transformation is a useful technique for trying to normalize a data set. In particular, taking the log or square root of a data set is often useful for data that exhibit moderate right skewness.

Another approach is to use techniques based on distributions other than the normal. For example, in reliability studies, the exponential, Weibull, and lognormal distributions are typically used as a basis for modeling rather than using the normal distribution. The probability plot correlation coefficient plot and the probability plot are useful tools for determining a good distributional model for the data.

Software

The skewness and kurtosis coefficients are available in most general purpose statistical software programs, including Dataplot.

1.3.5.12. Autocorrelation

Purpose: Detect Non-Randomness, Time Series Modeling

The autocorrelation (Box and Jenkins, 1976) function can be used for the following two purposes:

1. To detect non-randomness in data.
2. To identify an appropriate time series model if the data are not random.

Definition

Given measurements, $Y_1, Y_2, ..., Y_N$ at time $X_1, X_2, ..., X_N$, the lag k autocorrelation function is defined as

$$r_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Although the time variable, X, is not used in the formula for autocorrelation, the assumption is that the observations are equi-spaced.

Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times X_i and X_{i+k}.

When the autocorrelation is used to detect non-randomness, it is usually only the first (lag 1) autocorrelation that is of interest. When the autocorrelation is used to identify an appropriate time series model, the autocorrelations are usually plotted for many lags.

Sample Output

Dataplot generated the following autocorrelation output using the LEW.DAT data set:

```
THE LAG-ONE AUTOCORRELATION COEFFICIENT OF THE

200 OBSERVATIONS =  -0.3073048E+00

THE COMPUTED VALUE OF THE CONSTANT A  = -0.30730480E+00


lag     autocorrelation
 0.       1.00
 1.      -0.31
 2.      -0.74
 3.       0.77
 4.       0.21
 5.      -0.90
 6.       0.38
 7.       0.63
 8.      -0.77
 9.      -0.12
10.       0.82
11.      -0.40
12.      -0.55
13.       0.73
14.       0.07
15.      -0.76
16.       0.40
17.       0.48
18.      -0.70
19.      -0.03
20.       0.70
21.      -0.41
22.      -0.43
23.       0.67
24.       0.00
25.      -0.66
26.       0.42
27.       0.39
28.      -0.65
29.       0.03
30.       0.63
31.      -0.42
32.      -0.36
33.       0.64
34.      -0.05
35.      -0.60
36.       0.43
37.       0.32
38.      -0.64
39.       0.08
40.       0.58
```

```
41.     -0.45
42.     -0.28
43.      0.62
44.     -0.10
45.     -0.55
46.      0.45
47.      0.25
48.     -0.61
49.      0.14
```

Questions

The autocorrelation function can be used to answer the following questions

1. Was this sample data set generated from a random process?
2. Would a non-linear or time series model be a more appropriate model for these data than a simple constant plus error model?

Importance

Randomness is one of the key assumptions in determining if a univariate statistical process is in control. If the assumptions of constant location and scale, randomness, and fixed distribution are reasonable, then the univariate process can be modeled as:

$$Y_i = A_0 + E_i$$

where E_i is an error term.

If the randomness assumption is not valid, then a different model needs to be used. This will typically be either a time series model or a non-linear model (with time as the independent variable).

Related Techniques

Autocorrelation Plot
Run Sequence Plot
Lag Plot
Runs Test

Case Study

The heat flow meter data demonstrate the use of autocorrelation in determining if the data are from a random process.

The beam deflection data demonstrate the use of autocorrelation in developing a non-linear sinusoidal model.

Software

The autocorrelation capability is available in most general purpose statistical software programs, including Dataplot.

1.3.5.13. Runs Test for Detecting Non-randomness

Purpose: Detect Non-Randomness

The runs test (Bradley, 1968) can be used to decide if a data set is from a random process.

A run is defined as a series of increasing values or a series of decreasing values. The number of increasing, or decreasing, values is the length of the run. In a random data set, the probability that the (I+1)th value is larger or smaller than the Ith value follows a binomial distribution, which forms the basis of the runs test.

Typical Analysis and Test Statistics

The first step in the runs test is to compute the sequential differences (Y_i - Y_{i-1}). Positive values indicate an increasing value and negative values indicate a decreasing value. A runs test should include information such as the output shown below from Dataplot for the LEW.DAT data set. The output shows a table of:

1. runs of length exactly I for I = 1, 2, ..., 10
2. number of runs of length I
3. expected number of runs of length I
4. standard deviation of the number of runs of length I
5. a z-score where the z-score is defined to be

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

where $\bar{Y}$ is the sample mean and s is the sample standard deviation.

The z-score column is compared to a standard normal table. That is, at the 5% significance level, a z-score with an absolute value greater than 1.96 indicates non-randomness.

There are several alternative formulations of the runs test in the literature. For example, a series of coin tosses would record a series of heads and tails. A run of length r is r consecutive heads or r consecutive tails. To use the Dataplot RUNS command, you could code a sequence of the N = 10 coin tosses HHHHTTHTHH as

1 2 3 4 3 2 3 2 3 4

that is, a heads is coded as an increasing value and a tails is coded as a decreasing value.

Another alternative is to code values above the median as positive and values below the median as negative. There are other formulations as well. All of them can be converted to the Dataplot formulation. Just remember that it ultimately reduces to 2 choices. To use the Dataplot runs test, simply code one choice as an increasing value and the other as a decreasing value as in the heads/tails example above. If you are using other statistical software, you need to check the conventions used by that program.

Sample Output

Dataplot generated the following runs test output using the LEW.DAT data set:

```
                   RUNS UP

         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH EXACTLY I

 I          STAT     EXP(STAT)    SD(STAT)        Z

 1          18.0      41.7083      6.4900      -3.65
 2          40.0      18.2167      3.3444       6.51
 3           2.0       5.2125      2.0355      -1.58
 4           0.0       1.1302      1.0286      -1.10
 5           0.0       0.1986      0.4424      -0.45
 6           0.0       0.0294      0.1714      -0.17
 7           0.0       0.0038      0.0615      -0.06
 8           0.0       0.0004      0.0207      -0.02
 9           0.0       0.0000      0.0066      -0.01
10           0.0       0.0000      0.0020       0.00


         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH I OR MORE

 I          STAT     EXP(STAT)    SD(STAT)        Z

 1          60.0      66.5000      4.1972      -1.55
 2          42.0      24.7917      2.8083       6.13
```

```
   3          2.0      6.5750      2.1639      -2.11
   4          0.0      1.3625      1.1186      -1.22
   5          0.0      0.2323      0.4777      -0.49
   6          0.0      0.0337      0.1833      -0.18
   7          0.0      0.0043      0.0652      -0.07
   8          0.0      0.0005      0.0218      -0.02
   9          0.0      0.0000      0.0069      -0.01
  10          0.0      0.0000      0.0021       0.00


                    RUNS DOWN

          STATISTIC = NUMBER OF RUNS DOWN
              OF LENGTH EXACTLY I

   I          STAT     EXP(STAT)    SD(STAT)        Z

   1         25.0     41.7083      6.4900      -2.57
   2         35.0     18.2167      3.3444       5.02
   3          0.0      5.2125      2.0355      -2.56
   4          0.0      1.1302      1.0286      -1.10
   5          0.0      0.1986      0.4424      -0.45
   6          0.0      0.0294      0.1714      -0.17
   7          0.0      0.0038      0.0615      -0.06
   8          0.0      0.0004      0.0207      -0.02
   9          0.0      0.0000      0.0066      -0.01
  10          0.0      0.0000      0.0020       0.00


          STATISTIC = NUMBER OF RUNS DOWN
              OF LENGTH I OR MORE


   I          STAT     EXP(STAT)    SD(STAT)        Z

   1         60.0     66.5000      4.1972      -1.55
   2         35.0     24.7917      2.8083       3.63
   3          0.0      6.5750      2.1639      -3.04
   4          0.0      1.3625      1.1186      -1.22
   5          0.0      0.2323      0.4777      -0.49
   6          0.0      0.0337      0.1833      -0.18
   7          0.0      0.0043      0.0652      -0.07
   8          0.0      0.0005      0.0218      -0.02
   9          0.0      0.0000      0.0069      -0.01
  10          0.0      0.0000      0.0021       0.00
```

```
                    RUNS TOTAL = RUNS UP + RUNS DOWN

                  STATISTIC = NUMBER OF RUNS TOTAL
                       OF LENGTH EXACTLY I

     I          STAT     EXP(STAT)    SD(STAT)        Z

     1          43.0     83.4167      9.1783        -4.40
     2          75.0     36.4333      4.7298         8.15
     3           2.0     10.4250      2.8786        -2.93
     4           0.0      2.2603      1.4547        -1.55
     5           0.0      0.3973      0.6257        -0.63
     6           0.0      0.0589      0.2424        -0.24
     7           0.0      0.0076      0.0869        -0.09
     8           0.0      0.0009      0.0293        -0.03
     9           0.0      0.0001      0.0093        -0.01
    10           0.0      0.0000      0.0028         0.00


                  STATISTIC = NUMBER OF RUNS TOTAL
                        OF LENGTH I OR MORE

     I          STAT     EXP(STAT)    SD(STAT)        Z

     1         120.0    133.0000      5.9358        -2.19
     2          77.0     49.5833      3.9716         6.90
     3           2.0     13.1500      3.0602        -3.64
     4           0.0      2.7250      1.5820        -1.72
     5           0.0      0.4647      0.6756        -0.69
     6           0.0      0.0674      0.2592        -0.26
     7           0.0      0.0085      0.0923        -0.09
     8           0.0      0.0010      0.0309        -0.03
     9           0.0      0.0001      0.0098        -0.01
    10           0.0      0.0000      0.0030         0.00


                   LENGTH OF THE LONGEST RUN UP         =     3
                   LENGTH OF THE LONGEST RUN DOWN       =     2
                   LENGTH OF THE LONGEST RUN UP OR DOWN =     3

                   NUMBER OF POSITIVE DIFFERENCES =   104
                   NUMBER OF NEGATIVE DIFFERENCES =    95
                   NUMBER OF ZERO     DIFFERENCES =     0
```

Interpretation of Sample Output

Scanning the last column labeled "Z", we note that most of the z-scores for run lengths 1, 2, and 3 have an absolute value greater than 1.96. This is strong evidence that these data are in fact not random.

Output from other statistical software may look somewhat different from the above output.

Question

The runs test can be used to answer the following question:

- Were these sample data generated from a random process?

Importance

Randomness is one of the key assumptions in determining if a univariate statistical process is in control. If the assumptions of constant location and scale, randomness, and fixed distribution are reasonable, then the univariate process can be modeled as:

$$Y_i = A_0 + E_i$$

where E_i is an error term.

If the randomness assumption is not valid, then a different model needs to be used. This will typically be either a times series model or a non-linear model (with time as the independent variable).

Related Techniques

Autocorrelation
Run Sequence Plot
Lag Plot

Case Study

Heat flow meter data

Software

Most general purpose statistical software programs, including Dataplot, support a runs test.

1.3.5.14. Anderson-Darling Test

Purpose: Test for Distributional Adequacy

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions. We do not provide the tables of critical values in this Handbook (see Stephens 1974, 1976, 1977, and 1979) since this test is usually applied with a statistical software program that will print the relevant critical values.

The Anderson-Darling test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests.

Definition

The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution.

H_a: The data do not follow the specified distribution

Test Statistic: The Anderson-Darling test statistic is defined as

$$A^2 = -N - S$$

where

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\ln F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right]$$

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the *ordered* data.

Significance Level:	α
Critical Region:	The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published (Stephens, 1974, 1976, 1977, 1979) for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value. Note that for a given distribution, the Anderson-Darling statistic may be multiplied by a constant (which usually depends on the sample size, *n*). These constants are given in the various papers by Stephens. In the sample output below, this is the "adjusted Anderson-Darling" statistic. This is what should be compared against the critical values. Also, be aware that different constants (and therefore critical values) have been published. You just need to be aware of what constant was used for a given set of critical values (the needed constant is typically given with the critical values).

Sample Output

Dataplot generated the following output for the Anderson-Darling test. 1,000 random numbers were generated for a normal, double exponential, Cauchy, and lognormal distribution. In all four cases, the Anderson-Darling test was applied to test for a normal distribution. When the data were generated using a normal distribution, the test statistic was small and the hypothesis was accepted. When the data were generated using the double exponential, Cauchy, and lognormal distributions, the statistics were significant, and the hypothesis of an underlying normal distribution was rejected at significance levels of 0.10, 0.05, and 0.01.

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the Cauchy random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

```
          ***************************************
          **  anderson darling normal test y1  **
          ***************************************


                  ANDERSON-DARLING 1-SAMPLE TEST
                  THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

  1. STATISTICS:
        NUMBER OF OBSERVATIONS                =      1000
        MEAN                                  =   0.4359940E-02
```

```
        STANDARD DEVIATION                       =    1.001816

        ANDERSON-DARLING TEST STATISTIC VALUE =   0.2565918
        ADJUSTED TEST STATISTIC VALUE         =   0.2576117

     2. CRITICAL VALUES:
        90         % POINT    =   0.6560000
        95         % POINT    =   0.7870000
        97.5       % POINT    =   0.9180000
        99         % POINT    =    1.092000

     3. CONCLUSION (AT THE 5% LEVEL):
        THE DATA DO COME FROM A NORMAL DISTRIBUTION.


        ***************************************
        **  anderson darling normal test y2  **
        ***************************************


                ANDERSON-DARLING 1-SAMPLE TEST
                THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

     1. STATISTICS:
        NUMBER OF OBSERVATIONS                   =     1000
        MEAN                                     =   0.2034888E-01
        STANDARD DEVIATION                       =    1.321627

        ANDERSON-DARLING TEST STATISTIC VALUE =    5.826050
        ADJUSTED TEST STATISTIC VALUE         =    5.849208

     2. CRITICAL VALUES:
        90         % POINT    =   0.6560000
        95         % POINT    =   0.7870000
        97.5       % POINT    =   0.9180000
        99         % POINT    =    1.092000

     3. CONCLUSION (AT THE 5% LEVEL):
        THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.


        ***************************************
        **  anderson darling normal test y3  **
        ***************************************


                ANDERSON-DARLING 1-SAMPLE TEST
                THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

     1. STATISTICS:
        NUMBER OF OBSERVATIONS                   =     1000
        MEAN                                     =    1.503854
        STANDARD DEVIATION                       =    35.13059

        ANDERSON-DARLING TEST STATISTIC VALUE =    287.6429
        ADJUSTED TEST STATISTIC VALUE         =    288.7863
```

```
2. CRITICAL VALUES:
      90          % POINT    =   0.6560000
      95          % POINT    =   0.7870000
      97.5        % POINT    =   0.9180000
      99          % POINT    =    1.092000

3. CONCLUSION (AT THE 5% LEVEL):
      THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.


      ***************************************
      **  anderson darling normal test y4  **
      ***************************************


              ANDERSON-DARLING 1-SAMPLE TEST
              THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

1. STATISTICS:
      NUMBER OF OBSERVATIONS                =     1000
      MEAN                                  =    1.518372
      STANDARD DEVIATION                    =    1.719969

      ANDERSON-DARLING TEST STATISTIC VALUE =    83.06335
      ADJUSTED TEST STATISTIC VALUE         =    83.39352

2. CRITICAL VALUES:
      90          % POINT    =   0.6560000
      95          % POINT    =   0.7870000
      97.5        % POINT    =   0.9180000
      99          % POINT    =    1.092000

3. CONCLUSION (AT THE 5% LEVEL):
      THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.
```

Interpretation of the Sample Output

The output is divided into three sections.

1. The first section prints the number of observations and estimates for the location and scale parameters.
2. The second section prints the upper critical value for the Anderson-Darling test statistic distribution corresponding to various significance levels. The value in the first column, the confidence level of the test, is equivalent to $100(1-\alpha)$. We reject the null hypothesis at that significance level if the value of the Anderson-Darling test statistic printed in section one is greater than the critical value printed in the last column.
3. The third section prints the conclusion for a 95% test. For a different significance level, the appropriate conclusion can be drawn from the table printed in section two. For example, for $\alpha = 0.10$, we look at the row for 90% confidence and compare the critical value 1.062 to the

Anderson-Darling test statistic (for the normal data) 0.256. Since the test statistic is less than the critical value, we do not reject the null hypothesis at the α = 0.10 level.

As we would hope, the Anderson-Darling test accepts the hypothesis of normality for the normal random numbers and rejects it for the 3 non-normal cases.

The output from other statistical software programs may differ somewhat from the output above.

Questions

The Anderson-Darling test can be used to answer the following questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?

Importance

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the data follow a Weibull distribution.

There are many non-parametric and robust techniques that do not make strong distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than non-parametric and robust techniques. Therefore, if the distributional assumptions can be validated, they are generally preferred.

Related Techniques

Chi-Square goodness-of-fit Test
Kolmogorov-Smirnov Test
Shapiro-Wilk Normality Test
Probability Plot
Probability Plot Correlation Coefficient Plot

Case Study

Airplane glass failure time data.

Software

The Anderson-Darling goodness-of-fit test is available in some general purpose statistical software programs, including Dataplot.

1.3.5.15. Chi-Square Goodness-of-Fit Test

Purpose: Test for distributional adequacy

The chi-square test (Snedecor and Cochran, 1989) is used to test if a sample of data came from a population with a specific distribution.

An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The chi-square goodness-of-fit test is applied to binned data (i.e., data put into classes). This is actually not a restriction since for non-binned data you can simply calculate a histogram or frequency table before generating the chi-square test. However, the value of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

Additional discussion of the chi-square goodness-of-fit test is contained in the product and process comparisons chapter (chapter 7).

Definition

The chi-square test is defined for the hypothesis:

H_0:	The data follow a specified distribution.
H_a:	The data do not follow the specified distribution.

Test Statistic: For the chi-square goodness-of-fit computation, the data are divided into *k* bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{k}(O_i - E_i)^2/E_i$$

where O_i is the observed frequency for bin *i* and E_i is the expected frequency for bin *i*. The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l))$$

where F is the cumulative Distribution function for the distribution being tested, Y_u is the upper limit for class i, Y_l is the lower limit for class i, and *N* is the sample size.

This test is sensitive to the choice of bins. There is no optimal choice for the bin width (since the optimal bin width depends on the distribution). Most reasonable choices should produce similar, but not identical, results. Dataplot uses 0.3*s, where s is the sample standard deviation, for the class width. The lower and upper bins are at the sample mean plus and minus 6.0*s, respectively. For the chi-square approximation to be valid, the expected frequency should be at least 5. This test is not valid for small samples, and if some of the counts are less than five, you may need to combine some bins in the tails.

Significance Level: α.

Critical Region: The test statistic follows, approximately, a chi-square distribution with (*k* - *c*) degrees of freedom where *k* is the number of non-empty cells and *c* = the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution + 1. For example, for a 3-parameter Weibull distribution, *c* = 4.

Therefore, the hypothesis that the data are from a population with the specified distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha, k-c)}$$

where $\chi^2_{(\alpha, k-c)}$ is the chi-square percent point function with *k* - *c* degrees of freedom and a significance level of α.

In the above formulas for the critical regions, the Handbook follows the convention that χ^2_{α} is the upper critical value from the chi-square distribution and $\chi^2_{1-\alpha}$ is the lower critical value from the chi-square distribution. Note that this is the opposite of what is used in some texts and software programs. In particular, Dataplot uses the opposite convention.

Sample Output

Dataplot generated the following output for the chi-square test where 1,000 random numbers were generated for the normal, double exponential, *t* with 3 degrees of freedom, and lognormal distributions. In all cases, the chi-square test was applied to test for a normal distribution. The test statistics show the characteristics of the test; when the data are from a normal distribution, the test statistic is small and the hypothesis is accepted; when the data are from the double exponential, *t*, and lognormal distributions, the statistics are significant and the hypothesis of an underlying normal distribution is rejected at significance levels of 0.10, 0.05, and 0.01.

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the *t* random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

```
      *************************************************
      **  normal chi-square goodness of fit test y1  **
      *************************************************


                  CHI-SQUARED GOODNESS-OF-FIT TEST

NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION:            NORMAL

SAMPLE:
   NUMBER OF OBSERVATIONS      =     1000
   NUMBER OF NON-EMPTY CELLS   =       24
   NUMBER OF PARAMETERS USED   =        0

TEST:
CHI-SQUARED TEST STATISTIC     =    17.52155
   DEGREES OF FREEDOM          =       23
   CHI-SQUARED CDF VALUE       =    0.217101

   ALPHA LEVEL         CUTOFF              CONCLUSION
           10%       32.00690               ACCEPT H0
            5%       35.17246               ACCEPT H0
            1%       41.63840               ACCEPT H0

      CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
      AND EXPECTED FREQUENCY
      WRITTEN TO FILE DPST1F.DAT

      *************************************************
      **  normal chi-square goodness of fit test y2  **
      *************************************************
```

```
                    CHI-SQUARED GOODNESS-OF-FIT TEST

 NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
 ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
 DISTRIBUTION:            NORMAL

 SAMPLE:
    NUMBER OF OBSERVATIONS      =     1000
    NUMBER OF NON-EMPTY CELLS   =       26
    NUMBER OF PARAMETERS USED   =        0

 TEST:
 CHI-SQUARED TEST STATISTIC     =    2030.784
    DEGREES OF FREEDOM          =       25
    CHI-SQUARED CDF VALUE       =    1.000000

    ALPHA LEVEL         CUTOFF              CONCLUSION
            10%       34.38158               REJECT H0
             5%       37.65248               REJECT H0
             1%       44.31411               REJECT H0

       CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
       AND EXPECTED FREQUENCY
       WRITTEN TO FILE DPST1F.DAT

       *************************************************
       **  normal chi-square goodness of fit test y3  **
       *************************************************


                    CHI-SQUARED GOODNESS-OF-FIT TEST

 NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
 ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
 DISTRIBUTION:            NORMAL

 SAMPLE:
    NUMBER OF OBSERVATIONS      =     1000
    NUMBER OF NON-EMPTY CELLS   =       25
    NUMBER OF PARAMETERS USED   =        0

 TEST:
 CHI-SQUARED TEST STATISTIC     =    103165.4
    DEGREES OF FREEDOM          =       24
    CHI-SQUARED CDF VALUE       =    1.000000

    ALPHA LEVEL         CUTOFF              CONCLUSION
            10%       33.19624               REJECT H0
             5%       36.41503               REJECT H0
```

```
                 1%       42.97982              REJECT H0

           CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
           AND EXPECTED FREQUENCY
           WRITTEN TO FILE DPST1F.DAT

           *************************************************
           **  normal chi-square goodness of fit test y4  **
           *************************************************


                       CHI-SQUARED GOODNESS-OF-FIT TEST

     NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
     ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
     DISTRIBUTION:            NORMAL

     SAMPLE:
        NUMBER OF OBSERVATIONS      =     1000
        NUMBER OF NON-EMPTY CELLS   =       10
        NUMBER OF PARAMETERS USED   =        0

     TEST:
     CHI-SQUARED TEST STATISTIC     =    1162098.
        DEGREES OF FREEDOM          =        9
        CHI-SQUARED CDF VALUE       =    1.000000

        ALPHA LEVEL         CUTOFF              CONCLUSION
                10%       14.68366               REJECT H0
                 5%       16.91898               REJECT H0
                 1%       21.66600               REJECT H0

           CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
           AND EXPECTED FREQUENCY
           WRITTEN TO FILE DPST1F.DAT
```

As we would hope, the chi-square test does not reject the normality hypothesis for the normal distribution data set and rejects it for the three non-normal cases.

Questions

The chi-square test can be used to answer the following types of questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?
- Are the data from a binomial distribution?

Importance	Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the distribution of the data follows a Weibull distribution. There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique, such as the sign test, that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumption can be confirmed, the parametric techniques are generally preferred. If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, a non-parametric or robust technique may be required.
Related Techniques	Anderson-Darling Goodness-of-Fit Test Kolmogorov-Smirnov Test Shapiro-Wilk Normality Test Probability Plots Probability Plot Correlation Coefficient Plot
Case Study	Airplane glass failure times data.
Software	Some general purpose statistical software programs, including Dataplot, provide a chi-square goodness-of-fit test for at least some of the common distributions.

1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test

Purpose: Test for Distributional Adequacy

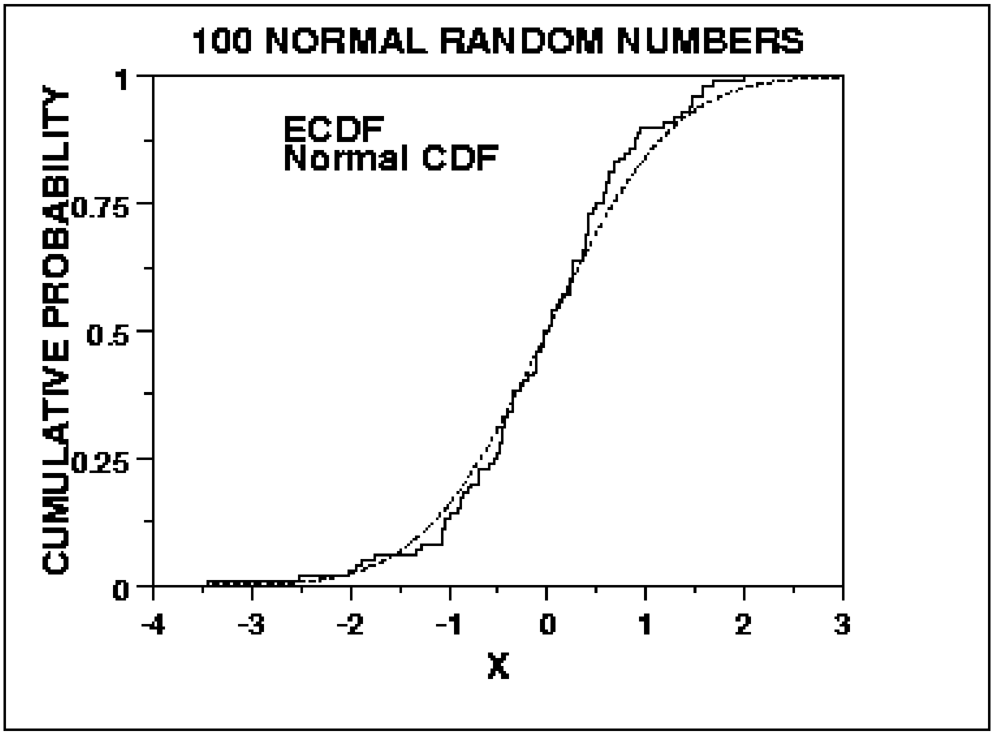
The Kolmogorov-Smirnov test (Chakravart, Laha, and Roy, 1967) is used to decide if a sample comes from a population with a specific distribution.

The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given N *ordered* data points $Y_1, Y_2, ..., Y_N$, the ECDF is defined as

$$E_N = n(i)/N$$

where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by $1/N$ at the value of each ordered data point.

The graph below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.



Characteristics and Limitations of the K-S Test

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important limitations:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

Due to limitations 2 and 3 above, many analysts prefer to use the Anderson-Darling goodness-of-fit test. However, the Anderson-Darling test is only available for a few specific distributions.

Definition

The Kolmogorov-Smirnov test is defined by:

H_0:	The data follow a specified distribution
H_a:	The data do not follow the specified distribution
Test Statistic:	The Kolmogorov-Smirnov test statistic is defined as $$D = \max_{1 \le i \le N} \left\| F(Y_i) - \frac{i}{N} \right\|$$ where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).
Significance Level:	α.
Critical Values:	The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated. We do not provide the K-S tables in the Handbook since software programs that perform a K-S test will provide the relevant critical values.

Sample Output

Dataplot generated the following output for the Kolmogorov-Smirnov test where 1,000 random numbers were generated for a normal, double exponential, *t* with 3 degrees of freedom, and lognormal distributions. In all cases, the Kolmogorov-Smirnov test was applied to test for a normal distribution. The Kolmogorov-Smirnov test accepts the normality hypothesis for the case of normal data and rejects it for the double exponential, *t*, and lognormal data with the exception of the double exponential data being significant at the 0.01 significance level.

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the *t* random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

```
*************************************************************
**  normal Kolmogorov-Smirnov goodness of fit test y1   **
*************************************************************
```

```
                  KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION:            NORMAL
   NUMBER OF OBSERVATIONS      =     1000

TEST:
KOLMOGOROV-SMIRNOV TEST STATISTIC     =   0.2414924E-01

   ALPHA LEVEL         CUTOFF              CONCLUSION
           10%       0.03858               ACCEPT H0
            5%       0.04301               ACCEPT H0
            1%       0.05155               ACCEPT H0

      *********************************************************
      **  normal Kolmogorov-Smirnov goodness of fit test y2  **
      *********************************************************


                  KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION:            NORMAL
   NUMBER OF OBSERVATIONS      =     1000

TEST:
KOLMOGOROV-SMIRNOV TEST STATISTIC     =   0.5140864E-01

   ALPHA LEVEL         CUTOFF              CONCLUSION
           10%       0.03858               REJECT H0
            5%       0.04301               REJECT H0
            1%       0.05155               ACCEPT H0

      *********************************************************
      **  normal Kolmogorov-Smirnov goodness of fit test y3  **
      *********************************************************


                  KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION:            NORMAL
   NUMBER OF OBSERVATIONS      =     1000

TEST:
KOLMOGOROV-SMIRNOV TEST STATISTIC     =   0.6119353E-01

   ALPHA LEVEL         CUTOFF              CONCLUSION
           10%       0.03858               REJECT H0
            5%       0.04301               REJECT H0
            1%       0.05155               REJECT H0

      *********************************************************
      **  normal Kolmogorov-Smirnov goodness of fit test y4  **
```

```
        ********************************************************

                    KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

 NULL HYPOTHESIS H0:      DISTRIBUTION FITS THE DATA
 ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
 DISTRIBUTION:            NORMAL
    NUMBER OF OBSERVATIONS      =     1000

 TEST:
 KOLMOGOROV-SMIRNOV TEST STATISTIC     =   0.5354889

    ALPHA LEVEL         CUTOFF              CONCLUSION
            10%       0.03858                REJECT H0
             5%       0.04301                REJECT H0
             1%       0.05155                REJECT H0
```

Questions

The Kolmogorov-Smirnov test can be used to answer the following types of questions:

- Are the data from a normal distribution?
- Are the data from a log-normal distribution?
- Are the data from a Weibull distribution?
- Are the data from an exponential distribution?
- Are the data from a logistic distribution?

Importance

Many statistical tests and procedures are based on specific distributional assumptions. The assumption of normality is particularly common in classical statistical tests. Much reliability modeling is based on the assumption that the data follow a Weibull distribution.

There are many non-parametric and robust techniques that are not based on strong distributional assumptions. By non-parametric, we mean a technique, such as the sign test, that is not based on a specific distributional assumption. By robust, we mean a statistical technique that performs well under a wide range of distributional assumptions. However, techniques based on specific distributional assumptions are in general more powerful than these non-parametric and robust techniques. By power, we mean the ability to detect a difference when that difference actually exists. Therefore, if the distributional assumptions can be confirmed, the parametric techniques are generally preferred.

If you are using a technique that makes a normality (or some other type of distributional) assumption, it is important to confirm that this assumption is in fact justified. If it is, the more powerful parametric techniques can be used. If the distributional assumption is not justified, using a non-parametric or robust technique may be required.

Related Techniques	Anderson-Darling goodness-of-fit Test Chi-Square goodness-of-fit Test Shapiro-Wilk Normality Test Probability Plots Probability Plot Correlation Coefficient Plot
Case Study	Airplane glass failure times data
Software	Some general purpose statistical software programs, including Dataplot, support the Kolmogorov-Smirnov goodness-of-fit test, at least for some of the more common distributions.

1.3.5.17. Grubbs' Test for Outliers

Purpose: Detection of Outliers

Grubbs' test (Grubbs 1969 and Stefansky 1972) is used to detect outliers in a univariate data set. It is based on the assumption of normality. That is, you should first verify that your data can be reasonably approximated by a normal distribution before applying the Grubbs' test.

Grubbs' test detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or less since it frequently tags most of the points as outliers.

Grubbs' test is also known as the maximum normed residual test.

Definition

Grubbs' test is defined for the hypothesis:

H_0: There are no outliers in the data set

H_a: There is at least one outlier in the data set

Test Statistic: The Grubbs' test statistic is defined as:

$$G = \frac{\max |Y_i - \bar{Y}|}{s}$$

where $\bar{Y}$ and s are the sample mean and standard deviation. The Grubbs test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation.

Significance Level: α.

Critical Region:

The hypothesis of no outliers is rejected if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/(2N),N-2)}}{N-2+t^2_{(\alpha/(2N),N-2)}}}$$

where $t_{(\alpha/(2N),N-2)}$ is the critical value of the *t*-distribution with (N-2) degrees of freedom and a significance level of $\alpha/(2N)$.

In the above formulas for the critical regions, the Handbook follows the convention that t_α is the upper critical value from the *t*-distribution and $t_{1-\alpha}$ is the lower critical value from the *t*-distribution. Note that this is the opposite of what is used in some texts and software programs. In particular, Dataplot uses the opposite convention.

Sample Output

Dataplot generated the following output for the ZARR13.DAT data set showing that Grubbs' test finds no outliers in the dataset:

```
*********************
**  grubbs test y  **
*********************


        GRUBBS TEST FOR OUTLIERS
        (ASSUMPTION: NORMALITY)

1. STATISTICS:
      NUMBER OF OBSERVATIONS      =      195
      MINIMUM                     =    9.196848
      MEAN                        =    9.261460
      MAXIMUM                     =    9.327973
   STANDARD DEVIATION          =   0.2278881E-01

   GRUBBS TEST STATISTIC       =    2.918673

2. PERCENT POINTS OF THE REFERENCE DISTRIBUTION
   FOR GRUBBS TEST STATISTIC
      0          % POINT    =          0.
      50         % POINT    =    2.984294
      75         % POINT    =    3.181226
      90         % POINT    =    3.424672
      95         % POINT    =    3.597898
      99         % POINT    =    3.970215

         37.59665      % POINT:        2.918673

3. CONCLUSION (AT THE 5% LEVEL):
      THERE ARE NO OUTLIERS.
```

Interpretation of Sample Output

The output is divided into three sections.

1. The first section prints the sample statistics used in the computation of the Grubbs' test and the value of the Grubbs' test statistic.
2. The second section prints the upper critical value for the Grubbs' test statistic distribution corresponding to various significance levels. The value in the first column, the confidence level of the test, is equivalent to 100(1-α). We reject the null hypothesis at that significance level if the value of the Grubbs' test statistic printed in section one is greater than the critical value printed in the last column.
3. The third section prints the conclusion for a 95% test. For a different significance level, the appropriate conclusion can be drawn from the table printed in section two. For example, for α = 0.10, we look at the row for 90% confidence and compare the critical value 3.42 to the Grubbs' test statistic 2.92. Since the test statistic is less than the critical value, we accept the null hypothesis at the α = 0.10 level.

Output from other statistical software may look somewhat different from the above output.

Questions

Grubbs' test can be used to answer the following questions:

1. Does the data set contain any outliers?
2. How many outliers does it contain?

Importance

Many statistical techniques are sensitive to the presence of outliers. For example, simple calculations of the mean and standard deviation may be distorted by a single grossly inaccurate data point.

Checking for outliers should be a routine part of any data analysis. Potential outliers should be examined to see if they are possibly erroneous. If the data point is in error, it should be corrected if possible and deleted if it is not possible. If there is no reason to believe that the outlying point is in error, it should not be deleted without careful consideration. However, the use of more robust techniques may be warranted. Robust techniques will often downweight the effect of outlying points without deleting them.

Related Techniques	Several graphical techniques can, and should, be used to detect outliers. A simple run sequence plot, a box plot, or a histogram should show any obviously outlying points. Run Sequence Plot Histogram Box Plot Normal Probability Plot Lag Plot
Case Study	Heat flow meter data.
Software	Some general purpose statistical software programs, including Dataplot, support the Grubbs' test.

1.3.5.18. Yates Analysis

Purpose: Estimate Factor Effects in a 2-Level Factorial Design

Full factorial and fractional factorial designs are common in designed experiments for engineering and scientific applications.

In these designs, each factor is assigned two levels. These are typically called the low and high levels. For computational purposes, the factors are scaled so that the low level is assigned a value of -1 and the high level is assigned a value of +1. These are also commonly referred to as "-" and "+".

A full factorial design contains all possible combinations of low/high levels for all the factors. A fractional factorial design contains a carefully chosen subset of these combinations. The criterion for choosing the subsets is discussed in detail in the process improvement chapter.

The Yates analysis exploits the special structure of these designs to generate least squares estimates for factor effects for all factors and all relevant interactions.

The mathematical details of the Yates analysis are given in chapter 10 of Box, Hunter, and Hunter (1978).

The Yates analysis is typically complemented by a number of graphical techniques such as the dex mean plot and the dex contour plot ("dex" represents "design of experiments"). This is demonstrated in the Eddy current case study.

Yates Order

Before performing a Yates analysis, the data should be arranged in "Yates order". That is, given *k* factors, the *k*th column consists of 2^{k-1} minus signs (i.e., the low level of the factor) followed by 2^{k-1} plus signs (i.e., the high level of the factor). For example, for a full factorial design with three factors, the design matrix is

```
- - -
+ - -
- + -
+ + -
- - +
+ - +
- + +
+ + +
```

Determining the Yates order for fractional factorial designs requires knowledge of the confounding structure of the fractional factorial design.

Yates Output

A Yates analysis generates the following output.

1. A factor identifier (from Yates order). The specific identifier will vary depending on the program used to generate the Yates analysis. Dataplot, for example, uses the following for a 3-factor model.

1 =	factor 1
2 =	factor 2
3 =	factor 3
12 =	interaction of factor 1 and factor 2
13 =	interaction of factor 1 and factor 3
23 =	interaction of factor 2 and factor 3
123 =	interaction of factors 1, 2, and 3

2. Least squares estimated factor effects ordered from largest in magnitude (most significant) to smallest in magnitude (least significant).

   That is, we obtain a ranked list of important factors.

3. A *t*-value for the individual factor effect estimates. The *t*-value is computed as

   $$t = \frac{e}{s_e}$$

   where *e* is the estimated factor effect and s_e is the standard deviation of the estimated factor effect.

4. The residual standard deviation that results from the model with the single term only. That is, the residual standard deviation from the model

   response = constant + 0.5 (X_i)

where X_i is the estimate of the ith factor or interaction effect.

5. The cumulative residual standard deviation that results from the model using the current term plus all terms preceding that term. That is,

   response = constant + 0.5 (all effect estimates down to and including the effect of interest)

   This consists of a monotonically decreasing set of residual standard deviations (indicating a better fit as the number of terms in the model increases). The first cumulative residual standard deviation is for the model

   response = constant

   where the constant is the overall mean of the response variable. The last cumulative residual standard deviation is for the model

   response = constant + 0.5*(all factor and interaction estimates)

   This last model will have a residual standard deviation of zero.

Sample Output

Dataplot generated the following Yates analysis output for the Eddy current data set:

```
(NOTE--DATA MUST BE IN STANDARD ORDER)
NUMBER OF OBSERVATIONS              =         8
NUMBER OF FACTORS                   =         3
NO REPLICATION CASE

PSEUDO-REPLICATION STAND. DEV.      =    0.20152531564E+00
PSEUDO-DEGREES OF FREEDOM           =         1
(THE PSEUDO-REP. STAND. DEV. ASSUMES ALL
3, 4, 5, ...-TERM INTERACTIONS ARE NOT REAL,
BUT MANIFESTATIONS OF RANDOM ERROR)

STANDARD DEVIATION OF A COEF.       =    0.14249992371E+00
(BASED ON PSEUDO-REP. ST. DEV.)

GRAND MEAN                          =    0.26587500572E+01
GRAND STANDARD DEVIATION            =    0.17410624027E+01

99% CONFIDENCE LIMITS (+-)          =    0.90710897446E+01
95% CONFIDENCE LIMITS (+-)          =    0.18106349707E+01
99.5% POINT OF T DISTRIBUTION       =    0.63656803131E+02
97.5% POINT OF T DISTRIBUTION       =    0.12706216812E+02

IDENTIFIER    EFFECT          T VALUE      RESSD:     RESSD:
                                           MEAN +     MEAN +
                                           TERM     CUM TERMS
----------------------------------------------------------
```

```
MEAN        2.65875                  1.74106    1.74106
   1        3.10250        21.8*     0.57272    0.57272
   2       -0.86750        -6.1      1.81264    0.30429
  23        0.29750         2.1      1.87270    0.26737
  13        0.24750         1.7      1.87513    0.23341
   3        0.21250         1.5      1.87656    0.19121
 123        0.14250         1.0      1.87876    0.18031
  12        0.12750         0.9      1.87912    0.00000
```

Interpretation of Sample Output

In summary, the Yates analysis provides us with the following ranked list of important factors along with the estimated effect estimate.

1. X1: effect estimate = 3.1025 ohms
2. X2: effect estimate = -0.8675 ohms
3. X2*X3: effect estimate = 0.2975 ohms
4. X1*X3: effect estimate = 0.2475 ohms
5. X3: effect estimate = 0.2125 ohms
6. X1*X2*X3: effect estimate = 0.1425 ohms
7. X1*X2: effect estimate = 0.1275 ohms

Model Selection and Validation

From the above Yates output, we can define the potential models from the Yates analysis. An important component of a Yates analysis is selecting the best model from the available potential models.

Once a tentative model has been selected, the error term should follow the assumptions for a univariate measurement process. That is, the model should be validated by analyzing the residuals.

Graphical Presentation

Some analysts may prefer a more graphical presentation of the Yates results. In particular, the following plots may be useful:

1. Ordered data plot
2. Ordered absolute effects plot
3. Cumulative residual standard deviation plot

Questions

The Yates analysis can be used to answer the following questions:

1. What is the ranked list of factors?
2. What is the goodness-of-fit (as measured by the residual standard deviation) for the various models?

Related Techniques

Multi-factor analysis of variance
Dex mean plot
Block plot
Dex contour plot

Case Study

The Yates analysis is demonstrated in the Eddy current case study.

Software

Many general purpose statistical software programs, including Dataplot, can perform a Yates analysis.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.5.18.1. Defining Models and Prediction Equations

Parameter Estimates Don't Change as Additional Terms Added

In most cases of least squares fitting, the model coefficients for previously added terms change depending on what was successively added. For example, the X1 coefficient might change depending on whether or not an X2 term was included in the model. This is **not** the case when the design is orthogonal, as is a 2^3 full factorial design. For orthogonal designs, the estimates for the previously included terms do not change as additional terms are added. This means the ranked list of effect estimates simultaneously serves as the least squares coefficient estimates for progressively more complicated models.

Yates Table

For convenience, we list the sample Yates output for the Eddy current data set here.

```
(NOTE--DATA MUST BE IN STANDARD ORDER)
NUMBER OF OBSERVATIONS              =        8
NUMBER OF FACTORS                   =        3
NO REPLICATION CASE

PSEUDO-REPLICATION STAND. DEV.      =    0.20152531564E+00
PSEUDO-DEGREES OF FREEDOM           =        1
(THE PSEUDO-REP. STAND. DEV. ASSUMES ALL
3, 4, 5, ...-TERM INTERACTIONS ARE NOT REAL,
BUT MANIFESTATIONS OF RANDOM ERROR)

STANDARD DEVIATION OF A COEF.       =    0.14249992371E+00
(BASED ON PSEUDO-REP. ST. DEV.)

GRAND MEAN                          =    0.26587500572E+01
GRAND STANDARD DEVIATION            =    0.17410624027E+01

99% CONFIDENCE LIMITS (+-)          =    0.90710897446E+01
95% CONFIDENCE LIMITS (+-)          =    0.18106349707E+01
99.5% POINT OF T DISTRIBUTION       =    0.63656803131E+02
```

```
97.5% POINT OF T DISTRIBUTION    =    0.12706216812E+02

IDENTIFIER    EFFECT        T VALUE      RESSD:      RESSD:
                                         MEAN +      MEAN +
                                         TERM      CUM TERMS
-----------------------------------------------------------
   MEAN       2.65875                    1.74106     1.74106
      1       3.10250        21.8*       0.57272     0.57272
      2      -0.86750        -6.1        1.81264     0.30429
     23       0.29750         2.1        1.87270     0.26737
     13       0.24750         1.7        1.87513     0.23341
      3       0.21250         1.5        1.87656     0.19121
    123       0.14250         1.0        1.87876     0.18031
     12       0.12750         0.9        1.87912     0.00000
```

The last column of the Yates table gives the residual standard deviation for 8 possible models, each with one more term than the previous model.

Potential Models

For this example, we can summarize the possible prediction equations using the second and last columns of the Yates table:

- $$\hat{Y} = 2.65875$$
  has a residual standard deviation of 1.74106 ohms. Note that this is the default model. That is, if no factors are important, the model is simply the overall mean.
- $$\hat{Y} = 2.65875 + 0.5(3.1025X1)$$
  has a residual standard deviation of 0.57272 ohms. (Here, X1 is either a +1 or -1, and similarly for the other factors and interactions (products).)
- $$\hat{Y} = 2.65875 + 0.5(3.1025X1 - 0.8675X2)$$
  has a residual standard deviation of 0.30429 ohms.
- $$\hat{Y} = 2.65875 + 0.5(3.1025X1 - 0.8675X2 + 0.2975X2 * X3)$$
  has a residual standard deviation of 0.26737 ohms.
- $$\begin{aligned}\hat{Y} = \; & 2.65875 + 0.5(3.1025X1 - 0.8675X2 + \\ & 0.2975X2 * X3 + 0.2475X1 * X3)\end{aligned}$$
  has a residual standard deviation of 0.23341 ohms
- $$\begin{aligned}\hat{Y} = \; & 2.65875 + 0.5(3.1025X1 - 0.8675X2 + \\ & 0.2975X2 * X3 + 0.2475X1 * X3 + 0.2125X3)\end{aligned}$$
  has a residual standard deviation of 0.19121 ohms.

$$\hat{Y} = 2.65875 + 0.5(3.1025X1 - 0.8675X2 + 0.2975X2*X3 + 0.2475X1*X3 + 0.2125X3 + 0.1425X1*X2*X3)$$

has a residual standard deviation of 0.18031 ohms.

- $$\hat{Y} = 2.65875 + 0.5(3.1025X1 - 0.8675X2 + 0.2975X2*X3 + 0.2475X1*X3 + 0.2125X3 + 0.1425X1*X2*X3 + 0.1275X1*X2)$$

  has a residual standard deviation of 0.0 ohms. Note that the model with all possible terms included will have a zero residual standard deviation. This will always occur with an unreplicated two-level factorial design.

Model Selection

The above step lists all the potential models. From this list, we want to select the most appropriate model. This requires balancing the following two goals.

1. We want the model to include all important factors.
2. We want the model to be parsimonious. That is, the model should be as simple as possible.

Note that the residual standard deviation alone is insufficient for determining the most appropriate model as it will always be decreased by adding additional factors. The next section describes a number of approaches for determining which factors (and interactions) to include in the model.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.5.18.2. Important Factors

Identify Important Factors

The Yates analysis generates a large number of potential models. From this list, we want to select the most appropriate model. This requires balancing the following two goals.

1. We want the model to include all important factors.
2. We want the model to be parsimonious. That is, the model should be as simple as possible.

In short, we want our model to include all the important factors and interactions and to omit the unimportant factors and interactions.

Seven criteria are utilized to define important factors. These seven criteria are not all equally important, nor will they yield identical subsets, in which case a consensus subset or a weighted consensus subset must be extracted. In practice, some of these criteria may not apply in all situations.

These criteria will be examined in the context of the Eddy current data set. The Yates Analysis page gave the sample Yates output for these data and the Defining Models and Predictions page listed the potential models from the Yates analysis.

In practice, not all of these criteria will be used with every analysis (and some analysts may have additional criteria). These critierion are given as useful guidelines. Mosts analysts will focus on those criteria that they find most useful.

Criteria for Including Terms in the Model

The seven criteria that we can use in determining whether to keep a factor in the model can be summarized as follows.

1. Effects: Engineering Significance
2. Effects: Order of Magnitude
3. Effects: Statistical Significance
4. Effects: Probability Plots
5. Averages: Youden Plot
6. Residual Standard Deviation: Engineering Significance
7. Residual Standard Deviation: Statistical Significance

The first four criteria focus on effect estimates with three numeric criteria and one graphical criteria. The fifth criteria focuses on averages. The last two criteria focus on the residual standard deviation of the model. We discuss each of these seven criteria in detail in the following sections.

The last section summarizes the conclusions based on all of the criteria.

Effects: Engineering Significance

The minimum engineering significant difference is defined as

$$|\hat{\beta}_i| > \Delta$$

where $|\hat{\beta}_i|$ is the absolute value of the parameter estimate (i.e., the effect) and Δ is the minimum engineering significant difference.

That is, declare a factor as "important" if the effect is greater than some a priori declared engineering difference. This implies that the engineering staff have in fact stated what a minimum effect will be. Oftentimes this is not the case. In the absence of an a priori difference, a good rough rule for the minimum engineering significant Δ is to keep only those factors whose effect is greater than, say, 10% of the current production average. In this case, let's say that the average detector has a sensitivity of 2.5 ohms. This would suggest that we would declare all factors whose effect is greater than 10% of 2.5 ohms = 0.25 ohm to be significant (from an engineering point of view).

Based on this minimum engineering significant difference criterion, we conclude that we should keep two terms: X1 and X2.

Effects: Order of Magnitude

The order of magnitude criterion is defined as

$$|\hat{\beta}_i| < 0.10 * max|\hat{\beta}_i|$$

That is, exclude any factor that is less than 10% of the maximum effect size. We may or may not keep the other factors. This criterion is neither engineering nor statistical, but it does offer some additional numerical insight. For the current example, the largest effect is from X1 (3.10250 ohms), and so 10% of that is 0.31 ohms, which suggests keeping all factors whose effects exceed 0.31 ohms.

Based on the order-of-magnitude criterion, we thus conclude that we should keep two terms: X1 and X2. A third term, X2*X3 (.29750), is just slightly under the cutoff level, so we may consider keeping it based on the other criterion.

Effects: Statistical Significance

Statistical significance is defined as

$$|\hat{\beta}_i| > 2sd(\hat{\beta}_i) = 2(\frac{2\sigma}{\sqrt{n}})$$

That is, declare a factor as important if its effect is more than 2 standard deviations away from 0 (0, by definition, meaning "no effect").

The "2" comes from normal theory (more specifically, a value of 1.96 yields a 95% confidence interval). More precise values would come from *t*-distribution theory.

The difficulty with this is that in order to invoke this criterion we need the standard deviation, σ, of an observation. This is problematic because

1. the engineer may not know σ;
2. the experiment might not have replication, and so a model-free estimate of σ is not obtainable;
3. obtaining an estimate of σ by assuming the sometimes- employed assumption of ignoring 3-term interactions and higher may be incorrect from an engineering point of view.

For the Eddy current example:

1. the engineer did **not** know σ;
2. the design (a 2^3 full factorial) did **not** have replication;

3. ignoring 3-term interactions and higher interactions leads to an estimate of σ based on omitting only a single term: the X1*X2*X3 interaction.

For the current example, if one assumes that the 3-term interaction is nil and hence represents a single drawing from a population centered at zero, then an estimate of the standard deviation of an effect is simply the estimate of the 3-factor interaction (0.1425). In the Dataplot output for our example, this is the effect estimate for the X1*X2*X3 interaction term (the EFFECT column for the row labeled "123"). Two standard deviations is thus 0.2850. For this example, the rule is thus to keep all $|\hat{\beta}_i|$ > 0.2850.

This results in keeping three terms: X1 (3.10250), X2 (-.86750), and X1*X2 (.29750).

Effects: Probability Plots

Probability plots can be used in the following manner.

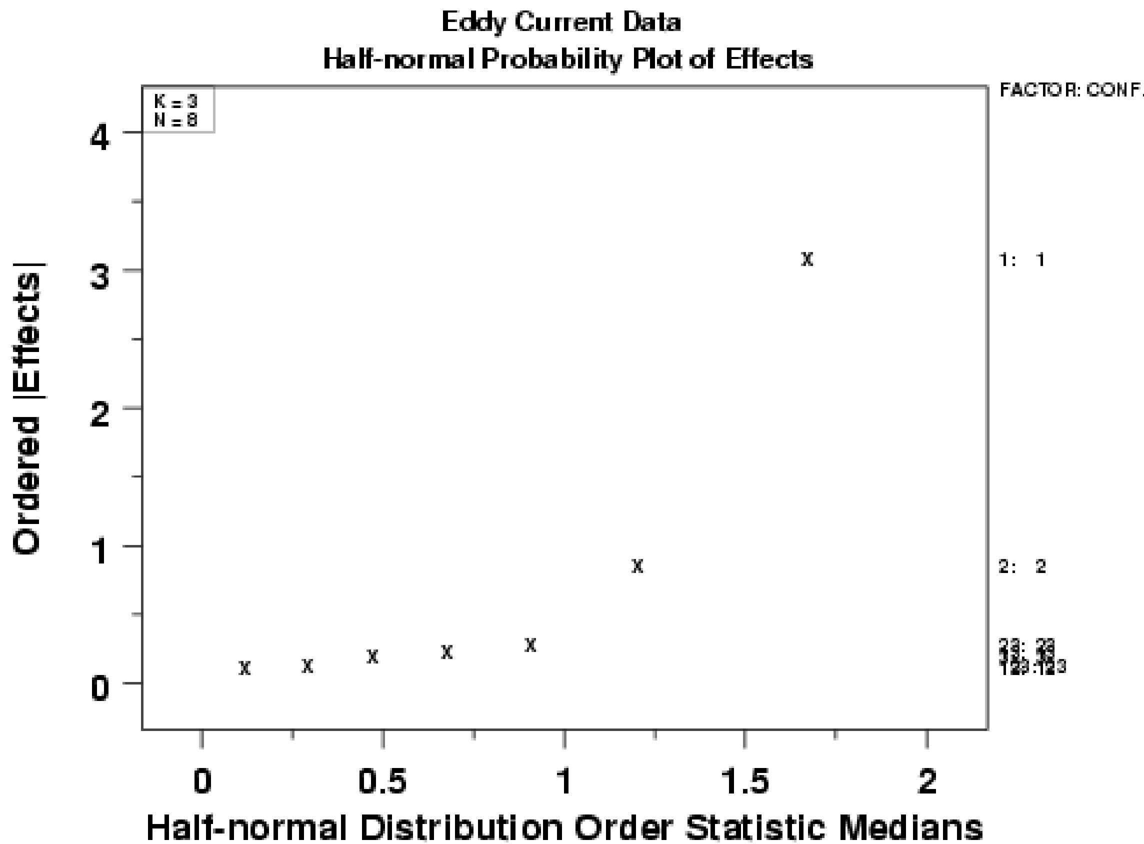
1. Normal Probability Plot: Keep a factor as "important" if it is well off the line through zero on a normal probability plot of the effect estimates.
2. Half-Normal Probability Plot: Keep a factor as "important" if it is well off the line near zero on a half-normal probability plot of the absolute value of effect estimates.

Both of these methods are based on the fact that the least squares estimates of effects for these 2-level orthogonal designs are simply the difference of averages and so the central limit theorem, loosely applied, suggests that (if no factor were important) the effect estimates should have approximately a normal distribution with mean zero and the absolute value of the estimates should have a half-normal distribution.

Since the half-normal probability plot is only concerned with effect magnitudes as opposed to signed effects (which are subject to the vagaries of how the initial factor codings +1 and -1 were assigned), the half-normal probability plot is preferred by some over the normal probability plot.

Normal Probablity Plot of Effects and Half-Normal Probability Plot of Effects

The following half-normal plot shows the normal probability plot of the effect estimates and the half-normal probability plot of the absolute value of the estimates for the Eddy current data.



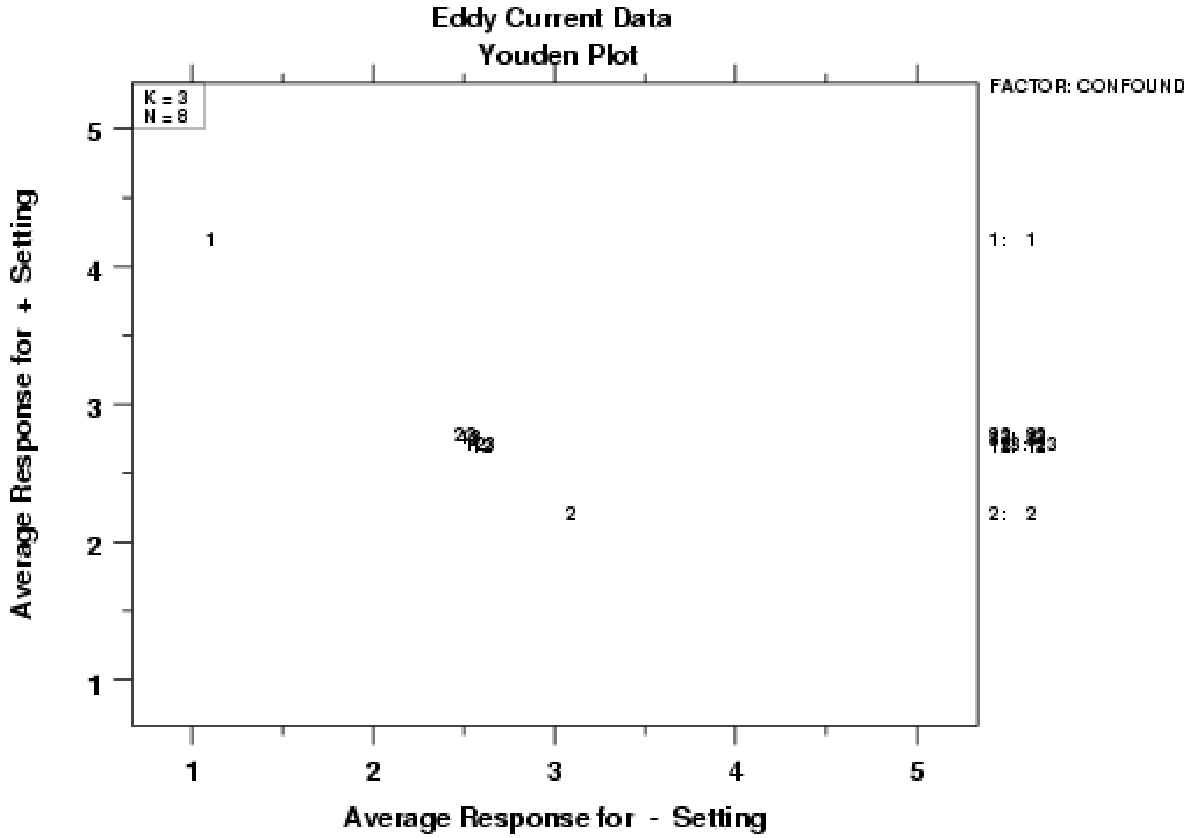
For the example at hand, both probability plots clearly show two factors displaced off the line, and from the third plot (with factor tags included), we see that those two factors are factor 1 and factor 2. All of the remaining five effects are behaving like random drawings from a normal distribution centered at zero, and so are deemed to be statistically non-significant. In conclusion, this rule keeps two factors: X1 (3.10250) and X2 (-.86750).

Effects: Youden Plot

A Youden plot can be used in the following way. Keep a factor as "important" if it is displaced away from the central-tendancy "bunch" in a Youden plot of high and low averages. By definition, a factor is important when its average response for the low (-1) setting is significantly different from its average response for the high (+1) setting. Conversely, if the low and high averages are about the same, then what difference does it make which setting to use and so why would such a factor be considered important? This fact in combination with the intrinsic benefits of the Youden plot for comparing pairs of items leads to the technique of generating a Youden plot of the low and high averages.

Youden Plot of Effect Estimatess

The following is the Youden plot of the effect estimatess for the Eddy current data.



For the example at hand, the Youden plot clearly shows a cluster of points near the grand average (2.65875) with two displaced points above (factor 1) and below (factor 2). Based on the Youden plot, we conclude to keep two factors: X1 (3.10250) and X2 (-.86750).

Residual Standard Deviation: Engineering Significance

This criterion is defined as

Residual Standard Deviation > Cutoff

That is, declare a factor as "important" if the cumulative model that includes the factor (and all larger factors) has a residual standard deviation smaller than an a priori engineering-specified minimum residual standard deviation.

This criterion is different from the others in that it is model focused. In practice, this criterion states that starting with the largest effect, we cumulatively keep adding terms to the model and monitor how the residual standard deviation for each progressively more complicated model becomes smaller. At some point, the cumulative model will become complicated enough and comprehensive enough that the resulting residual standard deviation will drop below the pre-specified engineering cutoff for the residual standard deviation. At that point, we stop adding terms and declare all of the model-included terms to be "important" and everything not in the model to be "unimportant".

This approach implies that the engineer has considered what a minimum residual standard deviation should be. In effect, this relates to what the engineer can tolerate for the magnitude of the typical residual (= difference between the raw data and the predicted value from the model).

In other words, how good does the engineer want the prediction equation to be. Unfortunately, this engineering specification has not always been formulated and so this criterion can become moot.

In the absence of a prior specified cutoff, a good rough rule for the minimum engineering residual standard deviation is to keep adding terms until the residual standard deviation just dips below, say, 5% of the current production average. For the Eddy current data, let's say that the average detector has a sensitivity of 2.5 ohms. Then this would suggest that we would keep adding terms to the model until the residual standard deviation falls below 5% of 2.5 ohms = 0.125 ohms.

Based on the minimum residual standard deviation criteria, and by scanning the far right column of the Yates table, we would conclude to keep the following terms:

1. X1 (with a cumulative residual standard deviation = 0.57272)
2. X2 (with a cumulative residual standard deviation = 0.30429)
3. X2*X3 (with a cumulative residual standard deviation = 0.26737)
4. X1*X3 (with a cumulative residual standard deviation = 0.23341)
5. X3 (with a cumulative residual standard deviation = 0.19121)
6. X1*X2*X3 (with a cumulative residual standard deviation = 0.18031)
7. X1*X2 (with a cumulative residual standard deviation = 0.00000)

Note that we must include **all** terms in order to drive the residual standard deviation below 0.125. Again, the 5% rule is a rough-and-ready rule that has no basis in engineering or statistics, but is simply a "numerics". Ideally, the engineer has a better cutoff for the residual standard deviation that is based on how well he/she wants the equation to peform in practice. If such a number were available, then for this criterion and data set we would select something less than the entire collection of terms.

Residual Standard Deviation: Statistical Significance

This criterion is defined as

Residual Standard Deviation > σ

where σ is the standard deviation of an observation under replicated conditions.

That is, declare a term as "important" until the cumulative model that includes the term has a residual standard deviation smaller than σ. In essence, we are allowing that we cannot demand a model fit any better than what we would obtain if we had replicated data; that is, we cannot demand that the residual standard deviation from any fitted model be any smaller than the (theoretical or actual) replication standard deviation. We can drive the fitted standard deviation down (by adding terms) until it achieves a value close to σ, but to attempt to drive it down further means that we are, in effect, trying to fit noise.

In practice, this criterion may be difficult to apply because

1. the engineer may not know σ;
2. the experiment might not have replication, and so a model-free estimate of σ is not obtainable.

For the current case study:

1. the engineer did **not** know σ;
2. the design (a 2^3 full factorial) did **not** have replication. The most common way of having replication in such designs is to have replicated center points at the center of the cube ((X1,X2,X3) = (0,0,0)).

Thus for this current case, this criteria could **not** be used to yield a subset of "important" factors.

Conclusions

In summary, the seven criteria for specifying "important" factors yielded the following for the Eddy current data:

1. Effects, Engineering Significance:	X1, X2
2. Effects, Numerically Significant:	X1, X2
3. Effects, Statistically Significant:	X1, X2, X2*X3
4. Effects, Probability Plots:	X1, X2
5. Averages, Youden Plot:	X1, X2
6. Residual SD, Engineering Significance:	all 7 terms
7. Residual SD, Statistical Significance:	not applicable

Such conflicting results are common. Arguably, the three most important criteria (listed in order of most important) are:

4. Effects, Probability Plots:	X1, X2
1. Effects, Engineering Significance:	X1, X2
3. Residual SD, Engineering Significance:	all 7 terms

Scanning all of the above, we thus declare the following consensus for the Eddy current data:

1. Important Factors: X1 and X2
2. Parsimonious Prediction Equation:

$$\hat{Y} = 2.65875 + 0.5(3.1025X1 - 0.8675X2)$$

(with a residual standard deviation of .30429 ohms)

Note that this is the initial model selection. We still need to perform model validation with a residual analysis.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6. Probability Distributions

Probability Distributions

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level.

Some practical uses of probability distributions are:

- To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
- For univariate data, it is often useful to determine a reasonable distributional model for the data.
- Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies with random numbers generated from using a specific probability distribution are often needed.

Table of Contents

1.3.6.1. What is a Probability Distribution

Discrete Distributions

The mathematical definition of a discrete probability function, p(x), is a function that satisfies the following properties.

1. The probability that x can take a specific value is p(x). That is

   $$P[X = x] = p(x) = p_x$$

2. p(x) is non-negative for all real x.
3. The sum of p(x) over all possible values of x is 1, that is

   $$\sum_j p_j = 1$$

   where j represents all possible values that x can have and p_j is the probability at x_j.

   One consequence of properties 2 and 3 is that 0 <= p(x) <= 1.

What does this actually mean? A discrete probability function is a function that can take a discrete number of values (not necessarily finite). This is most often the non-negative integers or some subset of the non-negative integers. There is no mathematical restriction that discrete probability functions only be defined at integers, but in practice this is usually what makes sense. For example, if you toss a coin 6 times, you can get 2 heads or 3 heads but not 2 1/2 heads. Each of the discrete values has a certain probability of occurrence that is between zero and one. That is, a discrete function that allows negative values or values greater than one is not a probability function. The condition that the probabilities sum to one means that at least one of the values has to occur.

Continuous Distributions

The mathematical definition of a continuous probability function, f(x), is a function that satisfies the following properties.

1. The probability that x is between two points a and b is

$$p[a \le x \le b] = \int_a^b f(x)dx$$

2. It is non-negative for all real x.
3. The integral of the probability function is one, that is

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

What does this actually mean? Since continuous probability functions are defined for an infinite number of points over a continuous interval, the probability at a single point is always zero. Probabilities are measured over intervals, not single points. That is, the area under the curve between two distinct points defines the probability for that interval. This means that the height of the probability function can in fact be greater than one. The property that the integral must equal one is equivalent to the property for discrete distributions that the sum of all the probabilities must equal one.

Probability Mass Functions Versus Probability Density Functions

Discrete probability functions are referred to as probability mass functions and continuous probability functions are referred to as probability density functions. The term probability functions covers both discrete and continuous distributions. When we are referring to probability functions in generic terms, we may use the term probability density functions to mean both discrete and continuous probability functions.

1.3.6.2. Related Distributions

Probability distributions are typically defined in terms of the probability density function. However, there are a number of probability functions used in applications.

Probability Density Function

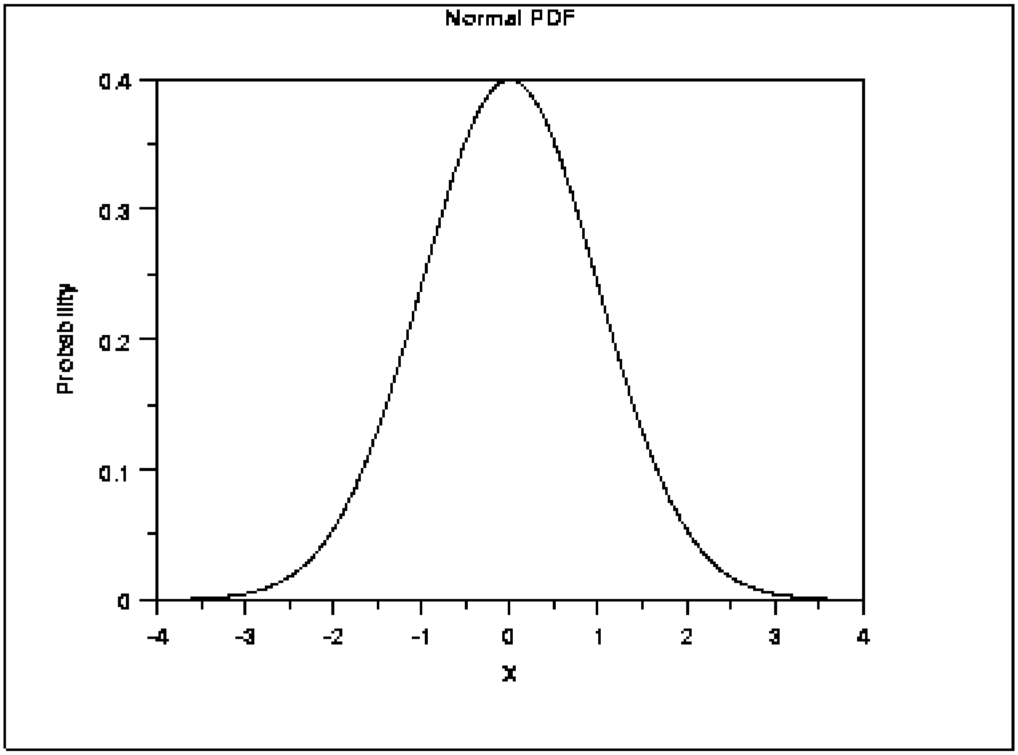
For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

$$\int_a^b f(x)dx = Pr[a \le X \le b]$$

For a discrete distribution, the pdf is the probability that the variate takes the value x.

$$f(x) = Pr[X = x]$$

The following is the plot of the normal probability density function.



Cumulative Distribution Function

The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x. That is

$$F(x) = Pr[X \le x] = \alpha$$

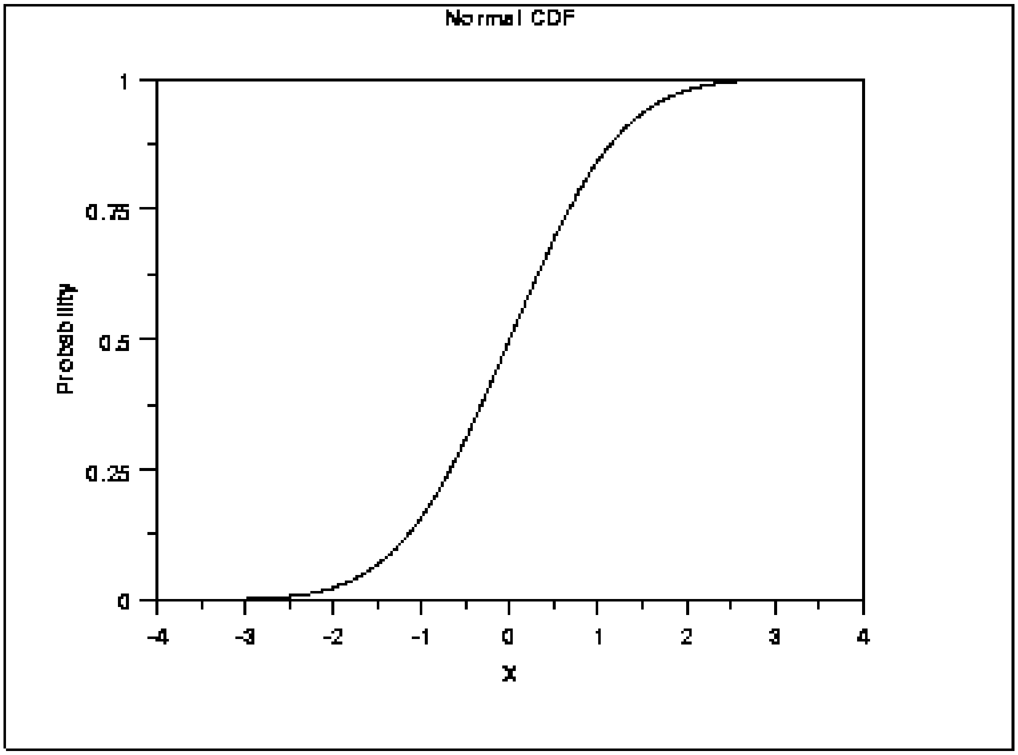
For a continuous distribution, this can be expressed mathematically as

$$F(x) = \int_{-\infty}^{x} f(\mu) d\mu$$

For a discrete distribution, the cdf can be expressed as

$$F(x) = \sum_{i=0}^{x} f(i)$$

The following is the plot of the normal cumulative distribution function.



The horizontal axis is the allowable domain for the given probability function. Since the vertical axis is a probability, it must fall between zero and one. It increases from zero to one as we go from left to right on the horizontal axis.

Percent Point Function

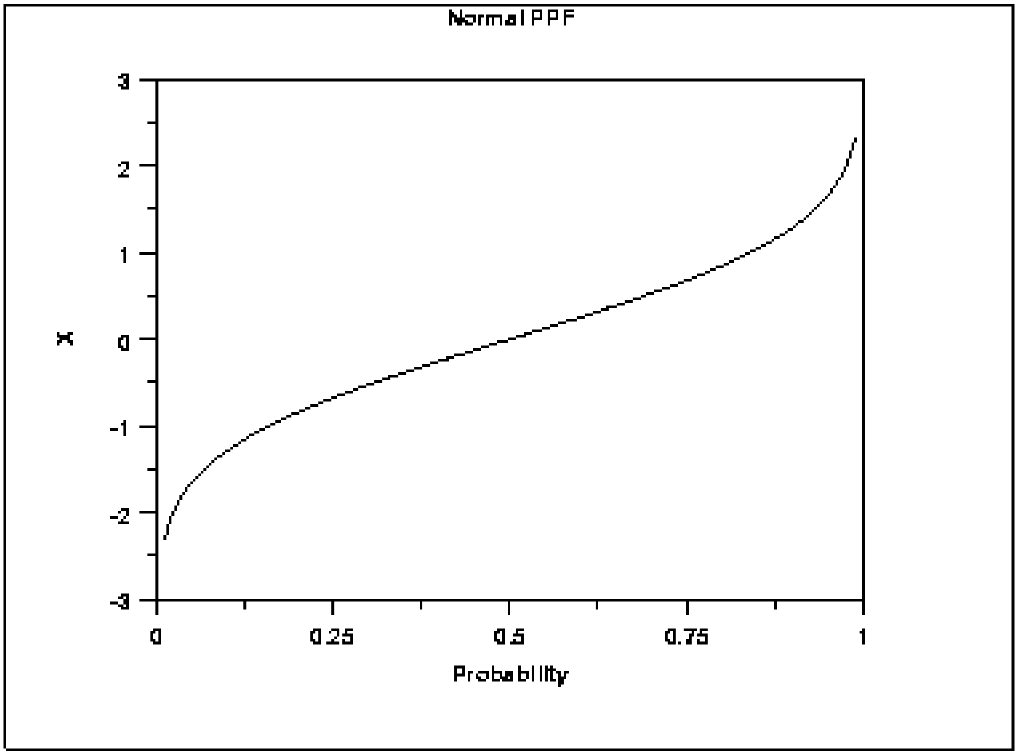
The percent point function (ppf) is the inverse of the cumulative distribution function. For this reason, the percent point function is also commonly referred to as the inverse distribution function. That is, for a distribution function we calculate the probability that the variable is less than or equal to x for a given x. For the percent point function, we start with the probability and compute the corresponding x for the cumulative distribution. Mathematically, this can be expressed as

$$Pr[X \leq G(\alpha)] = \alpha$$

or alternatively

$$x = G(\alpha) = G(F(x))$$

The following is the plot of the normal percent point function.



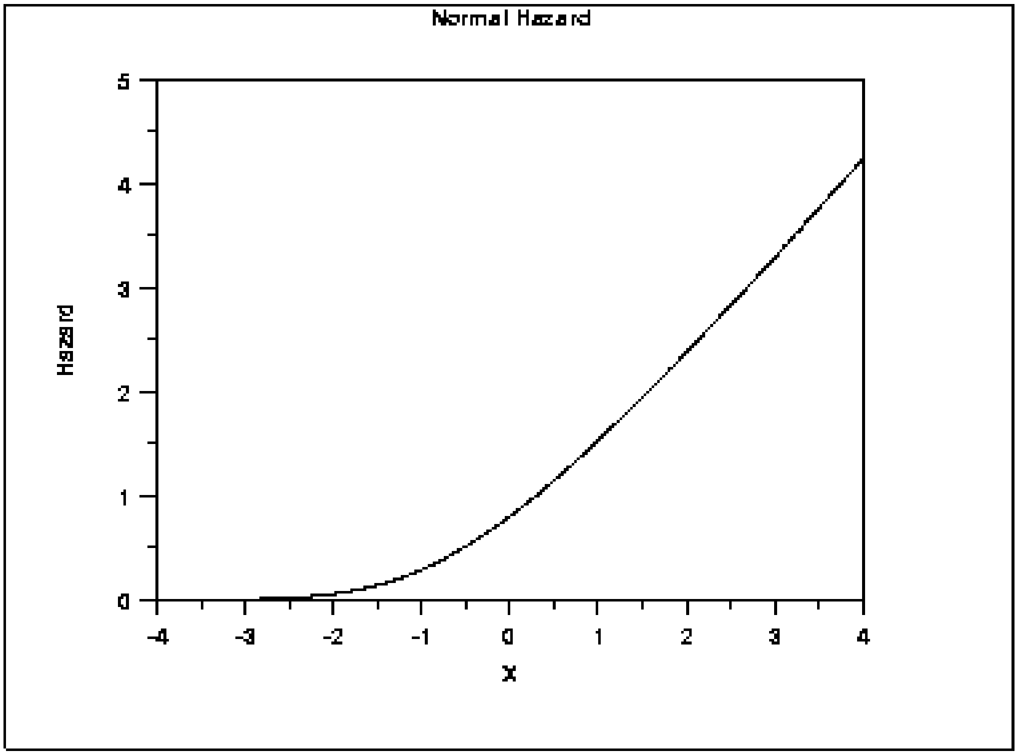
Since the horizontal axis is a probability, it goes from zero to one. The vertical axis goes from the smallest to the largest value of the cumulative distribution function.

Hazard Function

The hazard function is the ratio of the probability density function to the survival function, *S*(x).

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

The following is the plot of the normal distribution hazard function.



Hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the conditional failure density function rather than the hazard function.

Cumulative Hazard Function

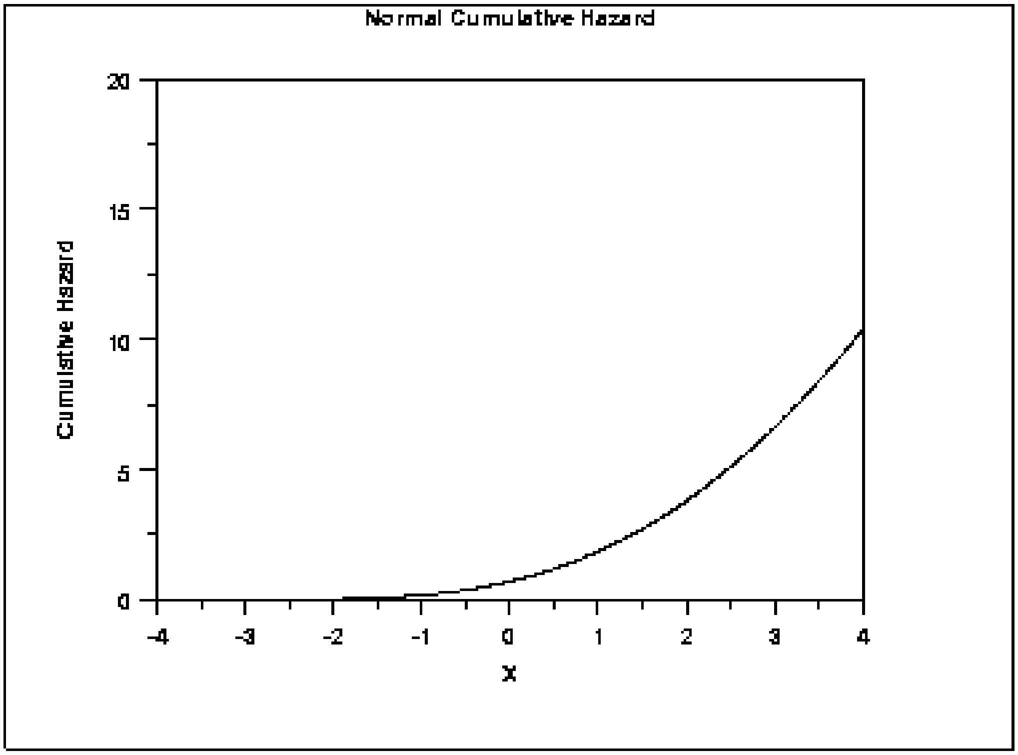
The cumulative hazard function is the integral of the hazard function. It can be interpreted as the probability of failure at time x given survival until time x.

$$H(x) = \int_{-\infty}^{x} h(\mu) d\mu$$

This can alternatively be expressed as

$$H(x) = -\ln(1 - F(x))$$

The following is the plot of the normal cumulative hazard function.



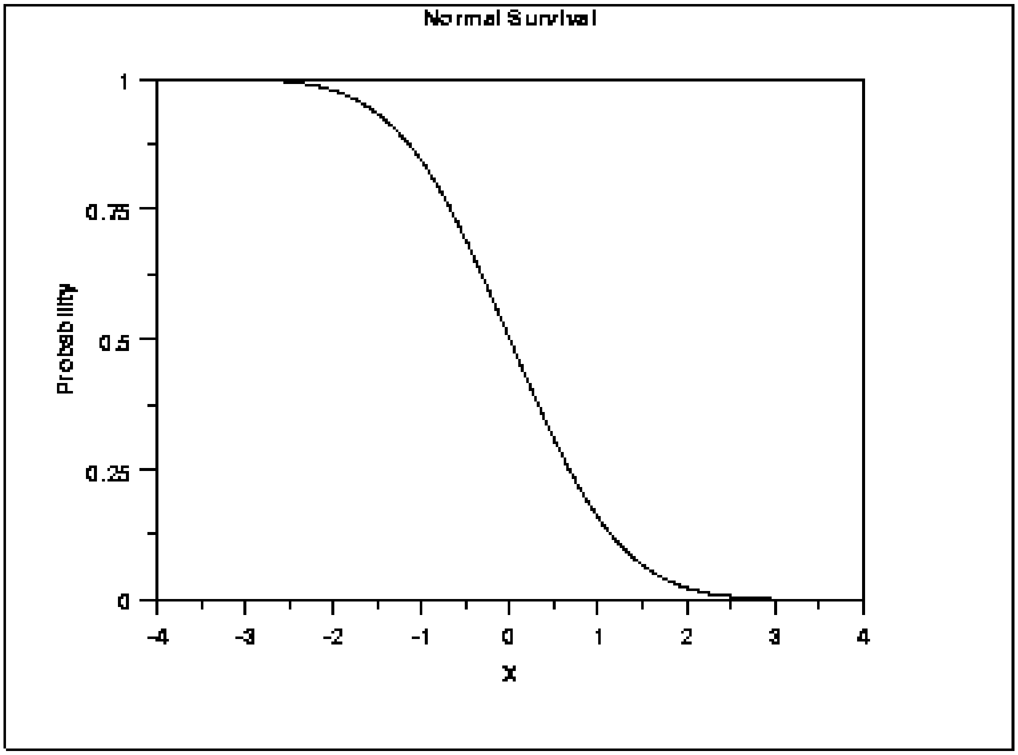
Cumulative hazard plots are most commonly used in reliability applications. Note that Johnson, Kotz, and Balakrishnan refer to this as the hazard function rather than the cumulative hazard function.

Survival Function

Survival functions are most often used in reliability and related fields. The survival function is the probability that the variate takes a value greater than x.

$$S(x) = Pr[X > x] = 1 - F(x)$$

The following is the plot of the normal distribution survival function.



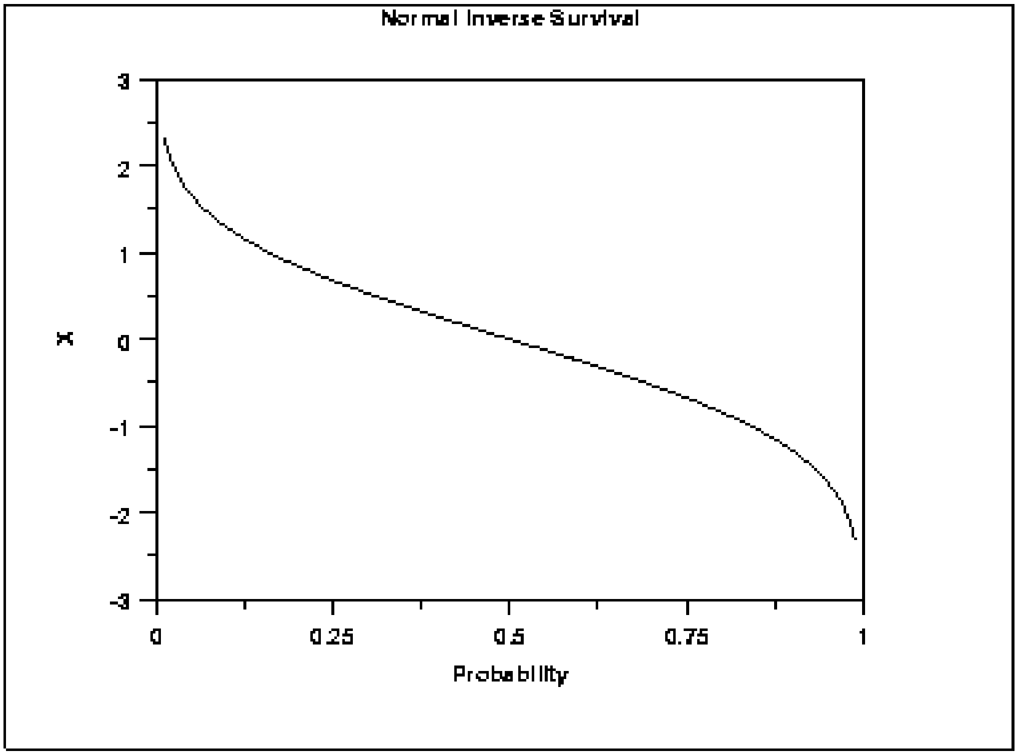
For a survival function, the y value on the graph starts at 1 and monotonically decreases to zero. The survival function should be compared to the cumulative distribution function.

Inverse Survival Function

Just as the percent point function is the inverse of the cumulative distribution function, the survival function also has an inverse function. The inverse survival function can be defined in terms of the percent point function.

$$Z(\alpha) = G(1 - \alpha)$$

The following is the plot of the normal distribution inverse survival function.



As with the percent point function, the horizontal axis is a probability. Therefore the horizontal axis goes from 0 to 1 regardless of the particular distribution. The appearance is similar to the percent point function. However, instead of going from the smallest to the largest value on the vertical axis, it goes from the largest to the smallest value.

1.3.6.3. Families of Distributions

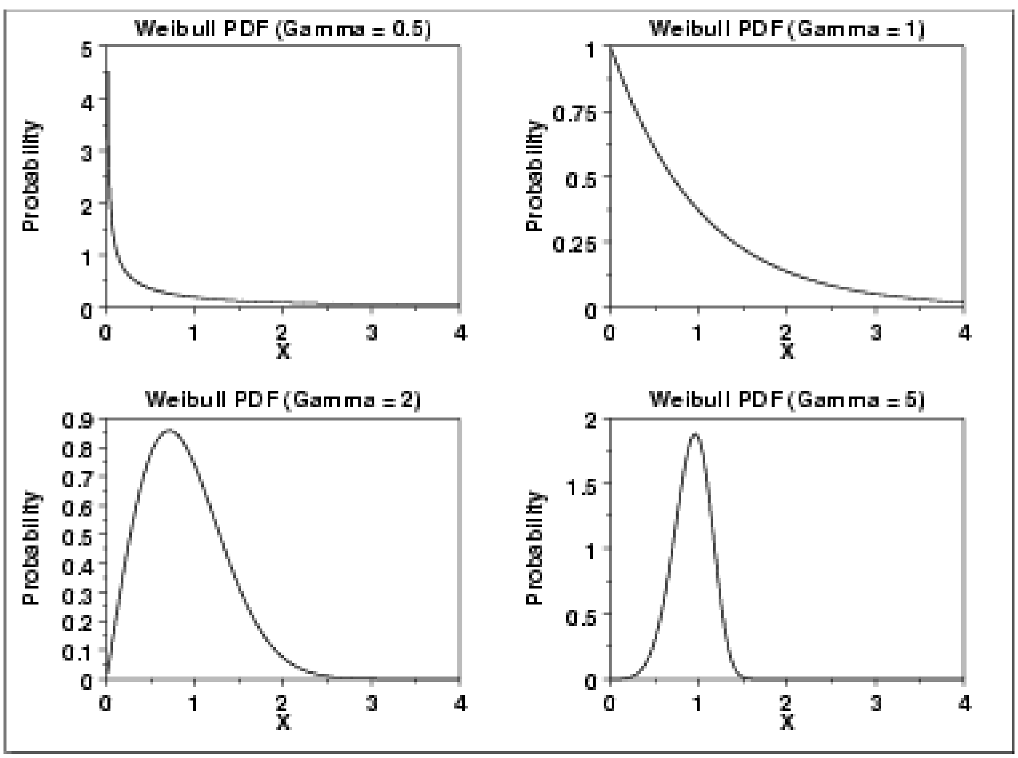
Shape Parameters

Many probability distributions are not a single distribution, but are in fact a family of distributions. This is due to the distribution having one or more shape parameters.

Shape parameters allow a distribution to take on a variety of shapes, depending on the value of the shape parameter. These distributions are particularly useful in modeling applications since they are flexible enough to model a variety of data sets.

Example: Weibull Distribution

The Weibull distribution is an example of a distribution that has a shape parameter. The following graph plots the Weibull pdf with the following values for the shape parameter: 0.5, 1.0, 2.0, and 5.0.



The shapes above include an exponential distribution, a right-skewed distribution, and a relatively symmetric distribution.

The Weibull distribution has a relatively simple distributional form. However, the shape parameter allows the Weibull to assume a wide variety of shapes. This combination of simplicity and flexibility in the shape of the Weibull distribution has made it an effective distributional model in reliability applications. This ability to model a wide variety of distributional shapes using a relatively simple distributional form is possible with many other distributional families as well.

PPCC Plots

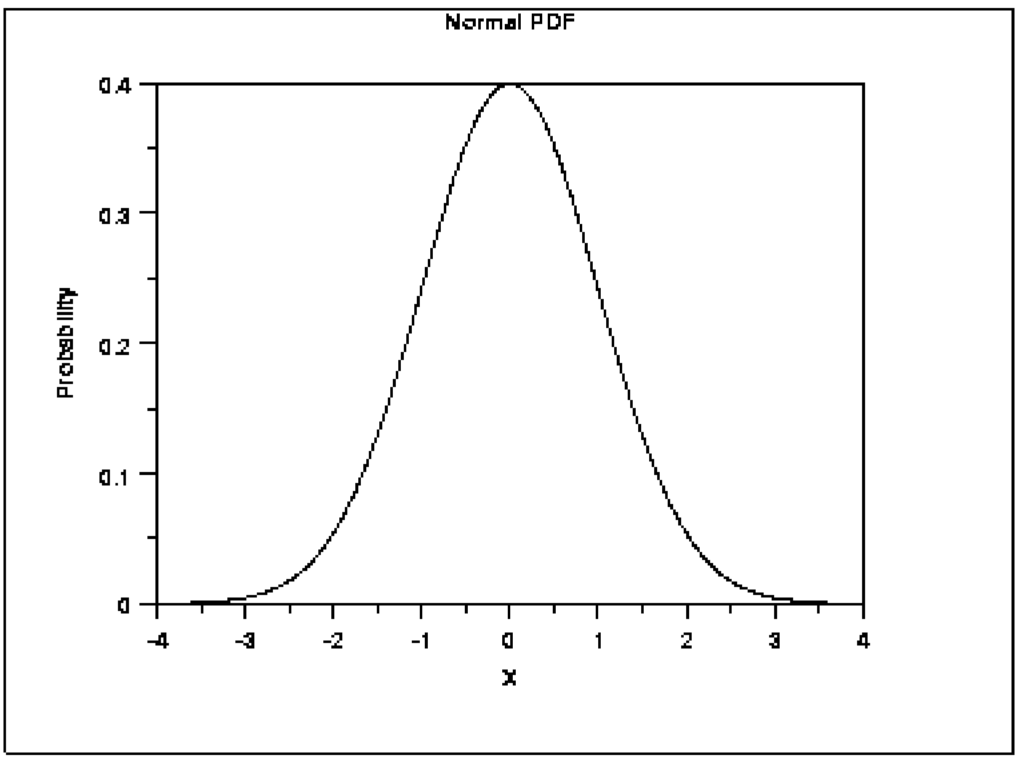
The PPCC plot is an effective graphical tool for selecting the member of a distributional family with a single shape parameter that best fits a given set of data.

1.3.6.4. Location and Scale Parameters

Normal PDF

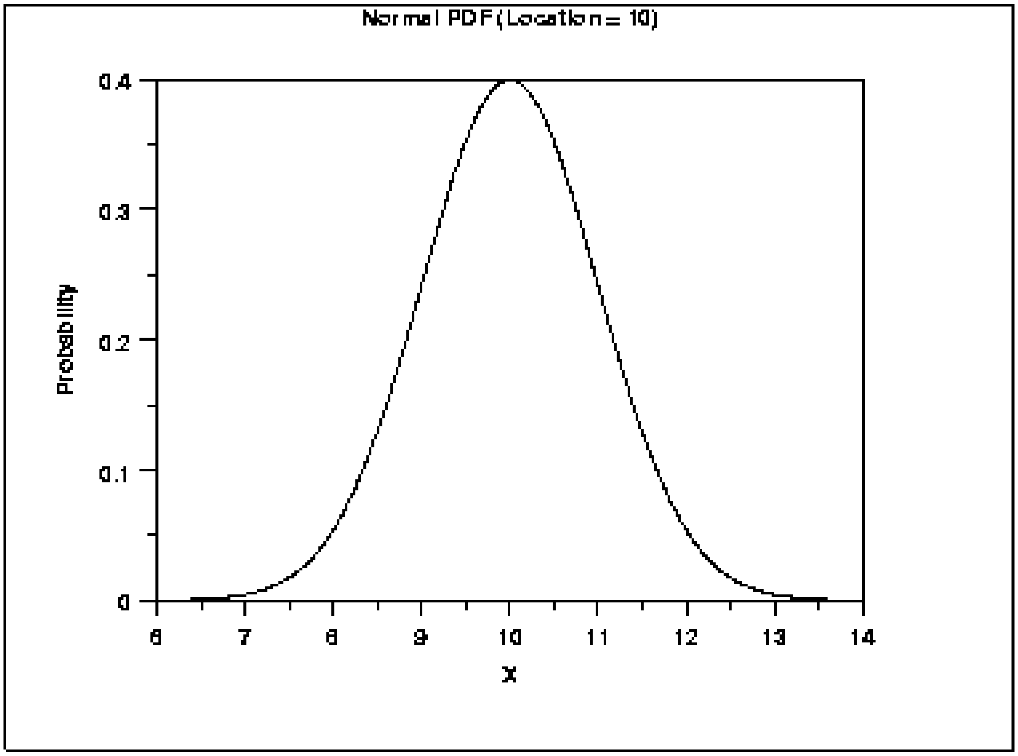
A probability distribution is characterized by location and scale parameters. Location and scale parameters are typically used in modeling applications.

For example, the following graph is the probability density function for the standard normal distribution, which has the location parameter equal to zero and scale parameter equal to one.



Location Parameter

The next plot shows the probability density function for a normal distribution with a location parameter of 10 and a scale parameter of 1.

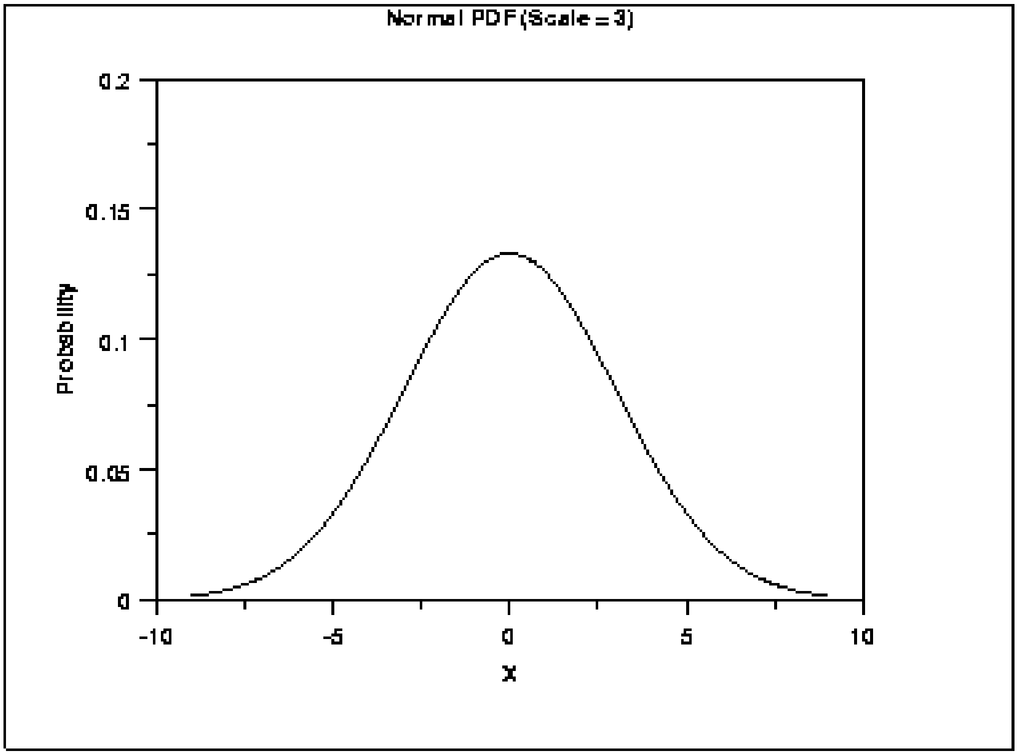


The effect of the location parameter is to translate the graph, relative to the standard normal distribution, 10 units to the right on the horizontal axis. A location parameter of -10 would have shifted the graph 10 units to the left on the horizontal axis.

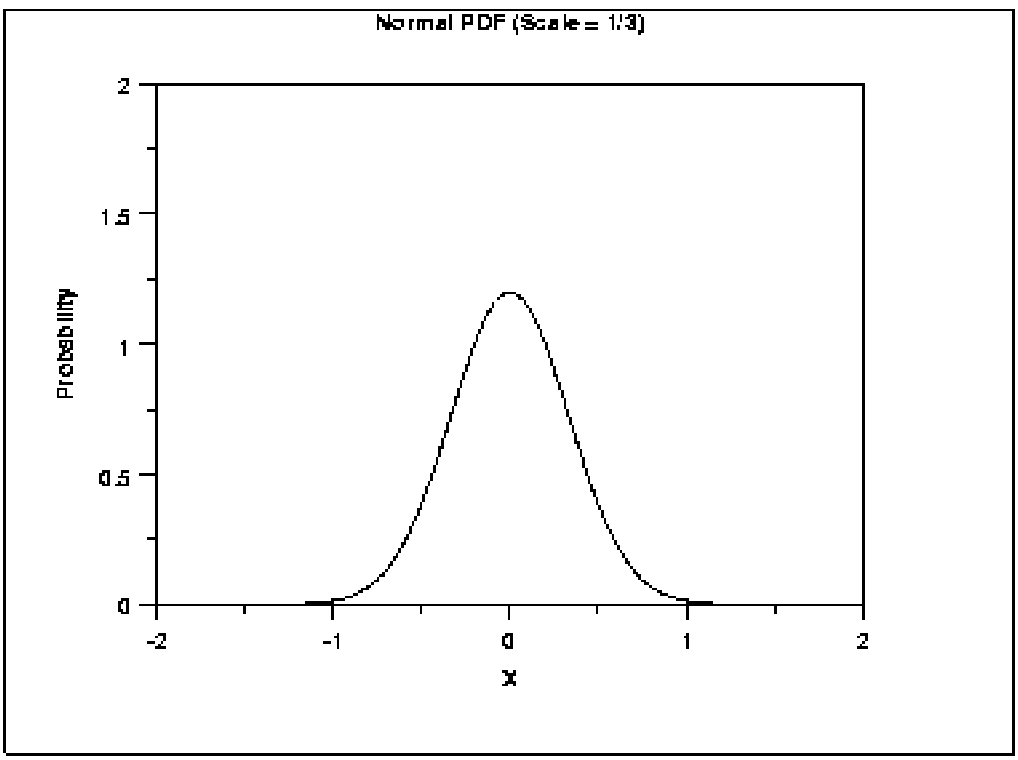
That is, a location parameter simply shifts the graph left or right on the horizontal axis.

Scale Parameter

The next plot has a scale parameter of 3 (and a location parameter of zero). The effect of the scale parameter is to stretch out the graph. The maximum y value is approximately 0.13 as opposed 0.4 in the previous graphs. The y value, i.e., the vertical axis value, approaches zero at about (+/-) 9 as opposed to (+/-) 3 with the first graph.



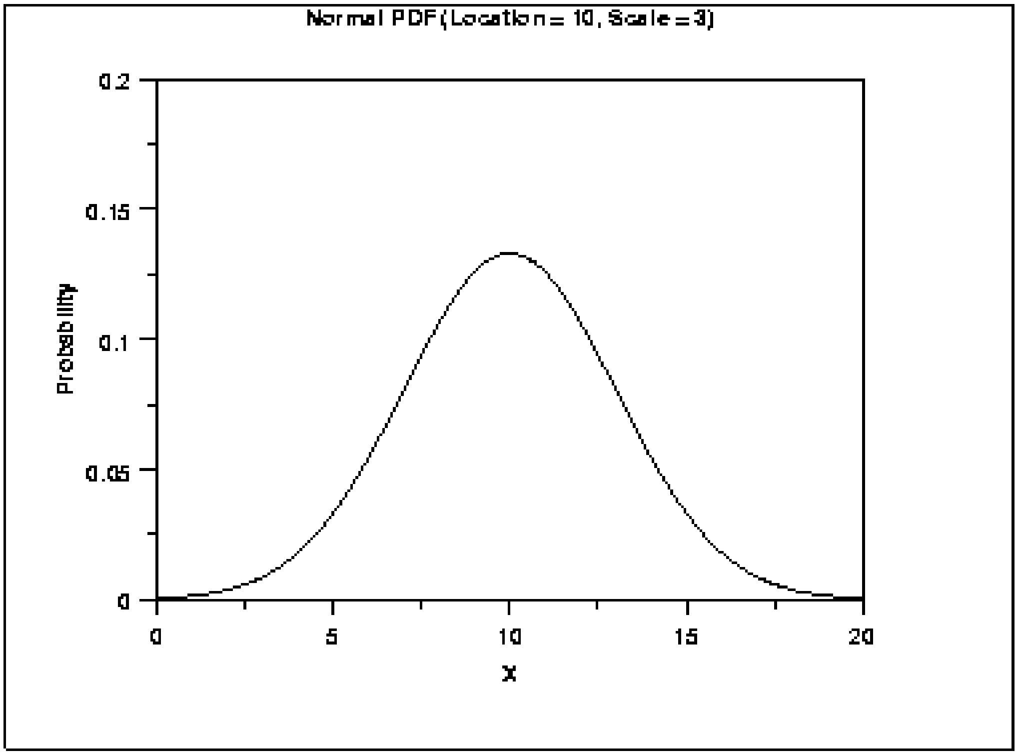
In contrast, the next graph has a scale parameter of 1/3 (=0.333). The effect of this scale parameter is to squeeze the pdf. That is, the maximum y value is approximately 1.2 as opposed to 0.4 and the y value is near zero at (+/-) 1 as opposed to (+/-) 3.



The effect of a scale parameter greater than one is to stretch the pdf. The greater the magnitude, the greater the stretching. The effect of a scale parameter less than one is to compress the pdf. The compressing approaches a spike as the scale parameter goes to zero. A scale parameter of 1 leaves the pdf unchanged (if the scale parameter is 1 to begin with) and non-positive scale parameters are not allowed.

Location and Scale Together

The following graph shows the effect of both a location and a scale parameter. The plot has been shifted right 10 units and stretched by a factor of 3.



Standard Form

The standard form of any distribution is the form that has location parameter zero and scale parameter one.

It is common in statistical software packages to only compute the standard form of the distribution. There are formulas for converting from the standard form to the form with other location and scale parameters. These formulas are independent of the particular probability distribution.

Formulas for Location and Scale Based on the Standard Form

The following are the formulas for computing various probability functions based on the standard form of the distribution. The parameter *a* refers to the location parameter and the parameter *b* refers to the scale parameter. Shape parameters are not included.

Cumulative Distribution Function	$F(x;a,b) = F((x-a)/b;0,1)$
Probability Density Function	$f(x;a,b) = (1/b)f((x-a)/b;0,1)$
Percent Point Function	$G(\alpha;a,b) = a + bG(\alpha;0,1)$
Hazard Function	$h(x;a,b) = (1/b)h((x-a)/b;0,1)$
Cumulative Hazard Function	$H(x;a,b) = H((x-a)/b;0,1)$
Survival Function	$S(x;a,b) = S((x-a)/b;0,1)$
Inverse Survival Function	$Z(\alpha;a,b) = a + bZ(\alpha;0,1)$
Random Numbers	$Y(a,b) = a + bY(0,1)$

Relationship to Mean and Standard Deviation

For the normal distribution, the location and scale parameters correspond to the mean and standard deviation, respectively. However, this is not necessarily true for other distributions. In fact, it is not true for most distributions.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6.5. Estimating the Parameters of a Distribution

Model a univariate data set with a probability distribution

One common application of probability distributions is modeling univariate data with a specific probability distribution. This involves the following two steps:

1. Determination of the "best-fitting" distribution.
2. Estimation of the parameters (shape, location, and scale parameters) for that distribution.

Various Methods

There are various methods, both numerical and graphical, for estimating the parameters of a probability distribution.

1. Method of moments
2. Maximum likelihood
3. Least squares
4. PPCC and probability plots

1.3.6.5.1. Method of Moments

Method of Moments

The method of moments equates sample moments to parameter estimates. When moment methods are available, they have the advantage of simplicity. The disadvantage is that they are often not available and they do not have the desirable optimality properties of maximum likelihood and least squares estimators.

The primary use of moment estimates is as starting values for the more precise maximum likelihood and least squares estimates.

Software

Most general purpose statistical software does not include explicit method of moments parameter estimation commands. However, when utilized, the method of moment formulas tend to be straightforward and can be easily implemented in most statistical software programs.

1.3.6.5.2. Maximum Likelihood

Maximum Likelihood

Maximum likelihood estimation begins with the mathematical expression known as a likelihood function of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data given the chosen probability model. This expression contains the unknown parameters. Those values of the parameter that maximize the sample likelihood are known as the maximum likelihood estimates.

The reliability chapter contains some examples of the likelihood functions for a few of the commonly used distributions in reliability analysis.

Advantages

The advantages of this method are:

- Maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations. For example, they can be applied in reliability analysis to censored data under various censoring models.
- Maximum likelihood methods have desirable mathematical and optimality properties. Specifically,
  1. They become minimum variance unbiased estimators as the sample size increases. By unbiased, we mean that if we take (a very large number of) random samples with replacement from a population, the average value of the parameter estimates will be theoretically exactly equal to the population value. By minimum variance, we mean that the estimator has the smallest variance, and thus the narrowest confidence interval, of all estimators of that type.
  2. They have approximate normal distributions and approximate sample variances that can be used to

generate confidence bounds and hypothesis tests for the parameters.

- Several popular statistical software packages provide excellent algorithms for maximum likelihood estimates for many of the commonly used distributions. This helps mitigate the computational complexity of maximum likelihood estimation.

Disadvantages

The disadvantages of this method are:

- The likelihood equations need to be specifically worked out for a given distribution and estimation problem. The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired.
- The numerical estimation is usually non-trivial. Except for a few cases where the maximum likelihood formulas are in fact simple, it is generally best to rely on high quality statistical software to obtain maximum likelihood estimates. Fortunately, high quality maximum likelihood software is becoming increasingly common.
- Maximum likelihood estimates can be heavily biased for small samples. The optimality properties may not apply for small samples.
- Maximum likelihood can be sensitive to the choice of starting values.

Software

Most general purpose statistical software programs support maximum likelihood estimation (MLE) in some form. MLE estimation can be supported in two ways.

1. A software program may provide a generic function minimization (or equivalently, maximization) capability. This is also referred to as function optimization. Maximum likelihood estimation is essentially a function optimization problem.

   This type of capability is particularly common in mathematical software programs.
2. A software program may provide MLE computations for a specific problem. For example, it may generate ML estimates for the parameters of a Weibull distribution.

   Statistical software programs will often provide ML estimates for many specific problems even when they do not support general function optimization.

The advantage of function minimization software is that it can be applied to many different MLE problems. The drawback is that you have to specify the maximum likelihood equations to the software. As the functions can be non-trivial, there is potential for error in entering the equations.

The advantage of the specific MLE procedures is that greater efficiency and better numerical stability can often be obtained by taking advantage of the properties of the specific estimation problem. The specific methods often return explicit confidence intervals. In addition, you do not have to know or specify the likelihood equations to the software. The disadvantage is that each MLE problem must be specifically coded.

Dataplot supports MLE for a limited number of distributions.

1.3.6.5.3. Least Squares

Least Squares

Non-linear least squares provides an alternative to maximum likelihood.

Advantages

The advantages of this method are:

- Non-linear least squares software may be available in many statistical software packages that do not support maximum likelihood estimates.
- It can be applied more generally than maximum likelihood. That is, if your software provides non-linear fitting and it has the ability to specify the probability function you are interested in, then you can generate least squares estimates for that distribution. This will allow you to obtain reasonable estimates for distributions even if the software does not provide maximum likelihood estimates.

Disadvantages

The disadvantages of this method are:

- It is not readily applicable to censored data.
- It is generally considered to have less desirable optimality properties than maximum likelihood.
- It can be quite sensitive to the choice of starting values.

Software

Non-linear least squares fitting is available in many general purpose statistical software programs. The macro developed for Dataplot can be adapted to many software programs that provide least squares estimation.

1.3.6.5.4. PPCC and Probability Plots

PPCC and Probability Plots

The PPCC plot can be used to estimate the shape parameter of a distribution with a single shape parameter. After finding the best value of the shape parameter, the probability plot can be used to estimate the location and scale parameters of a probability distribution.

Advantages

The advantages of this method are:

- It is based on two well-understood concepts.
    1. The linearity (i.e., straightness) of the probability plot is a good measure of the adequacy of the distributional fit.
    2. The correlation coefficient between the points on the probability plot is a good measure of the linearity of the probability plot.
- It is an easy technique to implement for a wide variety of distributions with a single shape parameter. The basic requirement is to be able to compute the percent point function, which is needed in the computation of both the probability plot and the PPCC plot.
- The PPCC plot provides insight into the sensitivity of the shape parameter. That is, if the PPCC plot is relatively flat in the neighborhood of the optimal value of the shape parameter, this is a strong indication that the fitted model will not be sensitive to small deviations, or even large deviations in some cases, in the value of the shape parameter.
- The maximum correlation value provides a method for comparing across distributions as well as identifying the best value of the shape parameter for a given distribution. For example, we could use the PPCC and probability fits for the Weibull, lognormal, and possibly several other distributions. Comparing the maximum correlation coefficient achieved for each distribution can help in selecting which is the best distribution to use.

Disadvantages

The disadvantages of this method are:

- It is limited to distributions with a single shape parameter.
- PPCC plots are not widely available in statistical software packages other than Dataplot (Dataplot provides PPCC plots for 40+ distributions). Probability plots are generally available. However, many statistical software packages only provide them for a limited number of distributions.
- Significance levels for the correlation coefficient (i.e., if the maximum correlation value is above a given value, then the distribution provides an adequate fit for the data with a given confidence level) have only been worked out for a limited number of distributions.

Case Study

The airplane glass failure time case study demonstrates the use of the PPCC and probability plots in finding the best distributional model and the parameter estimation of the distributional model.

Other Graphical Methods

For reliability applications, the hazard plot and the Weibull plot are alternative graphical methods that are commonly used to estimate parameters.

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions

1.3.6.6. Gallery of Distributions

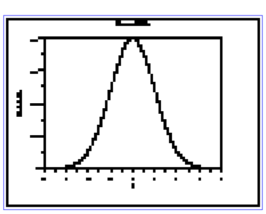
Gallery of Common Distributions

Detailed information on a few of the most common distributions is available below. There are a large number of distributions used in statistical applications. It is beyond the scope of this Handbook to discuss more than a few of these. Two excellent sources for additional detailed information on a large array of distributions are Johnson, Kotz, and Balakrishnan and Evans, Hastings, and Peacock. Equations for the probability functions are given for the standard form of the distribution. Formulas exist for defining the functions with location and scale parameters in terms of the standard form of the distribution.

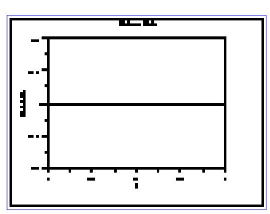
The sections on parameter estimation are restricted to the method of moments and maximum likelihood. This is because the least squares and PPCC and probability plot estimation procedures are generic. The maximum likelihood equations are not listed if they involve solving simultaneous equations. This is because these methods require sophisticated computer software to solve. Except where the maximum likelihood estimates are trivial, you should depend on a statistical software program to compute them. References are given for those who are interested.

Be aware that different sources may give formulas that are different from those shown here. In some cases, these are simply mathematically equivalent formulations. In other cases, a different parameterization may be used.

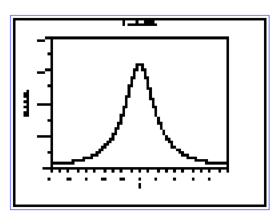
Continuous Distributions



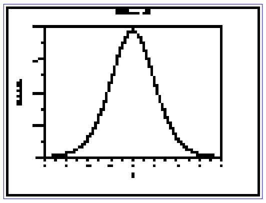
Normal Distribution

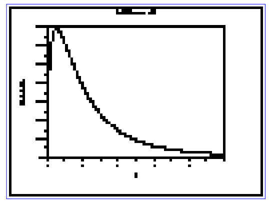


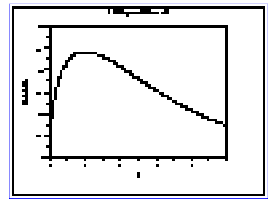
Uniform Distribution

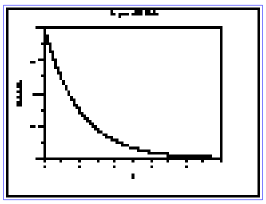


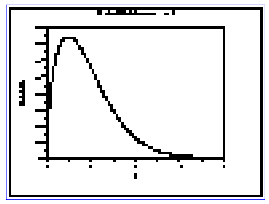
Cauchy Distribution

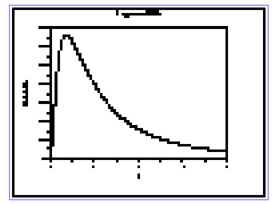

t Distribution

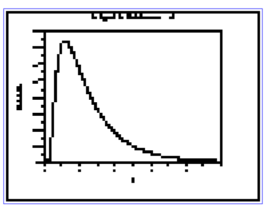

F Distribution

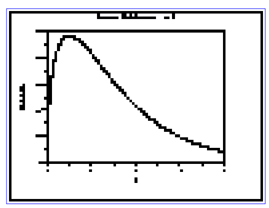

Chi-Square Distribution

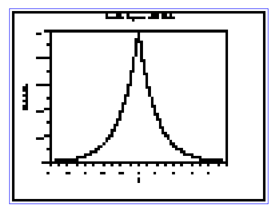

Exponential Distribution

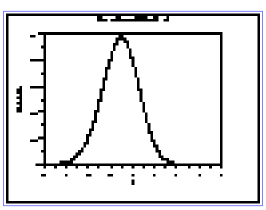

Weibull Distribution

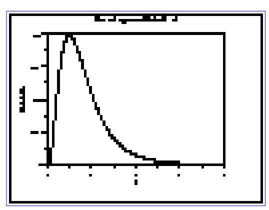

Lognormal Distribution

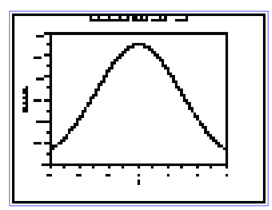

Fatigue Life Distribution

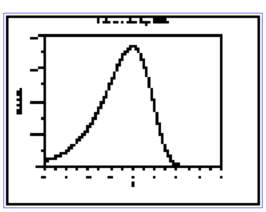

Gamma Distribution

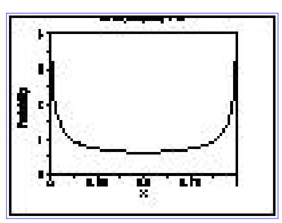

Double Exponential Distribution


Power Normal Distribution

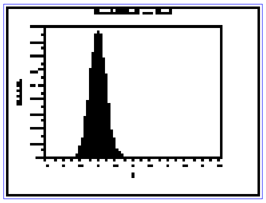

Power Lognormal Distribution


Tukey-Lambda Distribution

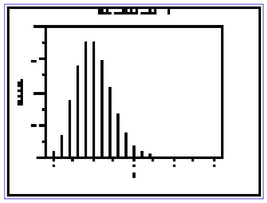

Extreme Value Type I Distribution


Beta Distribution

Discrete Distributions



[Binomial Distribution](#)



[Poisson Distribution](#)

1.3.6.6.1. Normal Distribution

Probability Density Function

The general formula for the probability density function of the normal distribution is

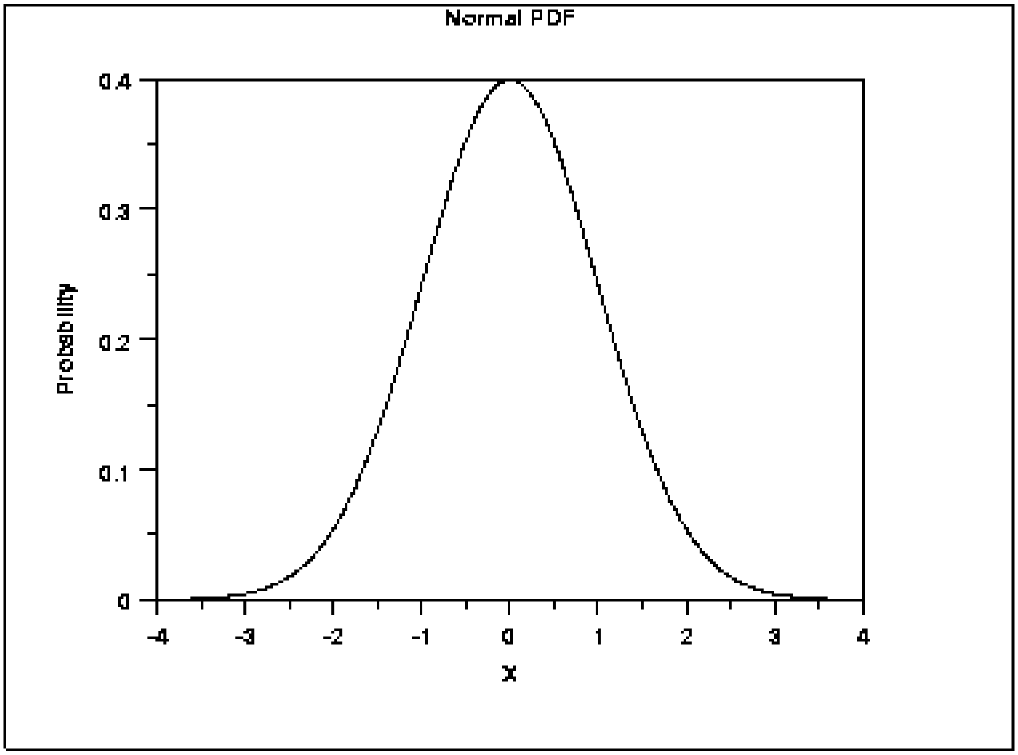
$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where μ is the location parameter and σ is the scale parameter. The case where μ = 0 and σ = 1 is called the **standard normal distribution**. The equation for the standard normal distribution is

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

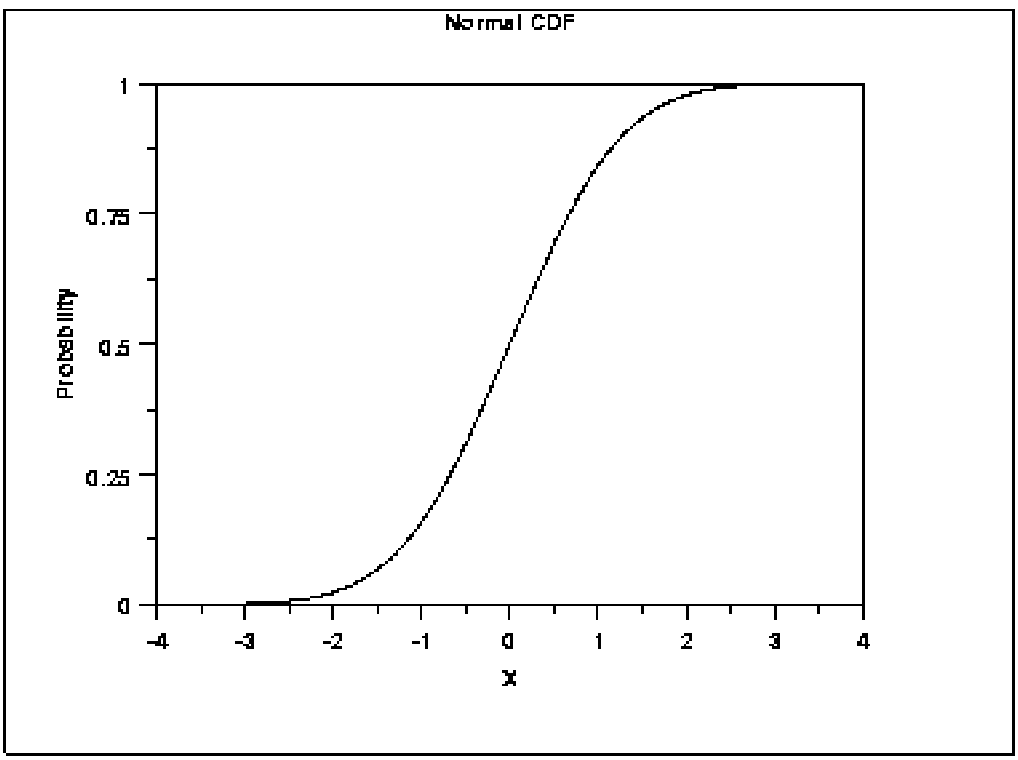
The following is the plot of the standard normal probability density function.



Cumulative Distribution Function

The formula for the cumulative distribution function of the normal distribution does not exist in a simple closed formula. It is computed numerically.

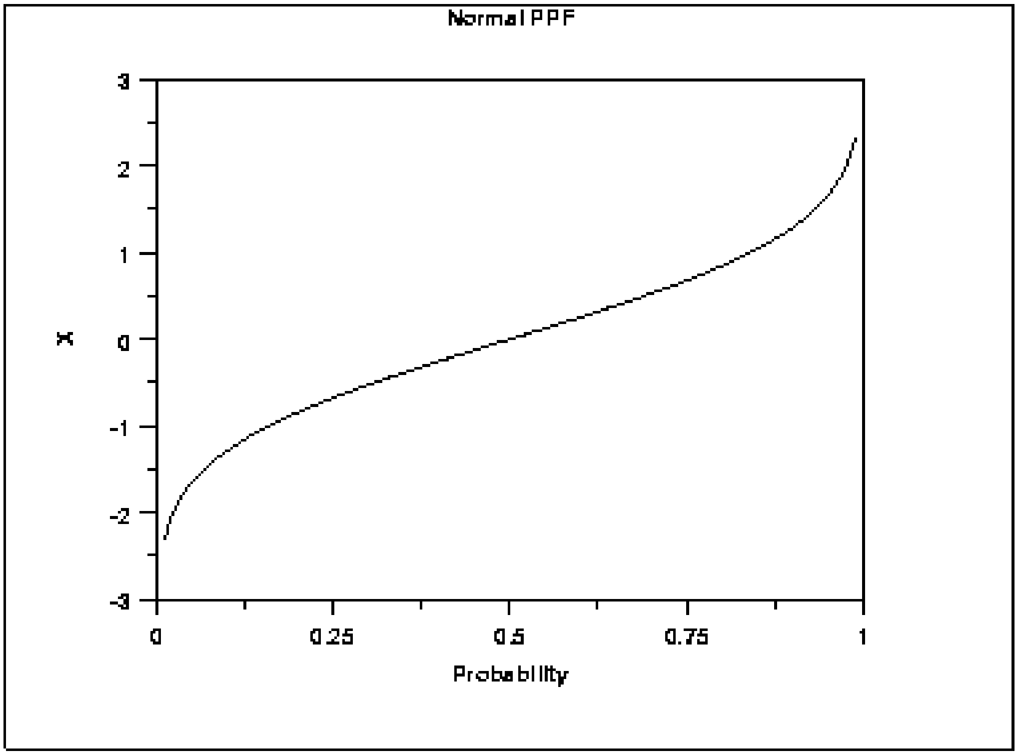
The following is the plot of the normal cumulative distribution function.



Percent Point Function

The formula for the percent point function of the normal distribution does not exist in a simple closed formula. It is computed numerically.

The following is the plot of the normal percent point function.



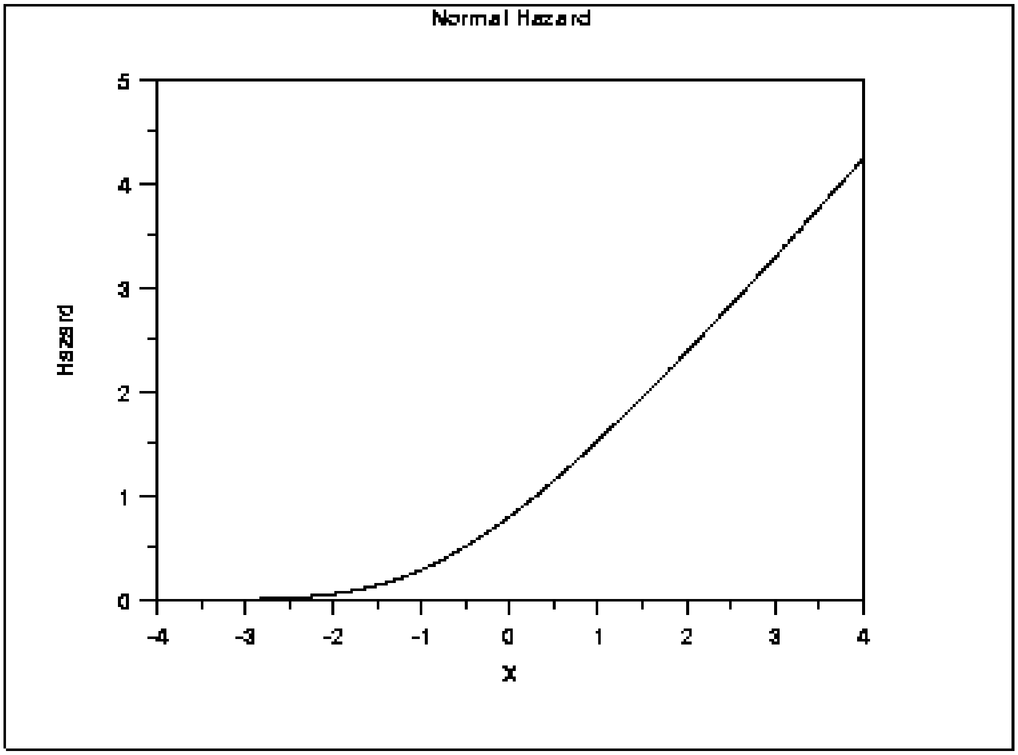
Hazard Function

The formula for the hazard function of the normal distribution is

$$h(x) = \frac{\phi(x)}{\Phi(-x)}$$

where Φ is the cumulative distribution function of the standard normal distribution and ϕ is the probability density function of the standard normal distribution.

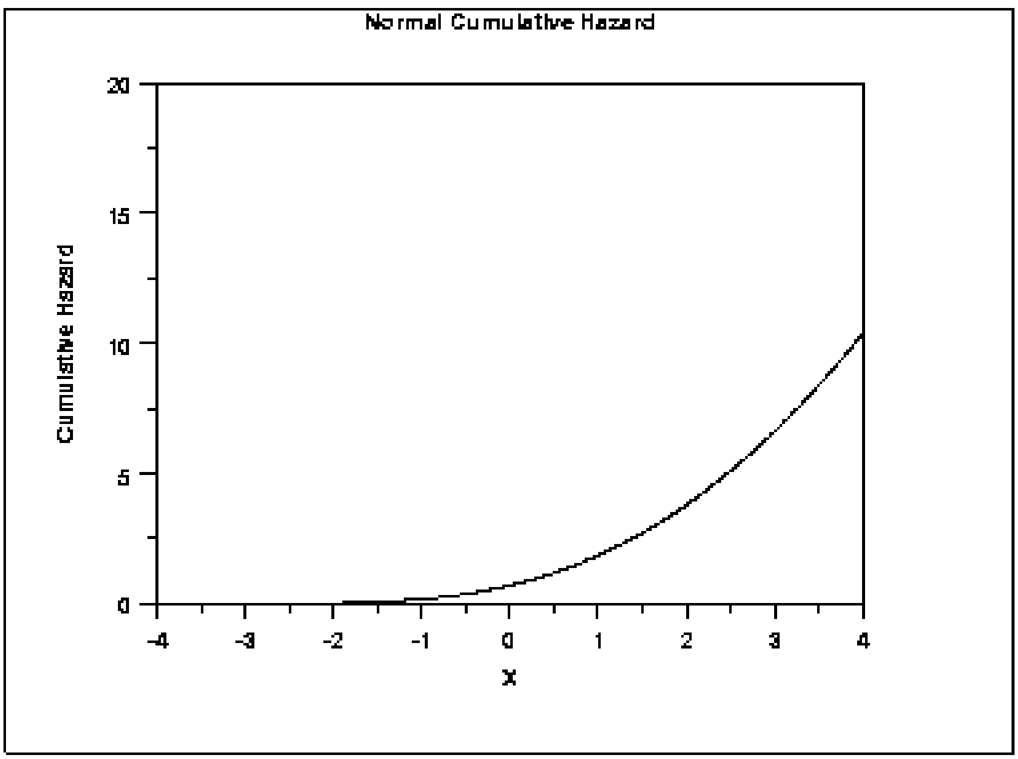
The following is the plot of the normal hazard function.



Cumulative Hazard Function

The normal cumulative hazard function can be computed from the normal cumulative distribution function.

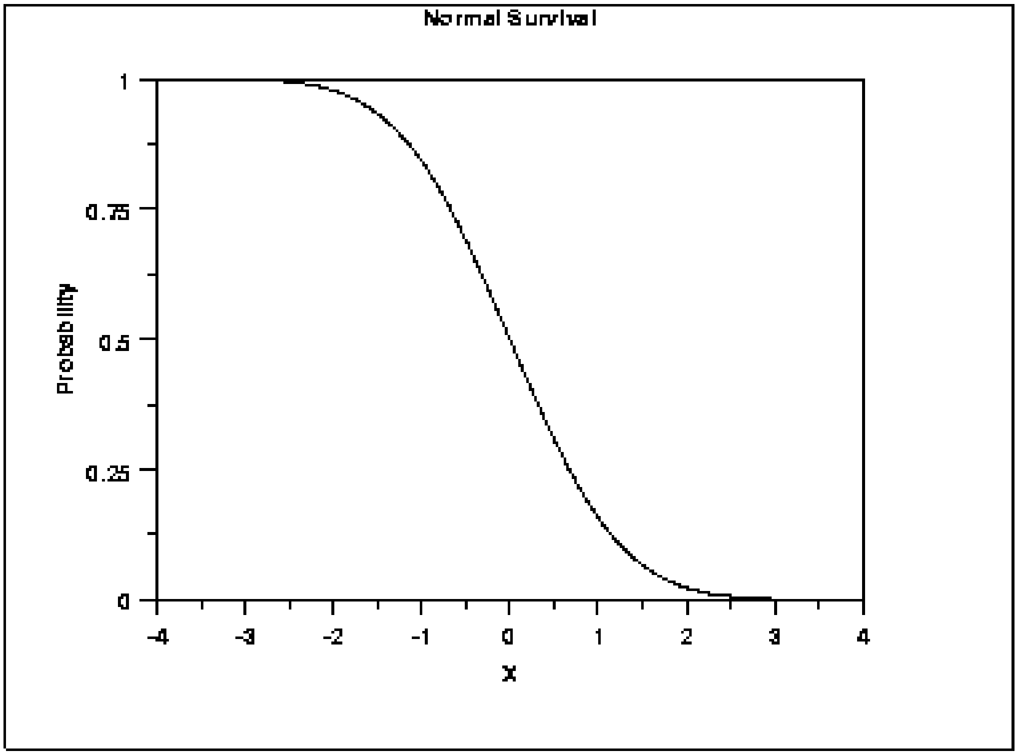
The following is the plot of the normal cumulative hazard function.



Survival Function

The normal survival function can be computed from the normal cumulative distribution function.

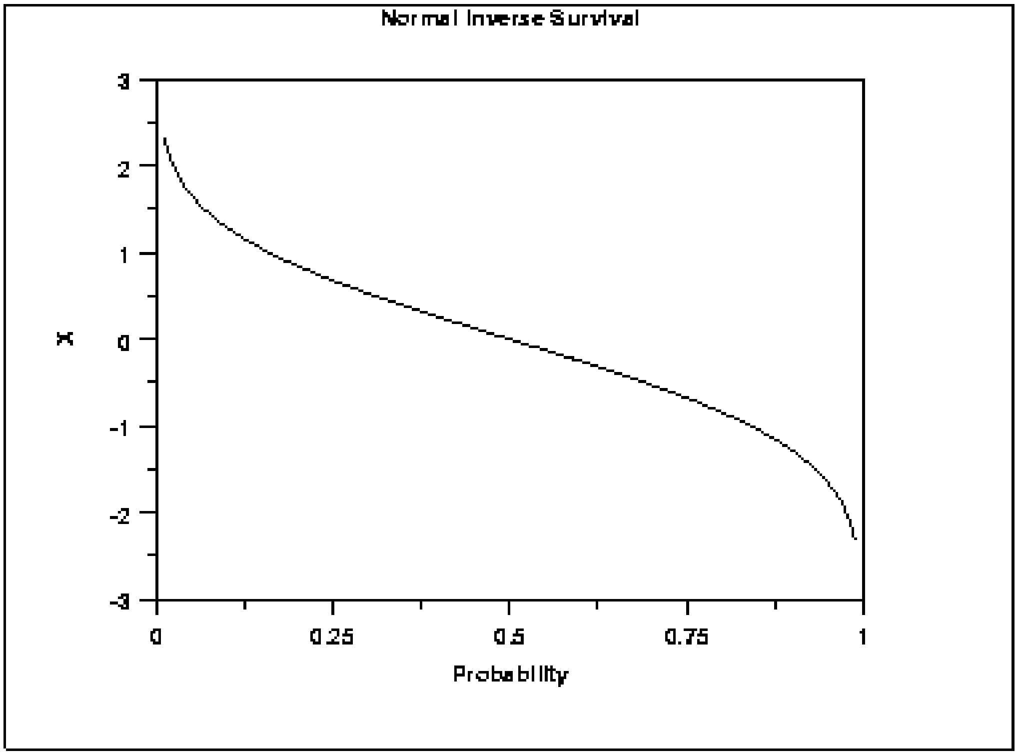
The following is the plot of the normal survival function.



Inverse Survival Function

The normal inverse survival function can be computed from the normal percent point function.

The following is the plot of the normal inverse survival function.



Common Statistics	Mean	The location parameter μ.
	Median	The location parameter μ.
	Mode	The location parameter μ.
	Range	Infinity in both directions.
	Standard Deviation	The scale parameter σ.
	Coefficient of Variation	σ/μ
	Skewness	0
	Kurtosis	3

Parameter Estimation

The location and scale parameters of the normal distribution can be estimated with the sample mean and sample standard deviation, respectively.

Comments

For both theoretical and practical reasons, the normal distribution is probably the most important distribution in statistics. For example,

- Many classical statistical tests are based on the assumption that the data follow a normal distribution. This assumption should be tested before applying these tests.
- In modeling applications, such as linear and non-linear regression, the error term is often assumed to follow a normal distribution with fixed location and scale.
- The normal distribution is used to find significance levels in many hypothesis tests and confidence intervals.

Theroretical Justification - Central Limit Theorem

The normal distribution is widely used. Part of the appeal is that it is well behaved and mathematically tractable. However, the central limit theorem provides a theoretical basis for why it has wide applicability.

The central limit theorem basically states that as the sample size (*N*) becomes large, the following occur:

1. The sampling distribution of the mean becomes approximately normal regardless of the distribution of the original variable.
2. The sampling distribution of the mean is centered at the population mean, μ, of the original variable. In addition, the standard deviation of the sampling distribution of the mean approaches $\sigma/\sqrt{N}$.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the normal distribution.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6.6.2. Uniform Distribution

Probability Density Function

The general formula for the probability density function of the uniform distribution is

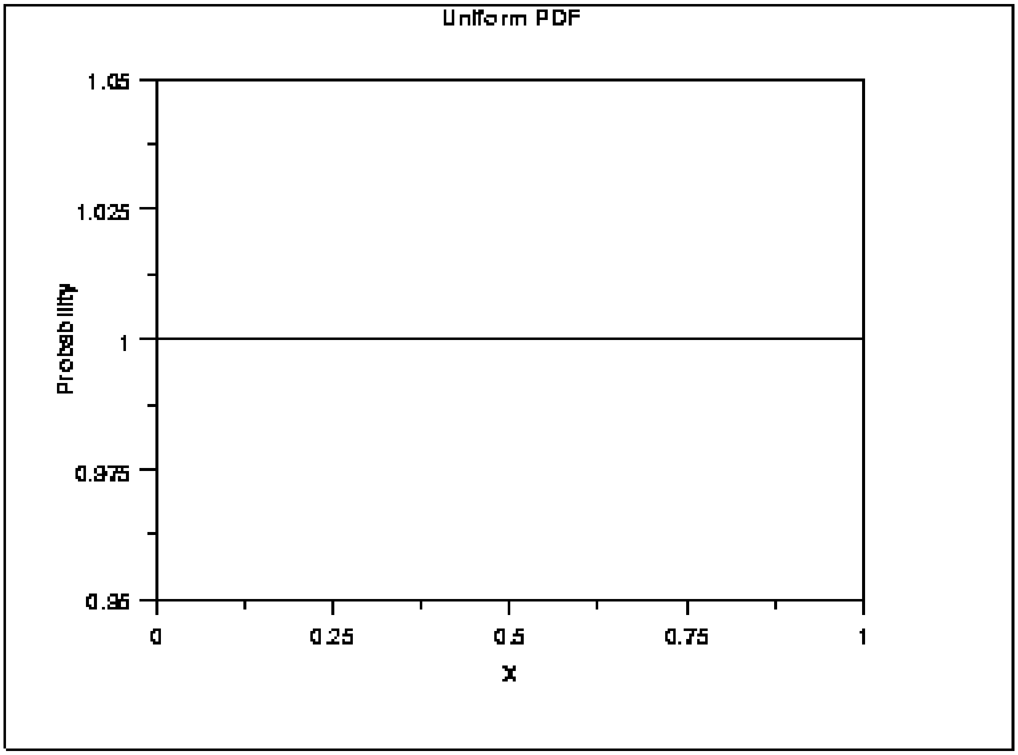
$$f(x) = \frac{1}{B - A} \qquad \mathrm{for}\ A \le x \le B$$

where A is the location parameter and (B - A) is the scale parameter. The case where A = 0 and B = 1 is called the **standard uniform distribution**. The equation for the standard uniform distribution is

$$f(x) = 1 \qquad \mathrm{for}\ 0 \le x \le 1$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the uniform probability density function.

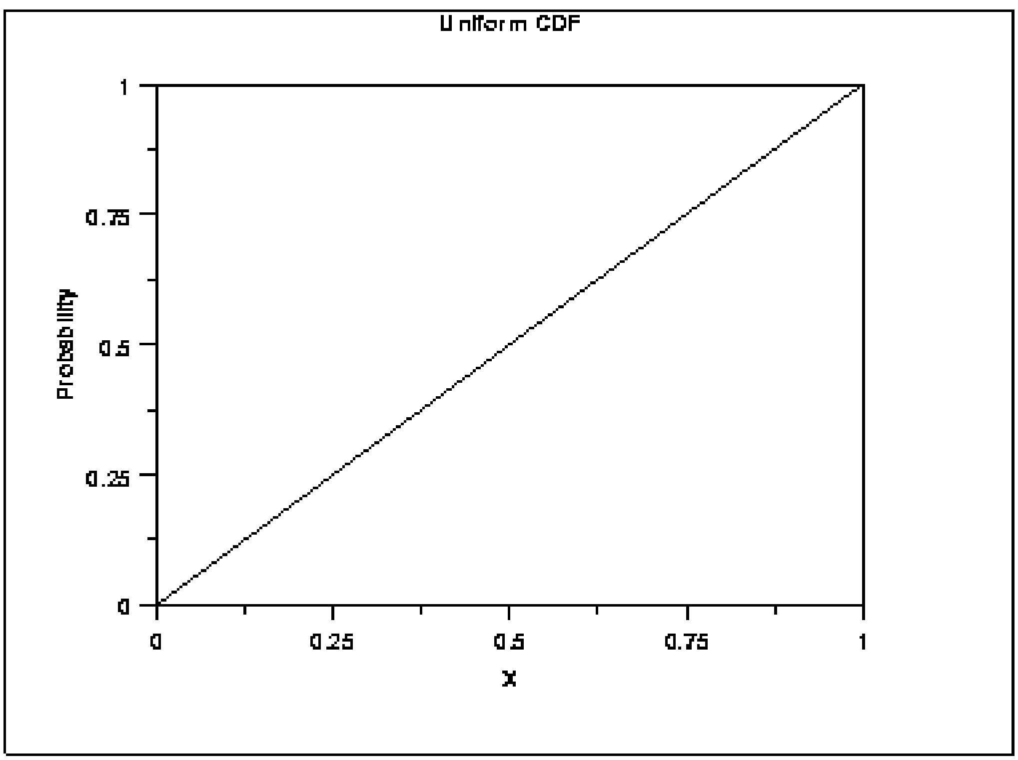


Cumulative Distribution Function

The formula for the cumulative distribution function of the uniform distribution is

$$F(x) = x \qquad \text{for } 0 \le x \le 1$$

The following is the plot of the uniform cumulative distribution function.

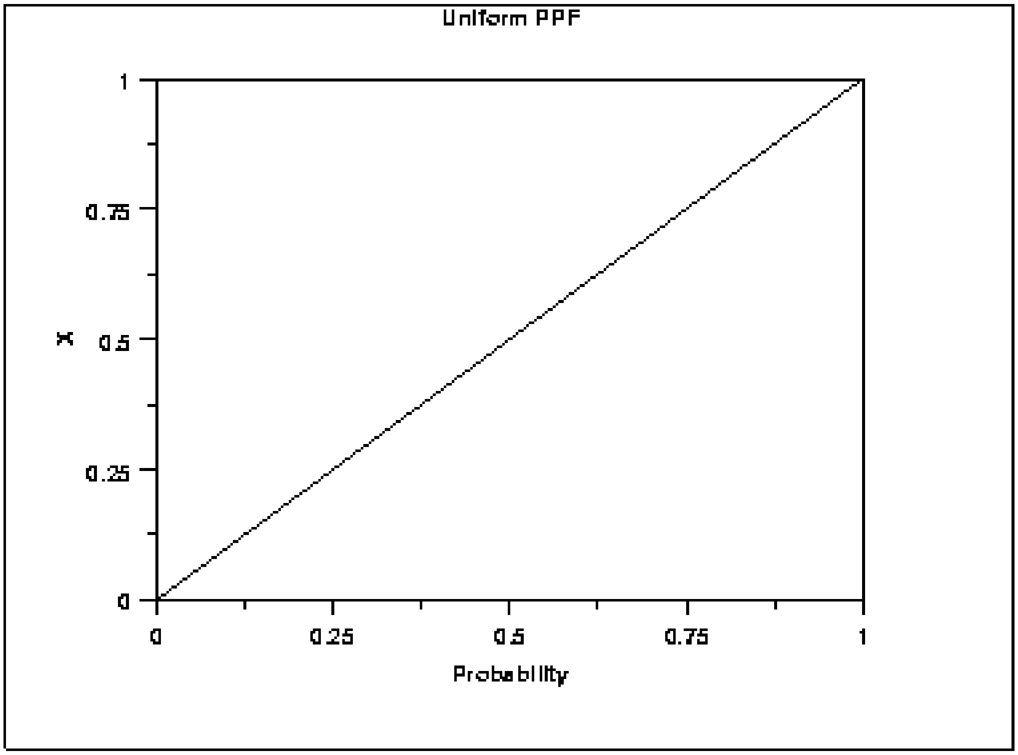


Percent Point Function

The formula for the percent point function of the uniform distribution is

$$G(p) = p \qquad \text{for } 0 \le p \le 1$$

The following is the plot of the uniform percent point function.

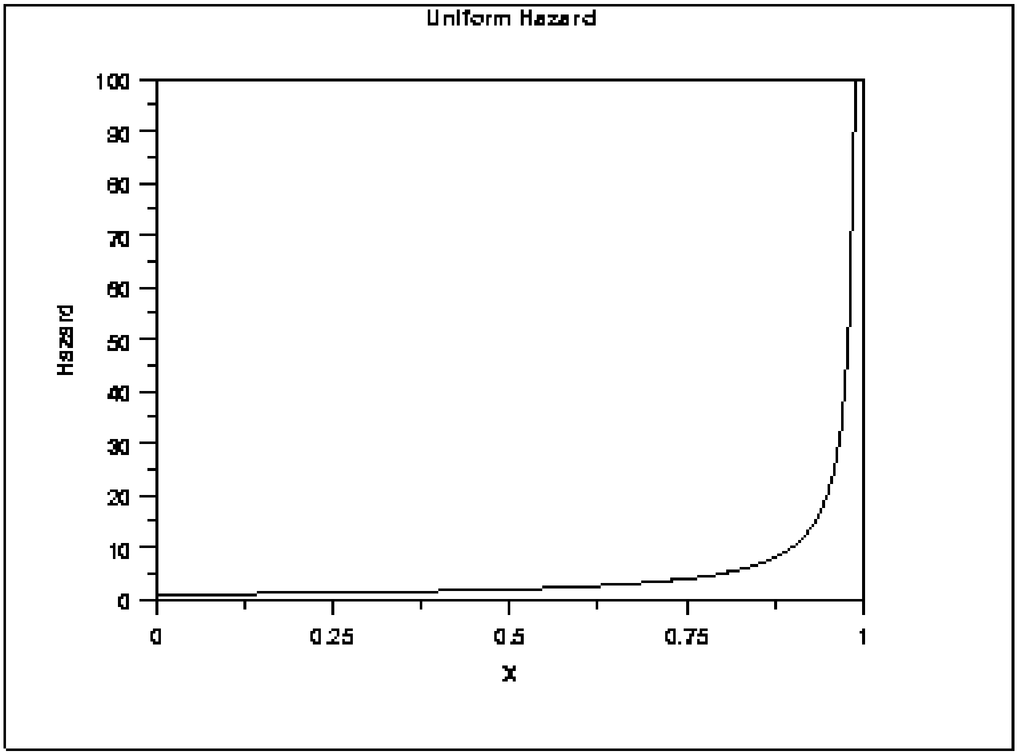


Hazard Function

The formula for the hazard function of the uniform distribution is

$$h(x) = \frac{1}{1-x} \qquad \text{for } 0 \le x < 1$$

The following is the plot of the uniform hazard function.

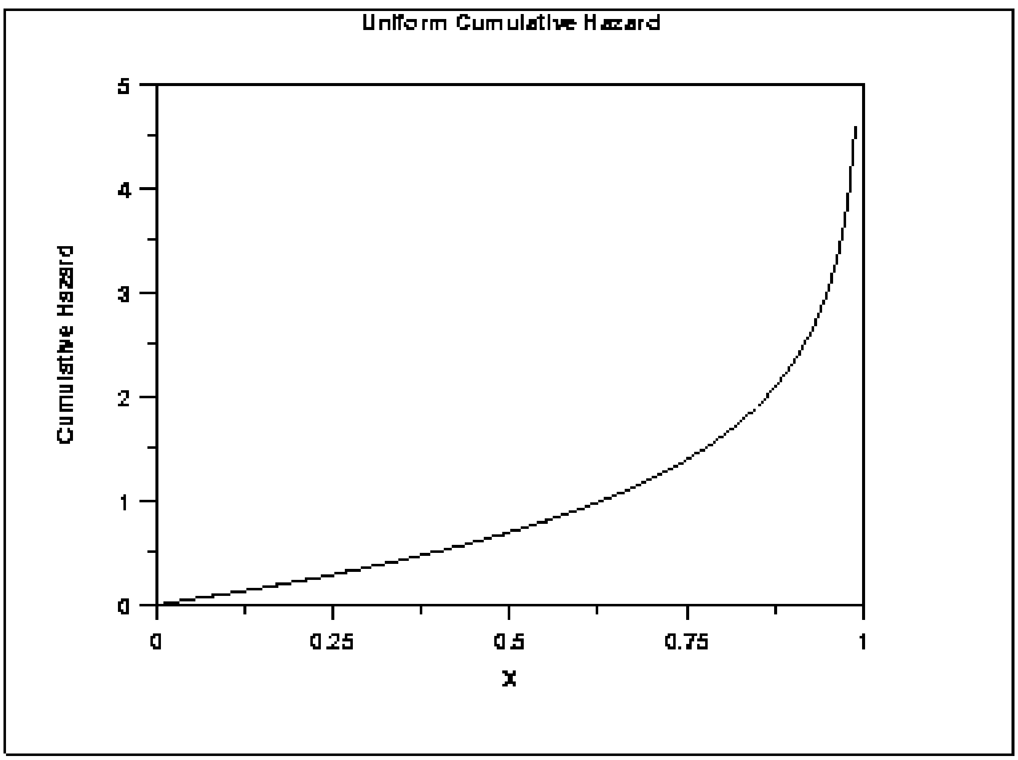


Cumulative Hazard Function

The formula for the cumulative hazard function of the uniform distribution is

$$H(x) = -ln(1-x) \qquad \text{for } 0 \le x < 1$$

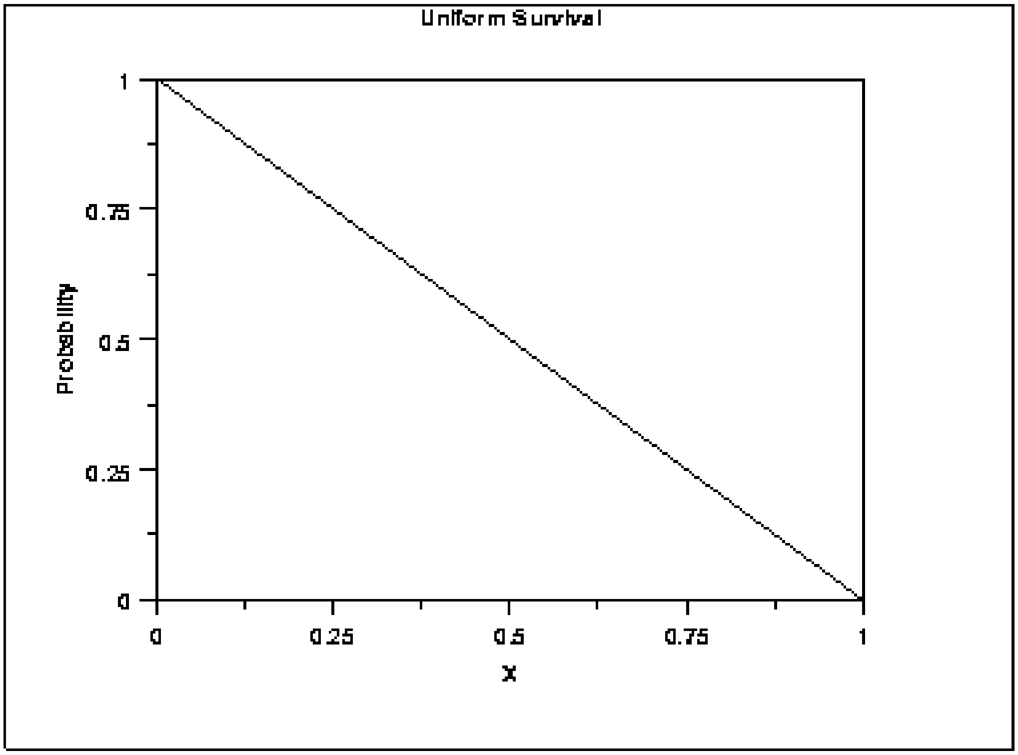
The following is the plot of the uniform cumulative hazard function.



Survival Function

The uniform survival function can be computed from the uniform cumulative distribution function.

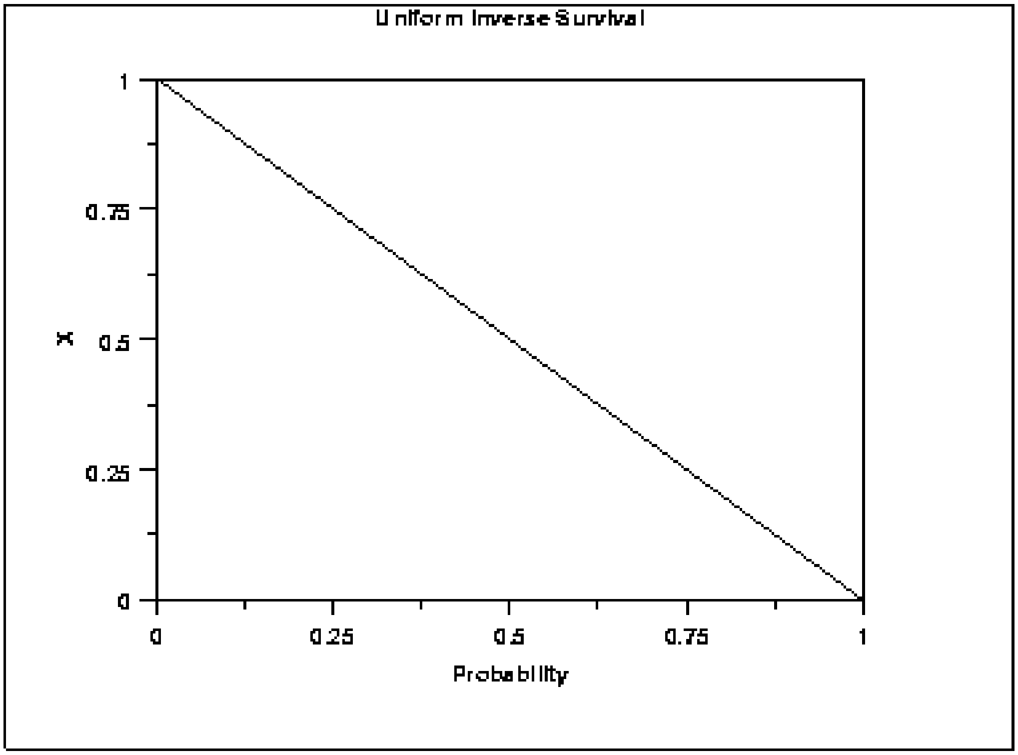
The following is the plot of the uniform survival function.



Inverse Survival Function

The uniform inverse survival function can be computed from the uniform percent point function.

The following is the plot of the uniform inverse survival function.



Common Statistics

Mean	(A + B)/2
Median	(A + B)/2
Range	B - A
Standard Deviation	$\sqrt{\frac{(B-A)^2}{12}}$
Coefficient of Variation	$\frac{(B-A)}{\sqrt{3}(B+A)}$
Skewness	0
Kurtosis	9/5

Parameter Estimation

The method of moments estimators for A and B are

$$A = \bar{x} - \sqrt{3}s$$
$$B = \bar{x} + \sqrt{3}s$$

The maximum likelihood estimators for A and B are

$$A = \text{midrange}(Y_1, Y_2, ..., Y_n) - 0.5[\text{range}(Y_1, Y_2, ..., Y_n]$$
$$B = \text{midrange}(Y_1, Y_2, ..., Y_n) + 0.5[\text{range}(Y_1, Y_2, ..., Y_n]$$

Comments	The uniform distribution defines equal probability over a given range for a continuous distribution. For this reason, it is important as a reference distribution. One of the most important applications of the uniform distribution is in the generation of random numbers. That is, almost all random number generators generate random numbers on the (0,1) interval. For other distributions, some transformation is applied to the uniform random numbers.
Software	Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the uniform distribution.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6.6.3. Cauchy Distribution

Probability Density Function

The general formula for the probability density function of the Cauchy distribution is

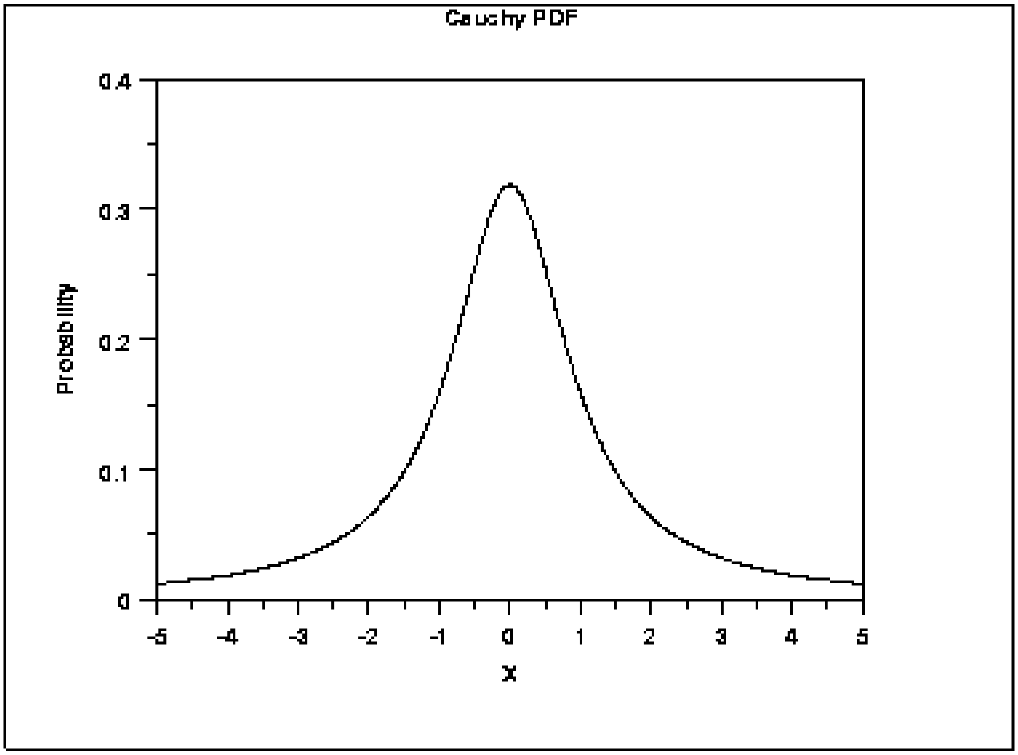
$$f(x) = \frac{1}{s\pi(1 + ((x - t)/s)^2)}$$

where ***t*** is the location parameter and ***s*** is the scale parameter. The case where ***t*** = 0 and ***s*** = 1 is called the **standard Cauchy distribution**. The equation for the standard Cauchy distribution reduces to

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the standard Cauchy probability density function.

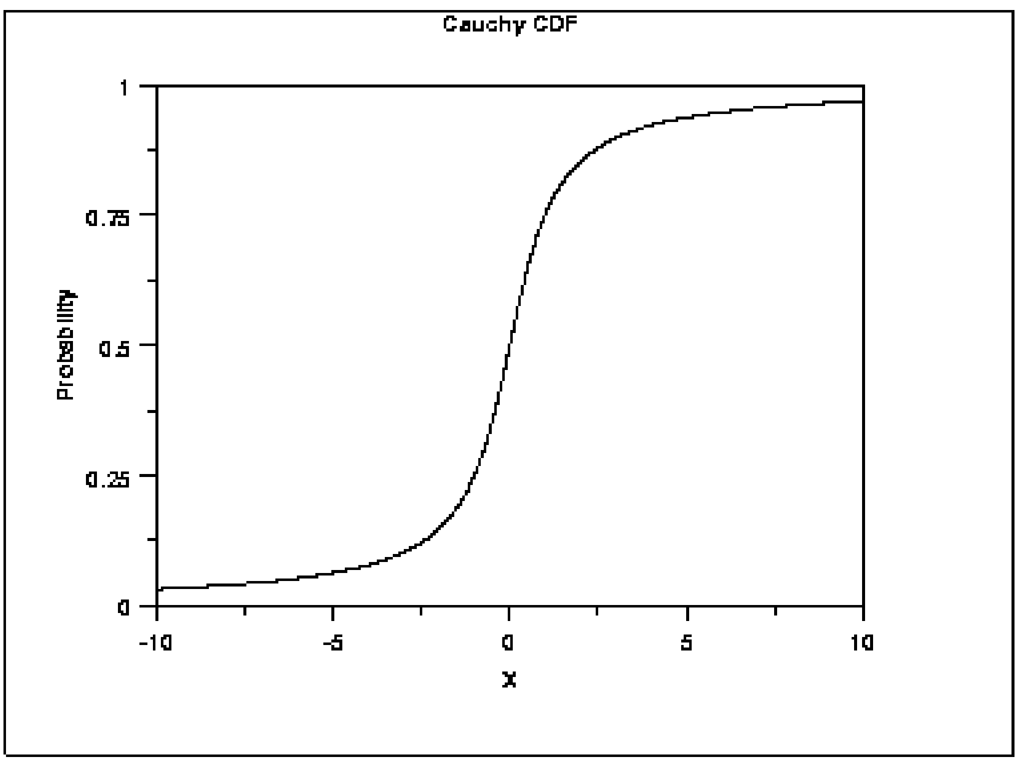


Cumulative Distribution Function

The formula for the cumulative distribution function for the Cauchy distribution is

$$F(x) = 0.5 + \frac{\arctan(x)}{\pi}$$

The following is the plot of the Cauchy cumulative distribution function.

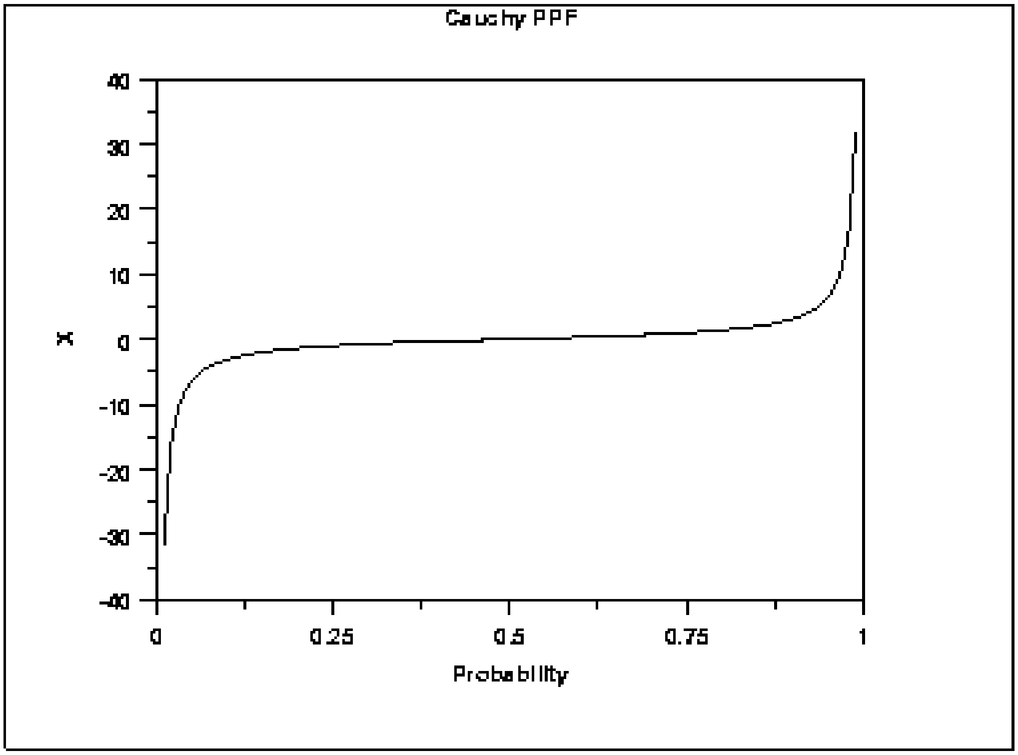


Percent Point Function

The formula for the percent point function of the Cauchy distribution is

$$G(p) = -\cot(\pi p)$$

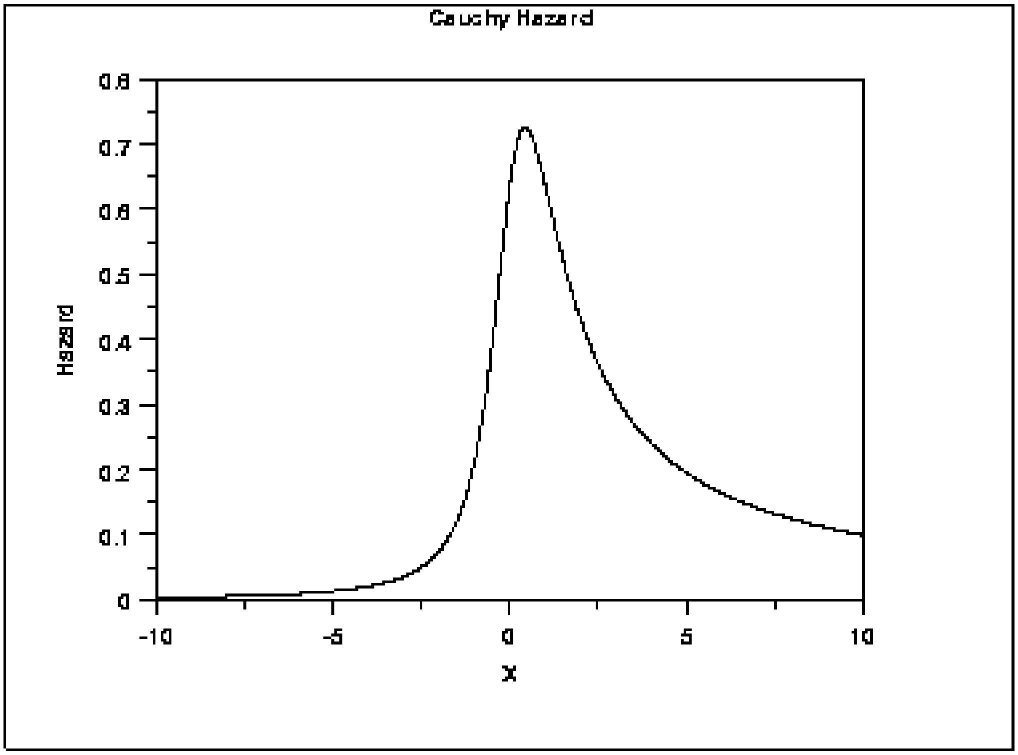
The following is the plot of the Cauchy percent point function.



Hazard Function

The Cauchy hazard function can be computed from the Cauchy probability density and cumulative distribution functions.

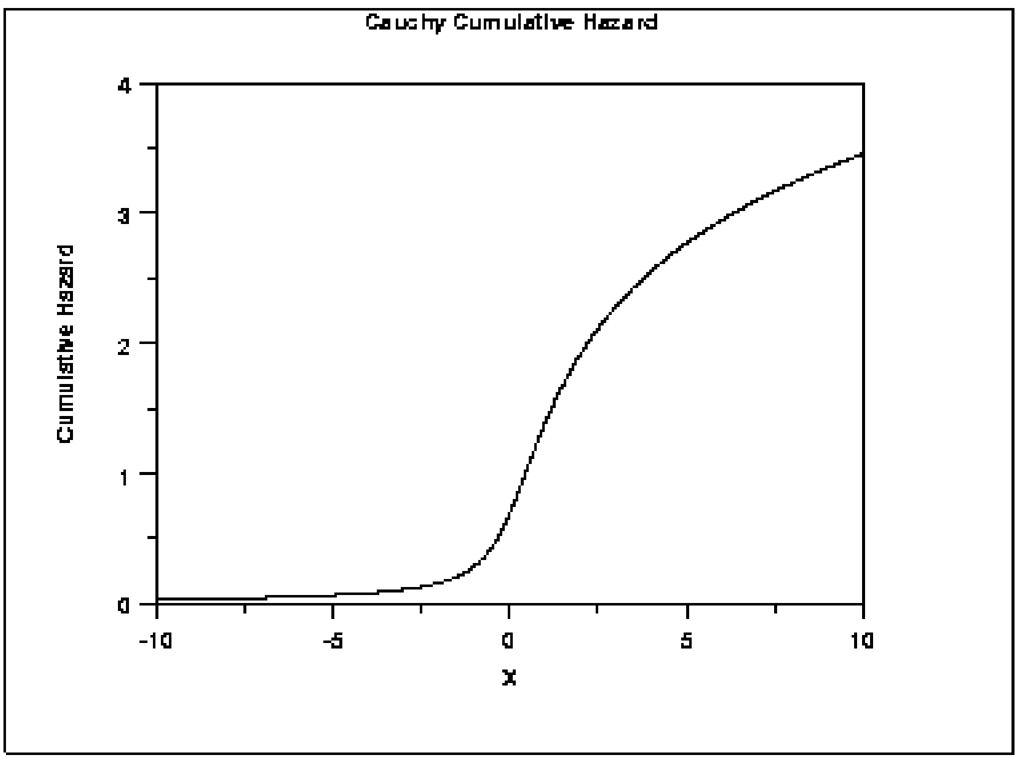
The following is the plot of the Cauchy hazard function.



Cumulative Hazard Function

The Cauchy cumulative hazard function can be computed from the Cauchy cumulative distribution function.

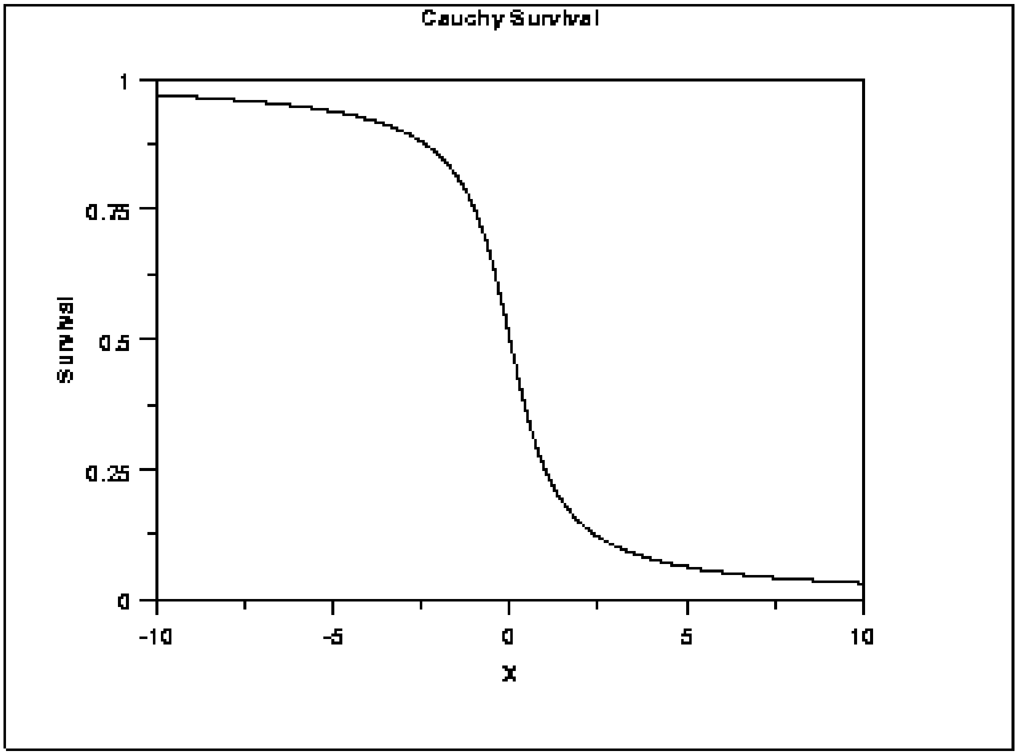
The following is the plot of the Cauchy cumulative hazard function.



Survival Function

The Cauchy survival function can be computed from the Cauchy cumulative distribution function.

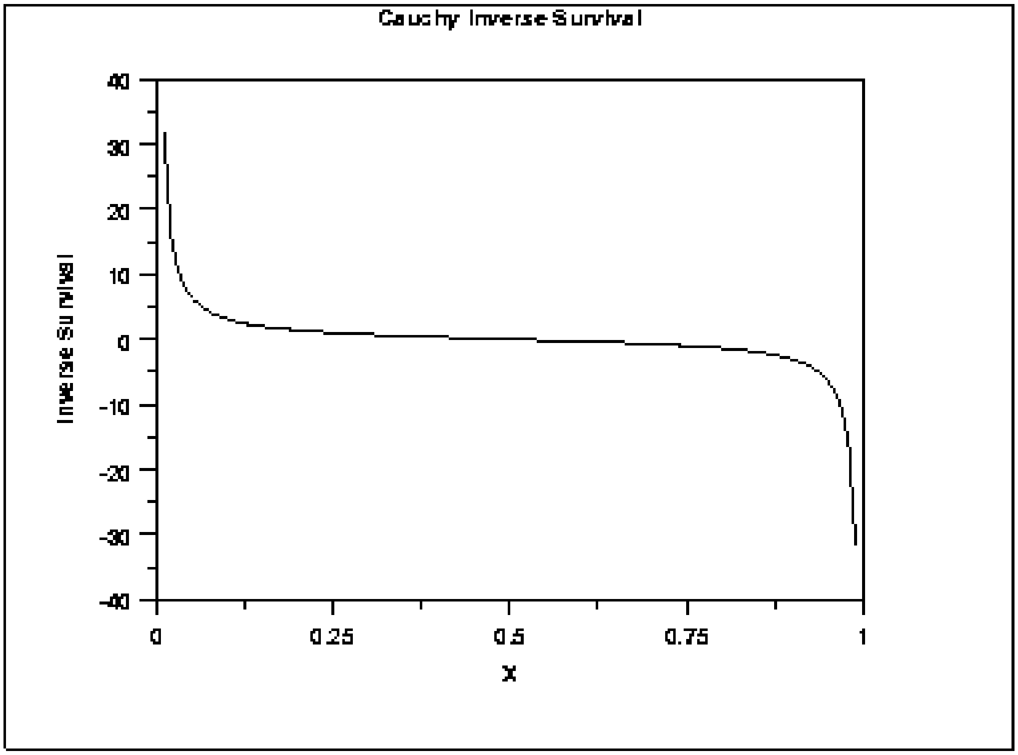
The following is the plot of the Cauchy survival function.



Inverse Survival Function

The Cauchy inverse survival function can be computed from the Cauchy percent point function.

The following is the plot of the Cauchy inverse survival function.



Common Statistics	Mean	The mean is undefined.
	Median	The location parameter t.
	Mode	The location parameter t.
	Range	Infinity in both directions.
	Standard Deviation	The standard deviation is undefined.
	Coefficient of Variation	The coefficient of variation is undefined.
	Skewness	The skewness is undefined.
	Kurtosis	The kurtosis is undefined.

Parameter Estimation

The likelihood functions for the Cauchy maximum likelihood estimates are given in chapter 16 of Johnson, Kotz, and Balakrishnan. These equations typically must be solved numerically on a computer.

Comments

The Cauchy distribution is important as an example of a pathological case. Cauchy distributions look similar to a normal distribution. However, they have much heavier tails. When studying hypothesis tests that assume normality, seeing how the tests perform on data from a Cauchy distribution is a good indicator of how sensitive the tests are to heavy-tail departures from normality. Likewise, it is a good check for robust techniques that are designed to work well under a wide variety of distributional assumptions.

The mean and standard deviation of the Cauchy distribution are undefined. The practical meaning of this is that collecting 1,000 data points gives no more accurate an estimate of the mean and standard deviation than does a single point.

Software

Many general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the Cauchy distribution.

1.3.6.6.4. t Distribution

Probability Density Function

The formula for the probability density function of the *t* distribution is

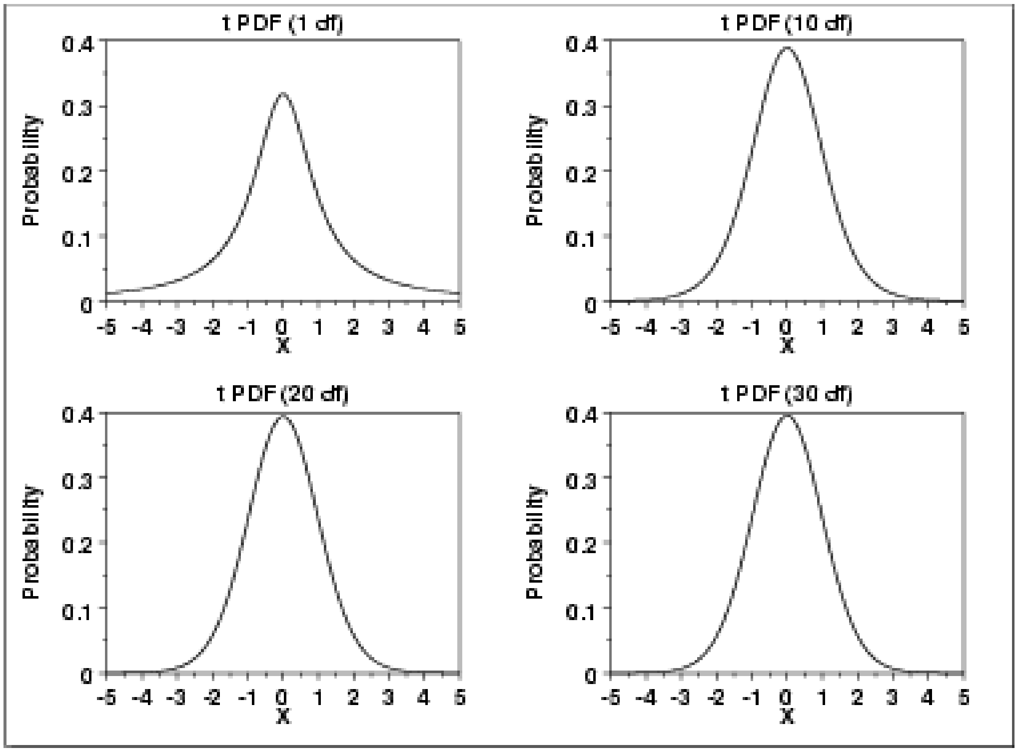
$$f(x) = \frac{\left(1 + \frac{x^2}{\nu}\right)^{\frac{-(\nu+1)}{2}}}{B(0.5, 0.5\nu)\sqrt{\nu}}$$

where B is the beta function and ν is a positive integer shape parameter. The formula for the beta function is

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

In a testing context, the *t* distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the *t* distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the *t* probability density function for 4 different values of the shape parameter.

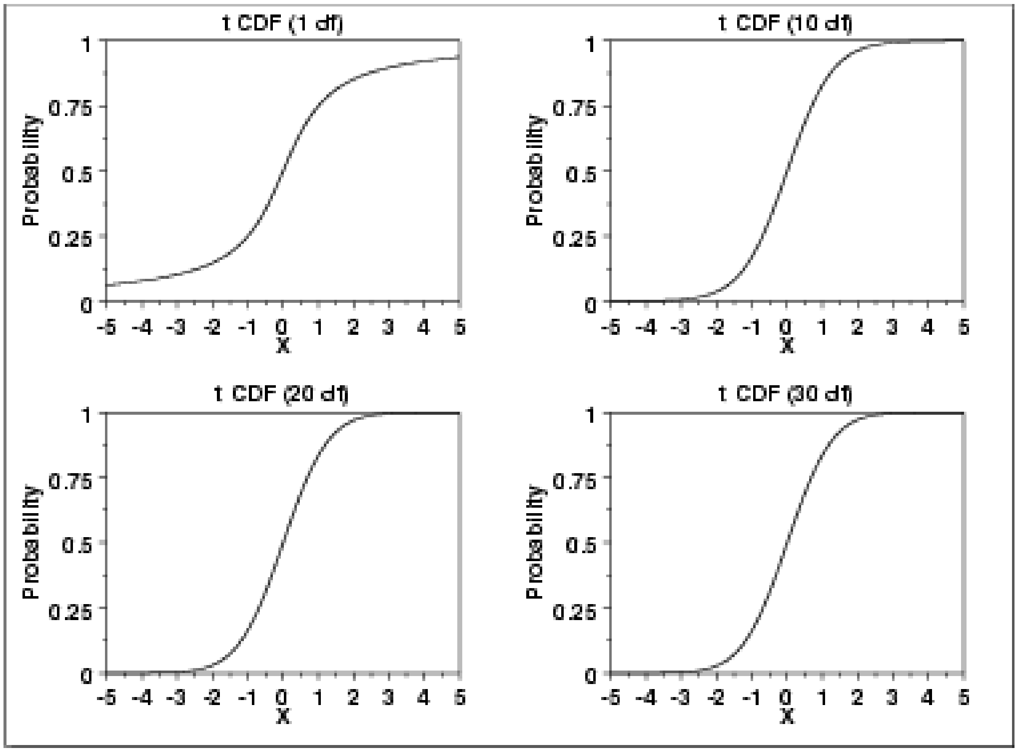


These plots all have a similar shape. The difference is in the heaviness of the tails. In fact, the *t* distribution with ν equal to 1 is a Cauchy distribution. The *t* distribution approaches a normal distribution as ν becomes large. The approximation is quite good for values of ν > 30.

Cumulative Distribution Function

The formula for the cumulative distribution function of the *t* distribution is complicated and is not included here. It is given in the Evans, Hastings, and Peacock book.

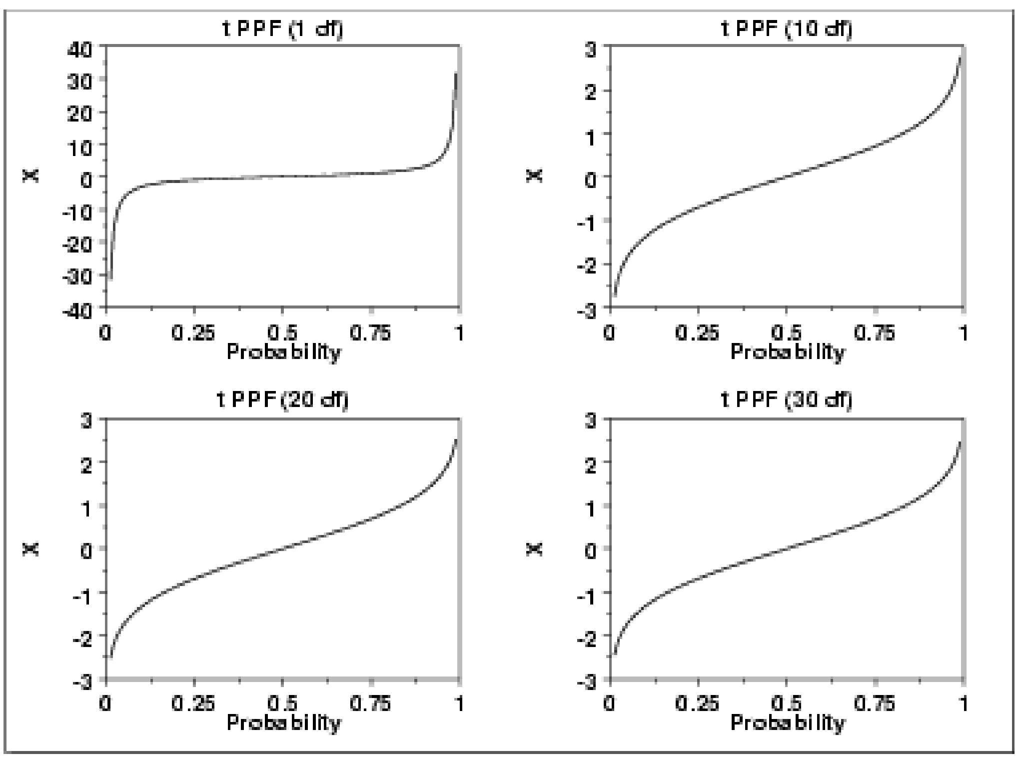
The following are the plots of the *t* cumulative distribution function with the same values of ν as the pdf plots above.



Percent Point Function

The formula for the percent point function of the *t* distribution does not exist in a simple closed form. It is computed numerically.

The following are the plots of the *t* percent point function with the same values of ν as the pdf plots above.



Other Probability Functions	Since the *t* distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

Mean	0 (It is undefined for ν equal to 1.)
Median	0
Mode	0
Range	Infinity in both directions.
Standard Deviation	$\sqrt{\frac{\nu}{(\nu - 2)}}$ It is undefined for ν equal to 1 or 2.
Coefficient of Variation	Undefined
Skewness	0. It is undefined for ν less than or equal to 3. However, the t distribution is symmetric in all cases.
Kurtosis	$\frac{3(\nu - 2)}{(\nu - 4)}$ It is undefined for ν less than or equal to 4.

Parameter Estimation	Since the *t* distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.
Comments	The *t* distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. The most common example is testing if data are consistent with the assumed process mean.
Software	Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the *t* distribution.

1.3.6.6.5. F Distribution

Probability Density Function

The F distribution is the ratio of two chi-square distributions with degrees of freedom ν_1 and ν_2, respectively, where each chi-square has first been divided by its degrees of freedom. The formula for the probability density function of the F distribution is

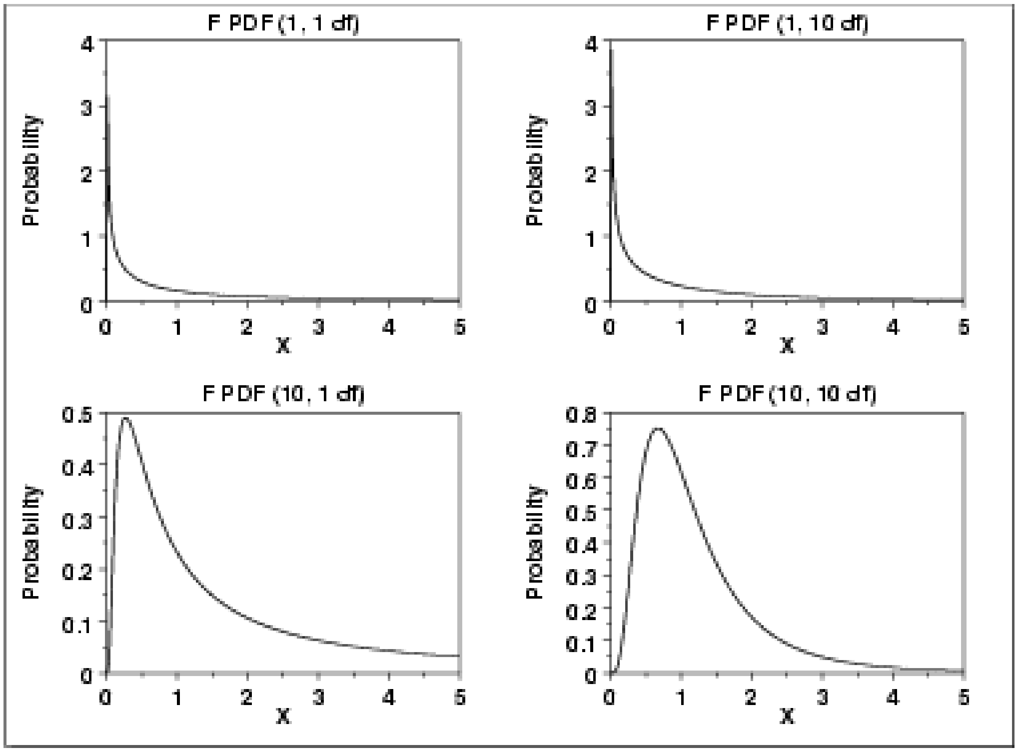
$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(1+\frac{\nu_1 x}{\nu_2})^{\frac{\nu_1+\nu_2}{2}}}$$

where ν_1 and ν_2 are the shape parameters and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

In a testing context, the F distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the F distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the F probability density function for 4 different values of the shape parameters.



Cumulative Distribution Function

The formula for the Cumulative distribution function of the F distribution is

$$F(x) = 1 - I_k(\frac{\nu_2}{2}, \frac{\nu_1}{2})$$

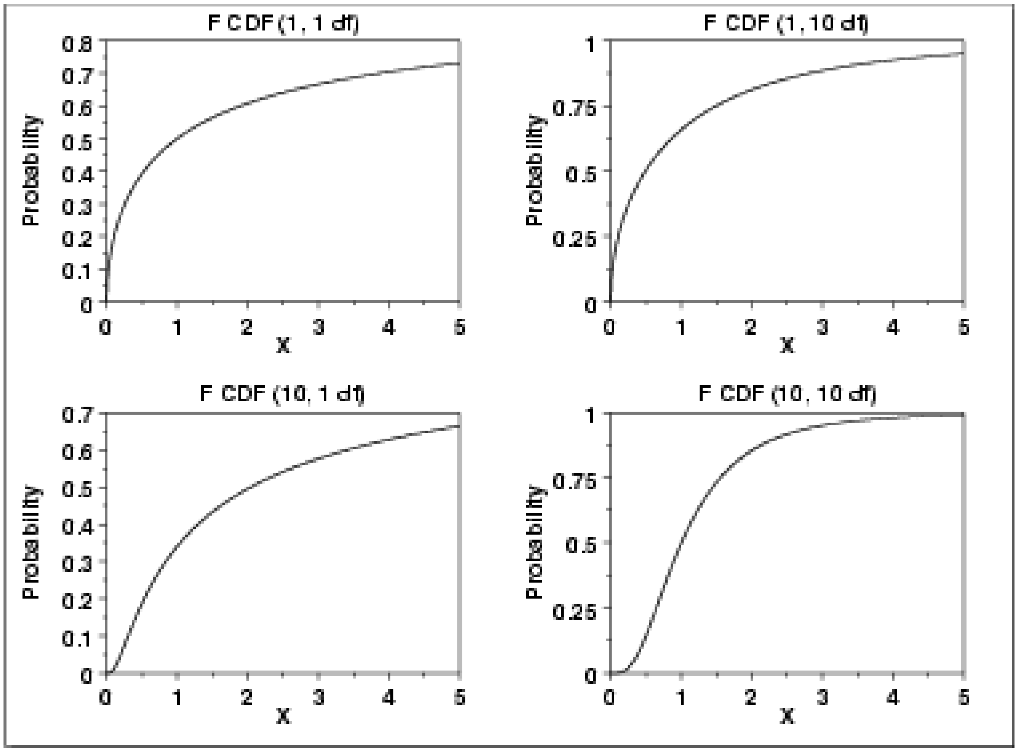
where k = ν_2/ (ν_2 + ν_1*x) and I_k is the incomplete beta function. The formula for the incomplete beta function is

$$I_k(x, \alpha, \beta) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt}{B(\alpha, \beta)}$$

where B is the beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

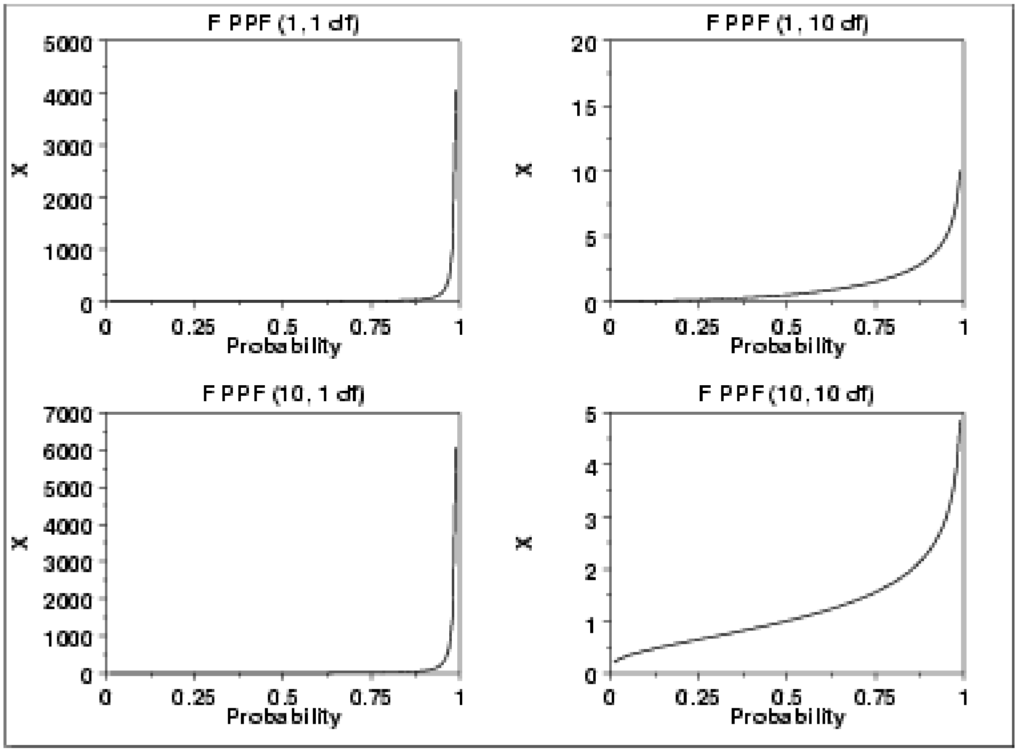
The following is the plot of the F cumulative distribution function with the same values of ν_1 and ν_2 as the pdf plots above.



Percent Point Function

The formula for the [percent point function](#) of the F distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the F percent point function with the same values of ν_1 and ν_2 as the pdf plots above.



Other Probability Functions

Since the F distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

The formulas below are for the case where the location parameter is zero and the scale parameter is one.

Mean	$\frac{\nu_2}{(\nu_2 - 2)}$	$\nu_2 > 2$
Mode	$\frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)}$	$\nu_1 > 2$
Range	0 to positive infinity	
Standard Deviation	$\sqrt{\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}}$	$\nu_2 > 4$
Coefficient of Variation	$\sqrt{\frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}}$	$\nu_2 > 4$
Skewness	$\frac{(2\nu_1 + \nu_2 - 2)\sqrt{8(\nu_2 - 4)}}{\sqrt{\nu_1}(\nu_2 - 6)\sqrt{(\nu_1 + \nu_2 - 2)}}$	$\nu_2 > 6$

Parameter Estimation

Since the F distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.

Comments

The F distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. Two common examples are the analysis of variance and the F test to determine if the variances of two populations are equal.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the F distribution.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6.6.6. Chi-Square Distribution

Probability Density Function

The chi-square distribution results when ν independent variables with standard normal distributions are squared and summed. The formula for the probability density function of the chi-square distribution is

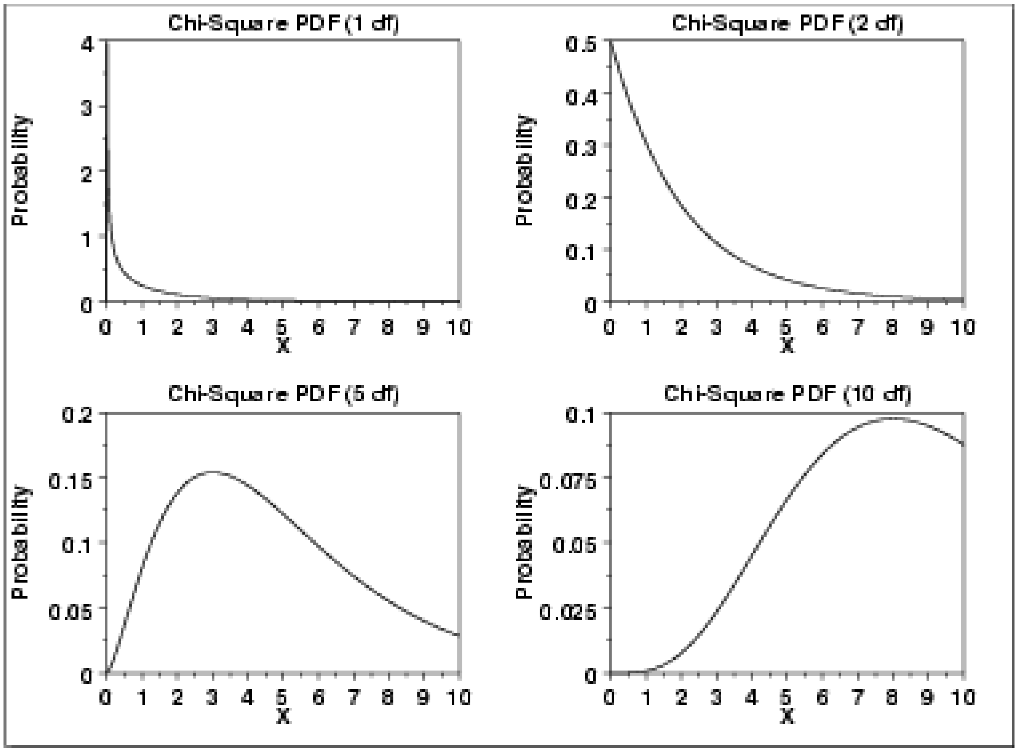
$$f(x) = \frac{e^{\frac{-x}{2}} x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} \qquad \mbox{for } x \ge 0$$

where ν is the shape parameter and Γ is the gamma function. The formula for the gamma function is

$$\Gamma(a) = \int_{0}^{\infty} t^{a-1}e^{-t}dt$$

In a testing context, the chi-square distribution is treated as a "standardized distribution" (i.e., no location or scale parameters). However, in a distributional modeling context (as with other probability distributions), the chi-square distribution itself can be transformed with a location parameter, μ, and a scale parameter, σ.

The following is the plot of the chi-square probability density function for 4 different values of the shape parameter.



Cumulative Distribution Function

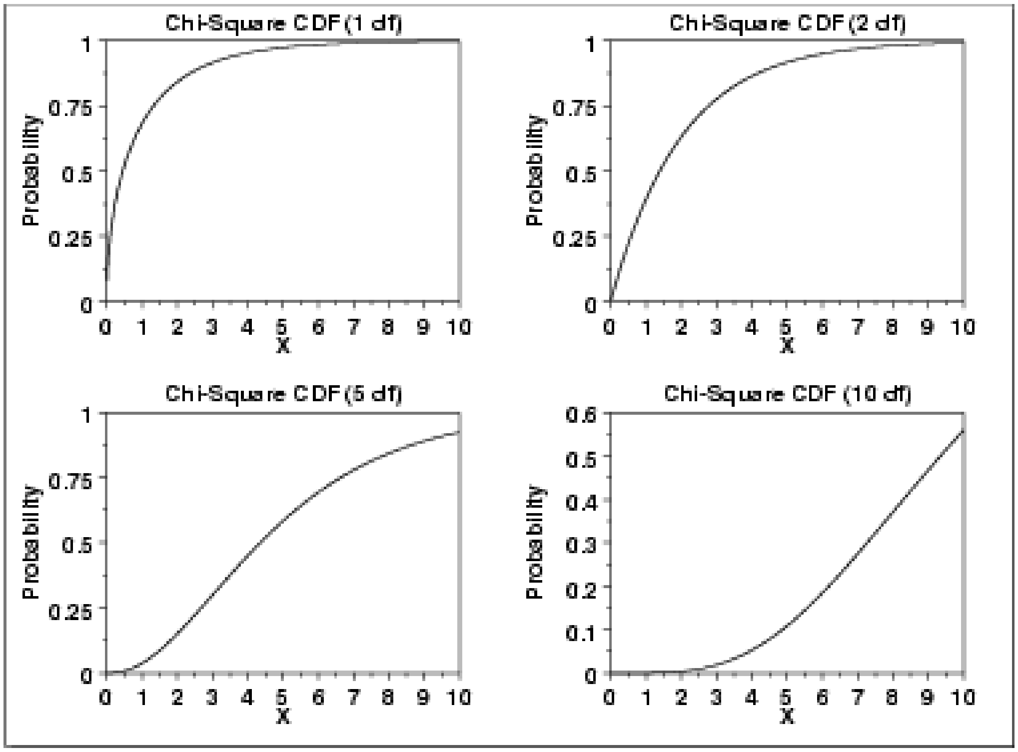
The formula for the cumulative distribution function of the chi-square distribution is

$$F(x) = \frac{\gamma(\frac{\nu}{2}, \frac{x}{2})}{\Gamma(\frac{\nu}{2})} \qquad \mbox{for } x \ge 0$$

where Γ is the gamma function defined above and γ is the incomplete gamma function. The formula for the incomplete gamma function is

$$\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} dt$$

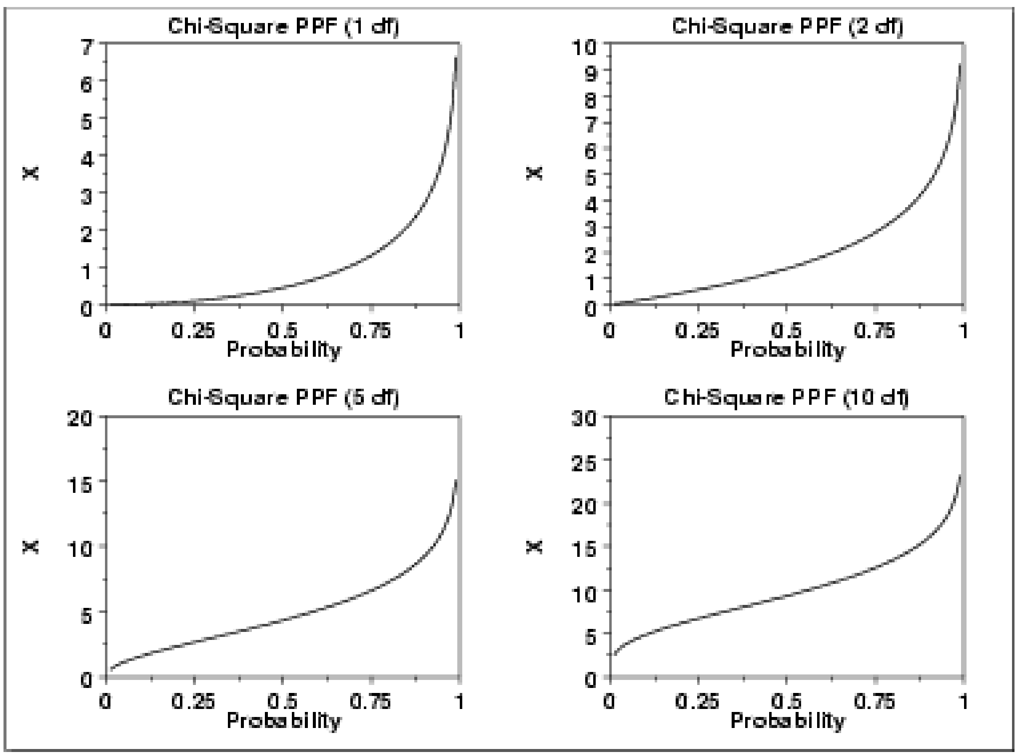
The following is the plot of the chi-square cumulative distribution function with the same values of ν as the pdf plots above.



Percent Point Function

The formula for the [percent point function](#) of the chi-square distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the chi-square percent point function with the same values of ν as the pdf plots above.



Other Probability Functions

Since the chi-square distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

Mean	ν
Median	approximately ν - 2/3 for large ν
Mode	$\nu - 2 \quad \text{for } \nu > 2$
Range	0 to positive infinity
Standard Deviation	$\sqrt{2\nu}$
Coefficient of Variation	$\sqrt{\frac{2}{\nu}}$
Skewness	$\frac{2^{1.5}}{\sqrt{\nu}}$
Kurtosis	$3 + \frac{12}{\nu}$

Parameter Estimation

Since the chi-square distribution is typically used to develop hypothesis tests and confidence intervals and rarely for modeling applications, we omit any discussion of parameter estimation.

Comments

The chi-square distribution is used in many cases for the critical regions for hypothesis tests and in determining confidence intervals. Two common examples are the chi-square test for independence in an ***RxC*** contingency table and the chi-square test to determine if the standard deviation of a population is equal to a pre-specified value.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the chi-square distribution.

1.3.6.6.7. Exponential Distribution

Probability Density Function

The general formula for the probability density function of the exponential distribution is

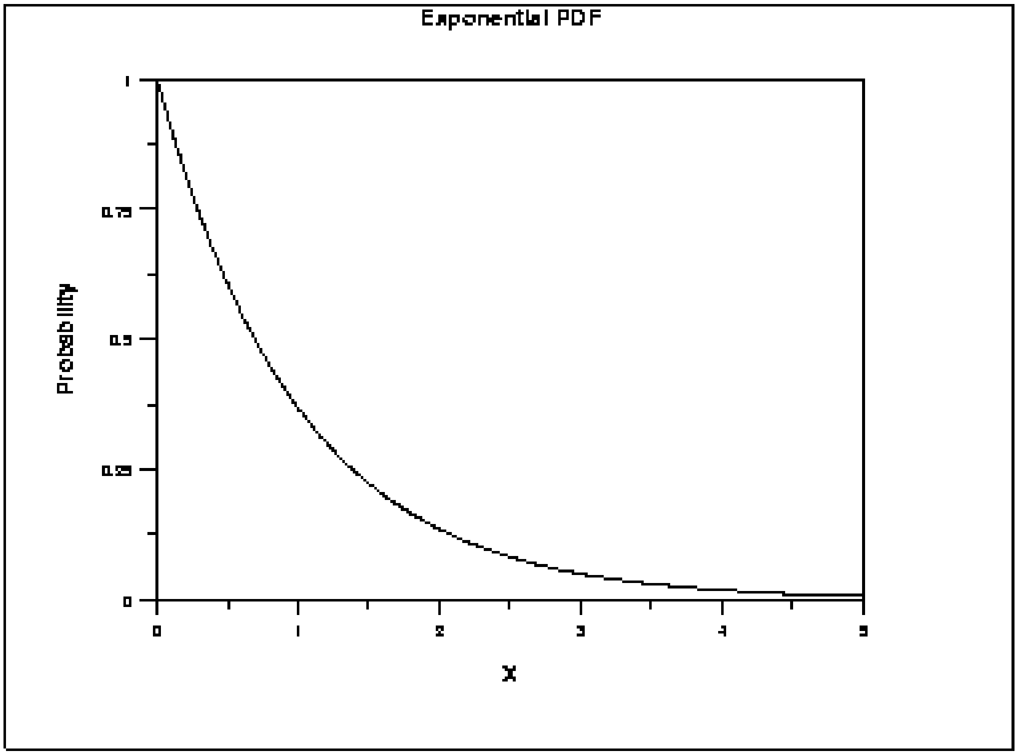
$$f(x) = \frac{1}{\beta} e^{-(x-\mu)/\beta} \qquad x \ge \mu; \beta > 0$$

where μ is the location parameter and β is the scale parameter (the scale parameter is often referred to as λ which equals $1/\beta$). The case where μ = 0 and β = 1 is called the **standard exponential distribution**. The equation for the standard exponential distribution is

$$f(x) = e^{-x} \qquad \mbox{for } x \ge 0$$

The general form of probability functions can be expressed in terms of the standard distribution. Subsequent formulas in this section are given for the 1-parameter (i.e., with scale parameter) form of the function.

The following is the plot of the exponential probability density function.

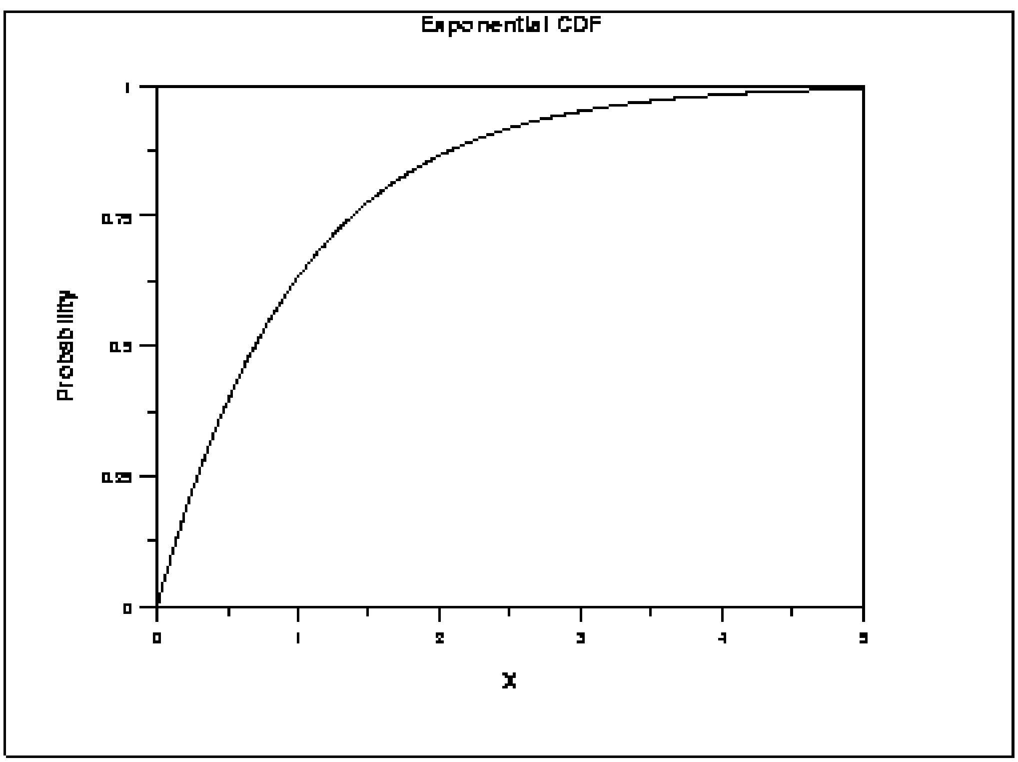


Cumulative Distribution Function

The formula for the cumulative distribution function of the exponential distribution is

$$F(x) = 1 - e^{-x/\beta} \qquad x \ge 0; \beta > 0$$

The following is the plot of the exponential cumulative distribution function.

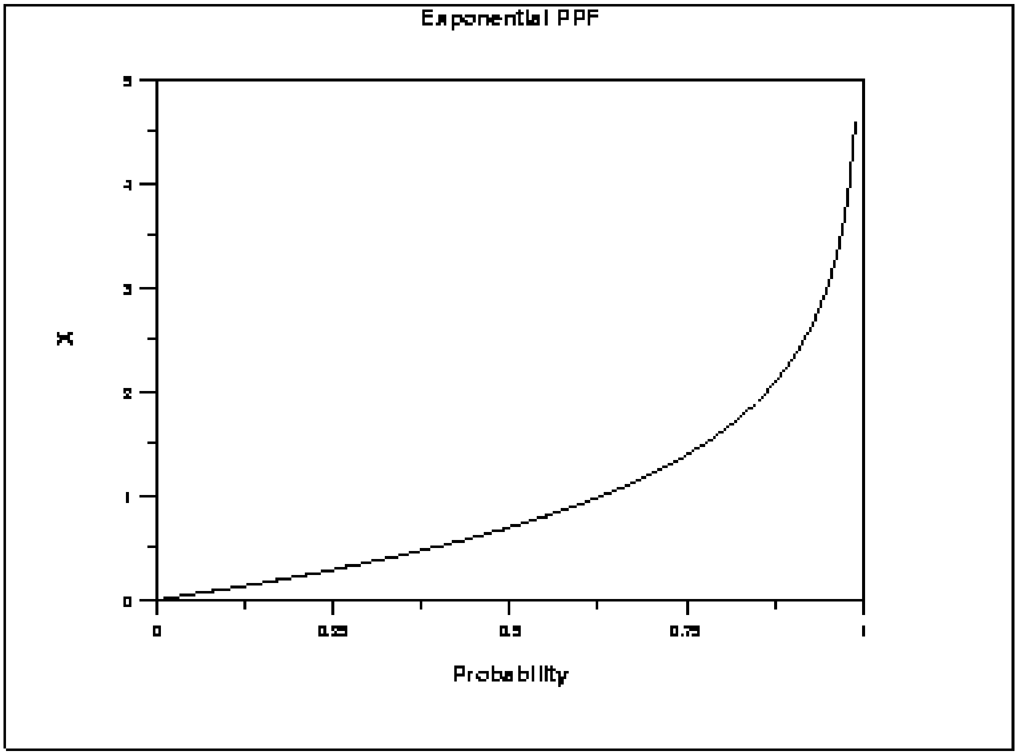


Percent Point Function

The formula for the percent point function of the exponential distribution is

$$G(p) = -\beta \ln(1 - p) \qquad 0 \le p < 1; \beta > 0$$

The following is the plot of the exponential percent point function.

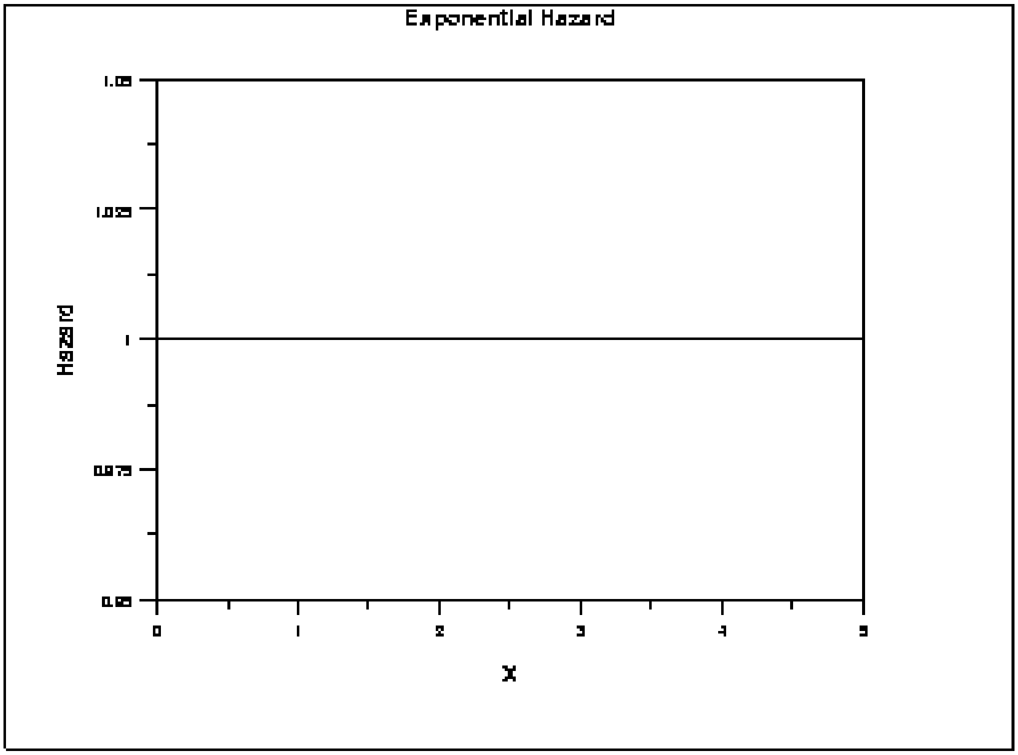


Hazard Function

The formula for the hazard function of the exponential distribution is

$$h(x) = \frac{1}{\beta} \qquad x \ge 0; \beta > 0$$

The following is the plot of the exponential hazard function.

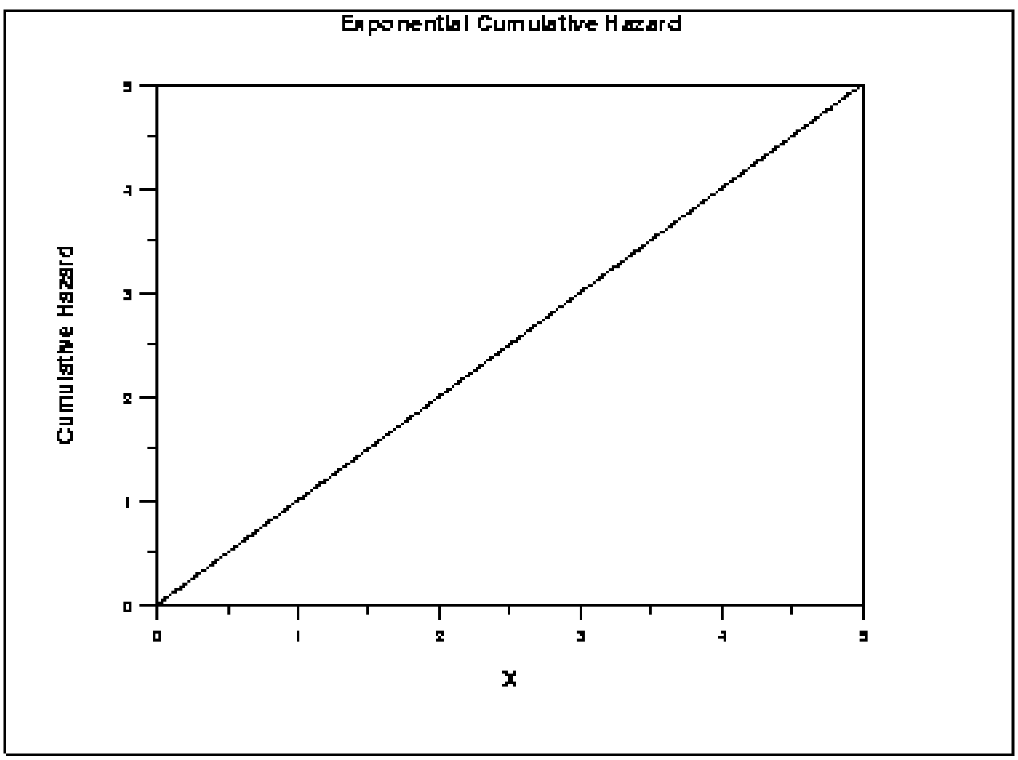


Cumulative Hazard Function

The formula for the cumulative hazard function of the exponential distribution is

$$H(x) = \frac{x}{\beta} \qquad x \ge 0; \beta > 0$$

The following is the plot of the exponential cumulative hazard function.

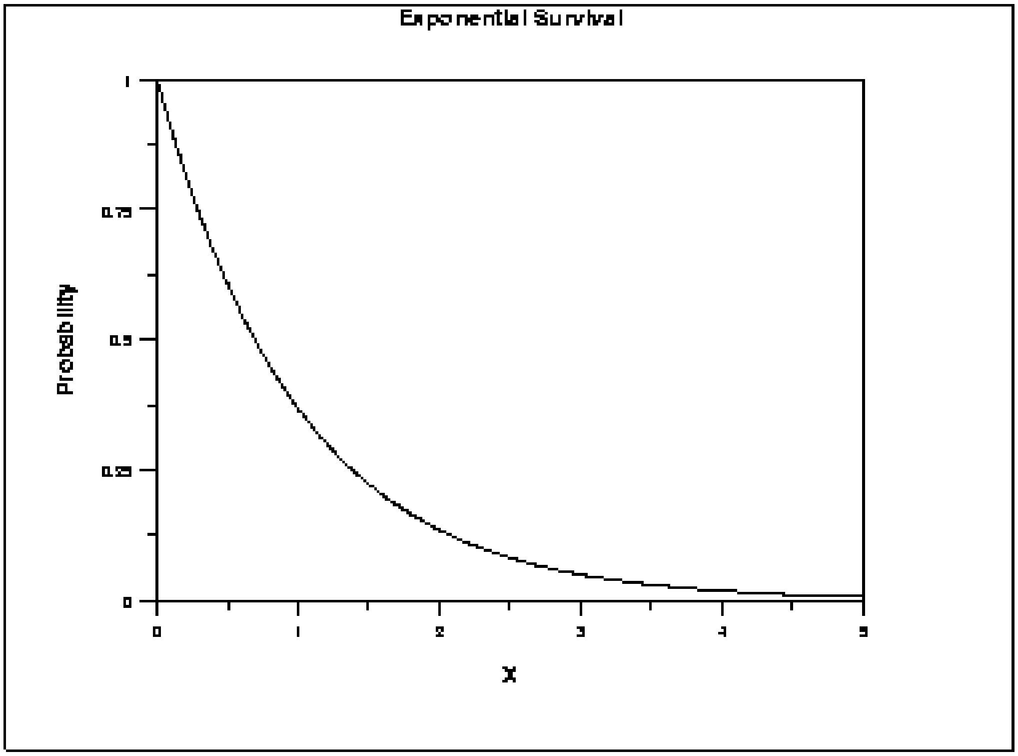


Survival Function

The formula for the survival function of the exponential distribution is

$$S(x) = e^{-x/\beta} \qquad x \ge 0; \beta > 0$$

The following is the plot of the exponential survival function.

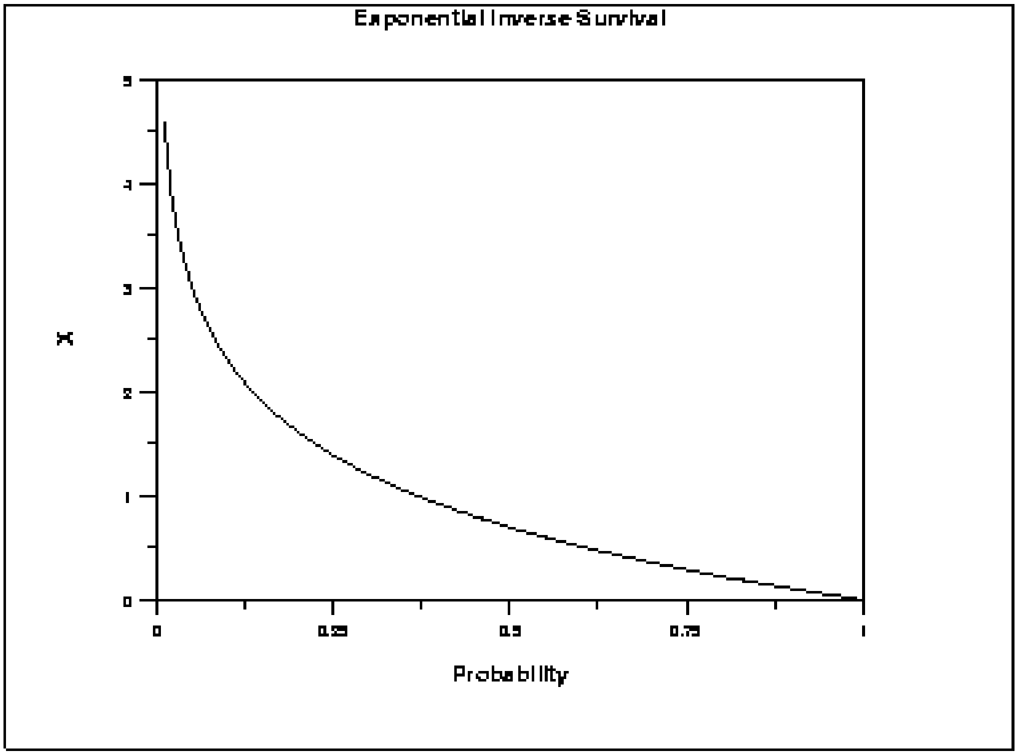


Inverse Survival Function

The formula for the inverse survival function of the exponential distribution is

$$Z(p) = -\beta \ln(p) \qquad 0 \le p < 1; \beta > 0$$

The following is the plot of the exponential inverse survival function.



Common Statistics	Mean	β
	Median	$\beta \ln 2$
	Mode	Zero
	Range	Zero to plus infinity
	Standard Deviation	β
	Coefficient of Variation	1
	Skewness	2
	Kurtosis	9

Parameter Estimation

For the full sample case, the maximum likelihood estimator of the scale parameter is the sample mean. Maximum likelihood estimation for the exponential distribution is discussed in the chapter on reliability (Chapter 8). It is also discussed in chapter 19 of Johnson, Kotz, and Balakrishnan.

Comments

The exponential distribution is primarily used in reliability applications. The exponential distribution is used to model data with a constant failure rate (indicated by the hazard plot which is simply equal to a constant).

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the exponential distribution.

1.3.6.6.8. Weibull Distribution

Probability Density Function

The formula for the probability density function of the general Weibull distribution is

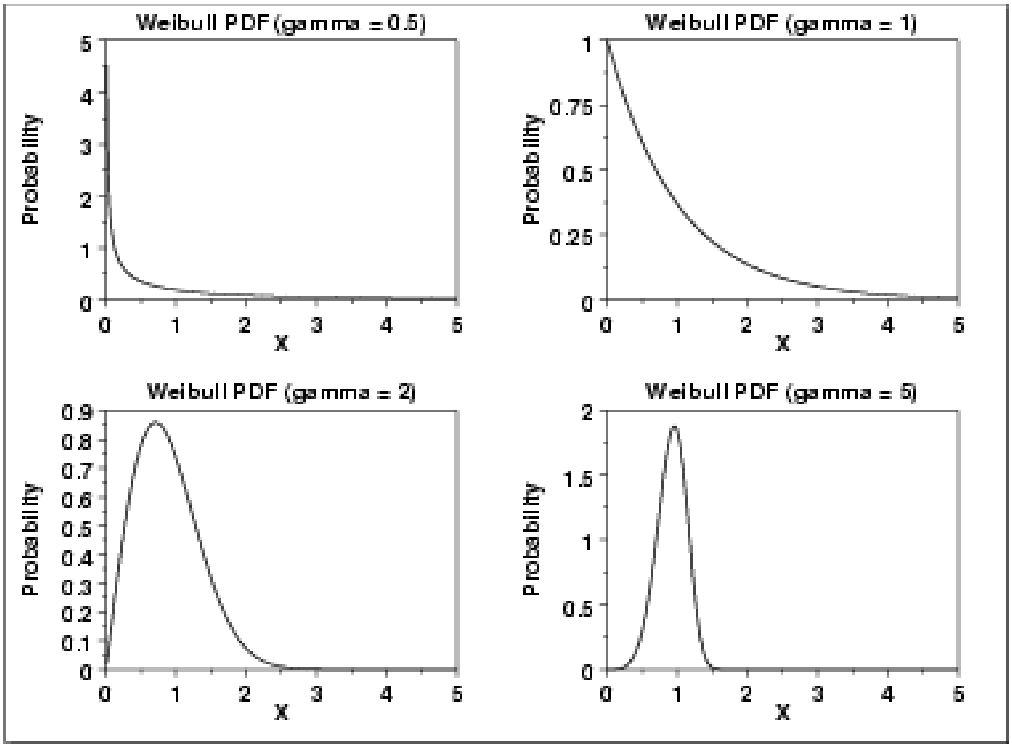
$$f(x) = \frac{\gamma}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{(\gamma-1)}\exp\left(-((x-\mu)/\alpha)^{\gamma}\right) \qquad x \ge \mu; \gamma, \alpha > 0$$

where γ is the shape parameter, μ is the location parameter and α is the scale parameter. The case where μ = 0 and α = 1 is called the **standard Weibull distribution**. The case where μ = 0 is called the 2-parameter Weibull distribution. The equation for the standard Weibull distribution reduces to

$$f(x) = \gamma x^{(\gamma-1)}\exp(-(x^{\gamma})) \qquad x \ge 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the Weibull probability density function.

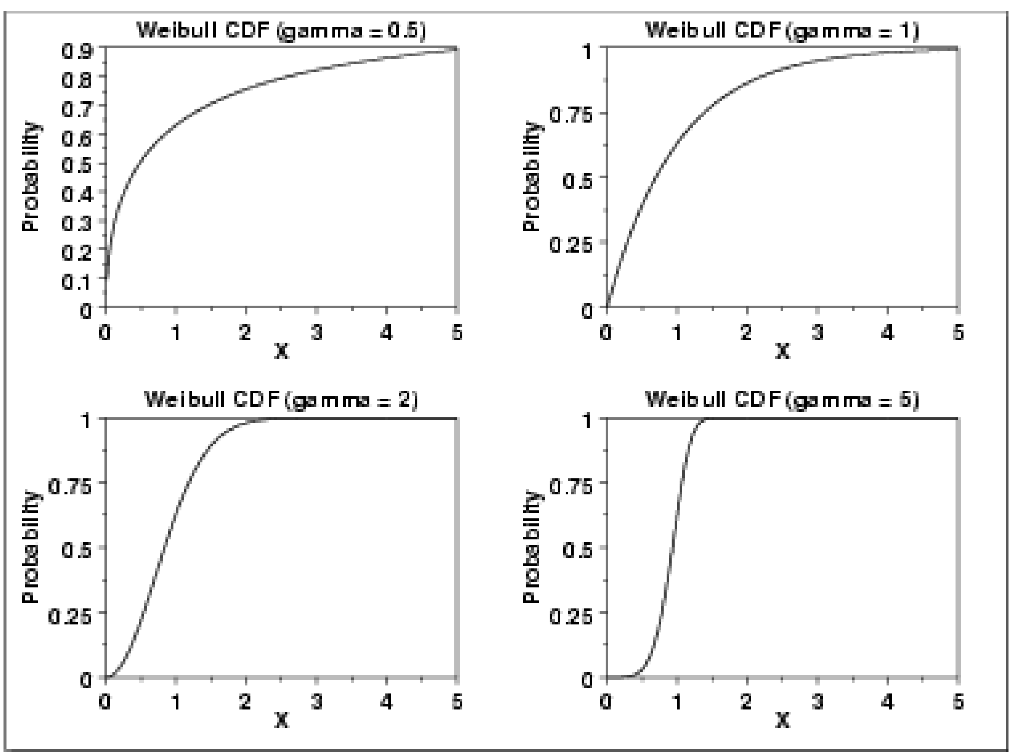


Cumulative Distribution Function

The formula for the cumulative distribution function of the Weibull distribution is

$$F(x) = 1 - e^{-(x^{\gamma})} \qquad x \ge 0; \gamma > 0$$

The following is the plot of the Weibull cumulative distribution function with the same values of γ as the pdf plots above.

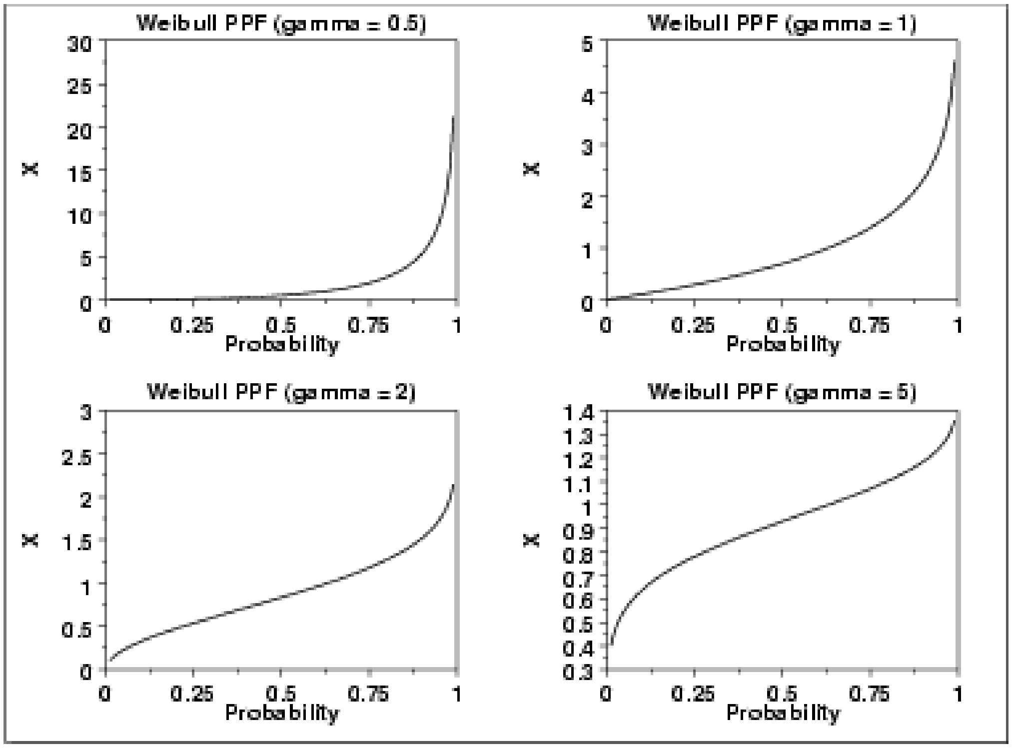


Percent Point Function

The formula for the percent point function of the Weibull distribution is

$$G(p) = (-\ln(1-p))^{1/\gamma} \qquad 0 \le p < 1; \gamma > 0$$

The following is the plot of the Weibull percent point function with the same values of γ as the pdf plots above.

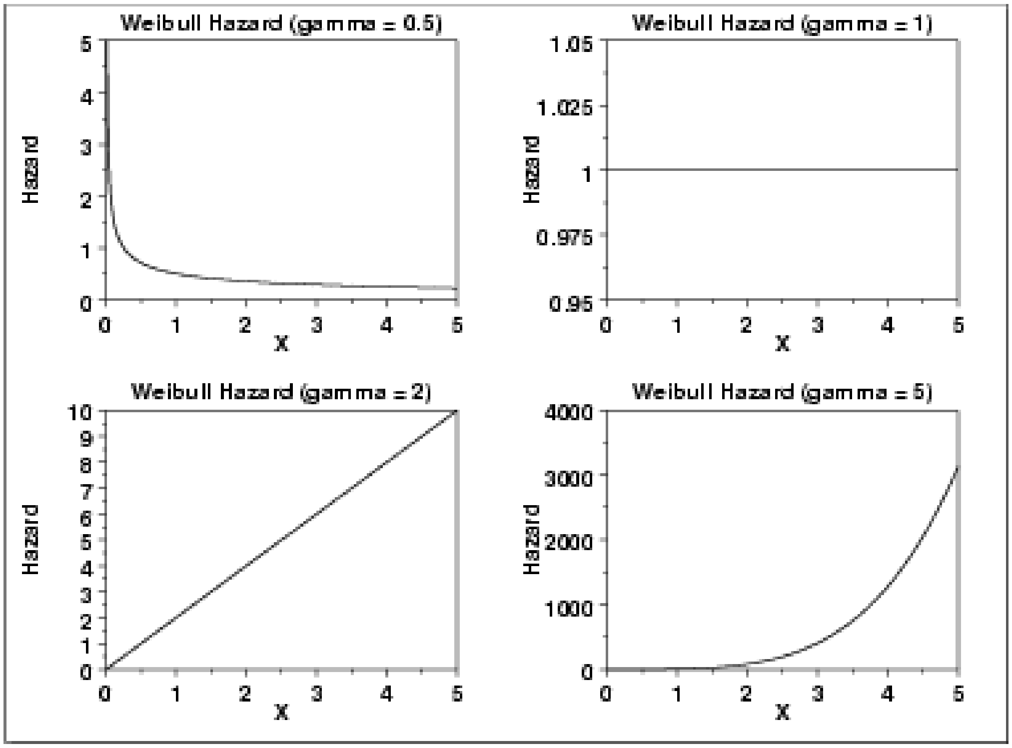


Hazard Function

The formula for the hazard function of the Weibull distribution is

$$h(x) = \gamma x^{(\gamma-1)} \qquad x \ge 0; \gamma > 0$$

The following is the plot of the Weibull hazard function with the same values of γ as the pdf plots above.

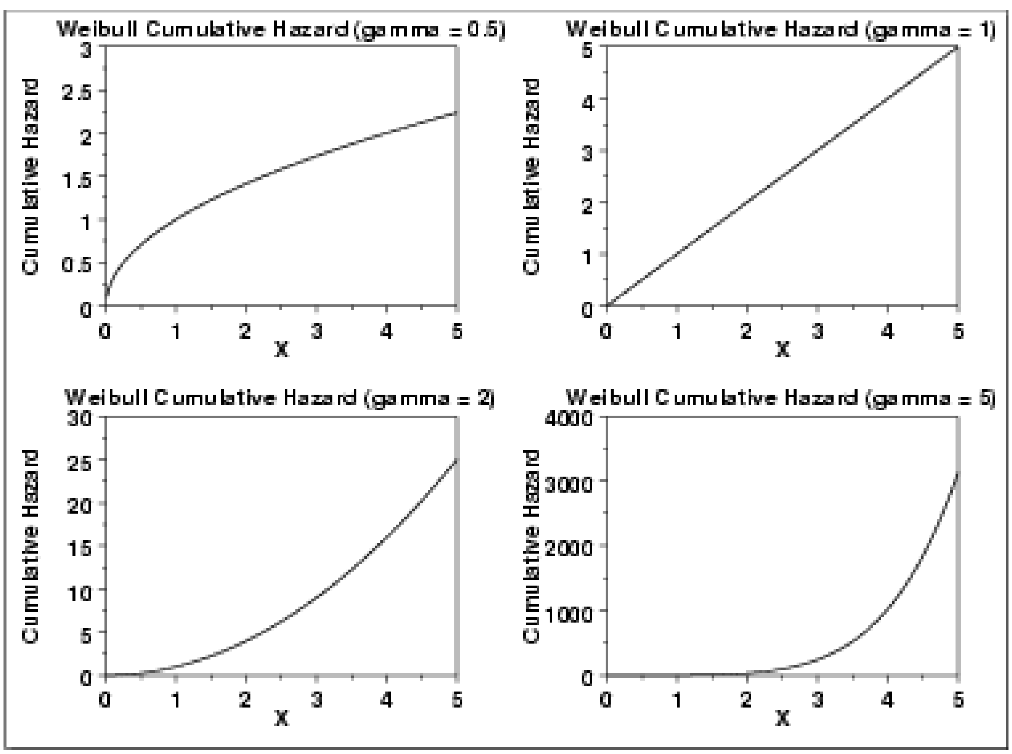


Cumulative Hazard Function

The formula for the cumulative hazard function of the Weibull distribution is

$$H(x) = x^{\gamma} \qquad x \ge 0; \gamma > 0$$

The following is the plot of the Weibull cumulative hazard function with the same values of γ as the pdf plots above.

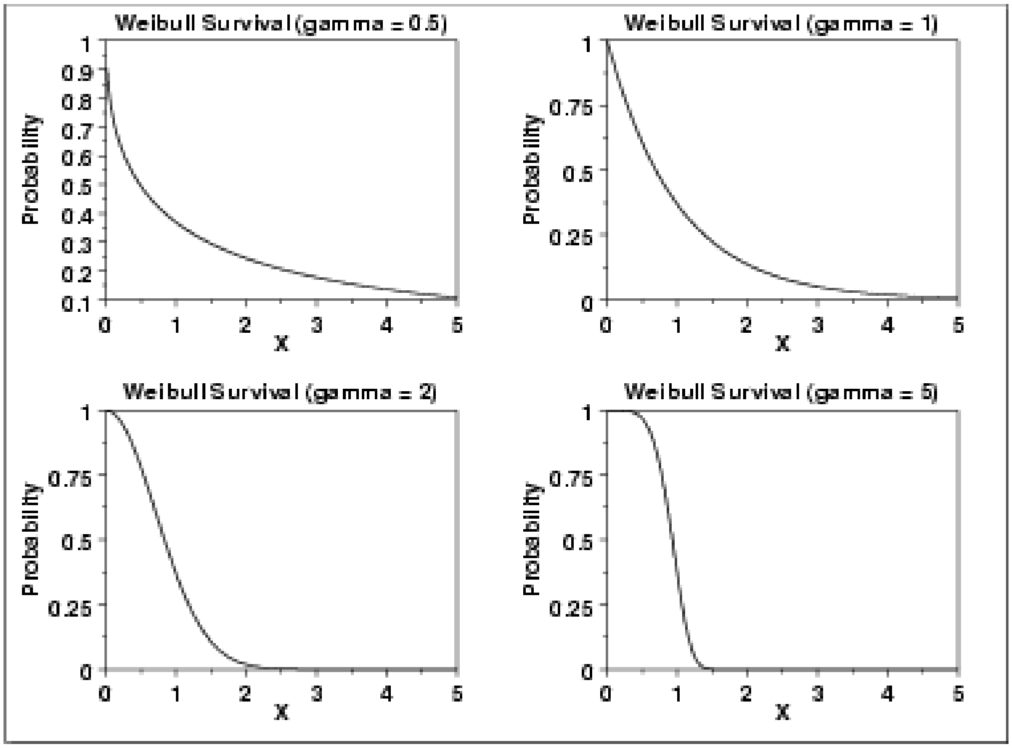


Survival Function

The formula for the survival function of the Weibull distribution is

$$S(x) = \exp -(x^{\gamma}) \qquad x \ge 0; \gamma > 0$$

The following is the plot of the Weibull survival function with the same values of γ as the pdf plots above.

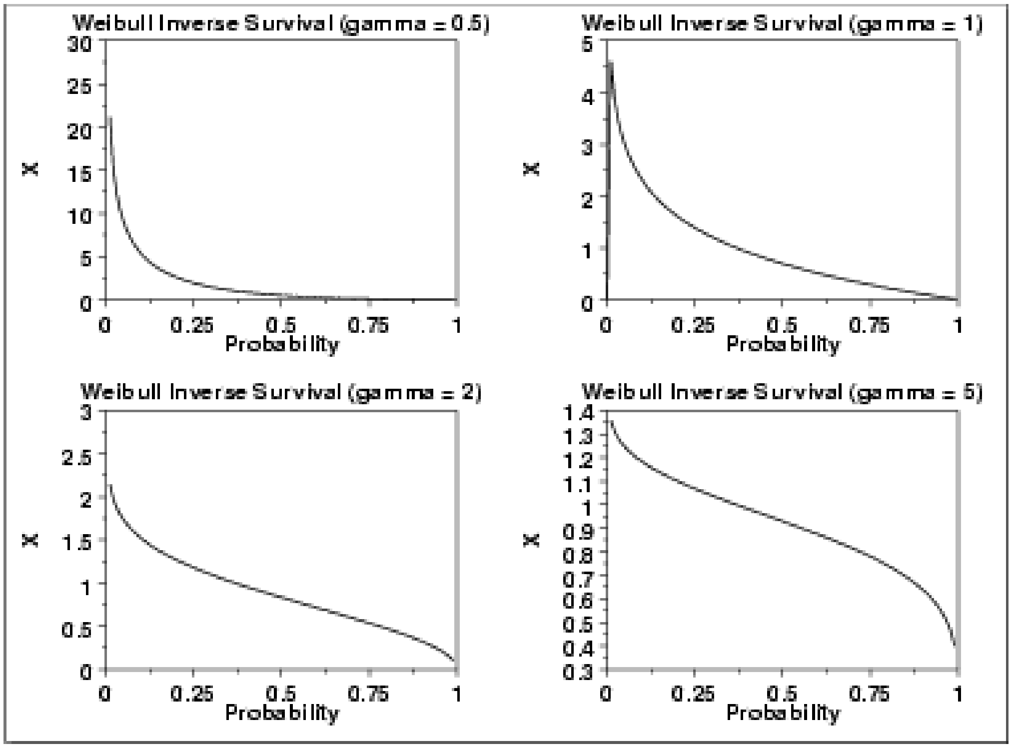


Inverse Survival Function

The formula for the inverse survival function of the Weibull distribution is

$$Z(p) = (-\ln(p))^{1/\gamma} \qquad 0 \le p < 1; \gamma > 0$$

The following is the plot of the Weibull inverse survival function with the same values of γ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	$\Gamma(\frac{\gamma+1}{\gamma})$ where Γ is the gamma function $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$
Median	$\ln(2)^{1/\gamma}$
Mode	$(1 - \frac{1}{\gamma})^{1/\gamma} \quad \gamma > 1$ $0 \quad \gamma \le 1$
Range	Zero to positive infinity.
Standard Deviation	$\sqrt{\Gamma(\frac{\gamma+2}{\gamma}) - (\Gamma(\frac{\gamma+1}{\gamma}))^2}$
Coefficient of Variation	$\sqrt{\frac{\Gamma(\frac{\gamma+2}{\gamma})}{(\Gamma(\frac{\gamma+1}{\gamma}))^2} - 1}$

Parameter Estimation	Maximum likelihood estimation for the Weibull distribution is discussed in the Reliability chapter (Chapter 8). It is also discussed in Chapter 21 of Johnson, Kotz, and Balakrishnan.
Comments	The Weibull distribution is used extensively in reliability applications to model failure times.
Software	Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the Weibull distribution.

1.3.6.6.9. Lognormal Distribution

Probability Density Function

A variable X is lognormally distributed if Y = LN(X) is normally distributed with "LN" denoting the natural logarithm. The general formula for the probability density function of the lognormal distribution is

$$f(x) = \frac{e^{-((\ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \qquad x \ge \theta; m, \sigma > 0$$

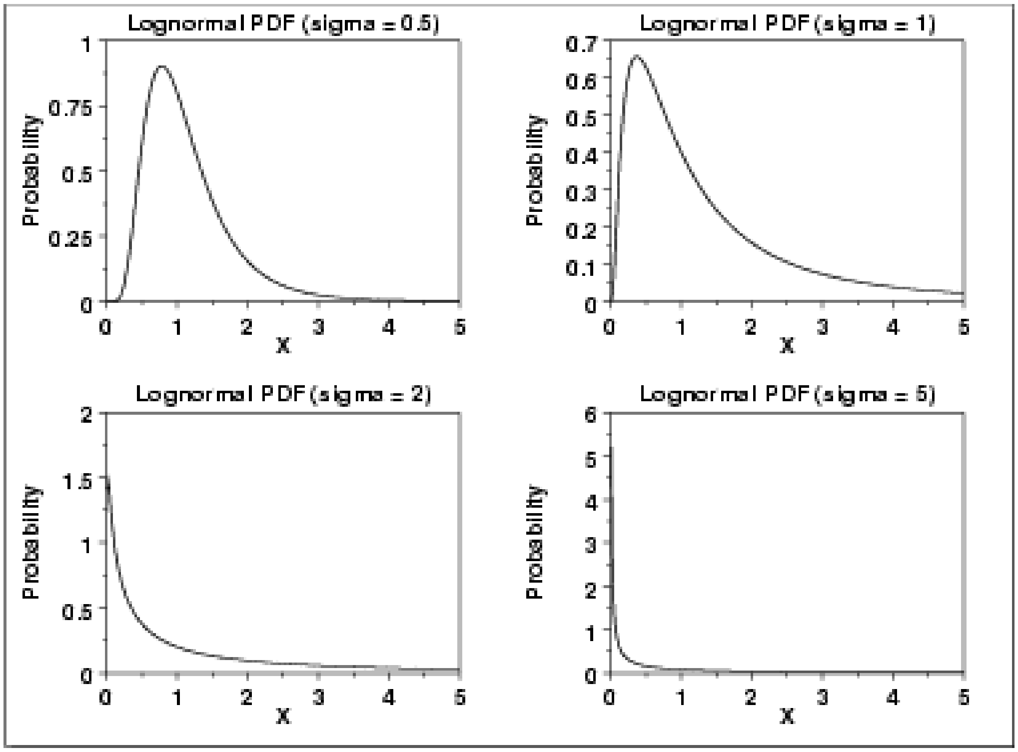
where σ is the shape parameter, θ is the location parameter and m is the scale parameter. The case where θ = 0 and m = 1 is called the **standard lognormal distribution**. The case where θ equals zero is called the 2-parameter lognormal distribution.

The equation for the standard lognormal distribution is

$$f(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}} \qquad x \ge 0; \sigma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the lognormal probability density function for four values of σ.



There are several common parameterizations of the lognormal distribution. The form given here is from Evans, Hastings, and Peacock.

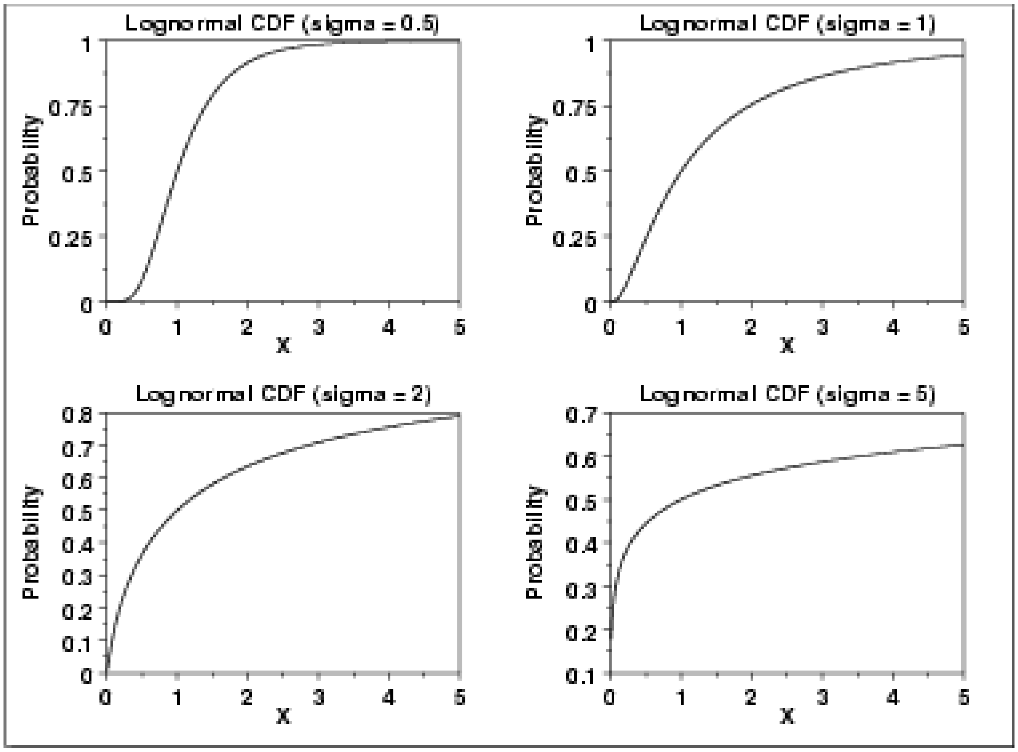
Cumulative Distribution Function

The formula for the cumulative distribution function of the lognormal distribution is

$$F(x) = \Phi\left(\frac{\ln(x)}{\sigma}\right) \qquad x \ge 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal cumulative distribution function with the same values of σ as the pdf plots above.



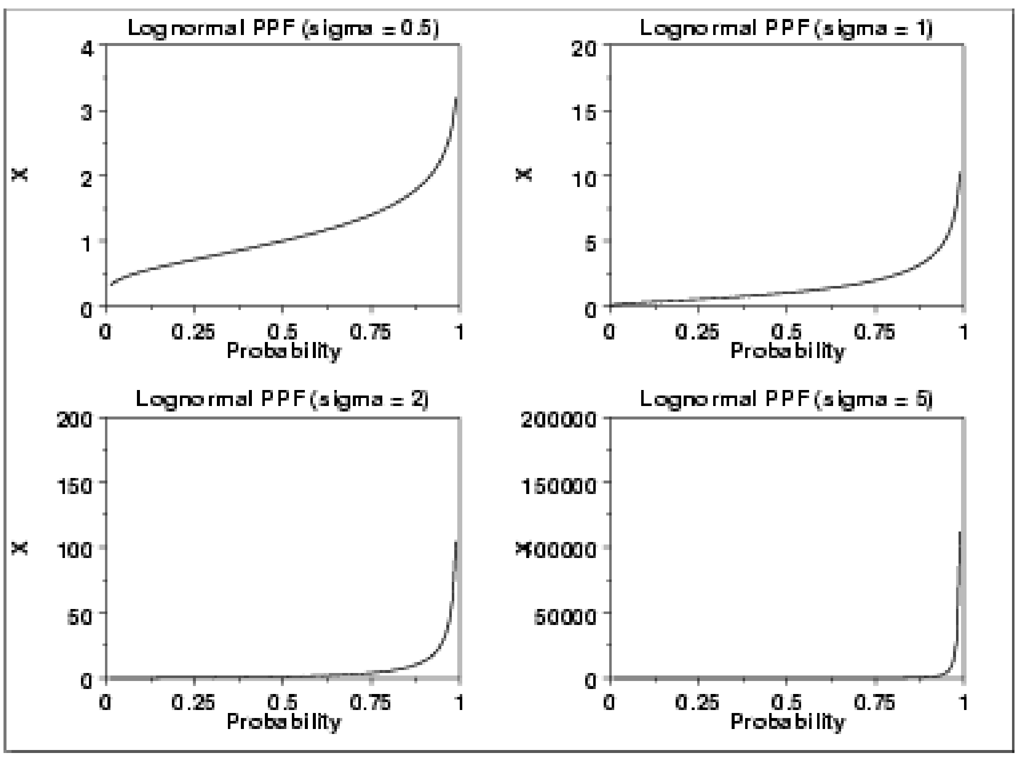
Percent Point Function

The formula for the percent point function of the lognormal distribution is

$$G(p) = \exp(\sigma \Phi^{-1}(p)) \qquad 0 \le p < 1; \sigma > 0$$

where Φ^{-1} is the percent point function of the normal distribution.

The following is the plot of the lognormal percent point function with the same values of σ as the pdf plots above.



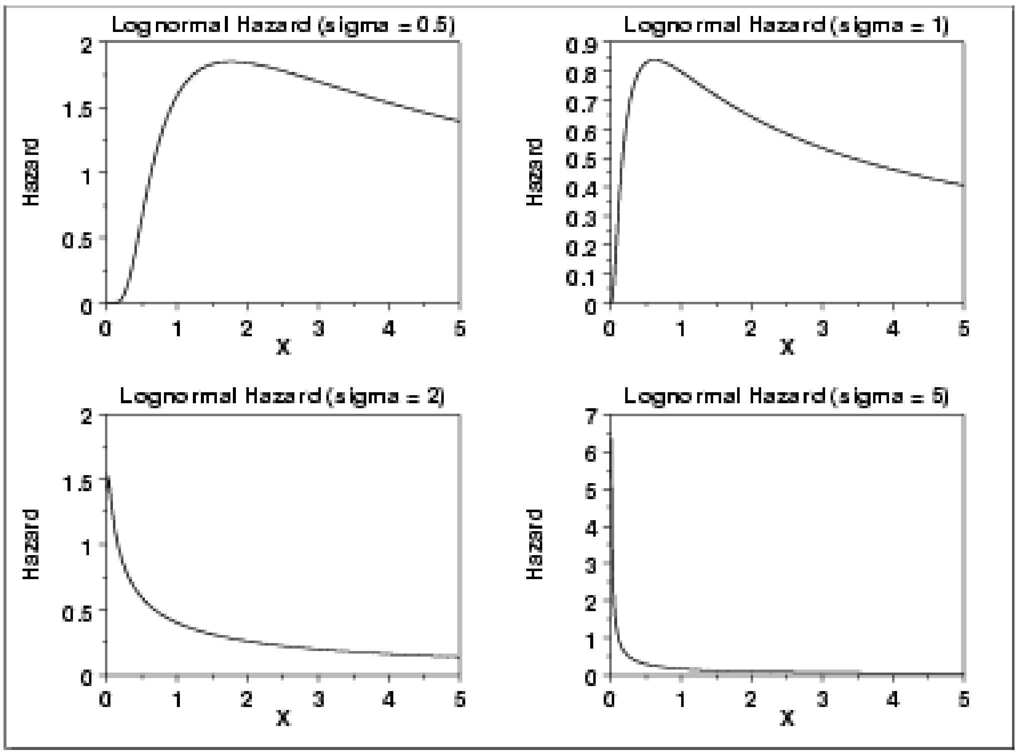
Hazard Function

The formula for the hazard function of the lognormal distribution is

$$h(x,\sigma) = \frac{\left(\frac{1}{x\sigma}\right)\phi\left(\frac{\ln x}{\sigma}\right)}{\Phi\left(\frac{-\ln x}{\sigma}\right)} \qquad x > 0; \sigma > 0$$

where ϕ is the probability density function of the normal distribution and Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal hazard function with the same values of σ as the pdf plots above.



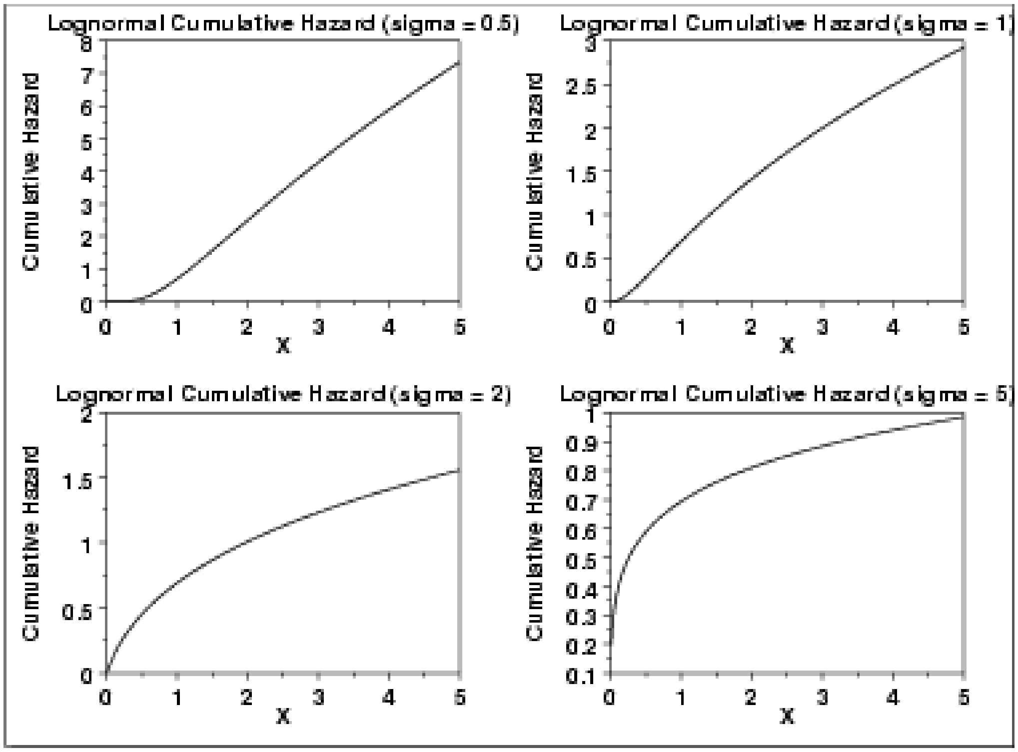
Cumulative Hazard Function

The formula for the cumulative hazard function of the lognormal distribution is

$$H(x) = \ln(1 - \Phi(\frac{\ln(x)}{\sigma})) \qquad x \ge 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal cumulative hazard function with the same values of σ as the pdf plots above.



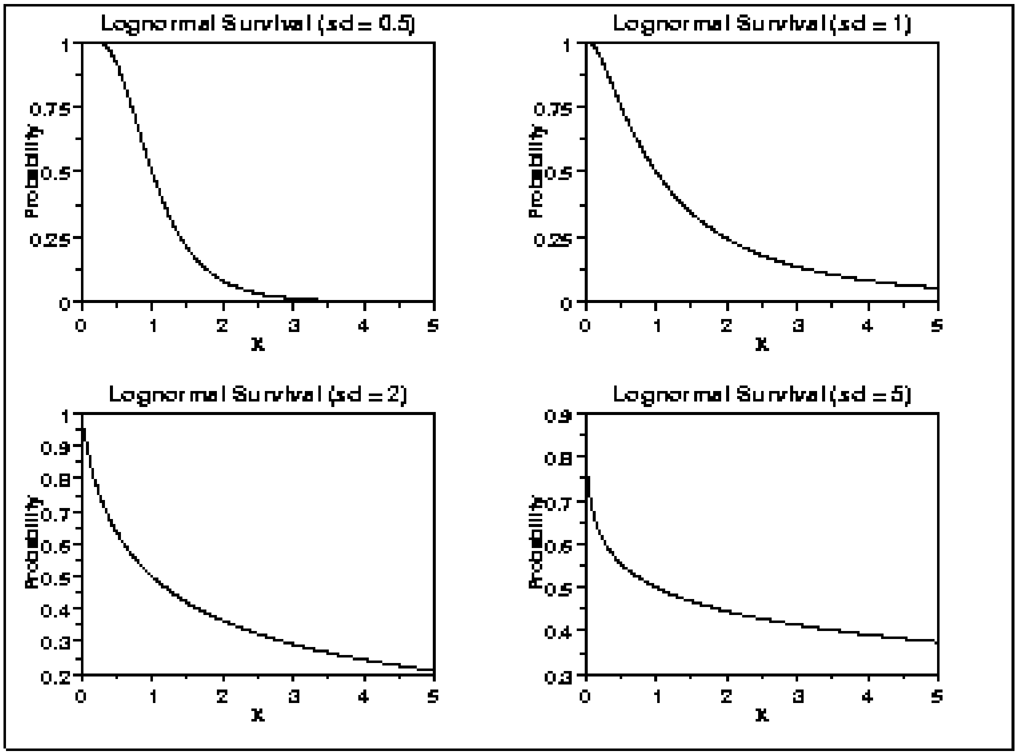
Survival Function

The formula for the survival function of the lognormal distribution is

$$S(x) = 1 - \Phi(\frac{\ln(x)}{\sigma}) \qquad x \ge 0; \sigma > 0$$

where Φ is the cumulative distribution function of the normal distribution.

The following is the plot of the lognormal survival function with the same values of σ as the pdf plots above.



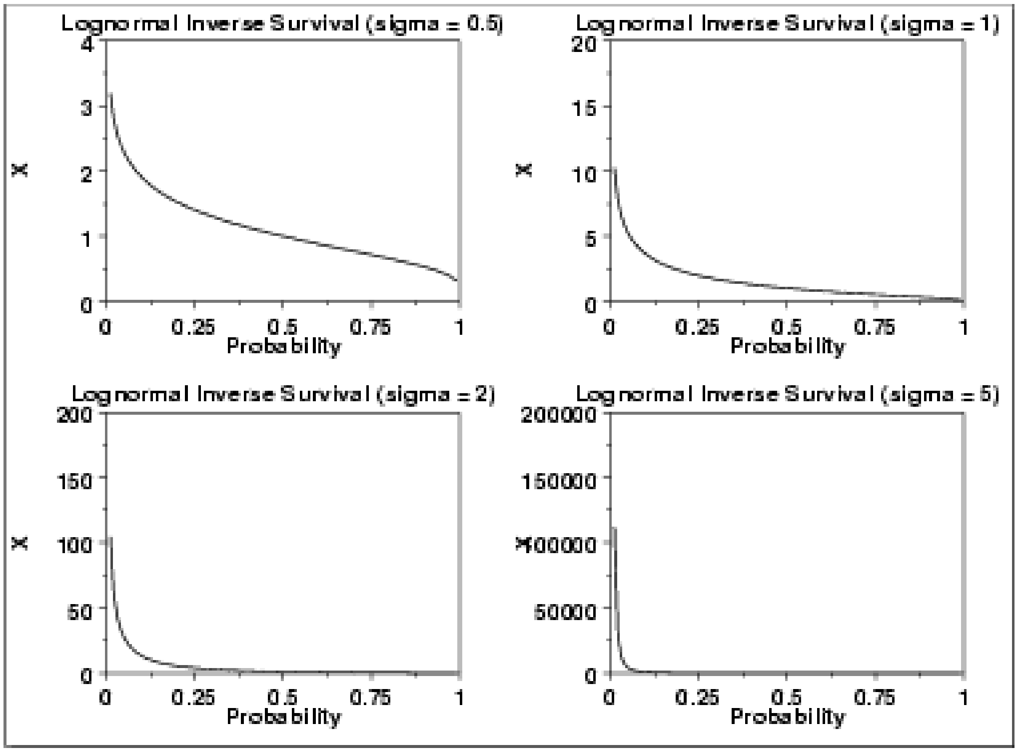
Inverse Survival Function

The formula for the inverse survival function of the lognormal distribution is

$$Z(p) = \exp(\sigma\Phi^{-1}(1-p)) \qquad 0 \le p < 1; \sigma > 0$$

where Φ^{-1} is the percent point function of the normal distribution.

The following is the plot of the lognormal inverse survival function with the same values of σ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	$e^{0.5\sigma^2}$
Median	Scale parameter ***m*** (= 1 if scale parameter not specified).
Mode	$\frac{1}{e^{\sigma^2}}$
Range	Zero to positive infinity
Standard Deviation	$\sqrt{e^{\sigma^2}(e^{\sigma^2}-1)}$
Skewness	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$
Kurtosis	$(e^{\sigma^2})^4 + 2(e^{\sigma^2})^3 + 3(e^{\sigma^2})^2 - 3$
Coefficient of Variation	$\sqrt{e^{\sigma^2}-1}$

Parameter Estimation

The maximum likelihood estimates for the scale parameter, ***m***, and the shape parameter, σ, are

$$\hat{m} = \exp \hat{\mu}$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(\ln(X_i) - \hat{\mu})^2}{N}}$$

where

$$\hat{\mu} = \frac{\sum_{i=1}^{N}\ln X_i}{N}$$

If the location parameter is known, it can be subtracted from the original data points before computing the maximum likelihood estimates of the shape and scale parameters.

Comments

The lognormal distribution is used extensively in reliability applications to model failure times. The lognormal and Weibull distributions are probably the most commonly used distributions in reliability applications.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the lognormal distribution.

1.3.6.6.10. Fatigue Life Distribution

Probability Density Function

The fatigue life distribution is also commonly known as the Birnbaum-Saunders distribution. There are several alternative formulations of the fatigue life distribution in the literature.

The general formula for the probability density function of the fatigue life distribution is

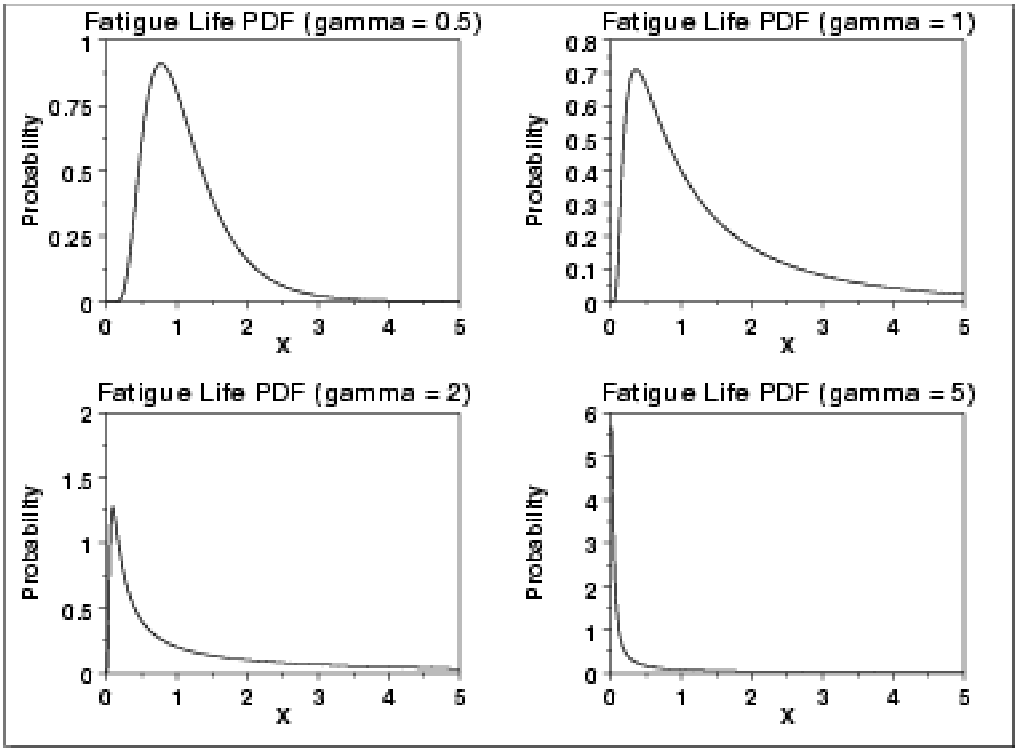
$$f(x) = \left(\frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)}\right)\phi\left(\frac{\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}}{\gamma}\right) \qquad x > \mu; \gamma, \beta > 0$$

where γ is the shape parameter, μ is the location parameter, β is the scale parameter, ϕ is the probability density function of the standard normal distribution, and Φ is the cumulative distribution function of the standard normal distribution. The case where μ = 0 and β = 1 is called the **standard fatigue life distribution**. The equation for the standard fatigue life distribution reduces to

$$f(x) = \left(\frac{\sqrt{x} + \sqrt{\frac{1}{x}}}{2\gamma x}\right)\phi\left(\frac{\sqrt{x} - \sqrt{\frac{1}{x}}}{\gamma}\right) \qquad x > 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the fatigue life probability density function.

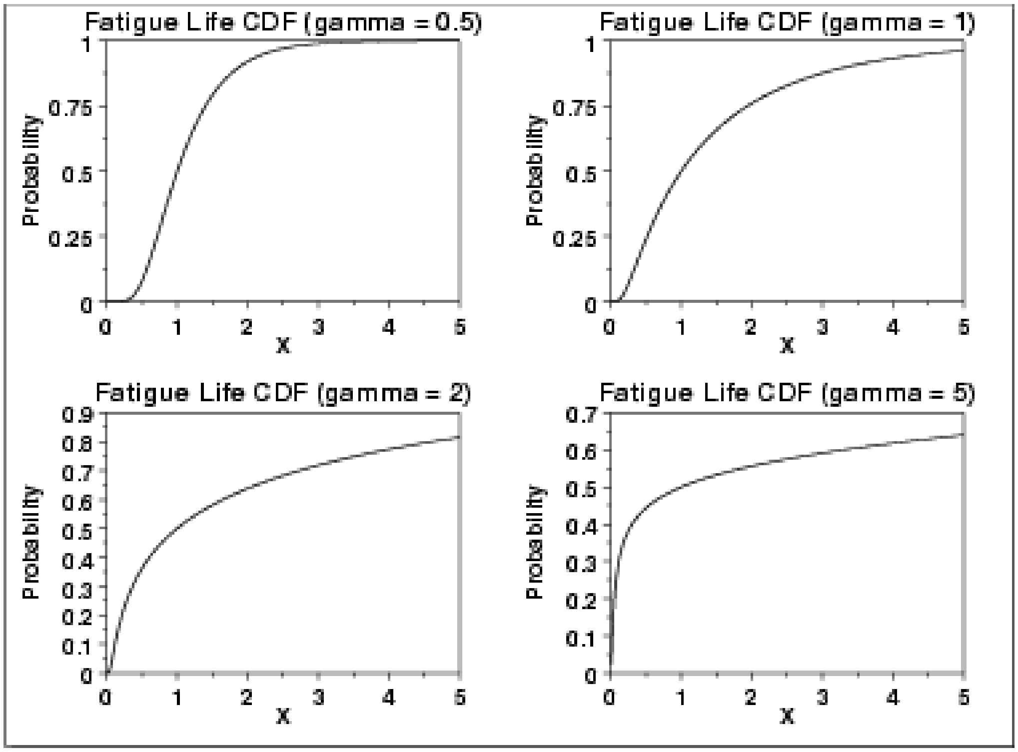


Cumulative Distribution Function

The formula for the cumulative distribution function of the fatigue life distribution is

$$F(x) = \Phi\left(\frac{\sqrt{x} - \sqrt{\frac{1}{x}}}{\gamma}\right) \qquad x > 0; \gamma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution. The following is the plot of the fatigue life cumulative distribution function with the same values of γ as the pdf plots above.

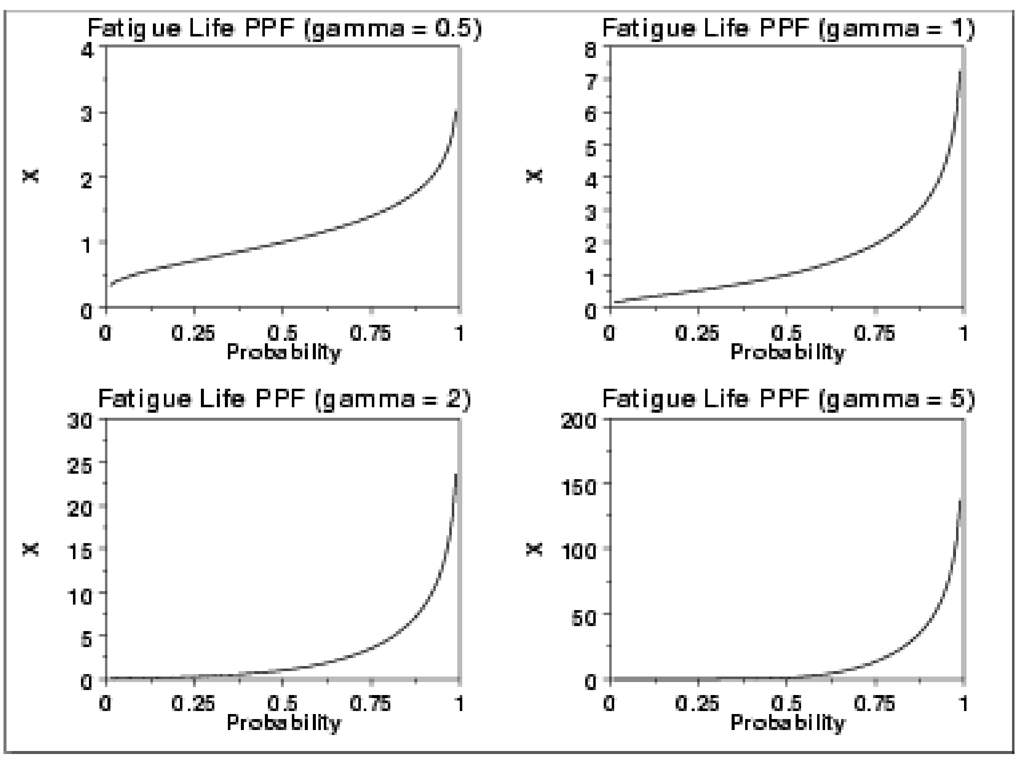


Percent Point Function

The formula for the percent point function of the fatigue life distribution is

$$G(p) = \frac{1}{4}\left[\gamma\Phi^{-1}(p) + \sqrt{4 + (\gamma\Phi^{-1}(p))^2}\right]^2$$

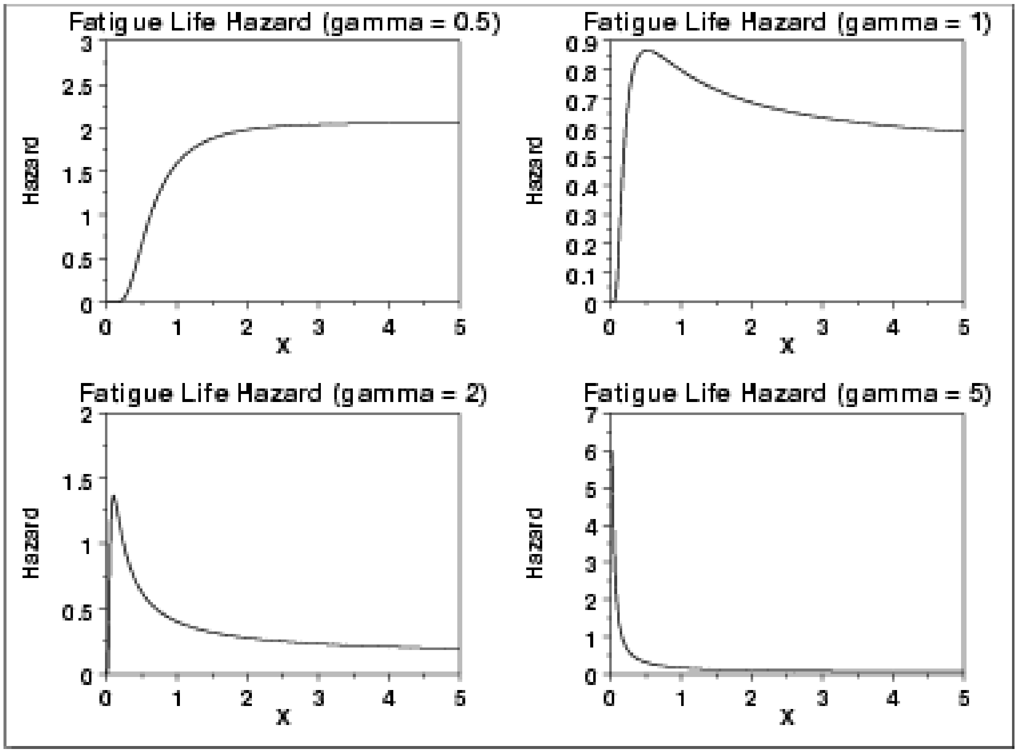
where Φ^{-1} is the percent point function of the standard normal distribution. The following is the plot of the fatigue life percent point function with the same values of γ as the pdf plots above.



Hazard Function

The fatigue life hazard function can be computed from the fatigue life probability density and cumulative distribution functions.

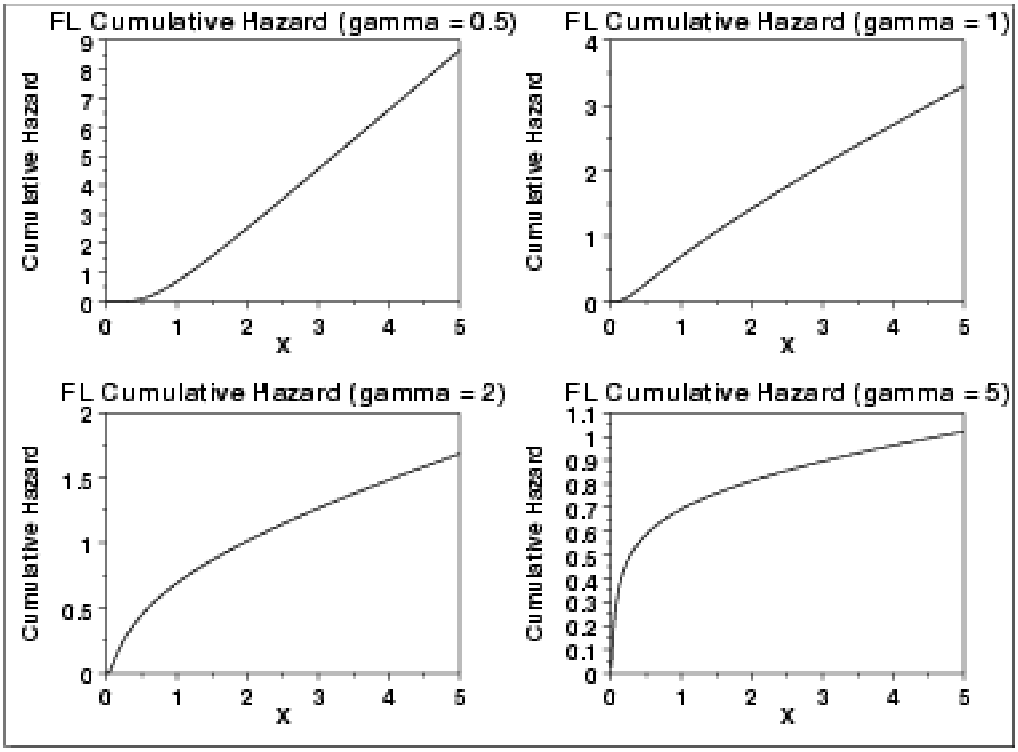
The following is the plot of the fatigue life hazard function with the same values of γ as the pdf plots above.



Cumulative Hazard Function

The fatigue life cumulative hazard function can be computed from the fatigue life cumulative distribution function.

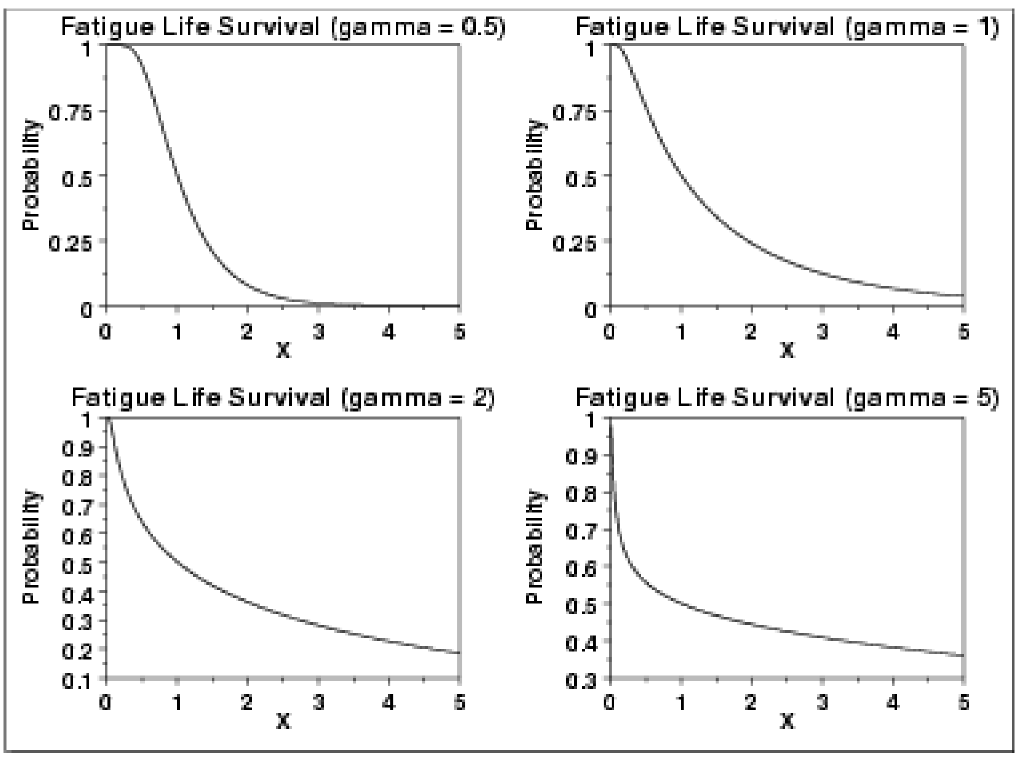
The following is the plot of the fatigue cumulative hazard function with the same values of γ as the pdf plots above.



Survival Function

The fatigue life [survival function](http://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm) can be computed from the fatigue life cumulative distribution function.

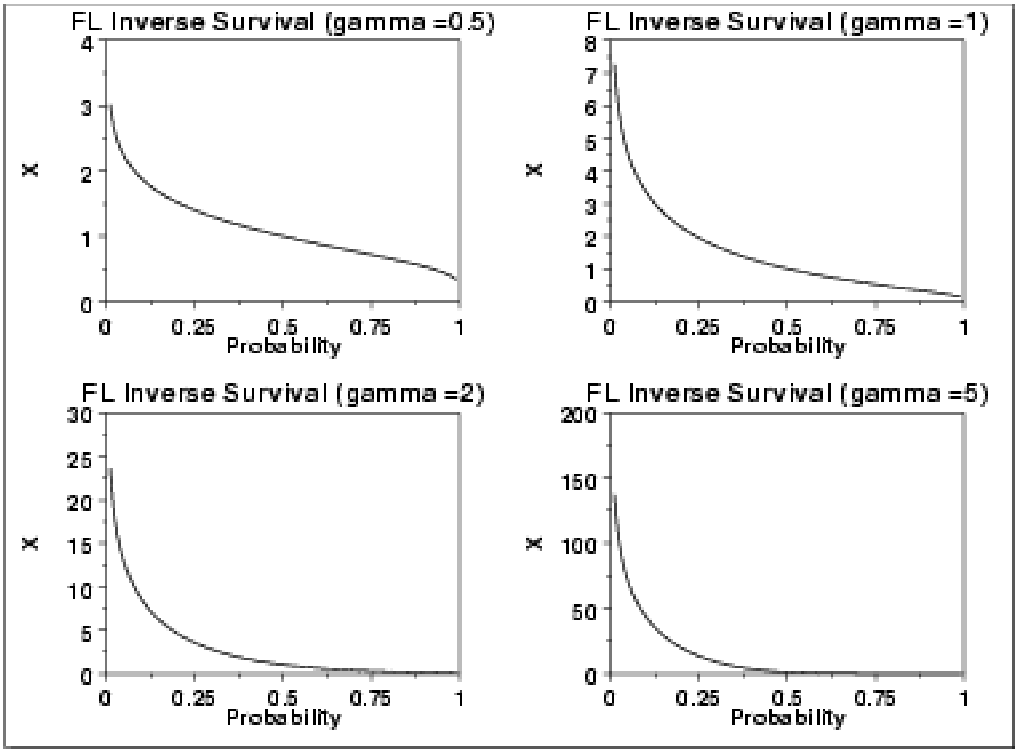
The following is the plot of the fatigue survival function with the same values of γ as the pdf plots above.



Inverse Survival Function

The fatigue life inverse survival function can be computed from the fatigue life percent point function.

The following is the plot of the gamma inverse survival function with the same values of γ as the pdf plots above.



Common Statistics

The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	$1+\frac{\gamma^2}{2}$
Range	Zero to positive infinity.
Standard Deviation	$\gamma\sqrt{1+\frac{5\gamma^2}{4}}$
Coefficient of Variation	$\frac{2+\gamma^2}{\gamma\sqrt{1+5\gamma^2}}$

Parameter Estimation

Maximum likelihood estimation for the fatigue life distribution is discussed in the Reliability chapter.

Comments

The fatigue life distribution is used extensively in reliability applications to model failure times.

Software

Some general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the fatigue life distribution. Support for this distribution is likely to be available for statistical programs that emphasize reliability applications.

1.3.6.6.11. Gamma Distribution

Probability Density Function

The general formula for the probability density function of the gamma distribution is

$$f(x) = \frac{\left(\frac{x-\mu}{\beta}\right)^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta\Gamma(\gamma)} \qquad x \ge \mu; \gamma, \beta > 0$$

where γ is the shape parameter, μ is the location parameter, β is the scale parameter, and Γ is the gamma function which has the formula

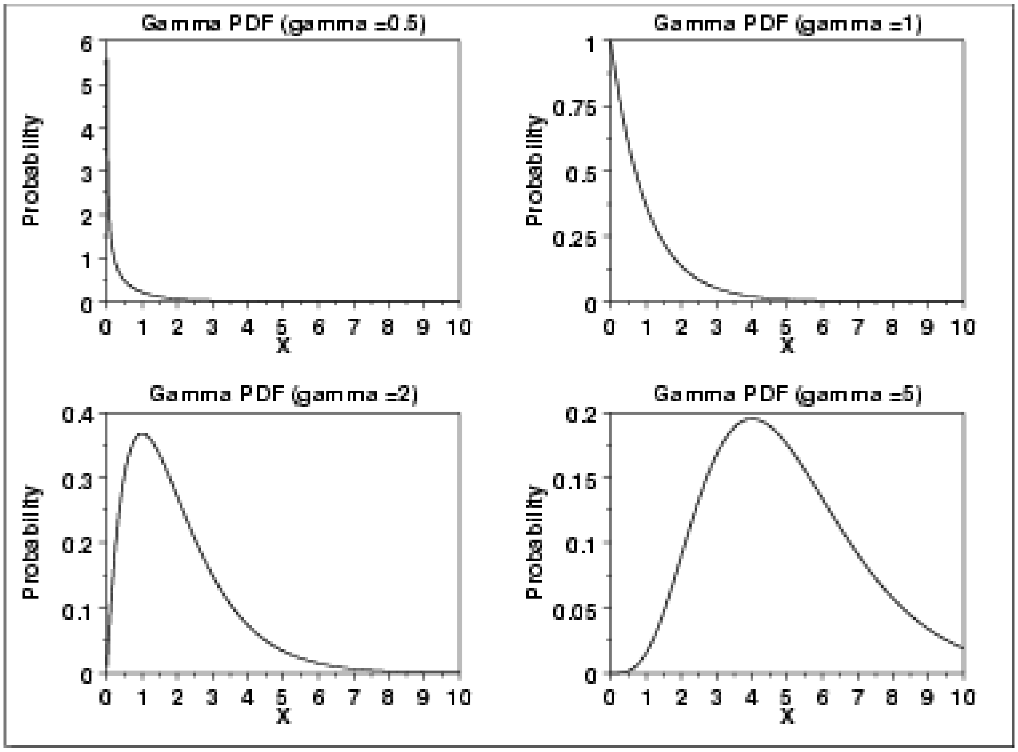
$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

The case where μ = 0 and β = 1 is called the **standard gamma distribution**. The equation for the standard gamma distribution reduces to

$$f(x) = \frac{x^{\gamma-1} e^{-x}}{\Gamma(\gamma)} \qquad x \ge 0; \gamma > 0$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the gamma probability density function.



Cumulative Distribution Function

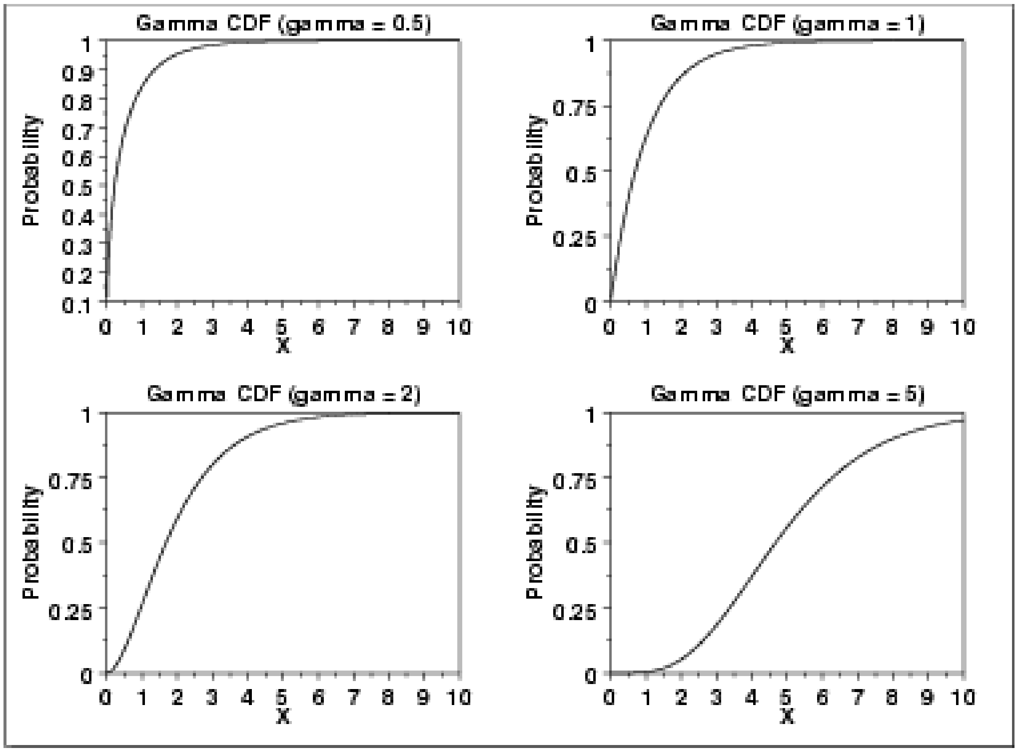
The formula for the cumulative distribution function of the gamma distribution is

$$F(x) = \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)} \qquad x \ge 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function. The incomplete gamma function has the formula

$$\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} dt$$

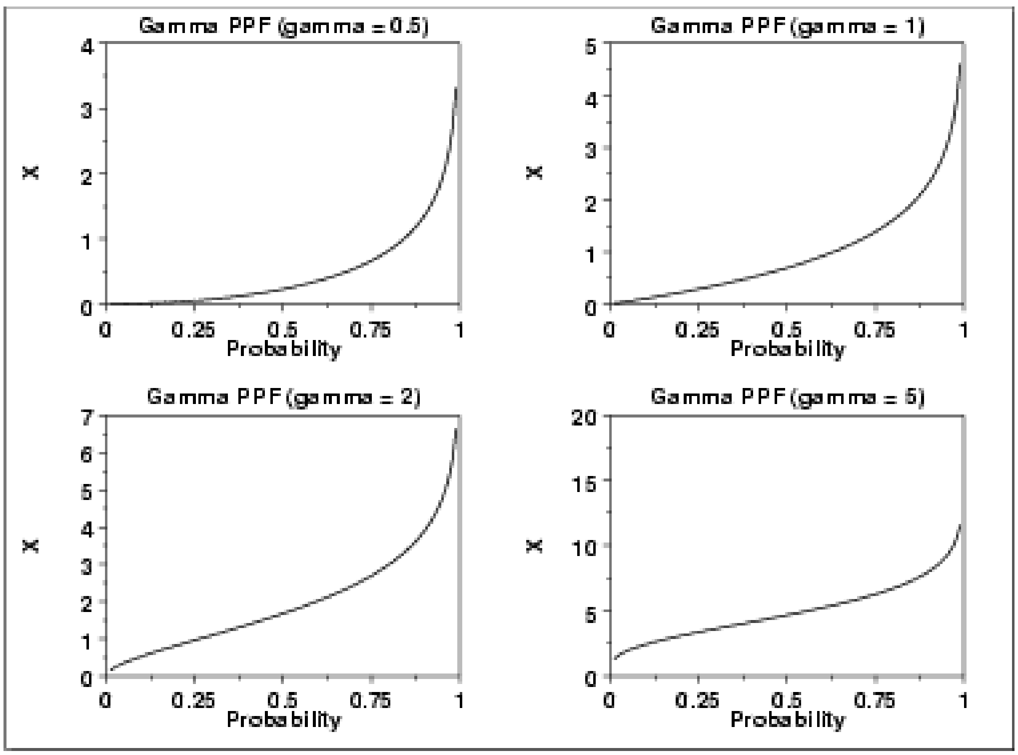
The following is the plot of the gamma cumulative distribution function with the same values of γ as the pdf plots above.



Percent Point Function

The formula for the [percent point function](http://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm#PPF) of the gamma distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the gamma percent point function with the same values of γ as the pdf plots above.

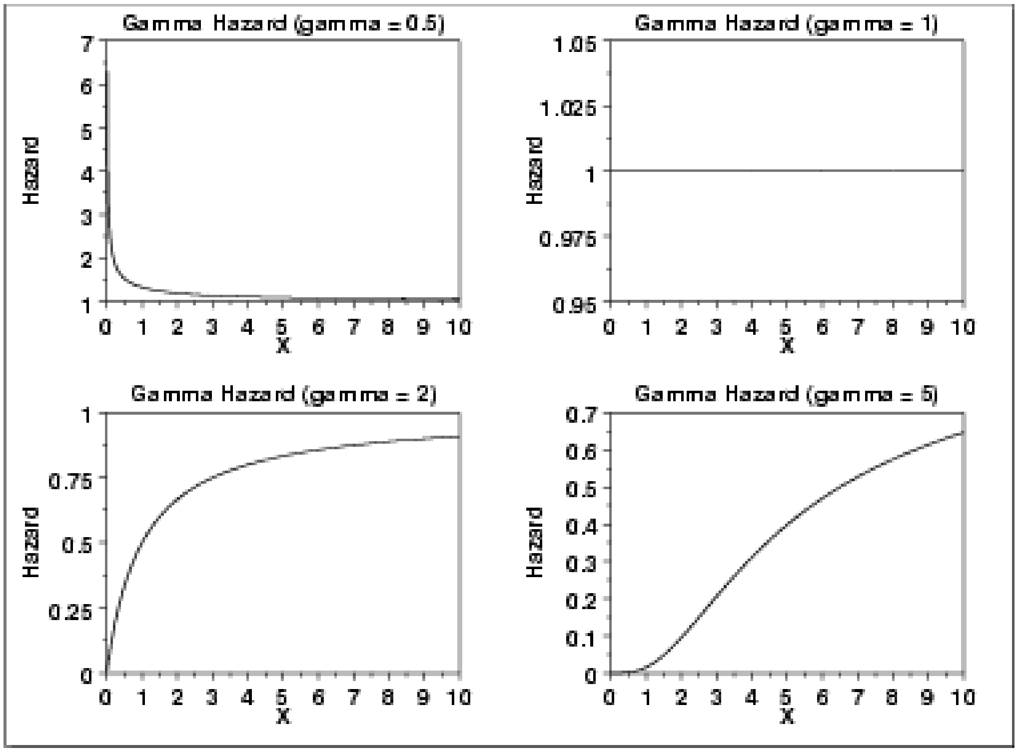


Hazard Function

The formula for the hazard function of the gamma distribution is

$$h(x) = \frac{x^{\gamma-1}e^{-x}}{\Gamma(\gamma) - \Gamma_x(\gamma)} \qquad x \ge 0; \gamma > 0$$

The following is the plot of the gamma hazard function with the same values of γ as the pdf plots above.



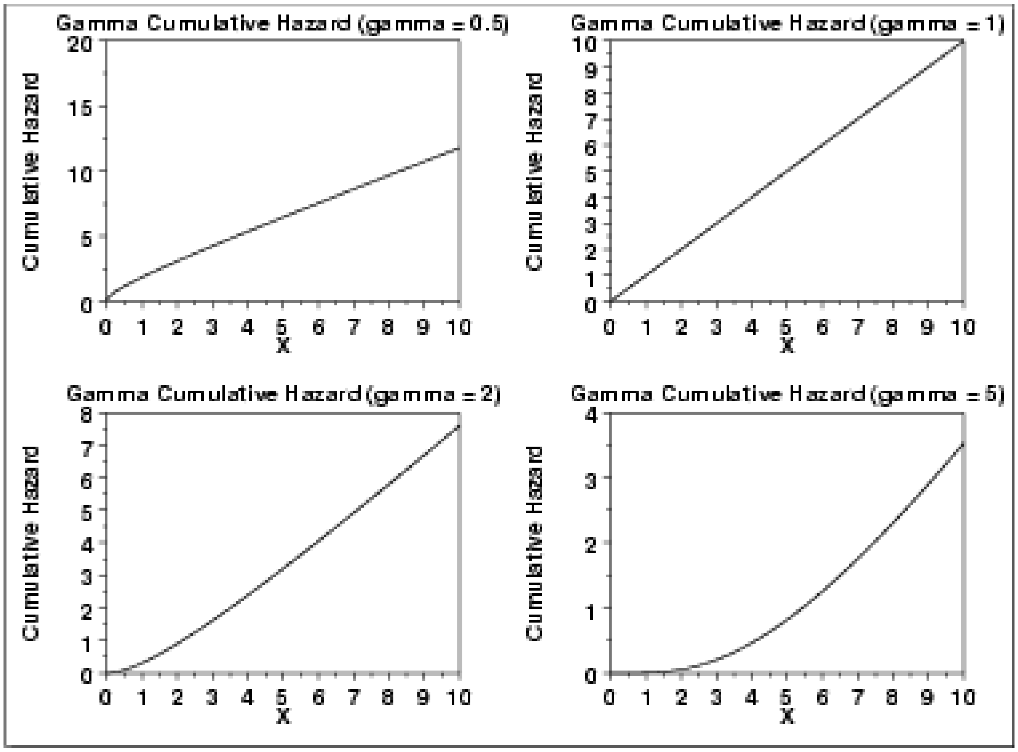
Cumulative Hazard Function

The formula for the cumulative hazard function of the gamma distribution is

$$H(x) = -\log\left(1 - \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)}\right) \qquad x \ge 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function defined above.

The following is the plot of the gamma cumulative hazard function with the same values of γ as the pdf plots above.



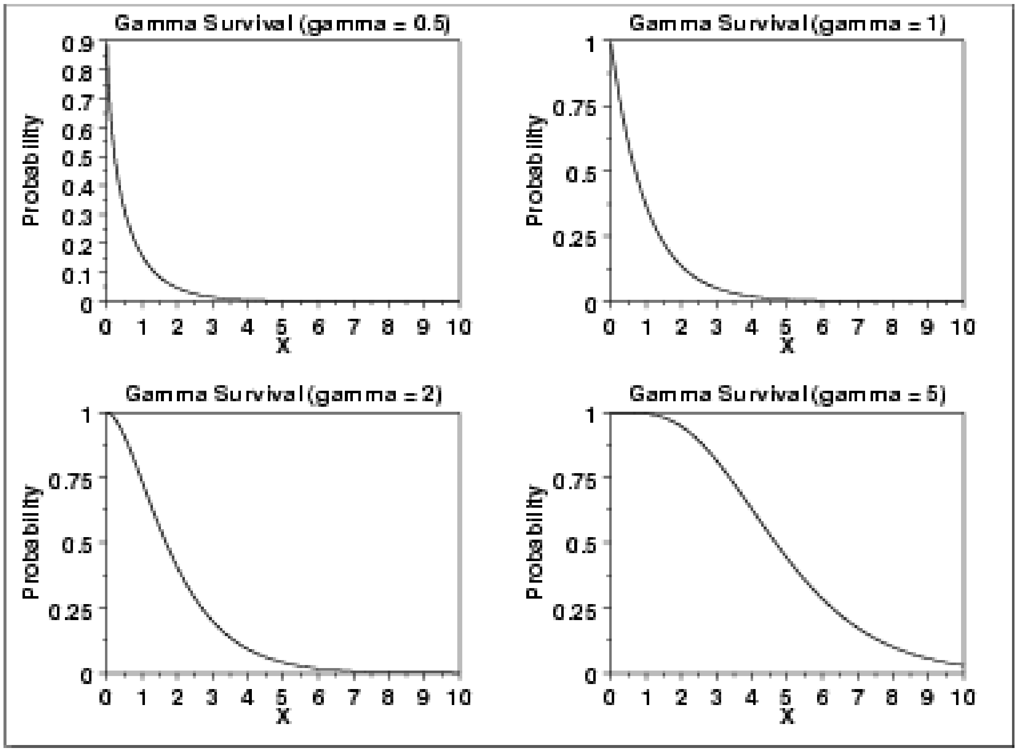
Survival Function

The formula for the survival function of the gamma distribution is

$$S(x) = 1 - \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)} \qquad x \ge 0; \gamma > 0$$

where Γ is the gamma function defined above and $\Gamma_x(a)$ is the incomplete gamma function defined above.

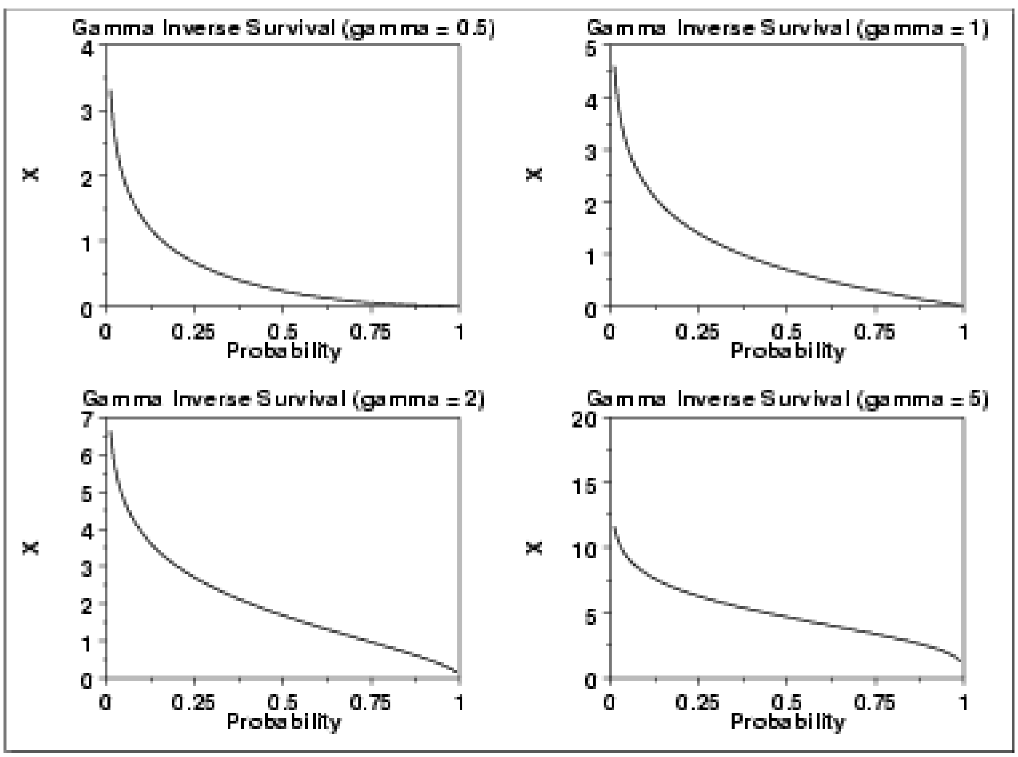
The following is the plot of the gamma survival function with the same values of γ as the pdf plots above.



Inverse Survival Function

The gamma inverse survival function does not exist in simple closed form. It is computed numberically.

The following is the plot of the gamma inverse survival function with the same values of γ as the pdf plots above.



Common Statistics	The formulas below are with the location parameter equal to zero and the scale parameter equal to one.

Mean	γ
Mode	$\gamma - 1 \qquad \gamma \geq 1$
Range	Zero to positive infinity.
Standard Deviation	$\sqrt{\gamma}$
Skewness	$\frac{2}{\sqrt{\gamma}}$
Kurtosis	$3 + \frac{6}{\gamma}$
Coefficient of Variation	$\frac{1}{\sqrt{\gamma}}$

Parameter Estimation

The method of moments estimators of the gamma distribution are

$$\hat{\gamma} = (\frac{\bar{x}}{s})^2$$

$$\hat{\beta} = \frac{s^2}{\bar{x}}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

The equations for the maximum likelihood estimation of the shape and scale parameters are given in Chapter 18 of Evans, Hastings, and Peacock and Chapter 17 of Johnson, Kotz, and Balakrishnan. These equations need to be solved numerically; this is typically accomplished by using statistical software packages.

Software

Some general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the gamma distribution.

1.3.6.6.12. Double Exponential Distribution

Probability Density Function

The general formula for the probability density function of the double exponential distribution is

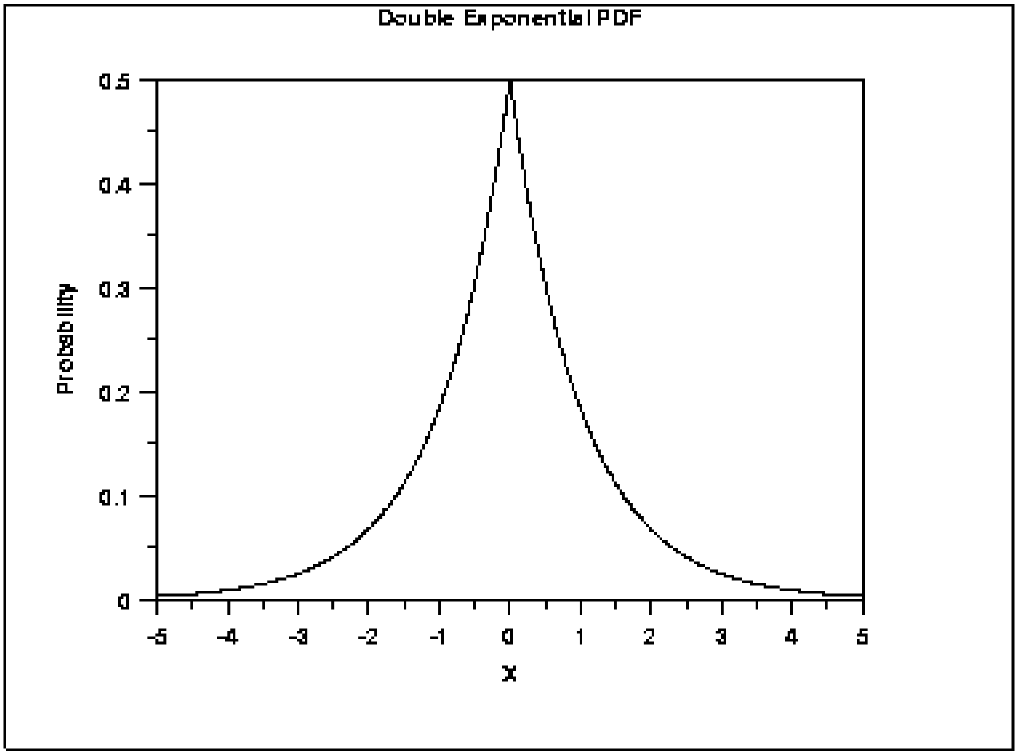
$$f(x) = \frac{e^{-\left|\frac{x-\mu}{\beta}\right|}}{2\beta}$$

where μ is the location parameter and β is the scale parameter. The case where μ = 0 and β = 1 is called the **standard double exponential distribution**. The equation for the standard double exponential distribution is

$$f(x) = \frac{e^{-|x|}}{2}$$

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the double exponential probability density function.

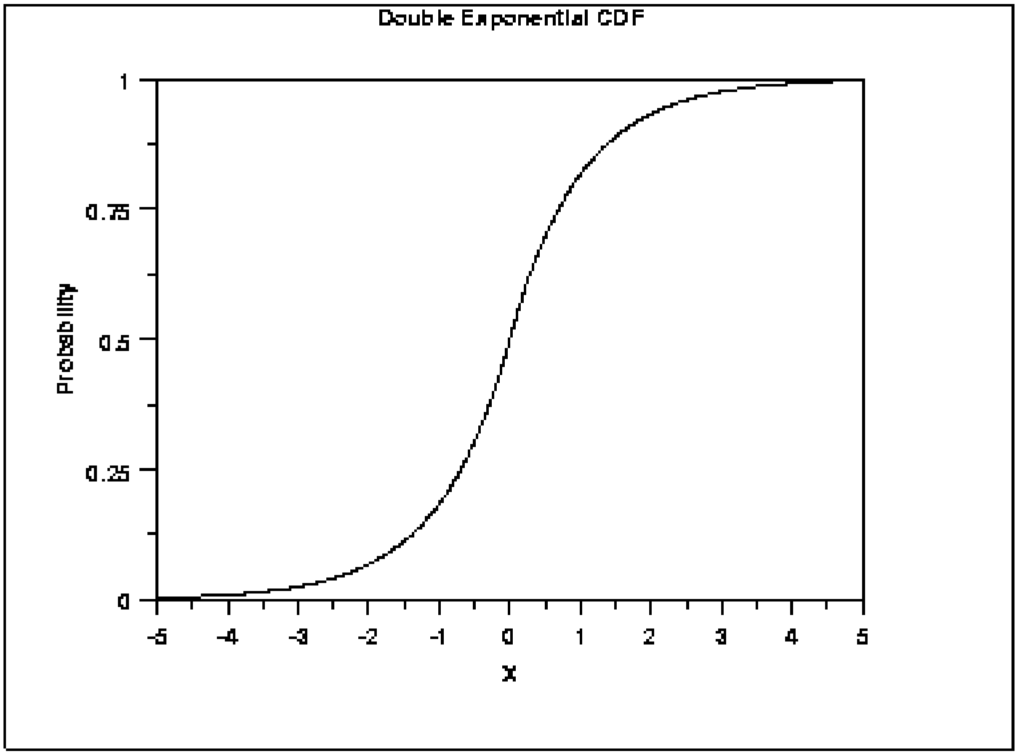


Cumulative Distribution Function

The formula for the cumulative distribution function of the double exponential distribution is

$$F(x) = \begin{cases} \frac{e^{x}}{2} & \text{for } x < 0 \\ 1 - \frac{e^{-x}}{2} & \text{for } x \ge 0 \end{cases}$$

The following is the plot of the double exponential cumulative distribution function.

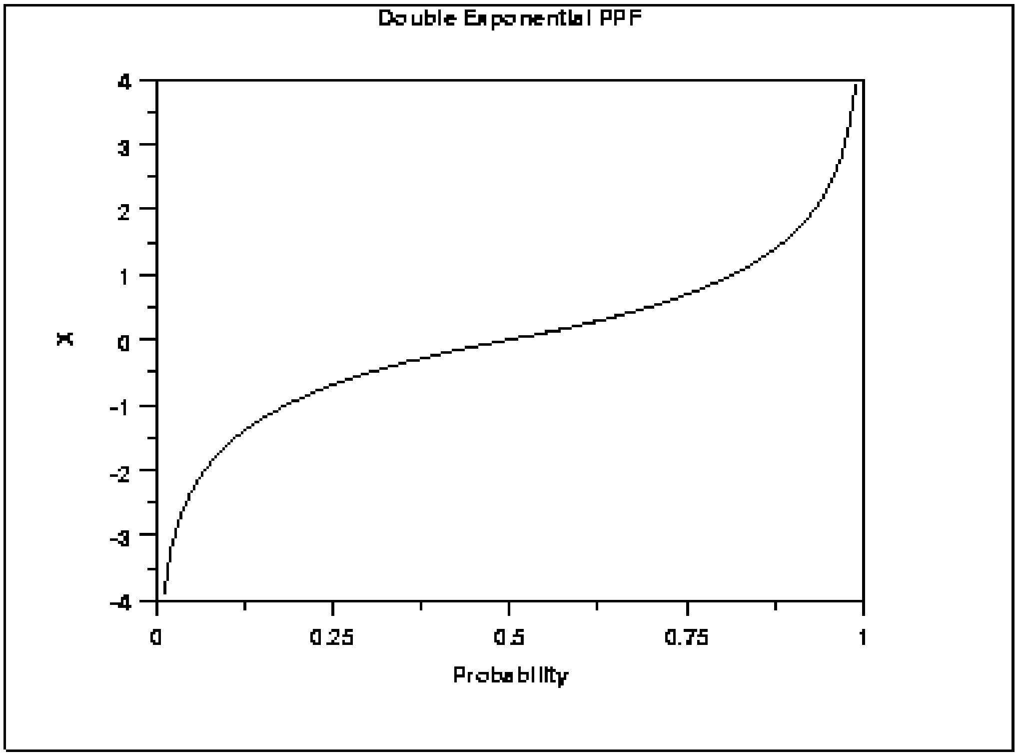


Percent Point Function

The formula for the percent point function of the double exponential distribution is

$$G(P) = \begin{cases} \log(2p) & \text{for } p \le 0.5 \\ -\log(2(1-p)) & \text{for } p > 0.5 \end{cases}$$

The following is the plot of the double exponential percent point function.

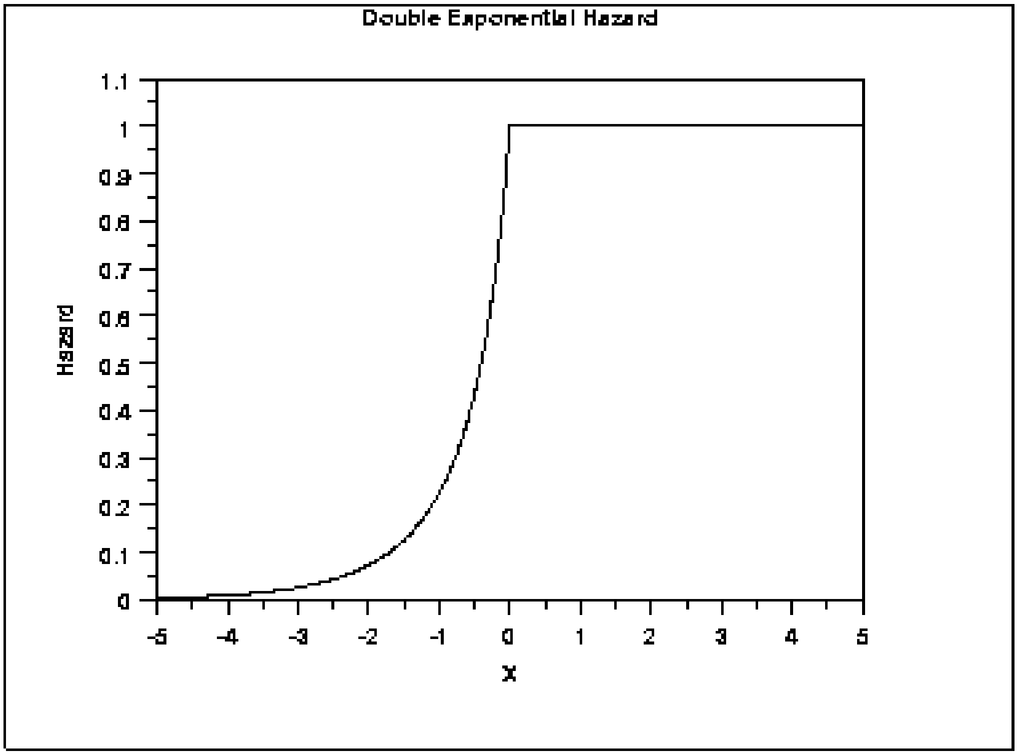


Hazard Function

The formula for the hazard function of the double exponential distribution is

$$h(x) = \begin{cases} \frac{e^{x}}{2-e^{x}} & \text{for } x < 0 \\ 1 & \text{for } x \ge 0 \end{cases}$$

The following is the plot of the double exponential hazard function.

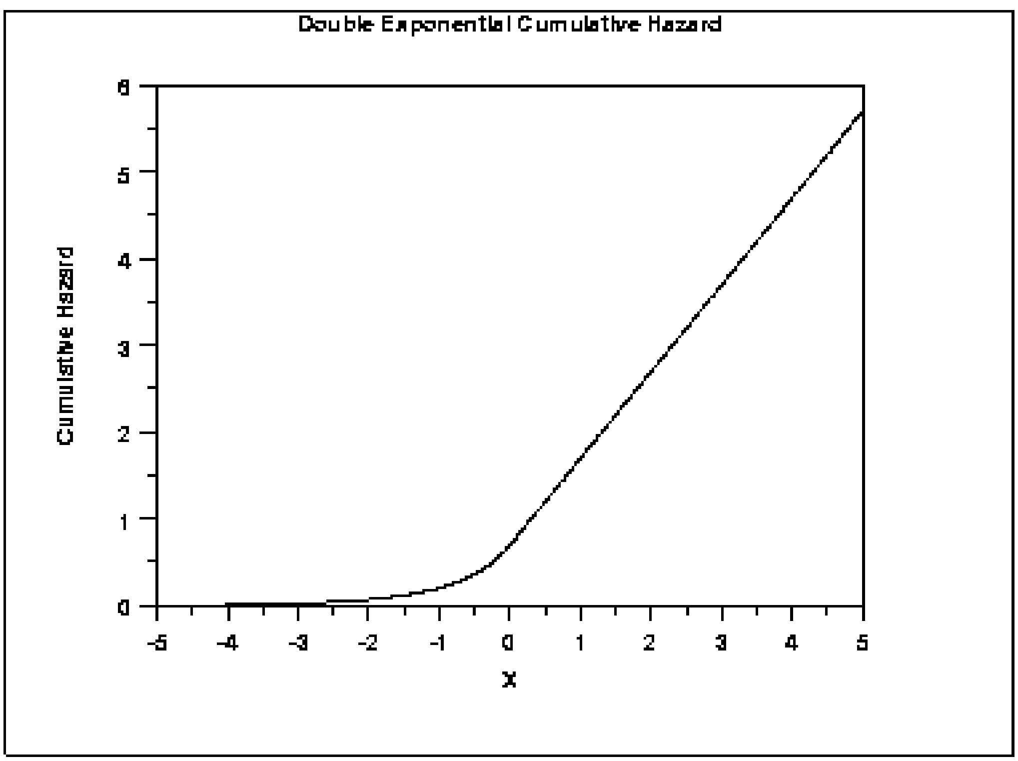


Cumulative Hazard Function

The formula for the cumulative hazard function of the double exponential distribution is

$$H(x) = \begin{array}{ll} -log(1 - \frac{e^{x}}{2}) & \text{for } x < 0 \\ x + \log(2) & \text{for } x \geq 0 \end{array}$$

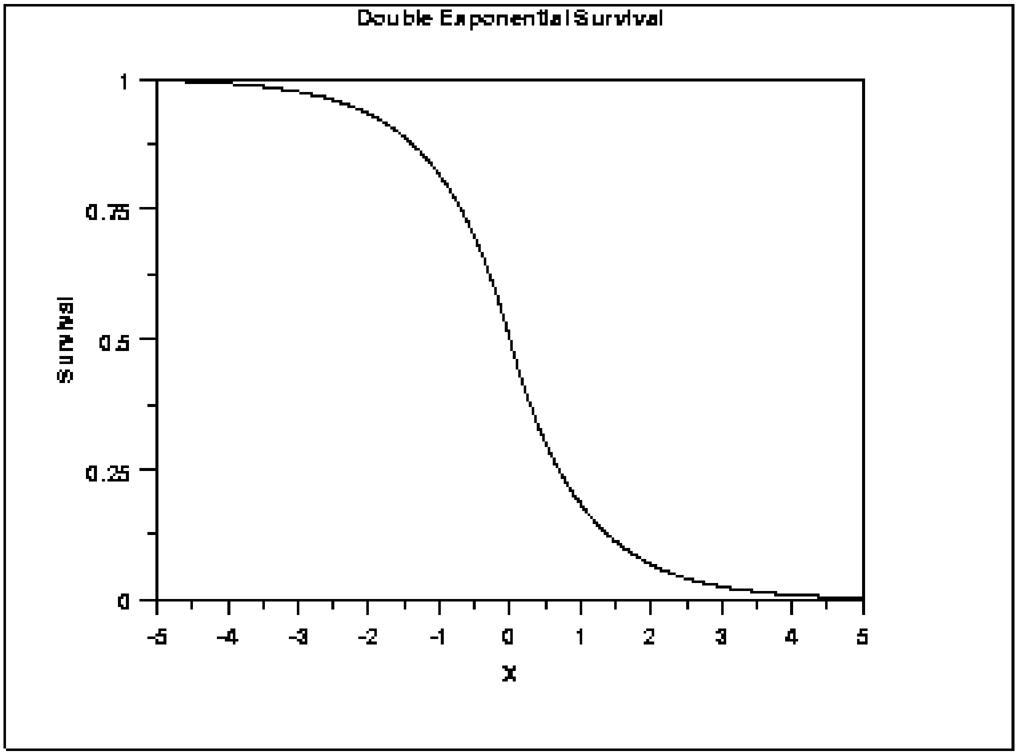
The following is the plot of the double exponential cumulative hazard function.



Survival Function

The double exponential survival function can be computed from the cumulative distribution function of the double exponential distribution.

The following is the plot of the double exponential survival function.

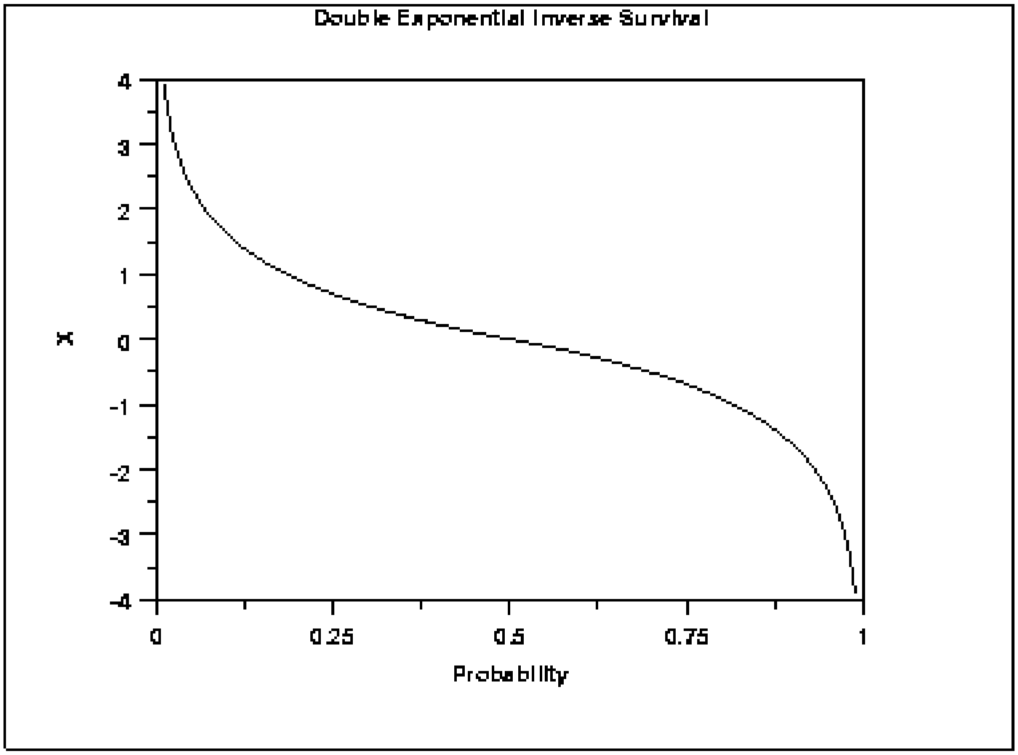


Inverse Survival Function

The formula for the inverse survival function of the double exponential distribution is

$$Z(P) = \begin{array}{ll} \log(2(1-p)) & \mbox{for } p \le 0.5 \\ -\log(2p) & \mbox{for } p > 0.5 \end{array}$$

The following is the plot of the double exponential inverse survival function.



Common Statistics	Mean	μ
	Median	μ
	Mode	μ
	Range	Negative infinity to positive infinity
	Standard Deviation	$\sqrt{2}\beta$
	Skewness	0
	Kurtosis	6
	Coefficient of Variation	$\sqrt{2}(\frac{\beta}{\mu})$

Parameter Estimation

The maximum likelihood estimators of the location and scale parameters of the double exponential distribution are

$$\hat{\mu} = \tilde{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} |X_i - \tilde{X}|}{N}$$

where $\tilde{X}$ is the sample median.

Software

Some general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the double exponential distribution.

1.3.6.6.13. Power Normal Distribution

Probability Density Function

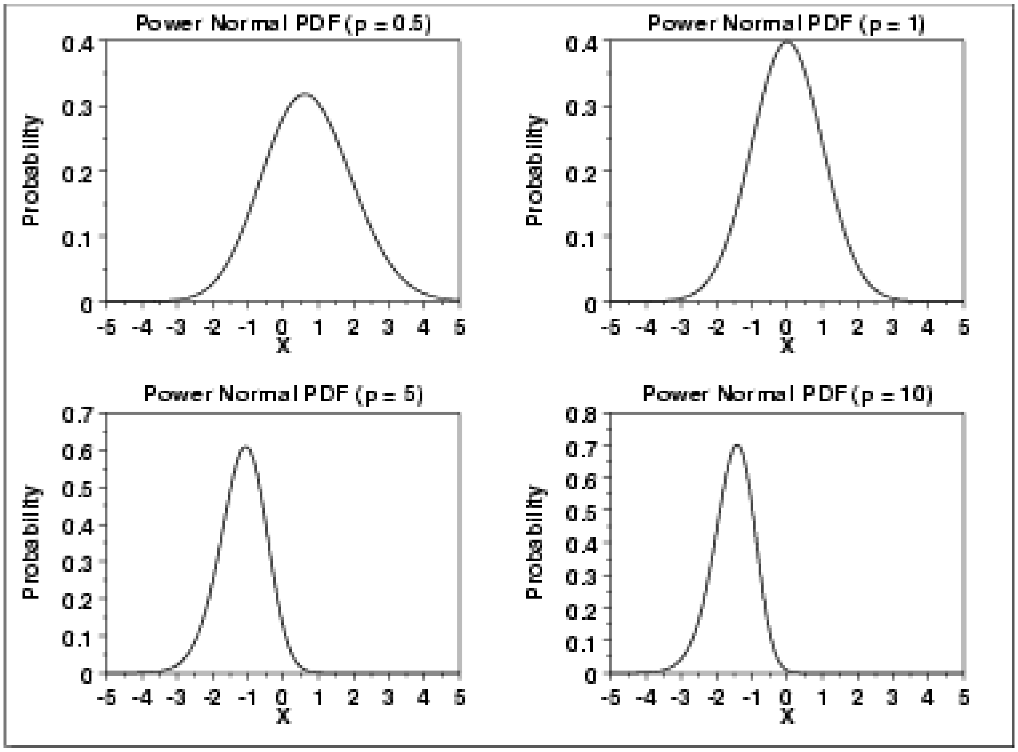
The formula for the probability density function of the standard form of the power normal distribution is

$$f(x,p) = p\phi(x)(\Phi(-x))^{p-1} \qquad x, p > 0$$

where p is the shape parameter (also referred to as the power parameter), Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

As with other probability distributions, the power normal distribution can be transformed with a location parameter, μ, and a scale parameter, σ. We omit the equation for the general form of the power normal distribution. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the power normal probability density function with four values of p.



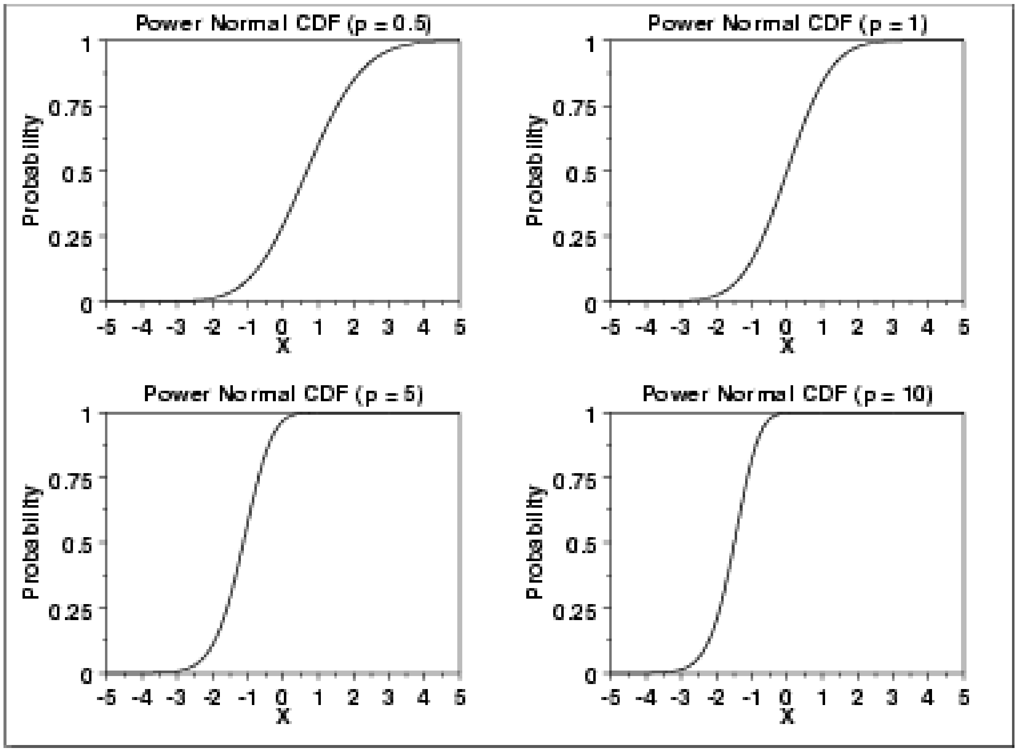
Cumulative Distribution Function

The formula for the cumulative distribution function of the power normal distribution is

$$F(x,p) = 1 - (\Phi(-x))^{p} \qquad x, p > 0$$

where Φ is the cumulative distribution function of the standard normal distribution.

The following is the plot of the power normal cumulative distribution function with the same values of ***p*** as the pdf plots above.



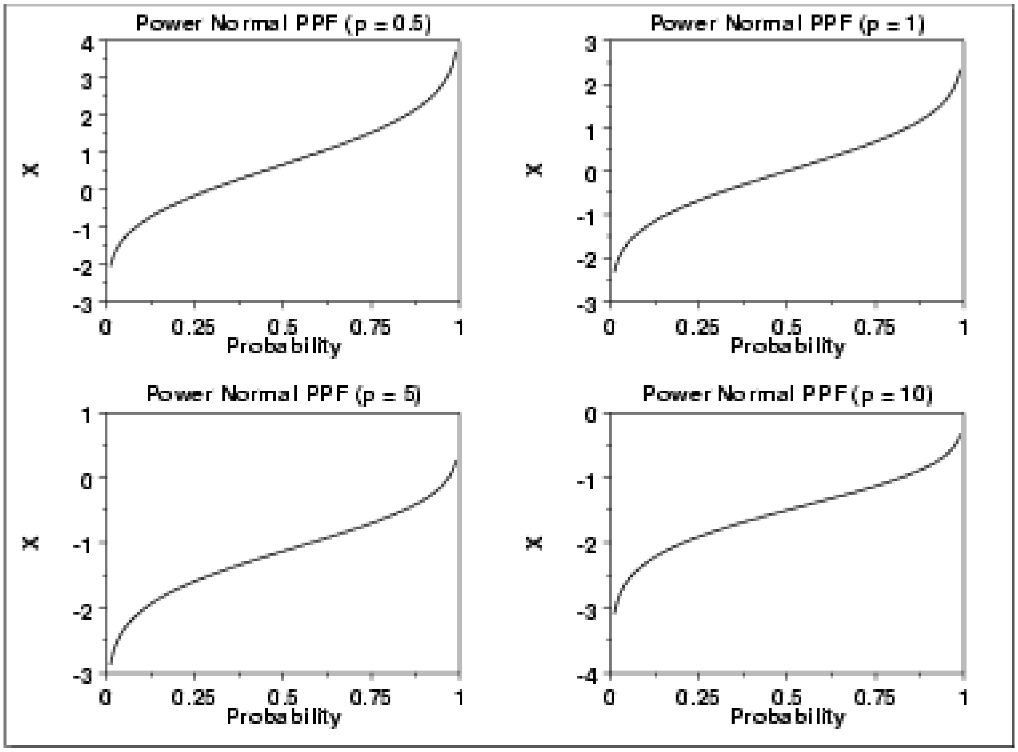
Percent Point Function

The formula for the percent point function of the power normal distribution is

$$G(f) = \Phi^{-1}(1 - (1 - f)^{1/p}) \qquad 0 < f < 1; p > 0$$

where Φ^{-1} is the percent point function of the standard normal distribution.

The following is the plot of the power normal percent point function with the same values of ***p*** as the pdf plots above.

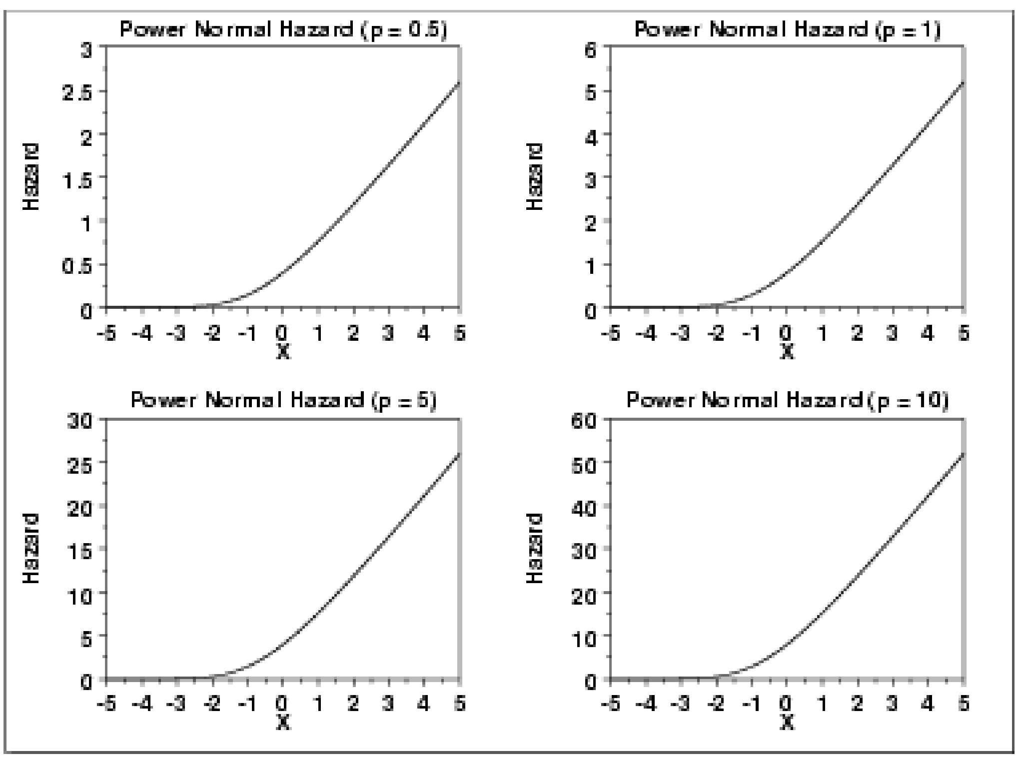


Hazard Function

The formula for the hazard function of the power normal distribution is

$$h(x,p) = \frac{p\phi(x)}{\Phi(-x)} \qquad x,p > 0$$

The following is the plot of the power normal hazard function with the same values of p as the pdf plots above.

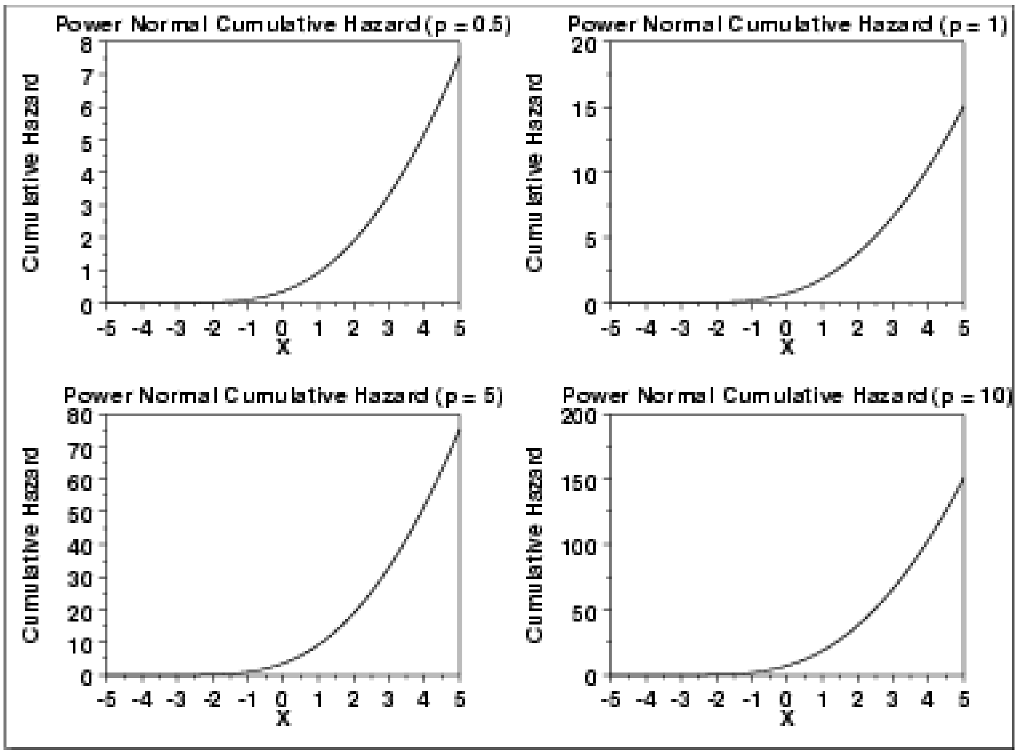


Cumulative Hazard Function

The formula for the cumulative hazard function of the power normal distribution is

$$H(x,p) = -\log\left((\Phi(-x))^p\right) \qquad x, p > 0$$

The following is the plot of the power normal cumulative hazard function with the same values of ***p*** as the pdf plots above.

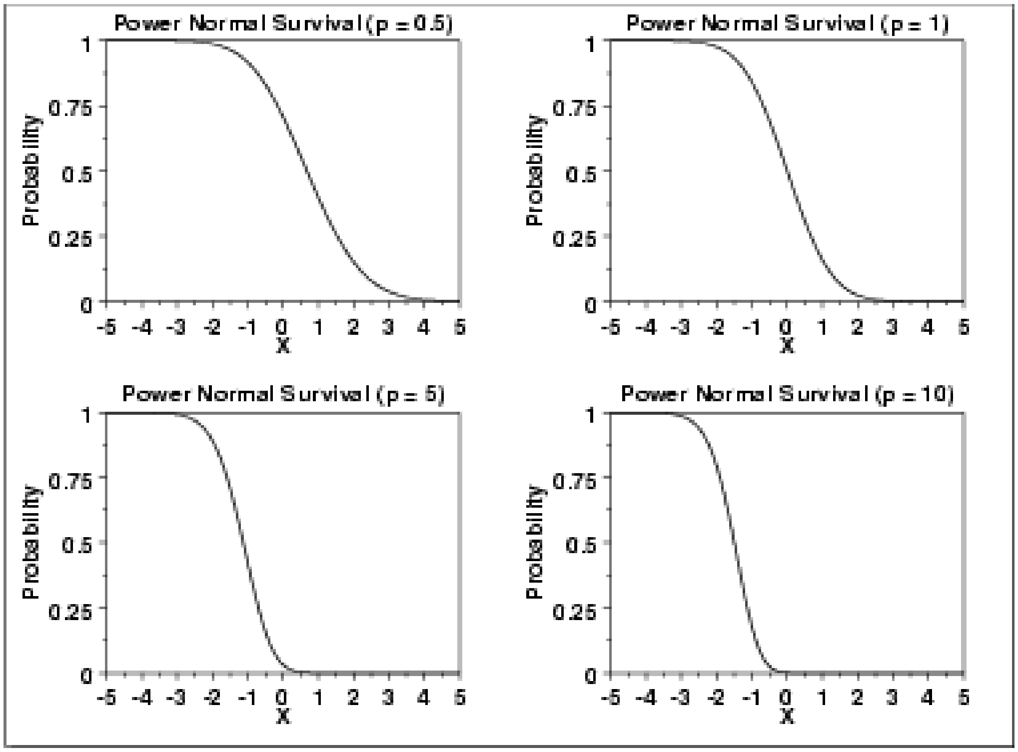


Survival Function

The formula for the survival function of the power normal distribution is

$$S(x,p) = (\Phi(-x))^p \qquad x, p > 0$$

The following is the plot of the power normal survival function with the same values of ***p*** as the pdf plots above.

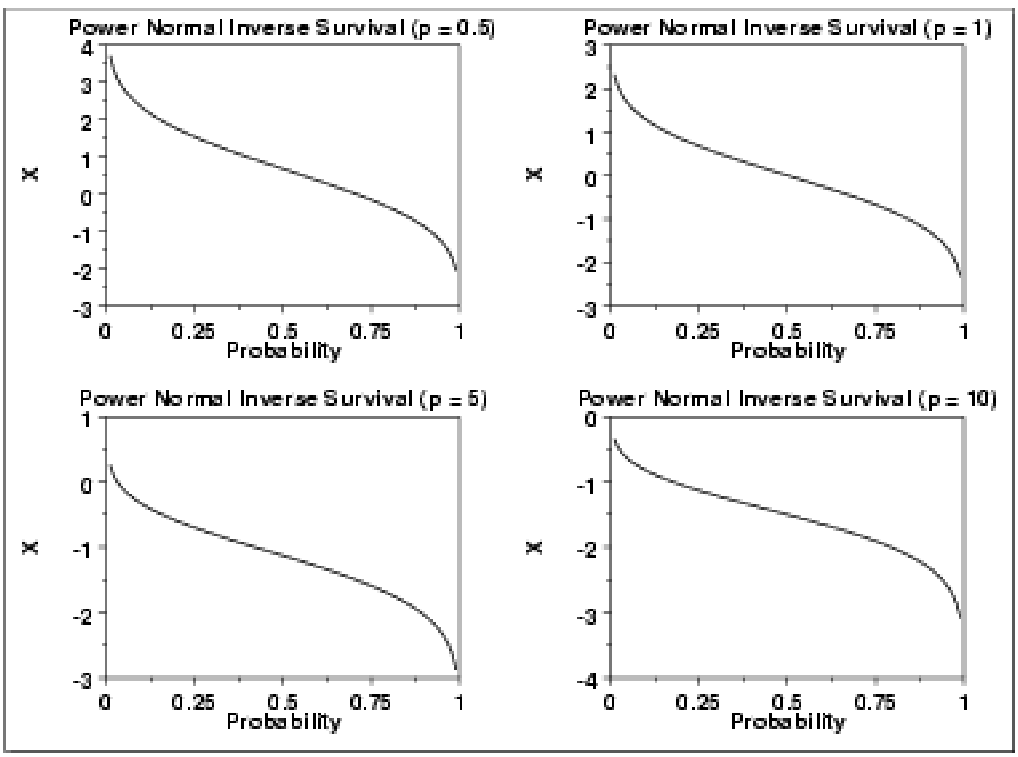


Inverse Survival Function

The formula for the inverse survival function of the power normal distribution is

$$Z(f) = \Phi^{-1}(1 - f^{1/p}) \qquad 0 < f < 1; p > 0$$

The following is the plot of the power normal inverse survival function with the same values of ***p*** as the pdf plots above.



Common Statistics	The statistics for the power normal distribution are complicated and require tables. Nelson discusses the mean, median, mode, and standard deviation of the power normal distribution and provides references to the appropriate tables.
Software	Most general purpose statistical software programs do not support the probability functions for the power normal distribution. Dataplot does support them.

1.3.6.6.14. Power Lognormal Distribution

Probability Density Function

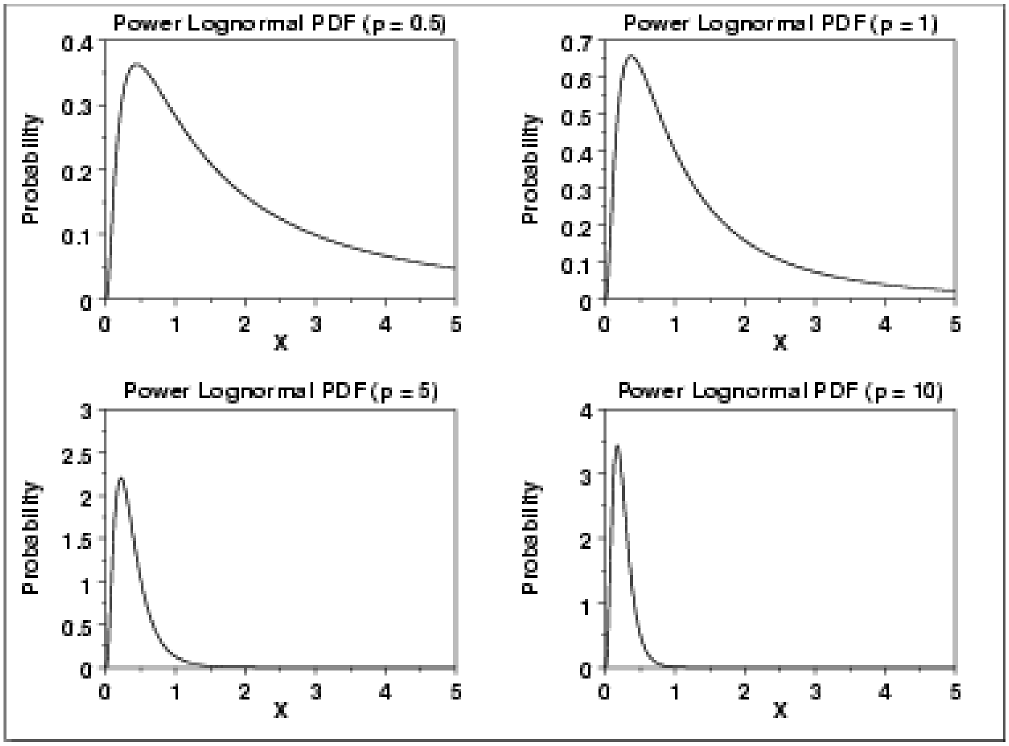
The formula for the probability density function of the standard form of the power lognormal distribution is

$$f(x,p,\sigma) = (\frac{p}{x\sigma})\phi(\frac{\log x}{\sigma})(\Phi(\frac{-\log x}{\sigma}))^{p-1} \qquad x,p,\sigma > 0$$

where p (also referred to as the power parameter) and σ are the shape parameters, Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

As with other probability distributions, the power lognormal distribution can be transformed with a location parameter, μ, and a scale parameter, B. We omit the equation for the general form of the power lognormal distribution. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the power lognormal probability density function with four values of p and σ set to 1.



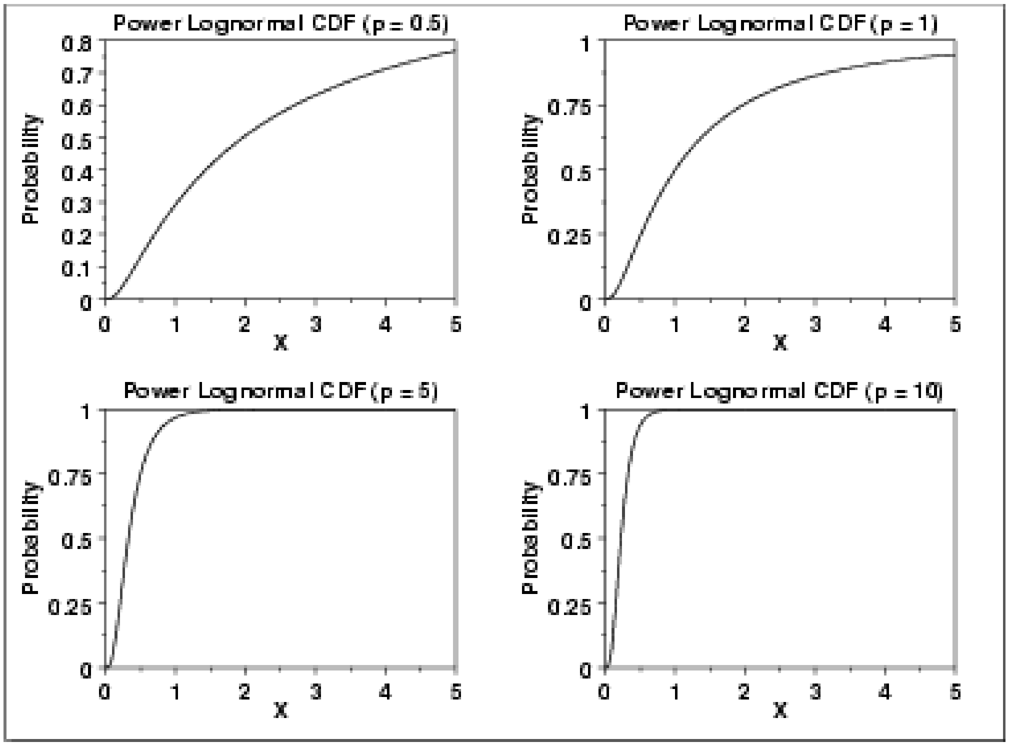
Cumulative Distribution Function

The formula for the cumulative distribution function of the power lognormal distribution is

$$F(x,p,\sigma) = 1 - (\Phi(\frac{-\log x}{\sigma}))^p \qquad x,p,\sigma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution.

The following is the plot of the power lognormal cumulative distribution function with the same values of ***p*** as the pdf plots above.



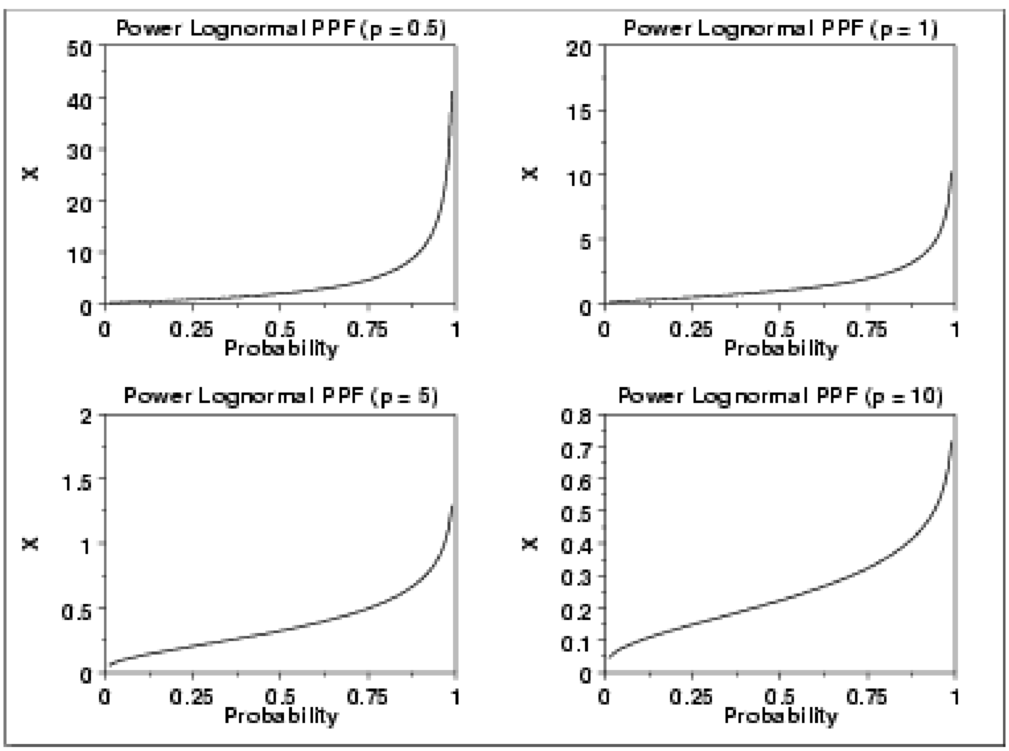
Percent Point Function

The formula for the percent point function of the power lognormal distribution is

$$G(f,p,\sigma) = \exp\left(\Phi^{-1}(1-(1-f)^{1/p})\sigma\right) \qquad 0 < p < 1; p, \sigma > 0$$

where Φ^{-1} is the percent point function of the standard normal distribution.

The following is the plot of the power lognormal percent point function with the same values of ***p*** as the pdf plots above.



Hazard Function

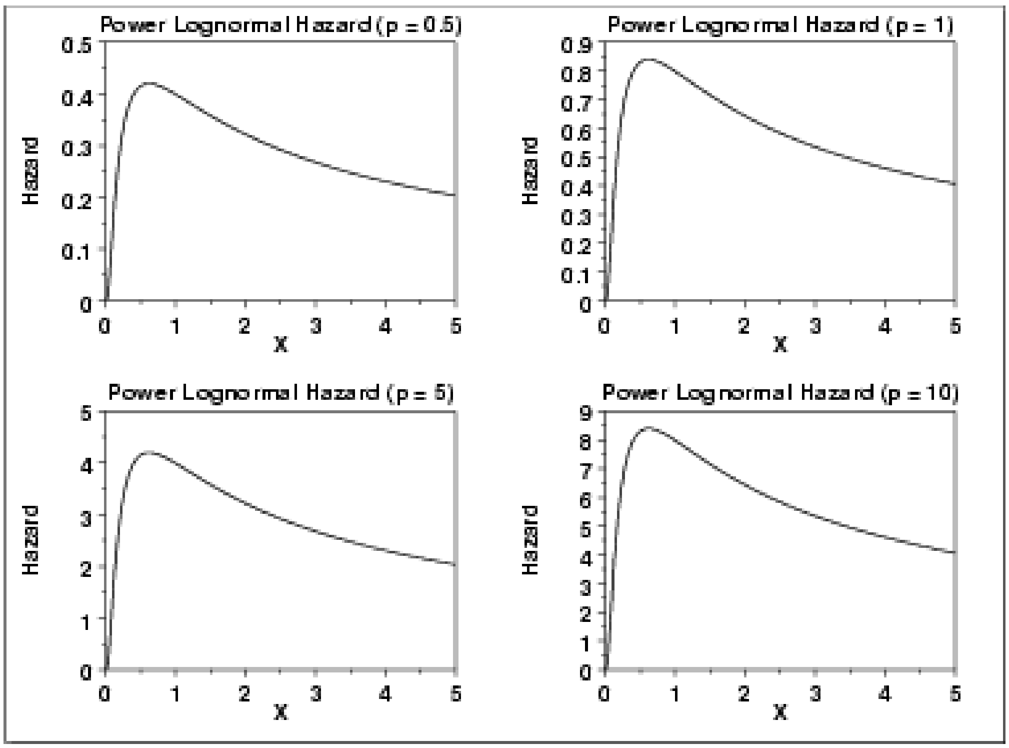
The formula for the hazard function of the power lognormal distribution is

$$h(x,p,\sigma) = \frac{p(\frac{1}{x\sigma})\phi(\frac{\log x}{\sigma})}{\Phi(\frac{-\log x}{\sigma})} \qquad x,p,\sigma > 0$$

where Φ is the cumulative distribution function of the standard normal distribution, and ϕ is the probability density function of the standard normal distribution.

Note that this is simply a multiple (p) of the lognormal hazard function.

The following is the plot of the power lognormal hazard function with the same values of ***p*** as the pdf plots above.

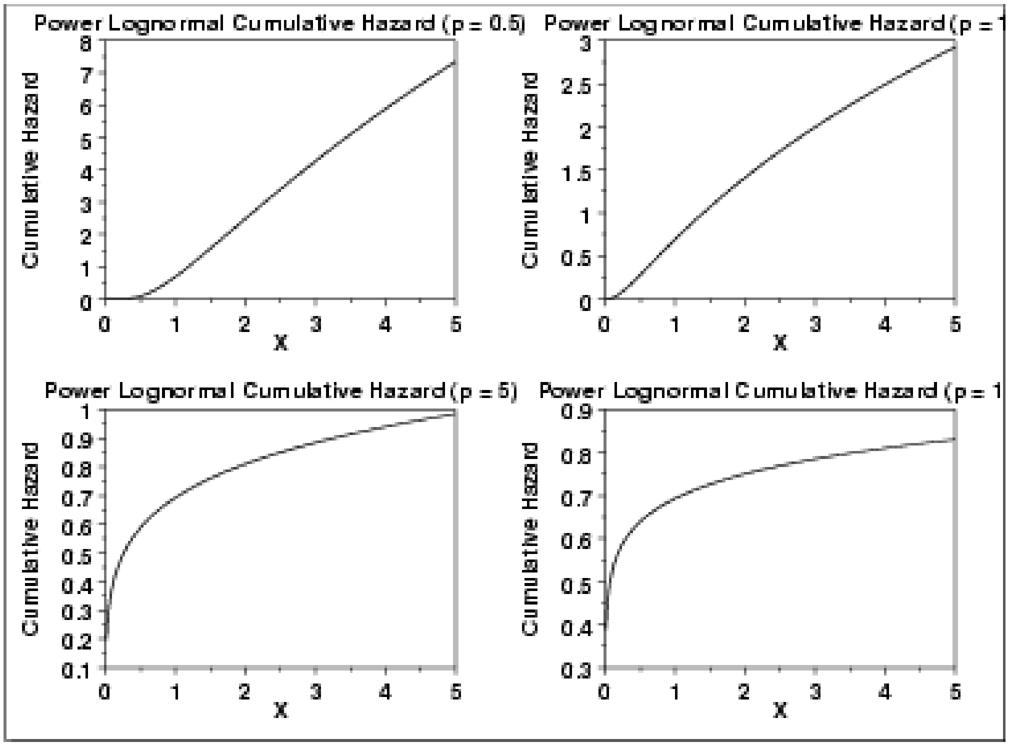


Cumulative Hazard Function

The formula for the cumulative hazard function of the power lognormal distribution is

$$H(x,p,\sigma) = -\log\left(\left(\Phi\left(\frac{-\log x}{\sigma}\right)\right)^p\right) \qquad x,p,\sigma > 0$$

The following is the plot of the power lognormal cumulative hazard function with the same values of ***p*** as the pdf plots above.

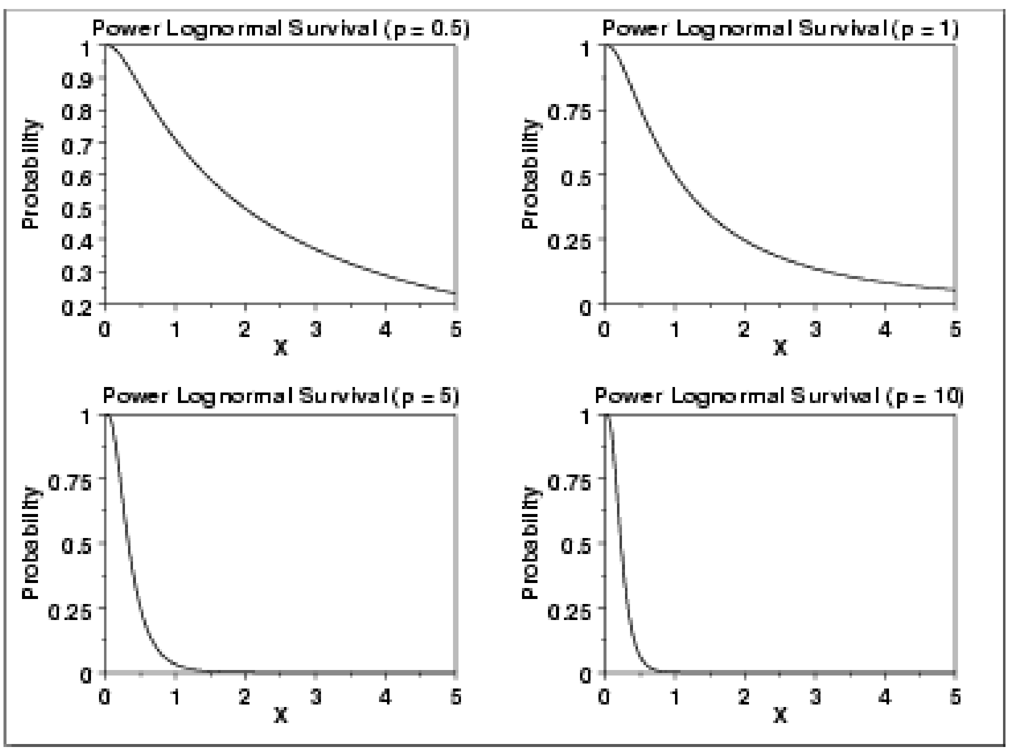


Survival Function

The formula for the survival function of the power lognormal distribution is

$$S(x,p,\sigma) = (\Phi(\frac{-\log x}{\sigma}))^p \qquad x,p,\sigma > 0$$

The following is the plot of the power lognormal survival function with the same values of ***p*** as the pdf plots above.

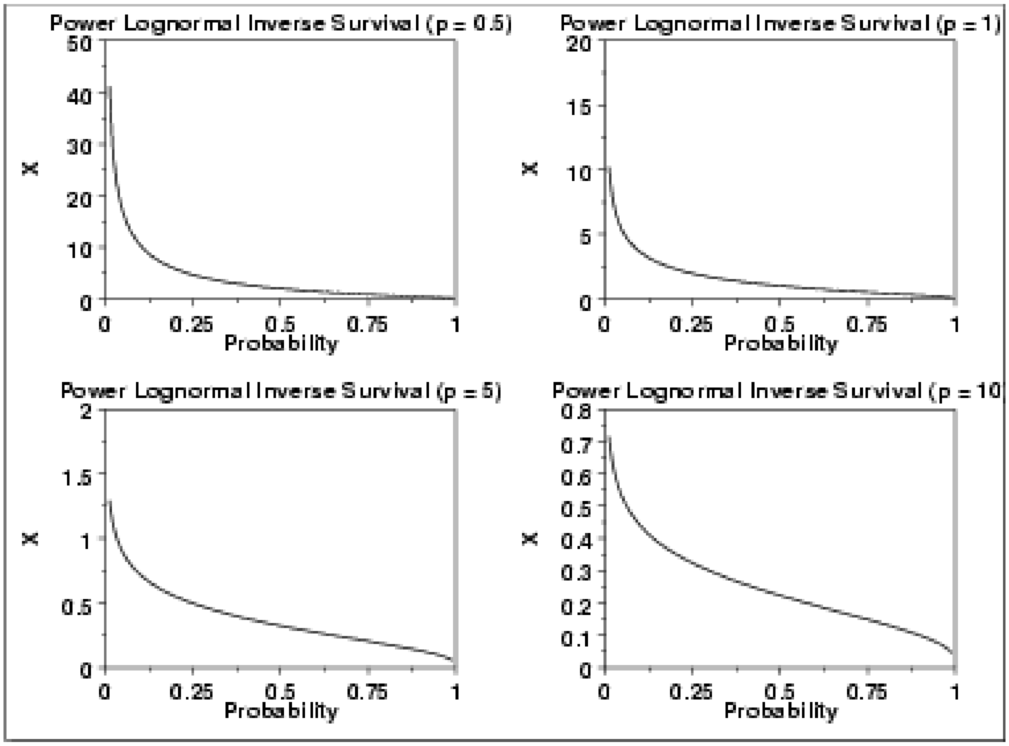


Inverse Survival Function

The formula for the inverse survival function of the power lognormal distribution is

$$Z(f,p,\sigma) = \exp\left(\Phi^{-1}(1 - f^{1/p})\sigma\right) \qquad 0 < p < 1; p, \sigma > 0$$

The following is the plot of the power lognormal inverse survival function with the same values of ***p*** as the pdf plots above.



Common Statistics

The statistics for the power lognormal distribution are complicated and require tables. Nelson discusses the mean, median, mode, and standard deviation of the power lognormal distribution and provides references to the appropriate tables.

Parameter Estimation

Nelson discusses maximum likelihood estimation for the power lognormal distribution. These estimates need to be performed with computer software. Software for maximum likelihood estimation of the parameters of the power lognormal distribution is not as readily available as for other reliability distributions such as the exponential, Weibull, and lognormal.

Software

Most general purpose statistical software programs do not support the probability functions for the power lognormal distribution. Dataplot does support them.

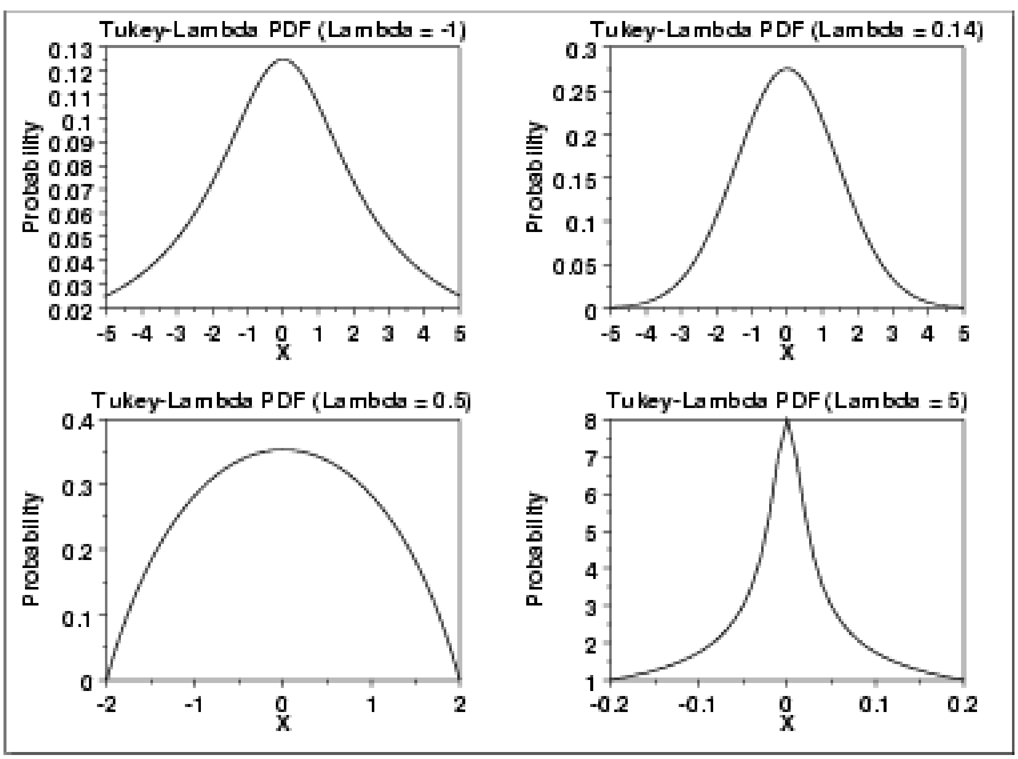
1.3.6.6.15. Tukey-Lambda Distribution

Probability Density Function

The Tukey-Lambda density function does not have a simple, closed form. It is computed numerically.

The Tukey-Lambda distribution has the shape parameter λ. As with other probability distributions, the Tukey-Lambda distribution can be transformed with a location parameter, μ, and a scale parameter, σ. Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

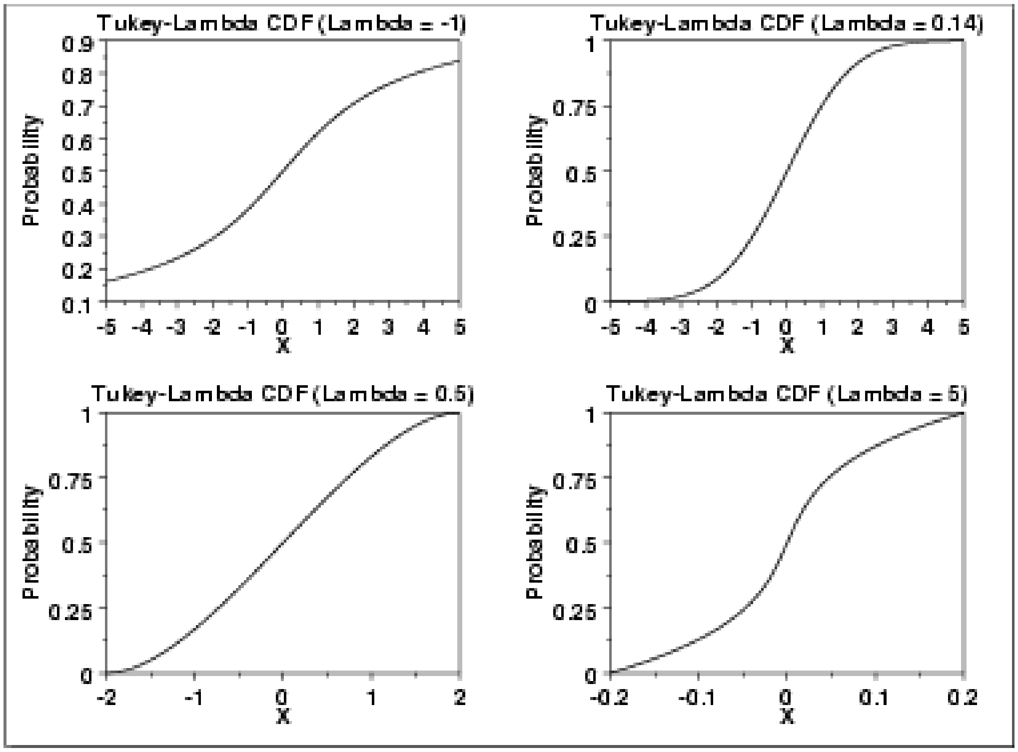
The following is the plot of the Tukey-Lambda probability density function for four values of λ.



Cumulative Distribution Function

The Tukey-Lambda distribution does not have a simple, closed form. It is computed numerically.

The following is the plot of the Tukey-Lambda cumulative distribution function with the same values of λ as the pdf plots above.

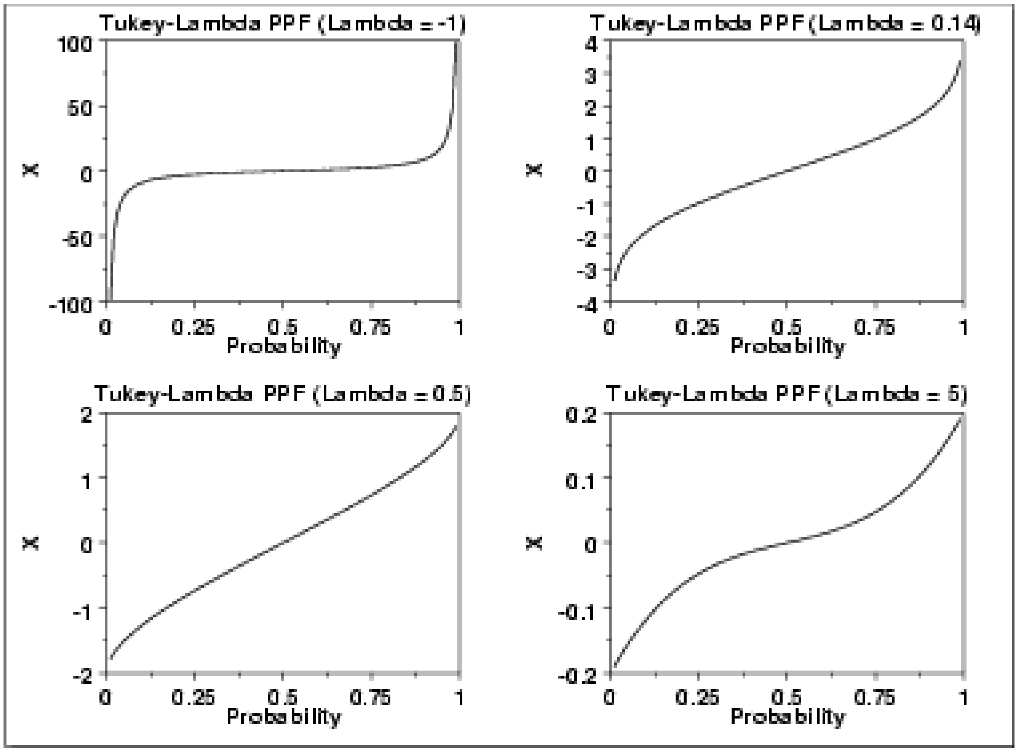


Percent Point Function

The formula for the percent point function of the standard form of the Tukey-Lambda distribution is

$$G(p) = \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda}$$

The following is the plot of the Tukey-Lambda percent point function with the same values of λ as the pdf plots above.



Other Probability Functions

The Tukey-Lambda distribution is typically used to identify an appropriate distribution (see the comments below) and not used in statistical models directly. For this reason, we omit the formulas, and plots for the hazard, cumulative hazard, survival, and inverse survival functions. We also omit the common statistics and parameter estimation sections.

Comments

The Tukey-Lambda distribution is actually a family of distributions that can approximate a number of common distributions. For example,

$\lambda = -1$	approximately Cauchy
$\lambda = 0$	exactly logistic
$\lambda = 0.14$	approximately normal
$\lambda = 0.5$	U-shaped
$\lambda = 1$	exactly uniform (from -1 to +1)

The most common use of this distribution is to generate a Tukey-Lambda PPCC plot of a data set. Based on the ppcc plot, an appropriate model for the data is suggested. For example, if the maximum correlation occurs for a value of λ at or near 0.14, then the data can be modeled with a normal distribution. Values of λ less than this imply a heavy-tailed distribution (with -1 approximating a Cauchy). That is, as the optimal value of λ goes from 0.14 to -1, increasingly heavy tails are implied. Similarly, as the optimal value of λ becomes greater than 0.14, shorter tails are implied.

As the Tukey-Lambda distribution is a symmetric distribution, the use of the Tukey-Lambda PPCC plot to determine a reasonable distribution to model the data only applies to symmetric distributuins. A histogram of the data should provide evidence as to whether the data can be reasonably modeled with a symmetric distribution.

Software

Most general purpose statistical software programs do not support the probability functions for the Tukey-Lambda distribution. Dataplot does support them.

1.3.6.6.16. Extreme Value Type I Distribution

Probability Density Function

The extreme value type I distribution has two forms. One is based on the smallest extreme and the other is based on the largest extreme. We call these the minimum and maximum cases, respectively. Formulas and plots for both cases are given. The extreme value type I distribution is also referred to as the Gumbel distribution.

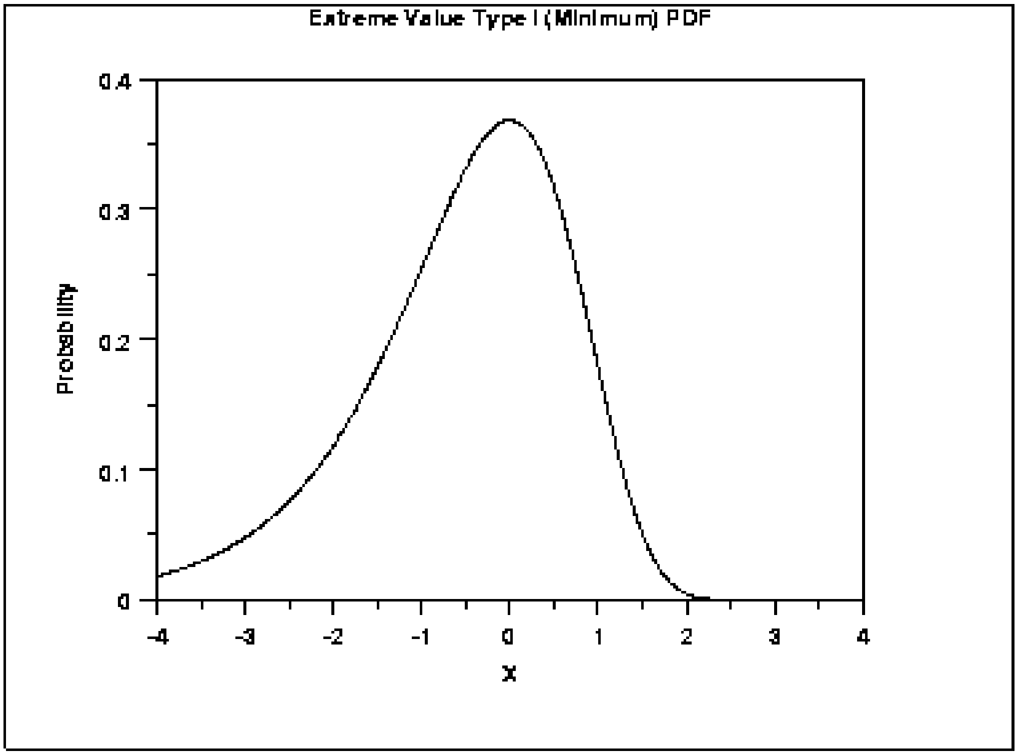
The general formula for the probability density function of the Gumbel (minimum) distribution is

$$f(x) = \frac{1}{\beta} e^{\frac{x-\mu}{\beta}} e^{-e^{\frac{x-\mu}{\beta}}}$$

where μ is the location parameter and β is the scale parameter. The case where μ = 0 and β = 1 is called the **standard Gumbel distribution**. The equation for the standard Gumbel distribution (minimum) reduces to

$$f(x) = e^{x} e^{-e^{x}}$$

The following is the plot of the Gumbel probability density function for the minimum case.



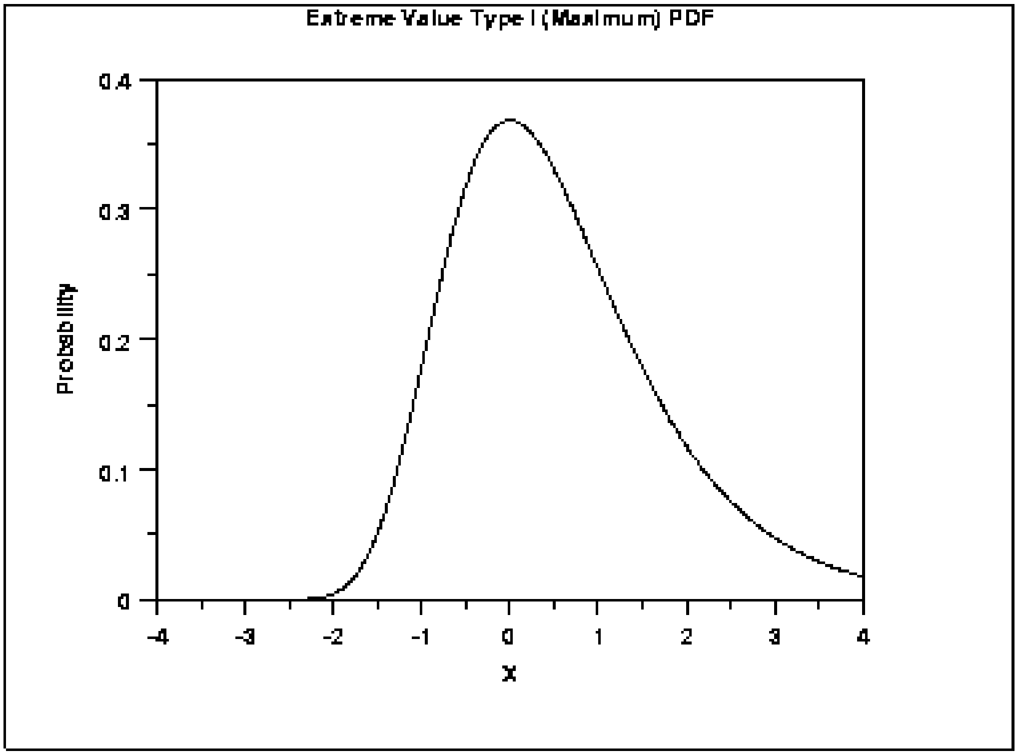
The general formula for the probability density function of the Gumbel (maximum) distribution is

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}}$$

where μ is the location parameter and β is the scale parameter. The case where μ = 0 and β = 1 is called the **standard Gumbel distribution**. The equation for the standard Gumbel distribution (maximum) reduces to

$$f(x) = e^{-x} e^{-e^{-x}}$$

The following is the plot of the Gumbel probability density function for the maximum case.



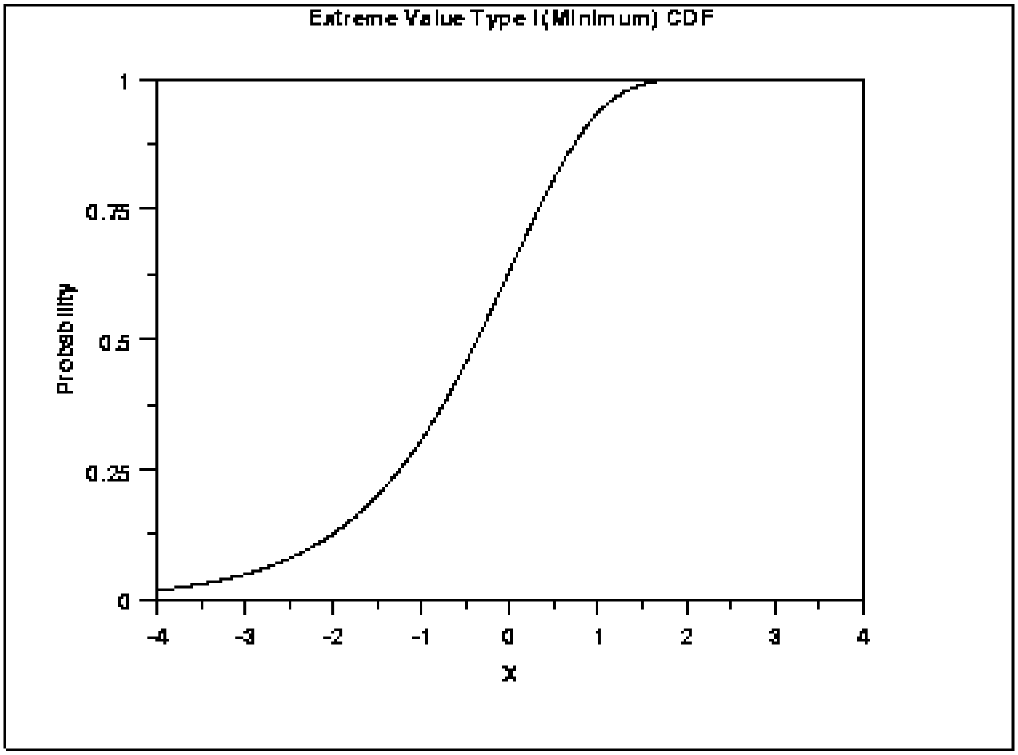
Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

Cumulative Distribution Function

The formula for the cumulative distribution function of the Gumbel distribution (minimum) is

$$F(x) = 1 - e^{-e^{x}}$$

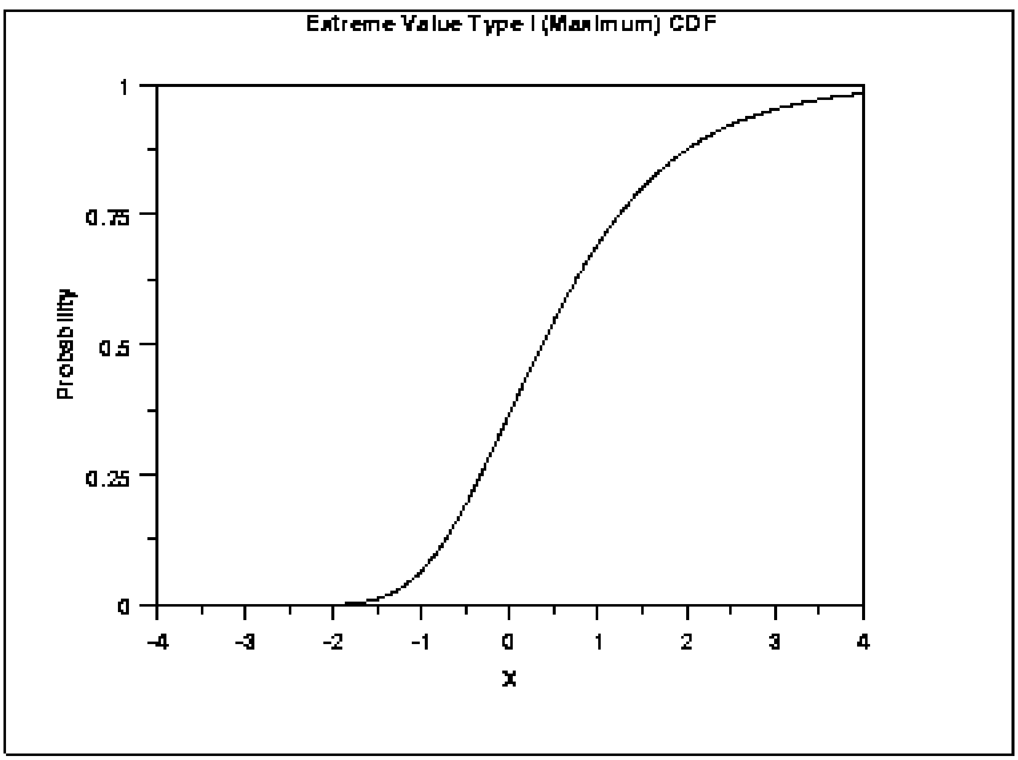
The following is the plot of the Gumbel cumulative distribution function for the minimum case.



The formula for the cumulative distribution function of the Gumbel distribution (maximum) is

$$F(x) = e^{-e^{-x}}$$

The following is the plot of the Gumbel cumulative distribution function for the maximum case.

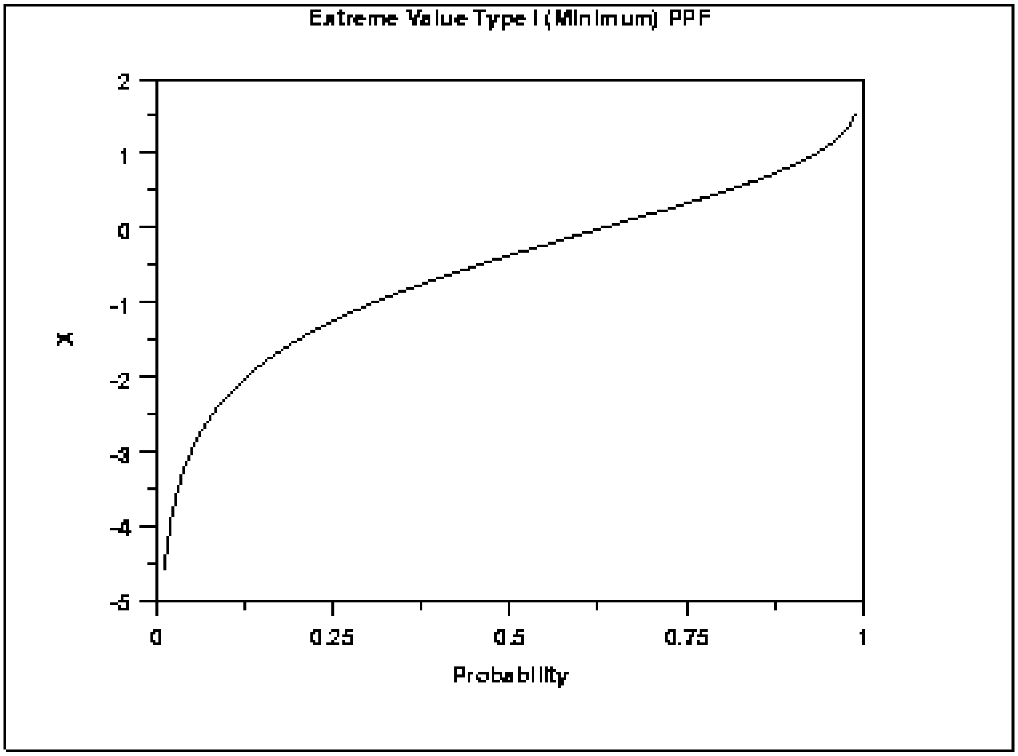


Percent Point Function

The formula for the percent point function of the Gumbel distribution (minimum) is

$$G(p) = \ln(\ln(\frac{1}{1-p}))$$

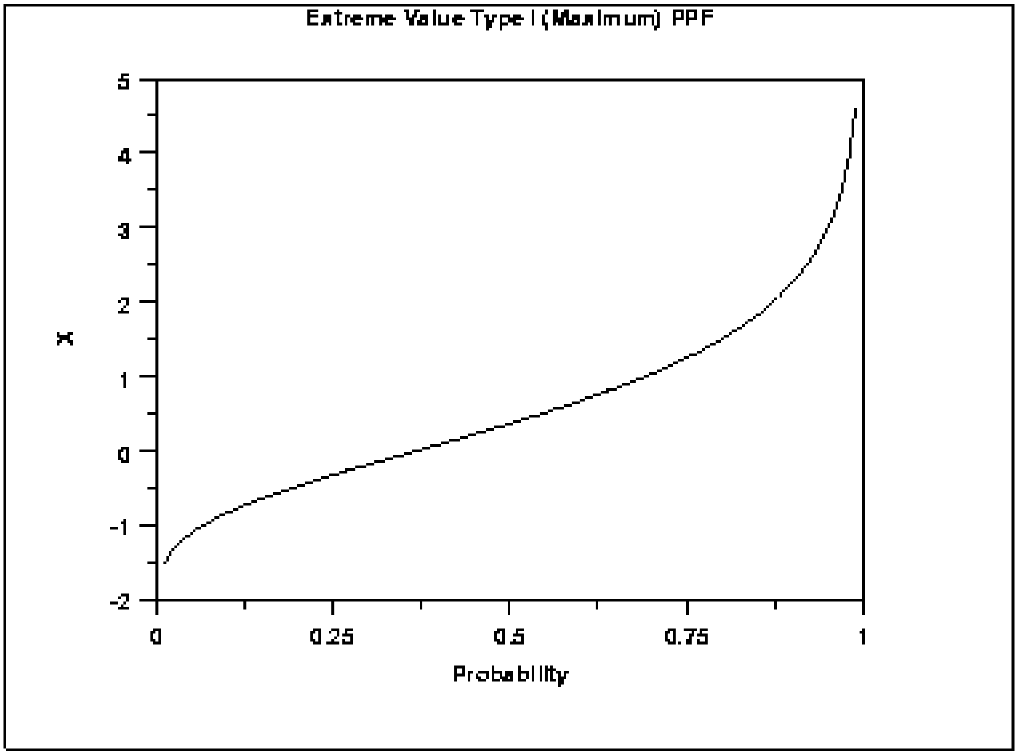
The following is the plot of the Gumbel percent point function for the minimum case.



The formula for the percent point function of the Gumbel distribution (maximum) is

$$G(p) = -\ln(\ln(\frac{1}{p}))$$

The following is the plot of the Gumbel percent point function for the maximum case.

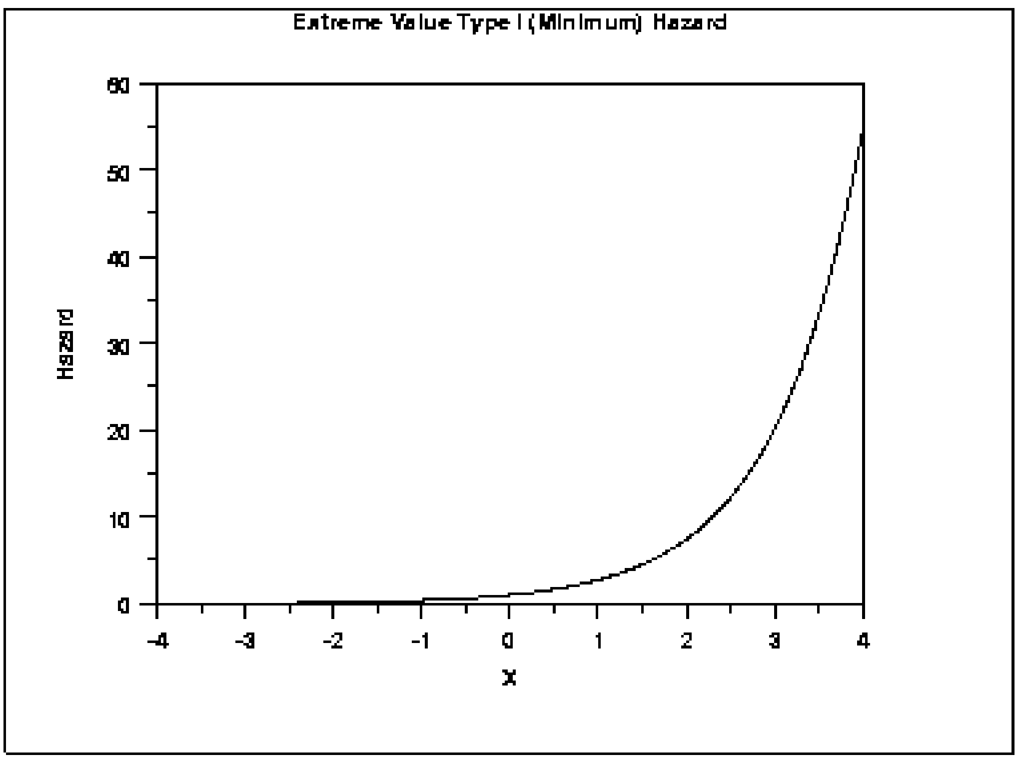


Hazard Function

The formula for the hazard function of the Gumbel distribution (minimum) is

$$h(x) = e^{x}$$

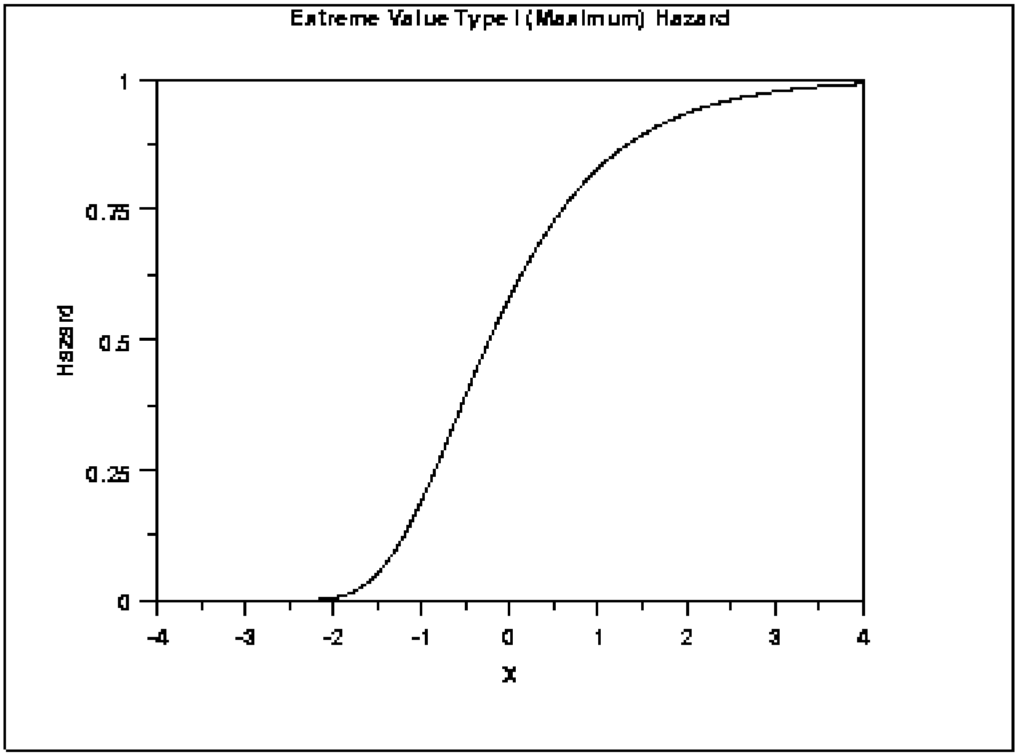
The following is the plot of the Gumbel hazard function for the minimum case.



The formula for the hazard function of the Gumbel distribution (maximum) is

$$h(x) = \frac{e^{-x}}{e^{e^{-x}} - 1}$$

The following is the plot of the Gumbel hazard function for the maximum case.

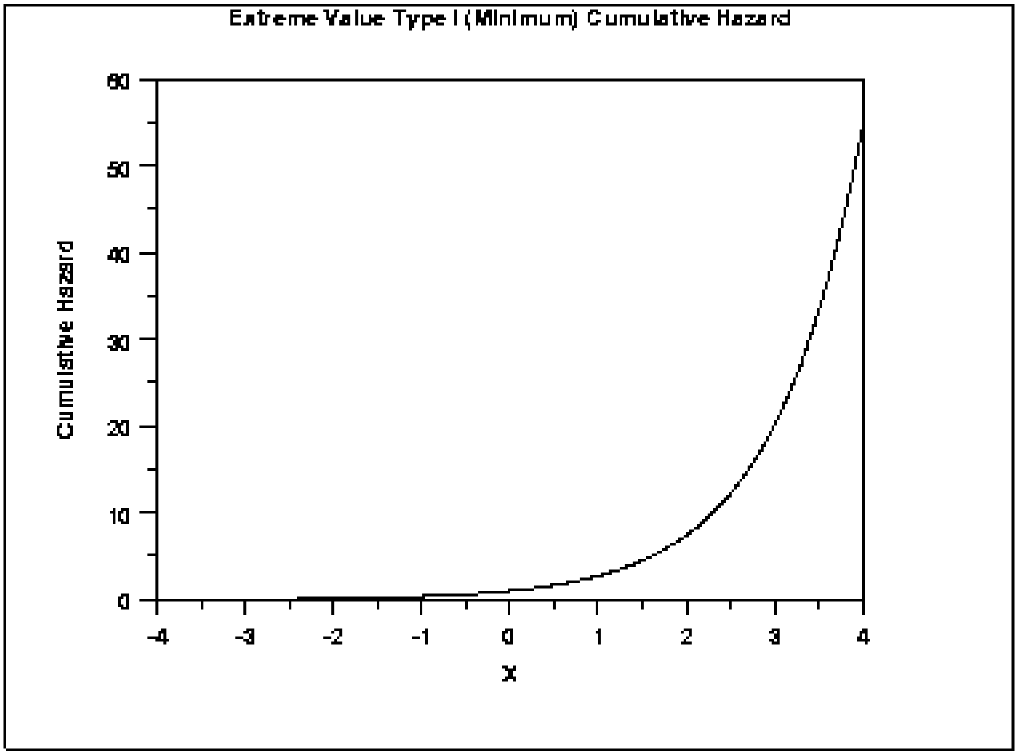


Cumulative Hazard Function

The formula for the cumulative hazard function of the Gumbel distribution (minimum) is

$$H(x) = e^{x}$$

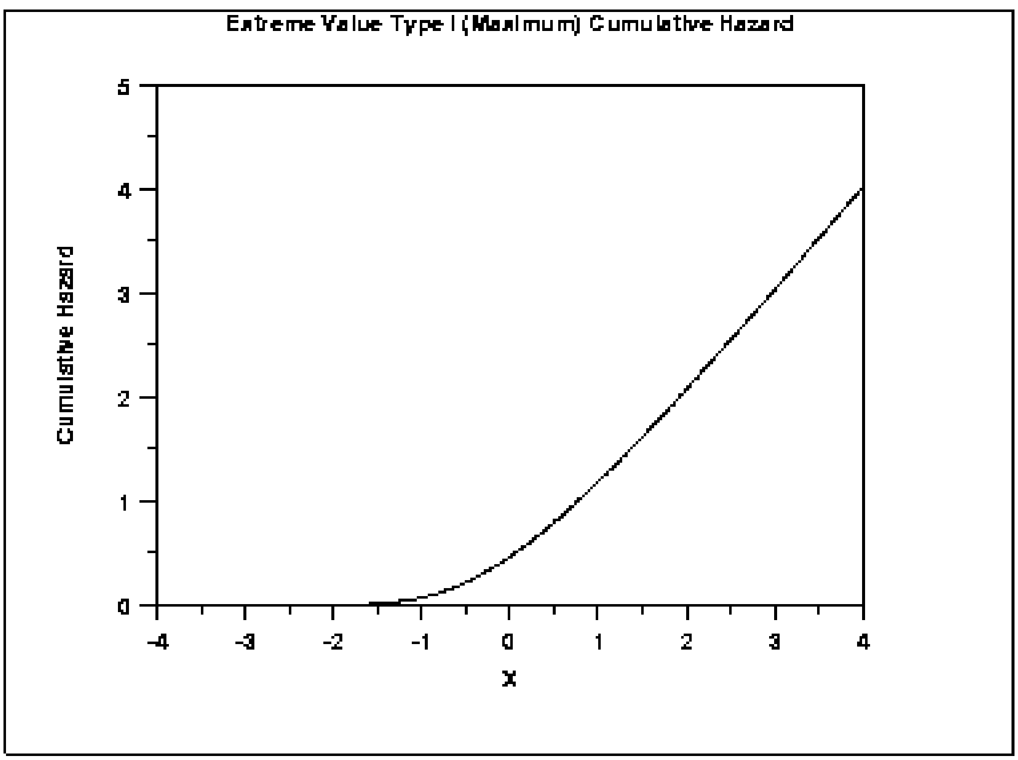
The following is the plot of the Gumbel cumulative hazard function for the minimum case.



The formula for the cumulative hazard function of the Gumbel distribution (maximum) is

$$H(x) = -\ln(1 - e^{-e^{-x}})$$

The following is the plot of the Gumbel cumulative hazard function for the maximum case.

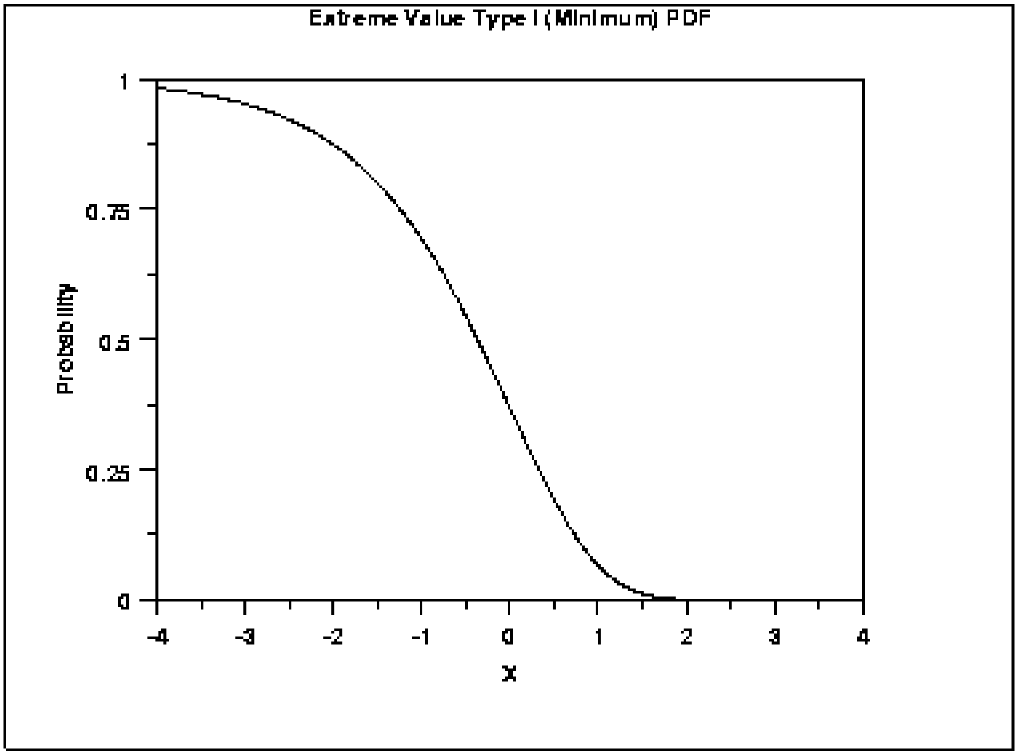


Survival Function

The formula for the survival function of the Gumbel distribution (minimum) is

$$S(x) = e^{-e^{x}}$$

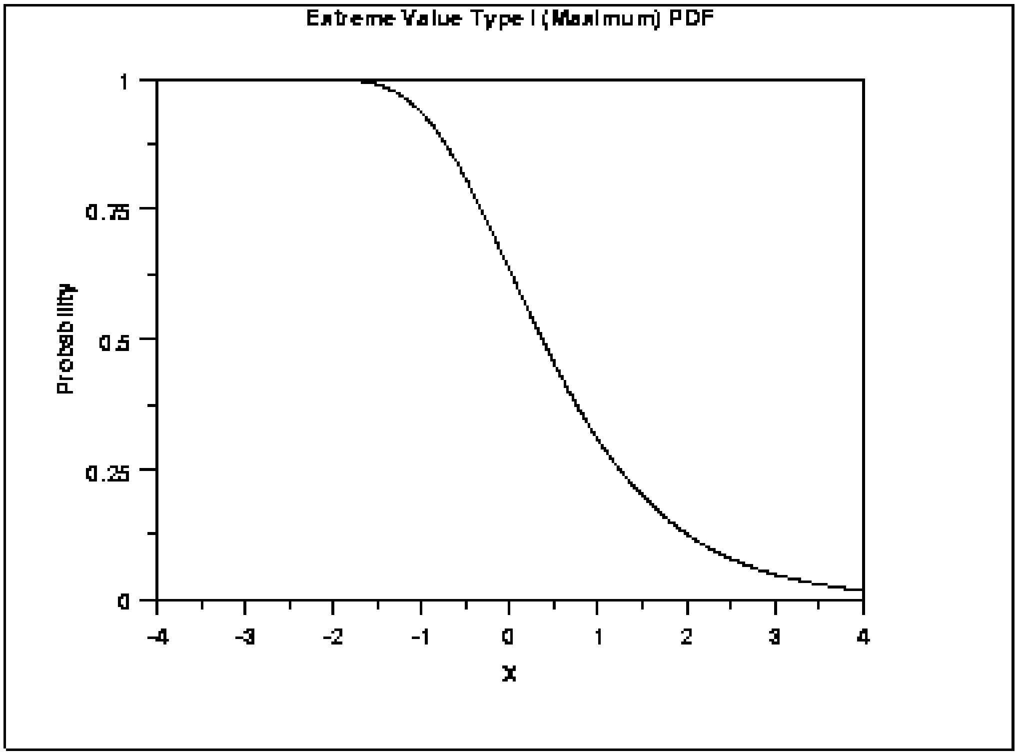
The following is the plot of the Gumbel survival function for the minimum case.



The formula for the survival function of the Gumbel distribution (maximum) is

$$S(x) = 1 - e^{-e^{-x}}$$

The following is the plot of the Gumbel survival function for the maximum case.

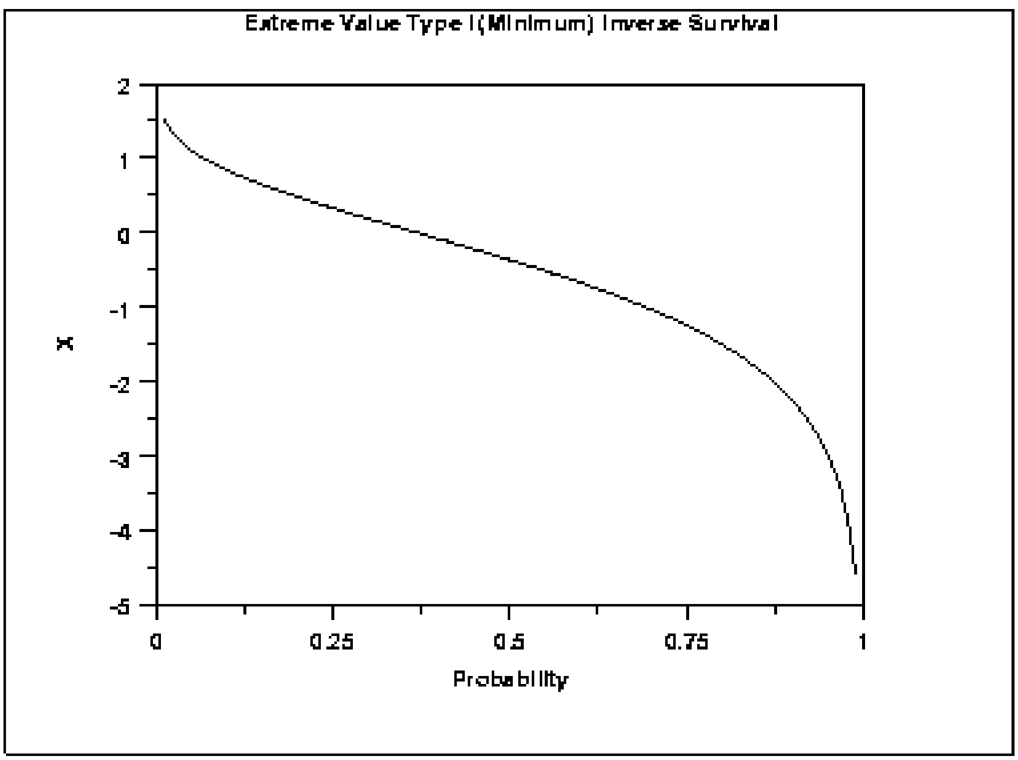


Inverse Survival Function

The formula for the [inverse survival function](eda362.htm#ISF) of the Gumbel distribution (minimum) is

$$Z(p) = \ln(\ln(\frac{1}{p}))$$

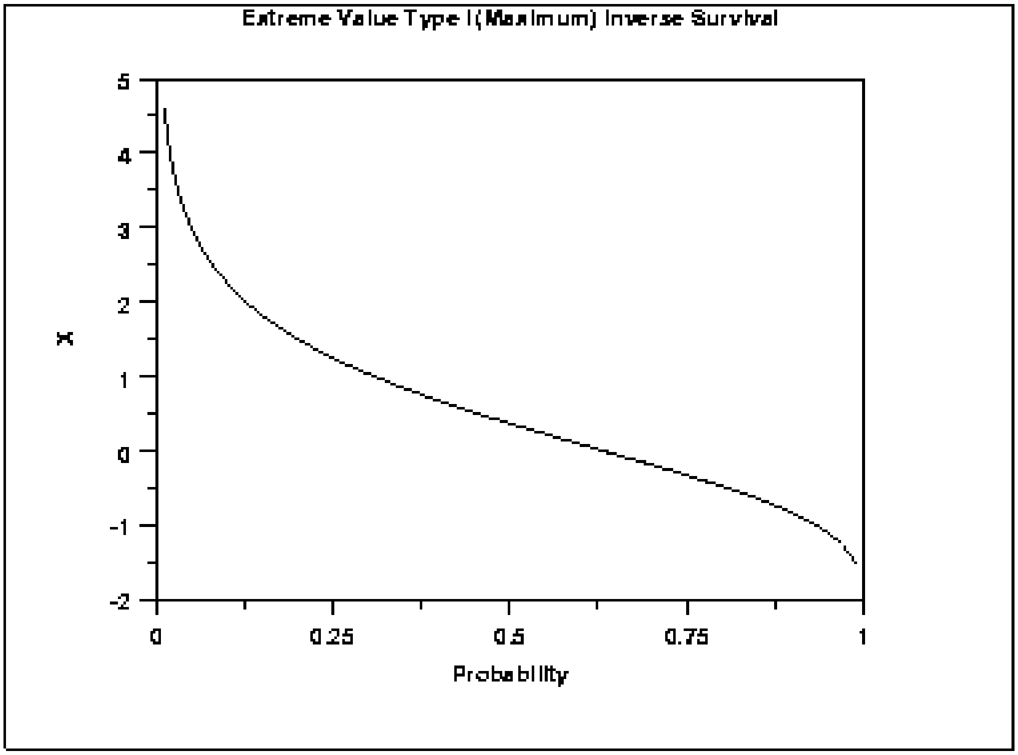
The following is the plot of the Gumbel inverse survival function for the minimum case.



The formula for the inverse survival function of the Gumbel distribution (maximum) is

$$Z(p) = -\ln(\ln(\frac{1}{1-p}))$$

The following is the plot of the Gumbel inverse survival function for the maximum case.



Common Statistics

The formulas below are for the maximum order statistic case.

Mean	$\mu + 0.5772\beta$ The constant 0.5772 is Euler's number.
Median	$\mu - \beta\ln(\ln(2))$
Mode	μ
Range	Negative infinity to positive infinity.
Standard Deviation	$\frac{\beta\pi}{\sqrt{6}}$
Skewness	1.13955
Kurtosis	5.4
Coefficient of Variation	$\frac{\beta\pi}{\sqrt{6}(\mu + 0.5772\beta)}$

Parameter Estimation

The method of moments estimators of the Gumbel (maximum) distribution are

$$\tilde{\beta} = \frac{s\sqrt{6}}{\pi}$$

$$\tilde{\mu} = \bar{X} - 0.5772\tilde{\beta}$$

where $\bar{X}$ and s are the sample mean and standard deviation, respectively.

The equations for the maximum likelihood estimation of the shape and scale parameters are discussed in Chapter 15 of Evans, Hastings, and Peacock and Chapter 22 of Johnson, Kotz, and Balakrishnan. These equations need to be solved numerically and this is typically accomplished by using statistical software packages.

Software

Some general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the extreme value type I distribution.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.3.6.6.17. Beta Distribution

Probability Density Function

The general formula for the probability density function of the beta distribution is

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \le x \le b; p, q > 0$$

where *p* and *q* are the shape parameters, *a* and *b* are the lower and upper bounds, respectively, of the distribution, and *B*(*p*,*q*) is the beta function. The beta function has the formula

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

The case where *a* = 0 and *b* = 1 is called the **standard beta distribution**. The equation for the standard beta distribution is

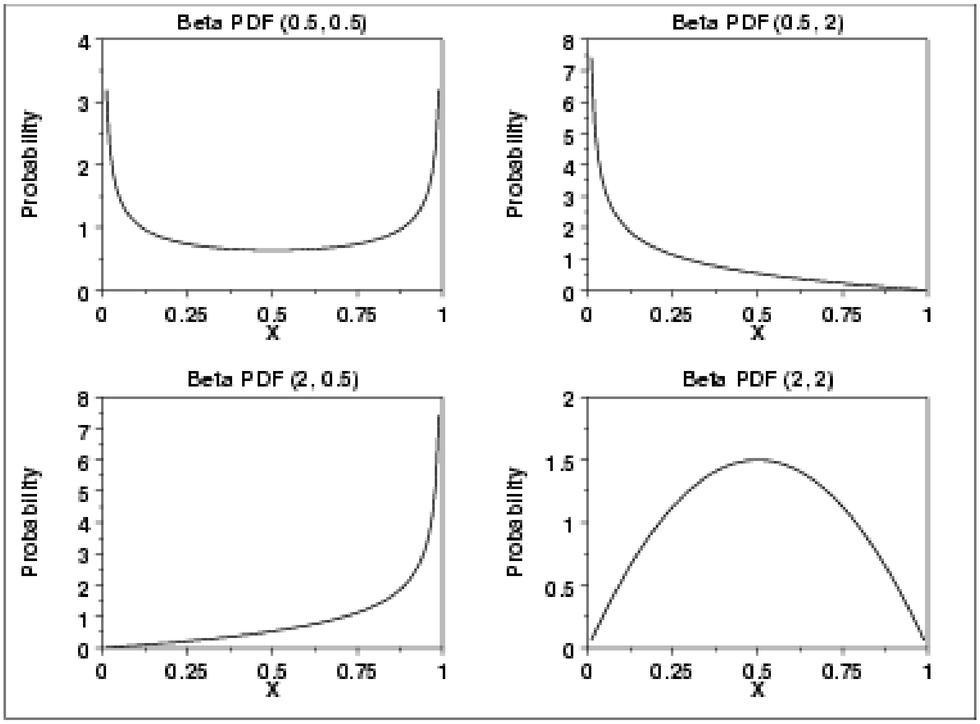
$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)} \qquad 0 \le x \le 1; p, q > 0$$

Typically we define the general form of a distribution in terms of location and scale parameters. The beta is different in that we define the general distribution in terms of the lower and upper bounds. However, the location and scale parameters can be defined in terms of the lower and upper limits as follows:

location = *a*
scale = *b* - *a*

Since the general form of probability functions can be expressed in terms of the standard distribution, all subsequent formulas in this section are given for the standard form of the function.

The following is the plot of the beta probability density function for four different values of the shape parameters.



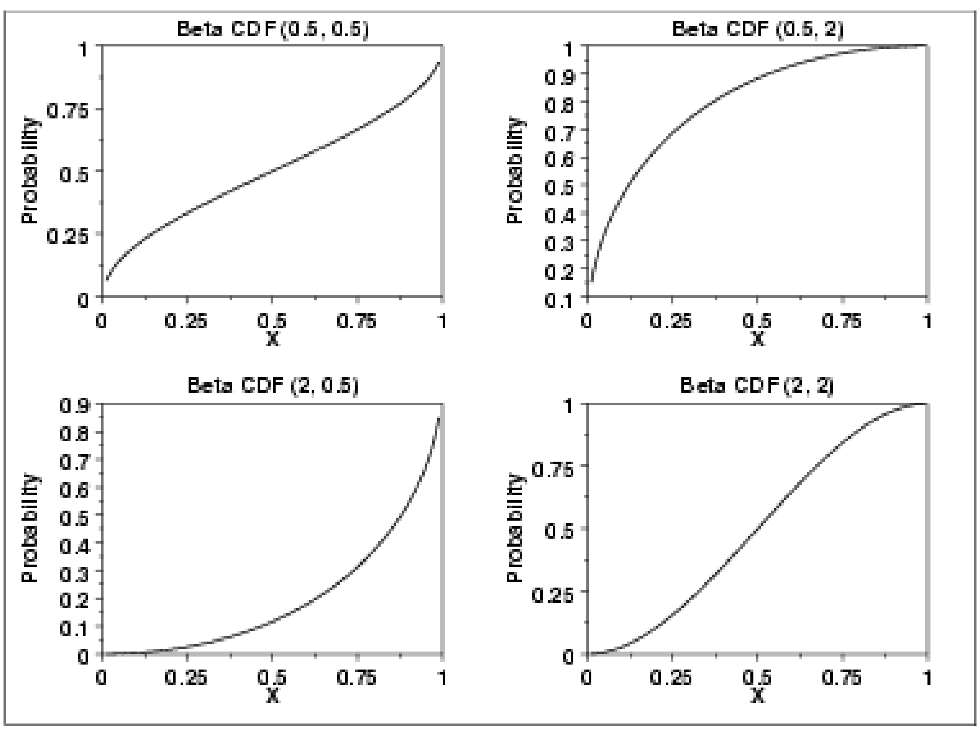
Cumulative Distribution Function

The formula for the cumulative distribution function of the beta distribution is also called the incomplete beta function ratio (commonly denoted by I_x) and is defined as

$$F(x) = I_x(p,q) = \frac{\int_0^x t^{p-1}(1-t)^{q-1}dt}{B(p,q)} \qquad 0 \le x \le 1; p, q > 0$$

where B is the beta function defined above.

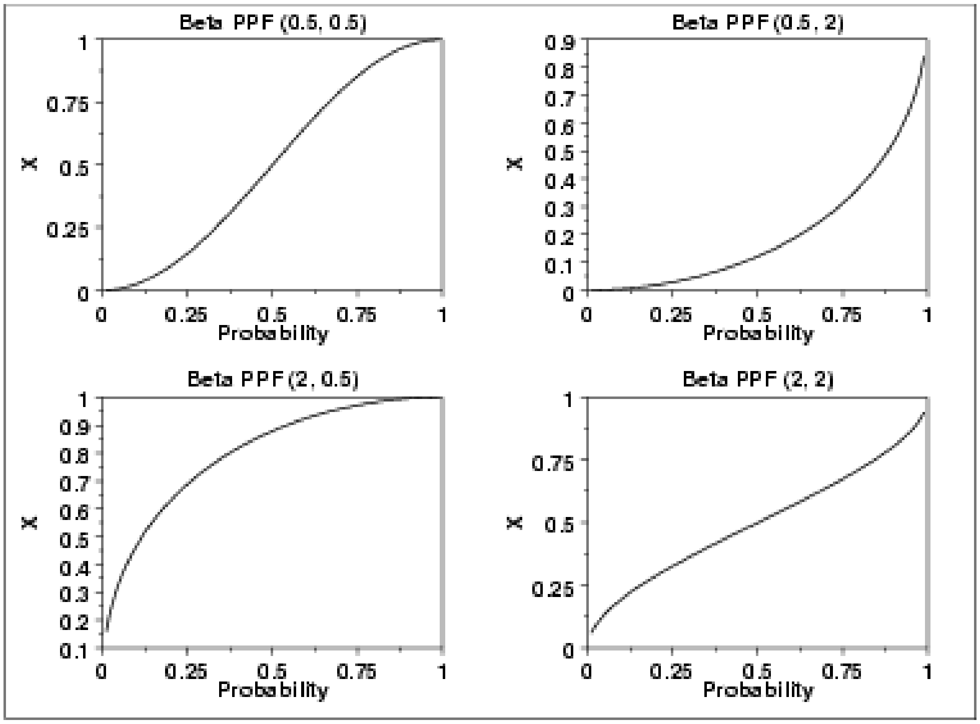
The following is the plot of the beta cumulative distribution function with the same values of the shape parameters as the pdf plots above.



Percent Point Function

The formula for the percent point function of the beta distribution does not exist in a simple closed form. It is computed numerically.

The following is the plot of the beta percent point function with the same values of the shape parameters as the pdf plots above.



Other Probability Functions

Since the beta distribution is not typically used for reliability applications, we omit the formulas and plots for the hazard, cumulative hazard, survival, and inverse survival probability functions.

Common Statistics

The formulas below are for the case where the lower limit is zero and the upper limit is one.

Mean	$\frac{p}{p+q}$
Mode	$\frac{p-1}{p+q-2} \qquad p, q > 1$
Range	0 to 1
Standard Deviation	$\sqrt{\frac{pq}{(P+q)^2(p+q+1)}}$
Coefficient of Variation	$\sqrt{\frac{q}{p(p+q+1)}}$
Skewness	$\frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

Parameter Estimation

First consider the case where *a* and *b* are assumed to be known. For this case, the method of moments estimates are

$$p = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

$$q = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

where $\bar{x}$ is the sample mean and s^2 is the sample variance. If *a* and *b* are not 0 and 1, respectively, then replace $\bar{x}$ with $\frac{\bar{x}-a}{b-a}$ and s^2 with $\frac{s^2}{(b-a)^2}$ in the above equations.

For the case when *a* and *b* are known, the maximum likelihood estimates can be obtained by solving the following set of equations

$$\psi(\hat{p}) - \psi(\hat{p}+\hat{q}) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{Y_i - a}{b-a}\right)$$

$$\psi(\hat{q}) - \psi(\hat{p}+\hat{q}) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{b - Y_i}{b-a}\right)$$

The maximum likelihood equations for the case when *a* and *b* are not known are given in pages 221-235 of Volume II of Johnson, Kotz, and Balakrishan.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the beta distribution.

1.3.6.6.18. Binomial Distribution

Probability Mass Function

The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. These outcomes are appropriately labeled "success" and "failure". The binomial distribution is used to obtain the probability of observing *x* successes in *N* trials, with the probability of success on a single trial denoted by *p*. The binomial distribution assumes that *p* is fixed for all trials.

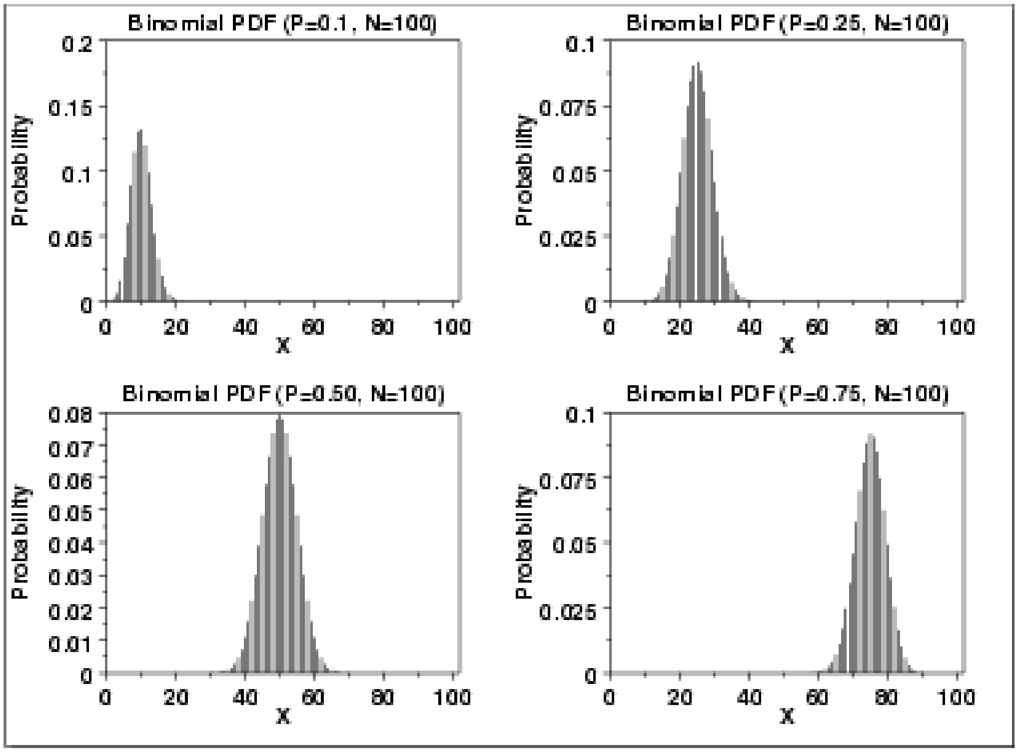
The formula for the binomial probability mass function is

$$P(x,p,n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \qquad \mbox{for } x = 0, 1, 2, \cdots, n$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The following is the plot of the binomial probability density function for four values of *p* and *n* = 100.

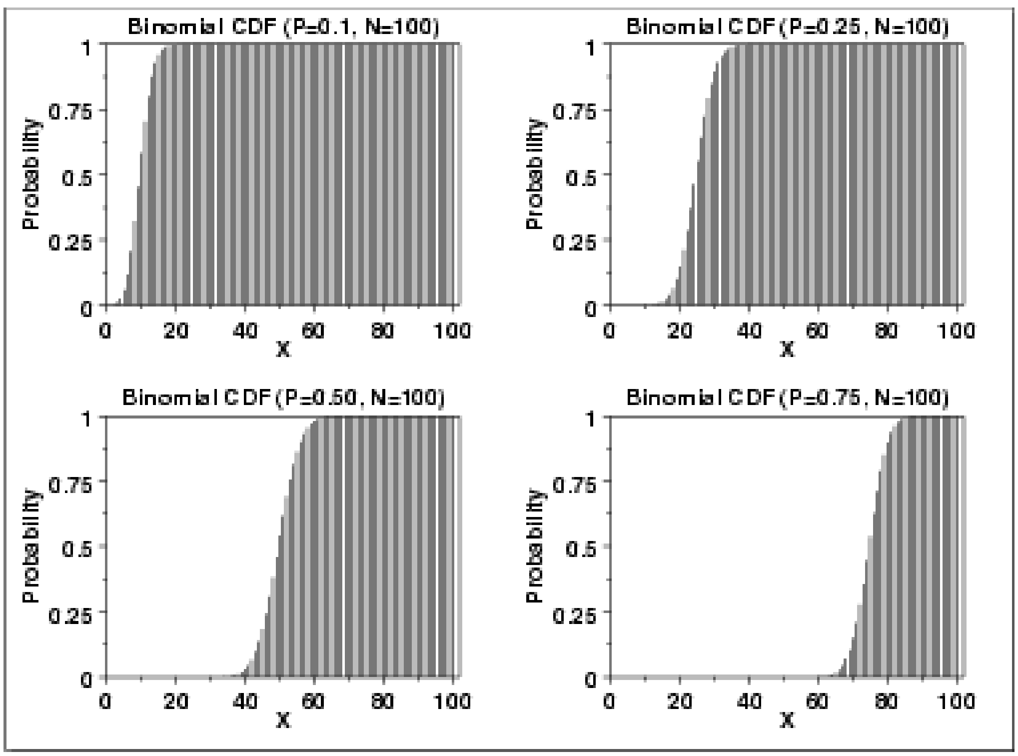


Cumulative Distribution Function

The formula for the binomial cumulative probability function is

$$F(x,p,n) = \sum_{i=0}^{x} \binom{n}{i} (p)^{i}(1-p)^{(n-i)}$$

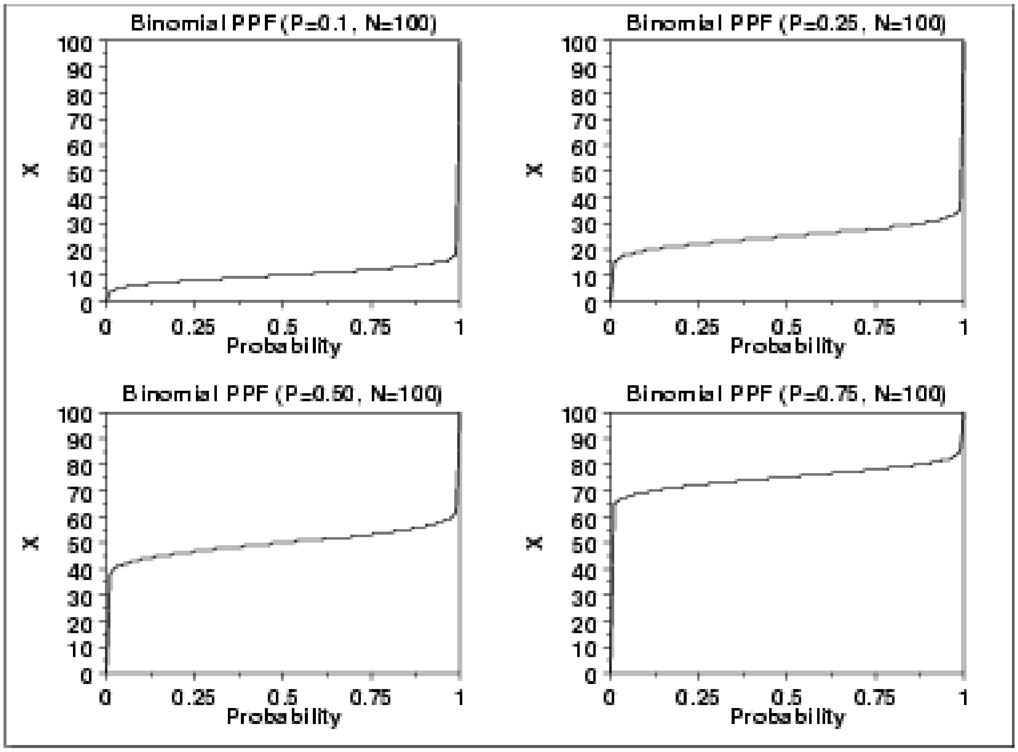
The following is the plot of the binomial cumulative distribution function with the same values of *p* as the pdf plots above.



Percent Point Function

The binomial percent point function does not exist in simple closed form. It is computed numerically. Note that because this is a discrete distribution that is only defined for integer values of *x*, the percent point function is not smooth in the way the percent point function typically is for a continuous distribution.

The following is the plot of the binomial percent point function with the same values of *p* as the pdf plots above.



Common Statistics

Mean	np
Mode	$p(n+1) - 1 \le x \le p(n+1)$
Range	0 to N
Standard Deviation	$\sqrt{np(1-p)}$
Coefficient of Variation	$\sqrt{\frac{(1-p)}{np}}$
Skewness	$\frac{(1-2p)}{\sqrt{np(1-p)}}$
Kurtosis	$3 - \frac{6}{n} + \frac{1}{np(1-p)}$

Comments

The binomial distribution is probably the most commonly used discrete distribution.

Parameter Estimation

The maximum likelihood estimator of *p* (*n* is fixed) is

$$\tilde{p} = \frac{x}{n}$$

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the binomial distribution.

1.3.6.6.19. Poisson Distribution

Probability Mass Function

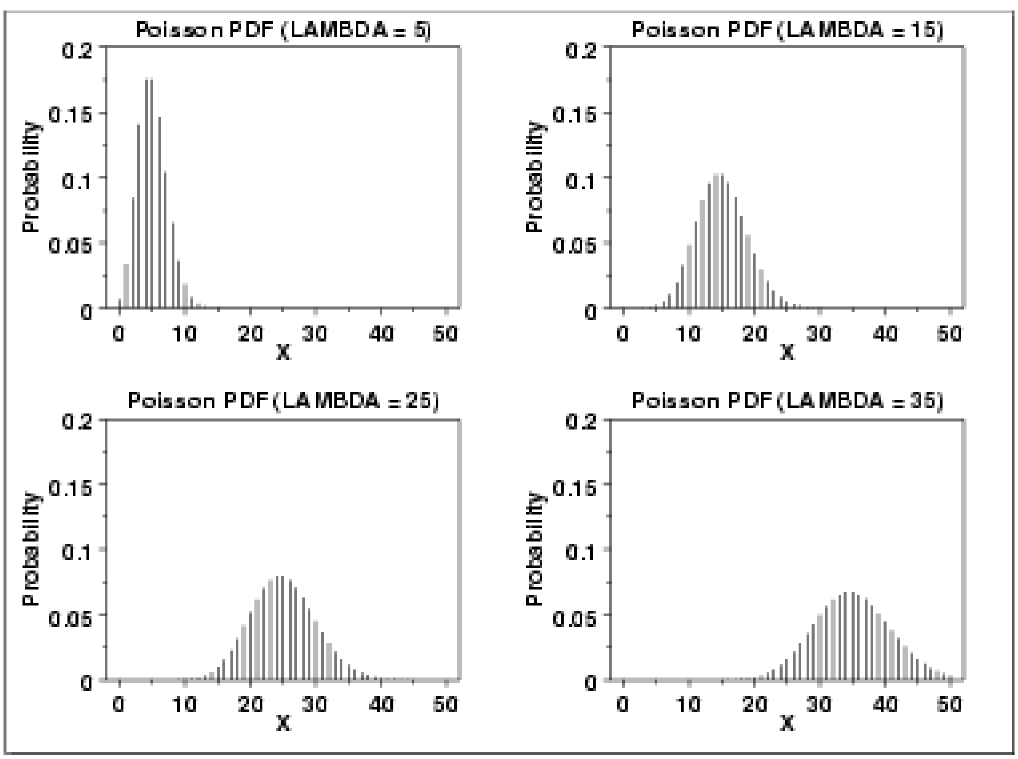
The Poisson distribution is used to model the number of events occurring within a given time interval.

The formula for the Poisson probability mass function is

$$p(x, \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \mbox{for } x = 0, 1, 2, \cdots$$

λ is the shape parameter which indicates the average number of events in the given time interval.

The following is the plot of the Poisson probability density function for four values of λ.

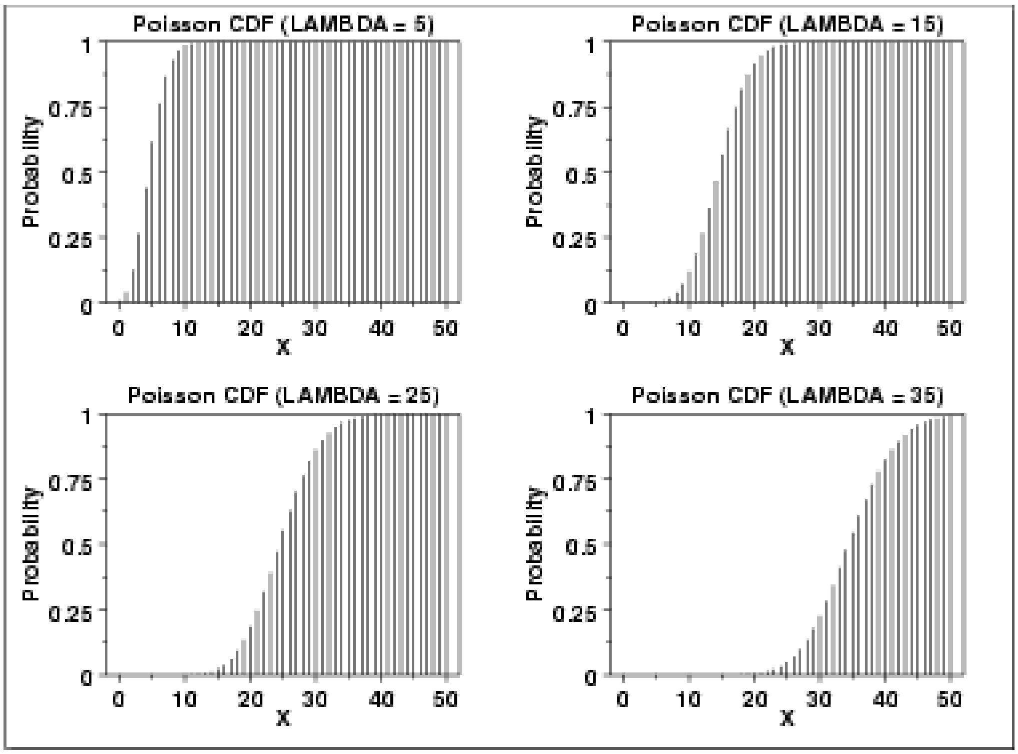


Cumulative Distribution Function

The formula for the Poisson cumulative probability function is

$$F(x,\lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^{i}}{i!}$$

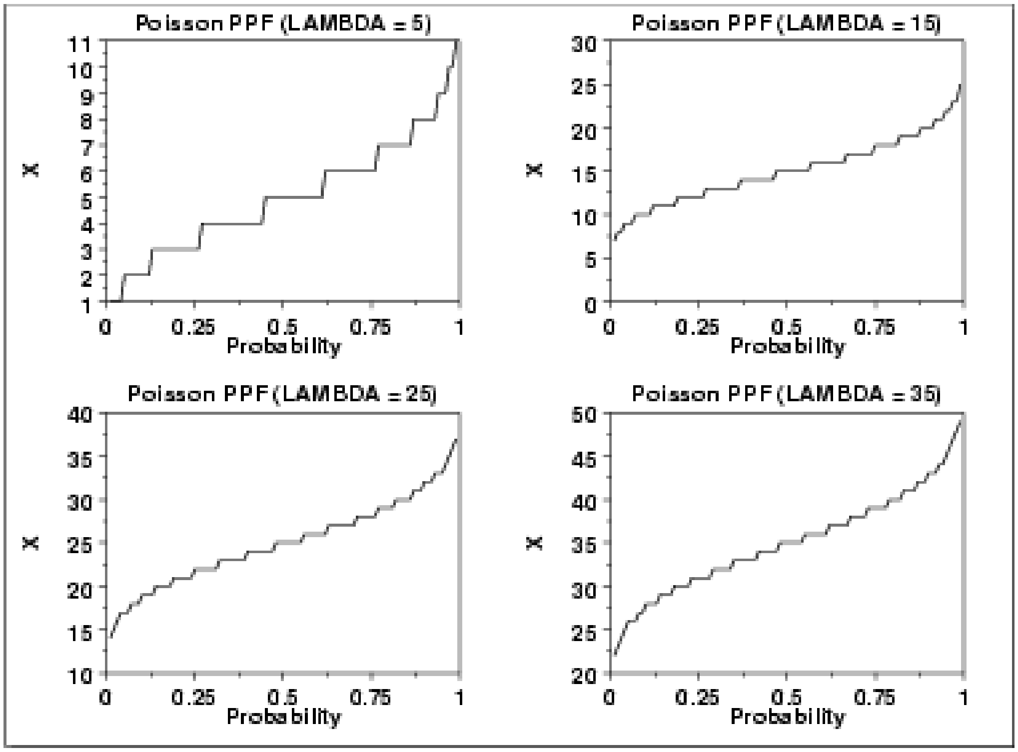
The following is the plot of the Poisson cumulative distribution function with the same values of λ as the pdf plots above.



Percent Point Function

The Poisson percent point function does not exist in simple closed form. It is computed numerically. Note that because this is a discrete distribution that is only defined for integer values of *x*, the percent point function is not smooth in the way the percent point function typically is for a continuous distribution.

The following is the plot of the Poisson percent point function with the same values of λ as the pdf plots above.



Common Statistics	Mean	λ
	Mode	For non-integer λ, it is the largest integer less than λ. For integer λ, x = λ and x = λ - 1 are both the mode.
	Range	0 to positive infinity
	Standard Deviation	$\sqrt{\lambda}$
	Coefficient of Variation	$\frac{1}{\sqrt{\lambda}}$
	Skewness	$\frac{1}{\sqrt{\lambda}}$
	Kurtosis	$3 + \frac{1}{\lambda}$

Parameter Estimation

The maximum likelihood estimator of λ is

$$\tilde{\lambda} = \bar{X}$$

where $\bar{X}$ is the sample mean.

Software

Most general purpose statistical software programs, including Dataplot, support at least some of the probability functions for the Poisson distribution.

1.3.6.7. Tables for Probability Distributions

Tables

Several commonly used tables for probability distributions can be referenced below.

The values from these tables can also be obtained from most general purpose statistical software programs. Most introductory statistics textbooks (e.g., Snedecor and Cochran) contain more extensive tables than are included here. These tables are included for convenience.

1. Cumulative distribution function for the standard normal distribution
2. Upper critical values of Student's t-distribution with ν degrees of freedom
3. Upper critical values of the F-distribution with ν_1 and ν_2 degrees of freedom
4. Upper critical values of the chi-square distribution with ν degrees of freedom
5. Critical values of t^* distribution for testing the output of a linear calibration line at 3 points
6. Upper critical values of the normal PPCC distribution

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions
1.3.6.7. Tables for Probability Distributions

1.3.6.7.1. Cumulative Distribution Function of the Standard Normal Distribution

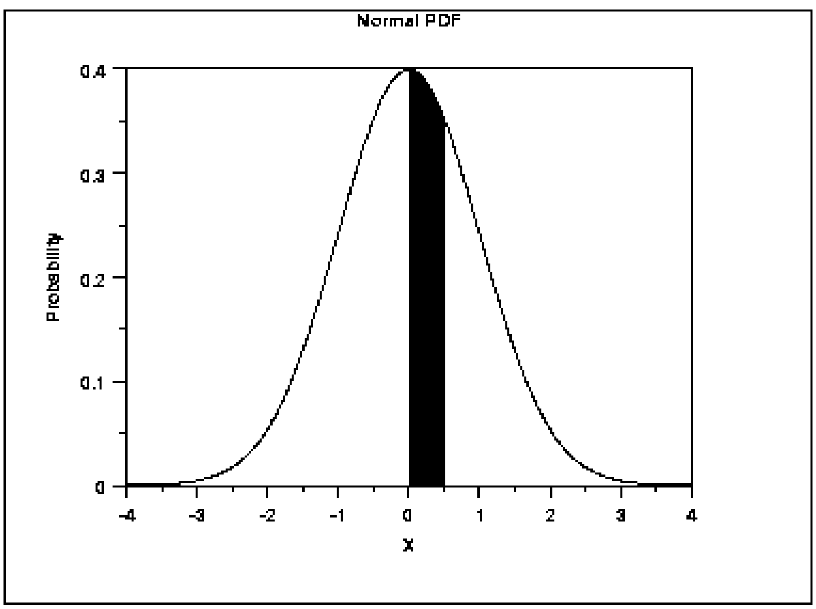
How to Use This Table

The table below contains the area under the standard normal curve from 0 to *z*. This can be used to compute the cumulative distribution function values for the standard normal distribution.

The table utilizes the symmetry of the normal distribution, so what in fact is given is

$$P[0 \leq x \leq |a|]$$

where *a* is the value of interest. This is demonstrated in the graph below for *a* = 0.5. The shaded area of the curve represents the probability that *x* is between 0 and *a*.



This can be clarified by a few simple examples.

1. What is the probability that *x* is less than or equal to 1.53? Look for 1.5 in the X column, go right to the 0.03 column to find the value 0.43699. Now add 0.5 (for the probability less than zero) to obtain the final result of 0.93699.
2. What is the probability that *x* is less than or equal to -1.53? For negative values, use the relationship

$$P[x \leq a] = 1 - P[x \leq |a|] \quad \text{for } x < 0$$

From the first example, this gives 1 - 0.93699 = 0.06301.

3. What is the probability that *x* is between -1 and 0.5? Look up the values for 0.5 (0.5 + 0.19146 = 0.69146) and -1 (1 - (0.5 + 0.34134) = 0.15866). Then subtract the results (0.69146 - 0.15866) to obtain the result 0.5328.

To use this table with a non-standard normal distribution (either the location parameter is not 0 or the scale parameter is not 1), standardize your value by subtracting the mean and dividing the result by the standard deviation. Then look up the value for this standardized value.

A few particularly important numbers derived from the table below, specifically numbers that are commonly used in significance tests, are summarized in the following table:

p	0.001	0.005	0.010	0.025	0.050	0.100
Z_p	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282

p	0.999	0.995	0.990	0.975	0.950	0.900
Z_p	+3.090	+2.576	+2.326	+1.960	+1.645	+1.282

These are critical values for the normal distribution.

Area under the Normal Curve from 0 to X

X	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189

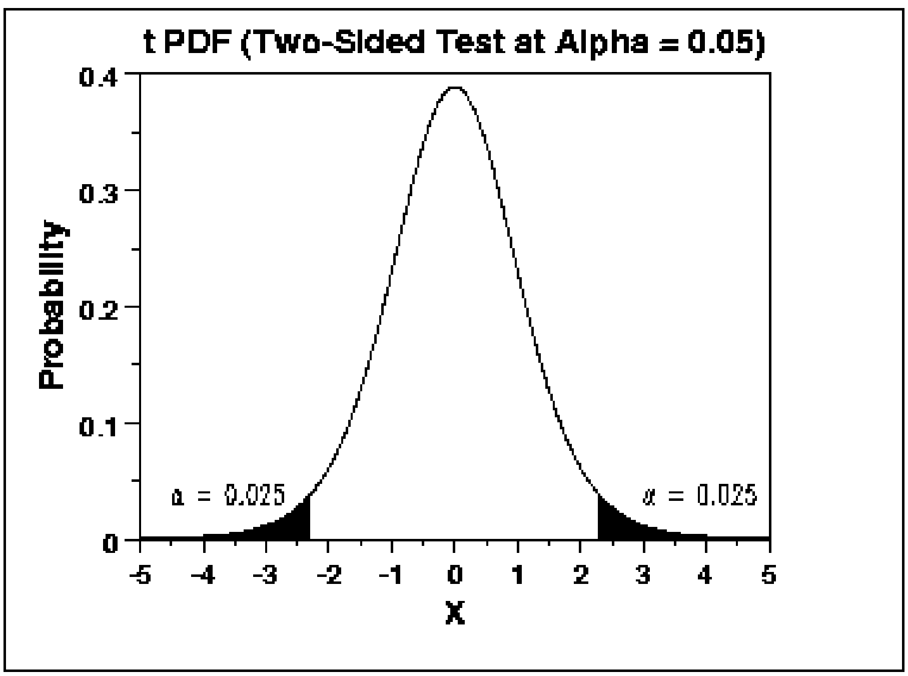
1.5	**0.43319**	**0.43448**	**0.43574**	**0.43699**	**0.43822**	**0.43943**	**0.44062**	**0.44179**	**0.44295**	**0.44408**
1.6	**0.44520**	**0.44630**	**0.44738**	**0.44845**	**0.44950**	**0.45053**	**0.45154**	**0.45254**	**0.45352**	**0.45449**
1.7	**0.45543**	**0.45637**	**0.45728**	**0.45818**	**0.45907**	**0.45994**	**0.46080**	**0.46164**	**0.46246**	**0.46327**
1.8	**0.46407**	**0.46485**	**0.46562**	**0.46638**	**0.46712**	**0.46784**	**0.46856**	**0.46926**	**0.46995**	**0.47062**
1.9	**0.47128**	**0.47193**	**0.47257**	**0.47320**	**0.47381**	**0.47441**	**0.47500**	**0.47558**	**0.47615**	**0.47670**
2.0	**0.47725**	**0.47778**	**0.47831**	**0.47882**	**0.47932**	**0.47982**	**0.48030**	**0.48077**	**0.48124**	**0.48169**
2.1	**0.48214**	**0.48257**	**0.48300**	**0.48341**	**0.48382**	**0.48422**	**0.48461**	**0.48500**	**0.48537**	**0.48574**
2.2	**0.48610**	**0.48645**	**0.48679**	**0.48713**	**0.48745**	**0.48778**	**0.48809**	**0.48840**	**0.48870**	**0.48899**
2.3	**0.48928**	**0.48956**	**0.48983**	**0.49010**	**0.49036**	**0.49061**	**0.49086**	**0.49111**	**0.49134**	**0.49158**
2.4	**0.49180**	**0.49202**	**0.49224**	**0.49245**	**0.49266**	**0.49286**	**0.49305**	**0.49324**	**0.49343**	**0.49361**
2.5	**0.49379**	**0.49396**	**0.49413**	**0.49430**	**0.49446**	**0.49461**	**0.49477**	**0.49492**	**0.49506**	**0.49520**
2.6	**0.49534**	**0.49547**	**0.49560**	**0.49573**	**0.49585**	**0.49598**	**0.49609**	**0.49621**	**0.49632**	**0.49643**
2.7	**0.49653**	**0.49664**	**0.49674**	**0.49683**	**0.49693**	**0.49702**	**0.49711**	**0.49720**	**0.49728**	**0.49736**
2.8	**0.49744**	**0.49752**	**0.49760**	**0.49767**	**0.49774**	**0.49781**	**0.49788**	**0.49795**	**0.49801**	**0.49807**
2.9	**0.49813**	**0.49819**	**0.49825**	**0.49831**	**0.49836**	**0.49841**	**0.49846**	**0.49851**	**0.49856**	**0.49861**
3.0	**0.49865**	**0.49869**	**0.49874**	**0.49878**	**0.49882**	**0.49886**	**0.49889**	**0.49893**	**0.49896**	**0.49900**
3.1	**0.49903**	**0.49906**	**0.49910**	**0.49913**	**0.49916**	**0.49918**	**0.49921**	**0.49924**	**0.49926**	**0.49929**
3.2	**0.49931**	**0.49934**	**0.49936**	**0.49938**	**0.49940**	**0.49942**	**0.49944**	**0.49946**	**0.49948**	**0.49950**
3.3	**0.49952**	**0.49953**	**0.49955**	**0.49957**	**0.49958**	**0.49960**	**0.49961**	**0.49962**	**0.49964**	**0.49965**
3.4	**0.49966**	**0.49968**	**0.49969**	**0.49970**	**0.49971**	**0.49972**	**0.49973**	**0.49974**	**0.49975**	**0.49976**
3.5	**0.49977**	**0.49978**	**0.49978**	**0.49979**	**0.49980**	**0.49981**	**0.49981**	**0.49982**	**0.49983**	**0.49983**
3.6	**0.49984**	**0.49985**	**0.49985**	**0.49986**	**0.49986**	**0.49987**	**0.49987**	**0.49988**	**0.49988**	**0.49989**
3.7	**0.49989**	**0.49990**	**0.49990**	**0.49990**	**0.49991**	**0.49991**	**0.49992**	**0.49992**	**0.49992**	**0.49992**
3.8	**0.49993**	**0.49993**	**0.49993**	**0.49994**	**0.49994**	**0.49994**	**0.49994**	**0.49995**	**0.49995**	**0.49995**
3.9	**0.49995**	**0.49995**	**0.49996**	**0.49996**	**0.49996**	**0.49996**	**0.49996**	**0.49996**	**0.49997**	**0.49997**
4.0	**0.49997**	**0.49997**	**0.49997**	**0.49997**	**0.49997**	**0.49997**	**0.49998**	**0.49998**	**0.49998**	**0.49998**

1.3.6.7.2. Upper Critical Values of the Student's-t Distribution

How to Use This Table

This table contains the upper critical values of the Student's *t*-distribution. The upper critical values are computed using the percent point function. Due to the symmetry of the *t*-distribution, this table can be used for both 1-sided (lower and upper) and 2-sided tests using the appropriate value of α.

The significance level, α, is demonstrated with the graph below which plots a t distribution with 10 degrees of freedom. The most commonly used significance level is α = 0.05. For a two-sided test, we compute the percent point function at $\alpha/2$ (0.025). If the absolute value of the test statistic is greater than the upper critical value (0.025), then we reject the null hypothesis. Due to the symmetry of the *t*-distribution, we only tabulate the upper critical values in the table below.



Given a specified value for α:

1. For a two-sided test, find the column corresponding to $\alpha/2$ and reject the null hypothesis if the absolute value of the test statistic is greater than the value of $t_{\alpha/2}$ in the table below.
2. For an upper one-sided test, find the column corresponding to α and reject the null hypothesis if the test statistic is greater than the tabled value.
3. For an lower one-sided test, find the column corresponding to α and reject the null hypothesis if the test statistic is less than the negative of the tabled value.

Upper critical values of Student's t distribution with ν degrees of freedom

	Probability of exceeding the critical value					
ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385

31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365
33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307
41.	1.303	1.683	2.020	2.421	2.701	3.301
42.	1.302	1.682	2.018	2.418	2.698	3.296
43.	1.302	1.681	2.017	2.416	2.695	3.291
44.	1.301	1.680	2.015	2.414	2.692	3.286
45.	1.301	1.679	2.014	2.412	2.690	3.281
46.	1.300	1.679	2.013	2.410	2.687	3.277
47.	1.300	1.678	2.012	2.408	2.685	3.273
48.	1.299	1.677	2.011	2.407	2.682	3.269
49.	1.299	1.677	2.010	2.405	2.680	3.265
50.	1.299	1.676	2.009	2.403	2.678	3.261
51.	1.298	1.675	2.008	2.402	2.676	3.258
52.	1.298	1.675	2.007	2.400	2.674	3.255
53.	1.298	1.674	2.006	2.399	2.672	3.251
54.	1.297	1.674	2.005	2.397	2.670	3.248
55.	1.297	1.673	2.004	2.396	2.668	3.245
56.	1.297	1.673	2.003	2.395	2.667	3.242
57.	1.297	1.672	2.002	2.394	2.665	3.239
58.	1.296	1.672	2.002	2.392	2.663	3.237
59.	1.296	1.671	2.001	2.391	2.662	3.234
60.	1.296	1.671	2.000	2.390	2.660	3.232
61.	1.296	1.670	2.000	2.389	2.659	3.229
62.	1.295	1.670	1.999	2.388	2.657	3.227
63.	1.295	1.669	1.998	2.387	2.656	3.225
64.	1.295	1.669	1.998	2.386	2.655	3.223
65.	1.295	1.669	1.997	2.385	2.654	3.220
66.	1.295	1.668	1.997	2.384	2.652	3.218
67.	1.294	1.668	1.996	2.383	2.651	3.216
68.	1.294	1.668	1.995	2.382	2.650	3.214
69.	1.294	1.667	1.995	2.382	2.649	3.213
70.	1.294	1.667	1.994	2.381	2.648	3.211
71.	1.294	1.667	1.994	2.380	2.647	3.209

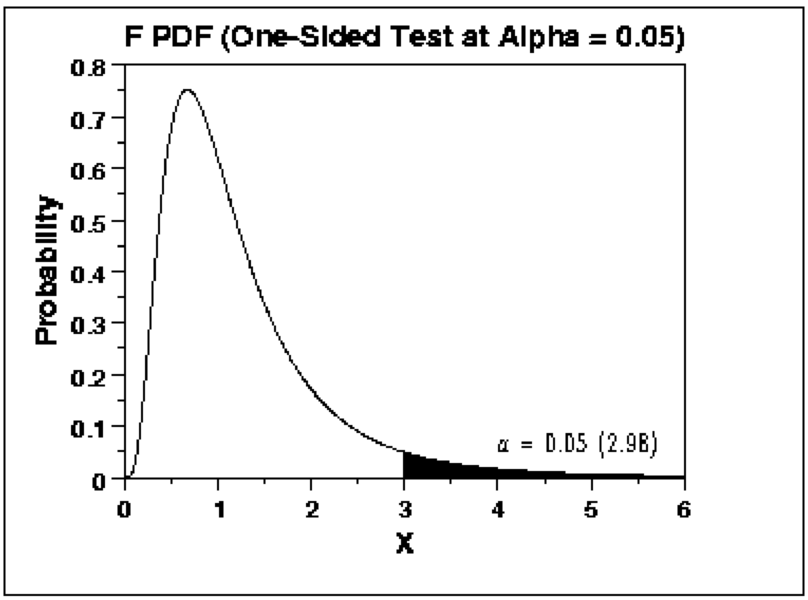
72.	1.293	1.666	1.993	2.379	2.646	3.207
73.	1.293	1.666	1.993	2.379	2.645	3.206
74.	1.293	1.666	1.993	2.378	2.644	3.204
75.	1.293	1.665	1.992	2.377	2.643	3.202
76.	1.293	1.665	1.992	2.376	2.642	3.201
77.	1.293	1.665	1.991	2.376	2.641	3.199
78.	1.292	1.665	1.991	2.375	2.640	3.198
79.	1.292	1.664	1.990	2.374	2.640	3.197
80.	1.292	1.664	1.990	2.374	2.639	3.195
81.	1.292	1.664	1.990	2.373	2.638	3.194
82.	1.292	1.664	1.989	2.373	2.637	3.193
83.	1.292	1.663	1.989	2.372	2.636	3.191
84.	1.292	1.663	1.989	2.372	2.636	3.190
85.	1.292	1.663	1.988	2.371	2.635	3.189
86.	1.291	1.663	1.988	2.370	2.634	3.188
87.	1.291	1.663	1.988	2.370	2.634	3.187
88.	1.291	1.662	1.987	2.369	2.633	3.185
89.	1.291	1.662	1.987	2.369	2.632	3.184
90.	1.291	1.662	1.987	2.368	2.632	3.183
91.	1.291	1.662	1.986	2.368	2.631	3.182
92.	1.291	1.662	1.986	2.368	2.630	3.181
93.	1.291	1.661	1.986	2.367	2.630	3.180
94.	1.291	1.661	1.986	2.367	2.629	3.179
95.	1.291	1.661	1.985	2.366	2.629	3.178
96.	1.290	1.661	1.985	2.366	2.628	3.177
97.	1.290	1.661	1.985	2.365	2.627	3.176
98.	1.290	1.661	1.984	2.365	2.627	3.175
99.	1.290	1.660	1.984	2.365	2.626	3.175
100.	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

1.3.6.7.3. Upper Critical Values of the F Distribution

How to Use This Table

This table contains the upper critical values of the F distribution. This table is used for one-sided F tests at the α = 0.05, 0.10, and 0.01 levels.

More specifically, a test statistic is computed with ν_1 and ν_2 degrees of freedom, and the result is compared to this table. For a one-sided test, the null hypothesis is rejected when the test statistic is greater than the tabled value. This is demonstrated with the graph of an F distribution with ν_1 = 10 and ν_2 = 10. The shaded area of the graph indicates the rejection region at the α significance level. Since this is a one-sided test, we have α probability in the upper tail of exceeding the critical value and zero in the lower tail. Because the F distribution is asymmetric, a two-sided test requires a set of of tables (not included here) that contain the rejection regions for both the lower and upper tails.



Contents

The following tables for ν_2 from 1 to 100 are included:

1. One sided, 5% significance level, ν_1 = 1 - 10
2. One sided, 5% significance level, ν_1 = 11 - 20
3. One sided, 10% significance level, ν_1 = 1 - 10
4. One sided, 10% significance level, ν_1 = 11 - 20
5. One sided, 1% significance level, ν_1 = 1 - 10
6. One sided, 1% significance level, ν_1 = 11 - 20

Upper critical values of the F distribution

for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

5% significance level

$$F_{.05}(\nu_1, \nu_2)$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.882	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544

16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091

39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118	2.071
42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112	2.065
43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106	2.059
44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101	2.054
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049
46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091	2.044
47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086	2.039
48	4.043	3.191	2.798	2.565	2.409	2.295	2.207	2.138	2.082	2.035
49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077	2.030
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
51	4.030	3.179	2.786	2.553	2.397	2.283	2.195	2.126	2.069	2.022
52	4.027	3.175	2.783	2.550	2.393	2.279	2.192	2.122	2.066	2.018
53	4.023	3.172	2.779	2.546	2.389	2.275	2.188	2.119	2.062	2.015
54	4.020	3.168	2.776	2.543	2.386	2.272	2.185	2.115	2.059	2.011
55	4.016	3.165	2.773	2.540	2.383	2.269	2.181	2.112	2.055	2.008
56	4.013	3.162	2.769	2.537	2.380	2.266	2.178	2.109	2.052	2.005
57	4.010	3.159	2.766	2.534	2.377	2.263	2.175	2.106	2.049	2.001
58	4.007	3.156	2.764	2.531	2.374	2.260	2.172	2.103	2.046	1.998
59	4.004	3.153	2.761	2.528	2.371	2.257	2.169	2.100	2.043	1.995
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
61	3.998	3.148	2.755	2.523	2.366	2.251	2.164	2.094	2.037	1.990

62	3.996	3.145	2.753	2.520	2.363	2.249	2.161	2.092	2.035	1.987
63	3.993	3.143	2.751	2.518	2.361	2.246	2.159	2.089	2.032	1.985
64	3.991	3.140	2.748	2.515	2.358	2.244	2.156	2.087	2.030	1.982
65	3.989	3.138	2.746	2.513	2.356	2.242	2.154	2.084	2.027	1.980
66	3.986	3.136	2.744	2.511	2.354	2.239	2.152	2.082	2.025	1.977
67	3.984	3.134	2.742	2.509	2.352	2.237	2.150	2.080	2.023	1.975
68	3.982	3.132	2.740	2.507	2.350	2.235	2.148	2.078	2.021	1.973
69	3.980	3.130	2.737	2.505	2.348	2.233	2.145	2.076	2.019	1.971
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
71	3.976	3.126	2.734	2.501	2.344	2.229	2.142	2.072	2.015	1.967
72	3.974	3.124	2.732	2.499	2.342	2.227	2.140	2.070	2.013	1.965
73	3.972	3.122	2.730	2.497	2.340	2.226	2.138	2.068	2.011	1.963
74	3.970	3.120	2.728	2.495	2.338	2.224	2.136	2.066	2.009	1.961
75	3.968	3.119	2.727	2.494	2.337	2.222	2.134	2.064	2.007	1.959
76	3.967	3.117	2.725	2.492	2.335	2.220	2.133	2.063	2.006	1.958
77	3.965	3.115	2.723	2.490	2.333	2.219	2.131	2.061	2.004	1.956
78	3.963	3.114	2.722	2.489	2.332	2.217	2.129	2.059	2.002	1.954
79	3.962	3.112	2.720	2.487	2.330	2.216	2.128	2.058	2.001	1.953
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
81	3.959	3.109	2.717	2.484	2.327	2.213	2.125	2.055	1.998	1.950
82	3.957	3.108	2.716	2.483	2.326	2.211	2.123	2.053	1.996	1.948
83	3.956	3.107	2.715	2.482	2.324	2.210	2.122	2.052	1.995	1.947
84	3.955	3.105	2.713	2.480	2.323	2.209	2.121	2.051	1.993	1.945

85	3.953	3.104	2.712	2.479	2.322	2.207	2.119	2.049	1.992	1.944
86	3.952	3.103	2.711	2.478	2.321	2.206	2.118	2.048	1.991	1.943
87	3.951	3.101	2.709	2.476	2.319	2.205	2.117	2.047	1.989	1.941
88	3.949	3.100	2.708	2.475	2.318	2.203	2.115	2.045	1.988	1.940
89	3.948	3.099	2.707	2.474	2.317	2.202	2.114	2.044	1.987	1.939
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
91	3.946	3.097	2.705	2.472	2.315	2.200	2.112	2.042	1.984	1.936
92	3.945	3.095	2.704	2.471	2.313	2.199	2.111	2.041	1.983	1.935
93	3.943	3.094	2.703	2.470	2.312	2.198	2.110	2.040	1.982	1.934
94	3.942	3.093	2.701	2.469	2.311	2.197	2.109	2.038	1.981	1.933
95	3.941	3.092	2.700	2.467	2.310	2.196	2.108	2.037	1.980	1.932
96	3.940	3.091	2.699	2.466	2.309	2.195	2.106	2.036	1.979	1.931
97	3.939	3.090	2.698	2.465	2.308	2.194	2.105	2.035	1.978	1.930
98	3.938	3.089	2.697	2.465	2.307	2.193	2.104	2.034	1.977	1.929
99	3.937	3.088	2.696	2.464	2.306	2.192	2.103	2.033	1.976	1.928
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927

ν_2 \ ν_1	11	12	13	14	15	16	17	18	19	20
1	242.983	243.906	244.690	245.364	245.950	246.464	246.918	247.323	247.686	248.013
2	19.405	19.413	19.419	19.424	19.429	19.433	19.437	19.440	19.443	19.446
3	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660

4	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803
5	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558
6	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874
7	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445
8	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150
9	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936
10	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774
11	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646
12	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544
13	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459
14	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388
15	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328
16	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276
17	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230
18	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191
19	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155
20	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124
21	2.283	2.250	2.222	2.197	2.176	2.156	2.139	2.123	2.109	2.096
22	2.259	2.226	2.198	2.173	2.151	2.131	2.114	2.098	2.084	2.071
23	2.236	2.204	2.175	2.150	2.128	2.109	2.091	2.075	2.061	2.048
24	2.216	2.183	2.155	2.130	2.108	2.088	2.070	2.054	2.040	2.027
25	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007
26	2.181	2.148	2.119	2.094	2.072	2.052	2.034	2.018	2.003	1.990

27	2.166	2.132	2.103	2.078	2.056	2.036	2.018	2.002	1.987	1.974
28	2.151	2.118	2.089	2.064	2.041	2.021	2.003	1.987	1.972	1.959
29	2.138	2.104	2.075	2.050	2.027	2.007	1.989	1.973	1.958	1.945
30	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932
31	2.114	2.080	2.051	2.026	2.003	1.983	1.965	1.948	1.933	1.920
32	2.103	2.070	2.040	2.015	1.992	1.972	1.953	1.937	1.922	1.908
33	2.093	2.060	2.030	2.004	1.982	1.961	1.943	1.926	1.911	1.898
34	2.084	2.050	2.021	1.995	1.972	1.952	1.933	1.917	1.902	1.888
35	2.075	2.041	2.012	1.986	1.963	1.942	1.924	1.907	1.892	1.878
36	2.067	2.033	2.003	1.977	1.954	1.934	1.915	1.899	1.883	1.870
37	2.059	2.025	1.995	1.969	1.946	1.926	1.907	1.890	1.875	1.861
38	2.051	2.017	1.988	1.962	1.939	1.918	1.899	1.883	1.867	1.853
39	2.044	2.010	1.981	1.954	1.931	1.911	1.892	1.875	1.860	1.846
40	2.038	2.003	1.974	1.948	1.924	1.904	1.885	1.868	1.853	1.839
41	2.031	1.997	1.967	1.941	1.918	1.897	1.879	1.862	1.846	1.832
42	2.025	1.991	1.961	1.935	1.912	1.891	1.872	1.855	1.840	1.826
43	2.020	1.985	1.955	1.929	1.906	1.885	1.866	1.849	1.834	1.820
44	2.014	1.980	1.950	1.924	1.900	1.879	1.861	1.844	1.828	1.814
45	2.009	1.974	1.945	1.918	1.895	1.874	1.855	1.838	1.823	1.808
46	2.004	1.969	1.940	1.913	1.890	1.869	1.850	1.833	1.817	1.803
47	1.999	1.965	1.935	1.908	1.885	1.864	1.845	1.828	1.812	1.798
48	1.995	1.960	1.930	1.904	1.880	1.859	1.840	1.823	1.807	1.793
49	1.990	1.956	1.926	1.899	1.876	1.855	1.836	1.819	1.803	1.789

50	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784
51	1.982	1.947	1.917	1.891	1.867	1.846	1.827	1.810	1.794	1.780
52	1.978	1.944	1.913	1.887	1.863	1.842	1.823	1.806	1.790	1.776
53	1.975	1.940	1.910	1.883	1.859	1.838	1.819	1.802	1.786	1.772
54	1.971	1.936	1.906	1.879	1.856	1.835	1.816	1.798	1.782	1.768
55	1.968	1.933	1.903	1.876	1.852	1.831	1.812	1.795	1.779	1.764
56	1.964	1.930	1.899	1.873	1.849	1.828	1.809	1.791	1.775	1.761
57	1.961	1.926	1.896	1.869	1.846	1.824	1.805	1.788	1.772	1.757
58	1.958	1.923	1.893	1.866	1.842	1.821	1.802	1.785	1.769	1.754
59	1.955	1.920	1.890	1.863	1.839	1.818	1.799	1.781	1.766	1.751
60	1.952	1.917	1.887	1.860	1.836	1.815	1.796	1.778	1.763	1.748
61	1.949	1.915	1.884	1.857	1.834	1.812	1.793	1.776	1.760	1.745
62	1.947	1.912	1.882	1.855	1.831	1.809	1.790	1.773	1.757	1.742
63	1.944	1.909	1.879	1.852	1.828	1.807	1.787	1.770	1.754	1.739
64	1.942	1.907	1.876	1.849	1.826	1.804	1.785	1.767	1.751	1.737
65	1.939	1.904	1.874	1.847	1.823	1.802	1.782	1.765	1.749	1.734
66	1.937	1.902	1.871	1.845	1.821	1.799	1.780	1.762	1.746	1.732
67	1.935	1.900	1.869	1.842	1.818	1.797	1.777	1.760	1.744	1.729
68	1.932	1.897	1.867	1.840	1.816	1.795	1.775	1.758	1.742	1.727
69	1.930	1.895	1.865	1.838	1.814	1.792	1.773	1.755	1.739	1.725
70	1.928	1.893	1.863	1.836	1.812	1.790	1.771	1.753	1.737	1.722
71	1.926	1.891	1.861	1.834	1.810	1.788	1.769	1.751	1.735	1.720
72	1.924	1.889	1.859	1.832	1.808	1.786	1.767	1.749	1.733	1.718

73	1.922	1.887	1.857	1.830	1.806	1.784	1.765	1.747	1.731	1.716
74	1.921	1.885	1.855	1.828	1.804	1.782	1.763	1.745	1.729	1.714
75	1.919	1.884	1.853	1.826	1.802	1.780	1.761	1.743	1.727	1.712
76	1.917	1.882	1.851	1.824	1.800	1.778	1.759	1.741	1.725	1.710
77	1.915	1.880	1.849	1.822	1.798	1.777	1.757	1.739	1.723	1.708
78	1.914	1.878	1.848	1.821	1.797	1.775	1.755	1.738	1.721	1.707
79	1.912	1.877	1.846	1.819	1.795	1.773	1.754	1.736	1.720	1.705
80	1.910	1.875	1.845	1.817	1.793	1.772	1.752	1.734	1.718	1.703
81	1.909	1.874	1.843	1.816	1.792	1.770	1.750	1.733	1.716	1.702
82	1.907	1.872	1.841	1.814	1.790	1.768	1.749	1.731	1.715	1.700
83	1.906	1.871	1.840	1.813	1.789	1.767	1.747	1.729	1.713	1.698
84	1.905	1.869	1.838	1.811	1.787	1.765	1.746	1.728	1.712	1.697
85	1.903	1.868	1.837	1.810	1.786	1.764	1.744	1.726	1.710	1.695
86	1.902	1.867	1.836	1.808	1.784	1.762	1.743	1.725	1.709	1.694
87	1.900	1.865	1.834	1.807	1.783	1.761	1.741	1.724	1.707	1.692
88	1.899	1.864	1.833	1.806	1.782	1.760	1.740	1.722	1.706	1.691
89	1.898	1.863	1.832	1.804	1.780	1.758	1.739	1.721	1.705	1.690
90	1.897	1.861	1.830	1.803	1.779	1.757	1.737	1.720	1.703	1.688
91	1.895	1.860	1.829	1.802	1.778	1.756	1.736	1.718	1.702	1.687
92	1.894	1.859	1.828	1.801	1.776	1.755	1.735	1.717	1.701	1.686
93	1.893	1.858	1.827	1.800	1.775	1.753	1.734	1.716	1.699	1.684
94	1.892	1.857	1.826	1.798	1.774	1.752	1.733	1.715	1.698	1.683
95	1.891	1.856	1.825	1.797	1.773	1.751	1.731	1.713	1.697	1.682

96	1.890	1.854	1.823	1.796	1.772	1.750	1.730	1.712	1.696	1.681
97	1.889	1.853	1.822	1.795	1.771	1.749	1.729	1.711	1.695	1.680
98	1.888	1.852	1.821	1.794	1.770	1.748	1.728	1.710	1.694	1.679
99	1.887	1.851	1.820	1.793	1.769	1.747	1.727	1.709	1.693	1.678
100	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676

Upper critical values of the F distribution

for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

10% significance level

$$F_{.10}(\nu_1, \nu_2)$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.858	60.195
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416

10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
31	2.875	2.482	2.270	2.136	2.042	1.973	1.920	1.877	1.842	1.812
32	2.869	2.477	2.263	2.129	2.036	1.967	1.913	1.870	1.835	1.805

33	2.864	2.471	2.258	2.123	2.030	1.961	1.907	1.864	1.828	1.799
34	2.859	2.466	2.252	2.118	2.024	1.955	1.901	1.858	1.822	1.793
35	2.855	2.461	2.247	2.113	2.019	1.950	1.896	1.852	1.817	1.787
36	2.850	2.456	2.243	2.108	2.014	1.945	1.891	1.847	1.811	1.781
37	2.846	2.452	2.238	2.103	2.009	1.940	1.886	1.842	1.806	1.776
38	2.842	2.448	2.234	2.099	2.005	1.935	1.881	1.838	1.802	1.772
39	2.839	2.444	2.230	2.095	2.001	1.931	1.877	1.833	1.797	1.767
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763
41	2.832	2.437	2.222	2.087	1.993	1.923	1.869	1.825	1.789	1.759
42	2.829	2.434	2.219	2.084	1.989	1.919	1.865	1.821	1.785	1.755
43	2.826	2.430	2.216	2.080	1.986	1.916	1.861	1.817	1.781	1.751
44	2.823	2.427	2.213	2.077	1.983	1.913	1.858	1.814	1.778	1.747
45	2.820	2.425	2.210	2.074	1.980	1.909	1.855	1.811	1.774	1.744
46	2.818	2.422	2.207	2.071	1.977	1.906	1.852	1.808	1.771	1.741
47	2.815	2.419	2.204	2.068	1.974	1.903	1.849	1.805	1.768	1.738
48	2.813	2.417	2.202	2.066	1.971	1.901	1.846	1.802	1.765	1.735
49	2.811	2.414	2.199	2.063	1.968	1.898	1.843	1.799	1.763	1.732
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
51	2.807	2.410	2.194	2.058	1.964	1.893	1.838	1.794	1.757	1.727
52	2.805	2.408	2.192	2.056	1.961	1.891	1.836	1.791	1.755	1.724
53	2.803	2.406	2.190	2.054	1.959	1.888	1.833	1.789	1.752	1.722
54	2.801	2.404	2.188	2.052	1.957	1.886	1.831	1.787	1.750	1.719
55	2.799	2.402	2.186	2.050	1.955	1.884	1.829	1.785	1.748	1.717

56	2.797	2.400	2.184	2.048	1.953	1.882	1.827		
1.782	1.746	1.715							
57	2.796	2.398	2.182	2.046	1.951	1.880	1.825		
1.780	1.744	1.713							
58	2.794	2.396	2.181	2.044	1.949	1.878	1.823		
1.779	1.742	1.711							
59	2.793	2.395	2.179	2.043	1.947	1.876	1.821		
1.777	1.740	1.709							
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819		
1.775	1.738	1.707							
61	2.790	2.392	2.176	2.039	1.944	1.873	1.818		
1.773	1.736	1.705							
62	2.788	2.390	2.174	2.038	1.942	1.871	1.816		
1.771	1.735	1.703							
63	2.787	2.389	2.173	2.036	1.941	1.870	1.814		
1.770	1.733	1.702							
64	2.786	2.387	2.171	2.035	1.939	1.868	1.813		
1.768	1.731	1.700							
65	2.784	2.386	2.170	2.033	1.938	1.867	1.811		
1.767	1.730	1.699							
66	2.783	2.385	2.169	2.032	1.937	1.865	1.810		
1.765	1.728	1.697							
67	2.782	2.384	2.167	2.031	1.935	1.864	1.808		
1.764	1.727	1.696							
68	2.781	2.382	2.166	2.029	1.934	1.863	1.807		
1.762	1.725	1.694							
69	2.780	2.381	2.165	2.028	1.933	1.861	1.806		
1.761	1.724	1.693							
70	2.779	2.380	2.164	2.027	1.931	1.860	1.804		
1.760	1.723	1.691							
71	2.778	2.379	2.163	2.026	1.930	1.859	1.803		
1.758	1.721	1.690							
72	2.777	2.378	2.161	2.025	1.929	1.858	1.802		
1.757	1.720	1.689							
73	2.776	2.377	2.160	2.024	1.928	1.856	1.801		
1.756	1.719	1.687							
74	2.775	2.376	2.159	2.022	1.927	1.855	1.800		
1.755	1.718	1.686							
75	2.774	2.375	2.158	2.021	1.926	1.854	1.798		
1.754	1.716	1.685							
76	2.773	2.374	2.157	2.020	1.925	1.853	1.797		
1.752	1.715	1.684							
77	2.772	2.373	2.156	2.019	1.924	1.852	1.796		
1.751	1.714	1.683							
78	2.771	2.372	2.155	2.018	1.923	1.851	1.795		
1.750	1.713	1.682							

79	2.770	2.371	2.154	2.017	1.922	1.850	1.794	1.749	1.712	1.681
80	2.769	2.370	2.154	2.016	1.921	1.849	1.793	1.748	1.711	1.680
81	2.769	2.369	2.153	2.016	1.920	1.848	1.792	1.747	1.710	1.679
82	2.768	2.368	2.152	2.015	1.919	1.847	1.791	1.746	1.709	1.678
83	2.767	2.368	2.151	2.014	1.918	1.846	1.790	1.745	1.708	1.677
84	2.766	2.367	2.150	2.013	1.917	1.845	1.790	1.744	1.707	1.676
85	2.765	2.366	2.149	2.012	1.916	1.845	1.789	1.744	1.706	1.675
86	2.765	2.365	2.149	2.011	1.915	1.844	1.788	1.743	1.705	1.674
87	2.764	2.365	2.148	2.011	1.915	1.843	1.787	1.742	1.705	1.673
88	2.763	2.364	2.147	2.010	1.914	1.842	1.786	1.741	1.704	1.672
89	2.763	2.363	2.146	2.009	1.913	1.841	1.785	1.740	1.703	1.671
90	2.762	2.363	2.146	2.008	1.912	1.841	1.785	1.739	1.702	1.670
91	2.761	2.362	2.145	2.008	1.912	1.840	1.784	1.739	1.701	1.670
92	2.761	2.361	2.144	2.007	1.911	1.839	1.783	1.738	1.701	1.669
93	2.760	2.361	2.144	2.006	1.910	1.838	1.782	1.737	1.700	1.668
94	2.760	2.360	2.143	2.006	1.910	1.838	1.782	1.736	1.699	1.667
95	2.759	2.359	2.142	2.005	1.909	1.837	1.781	1.736	1.698	1.667
96	2.759	2.359	2.142	2.004	1.908	1.836	1.780	1.735	1.698	1.666
97	2.758	2.358	2.141	2.004	1.908	1.836	1.780	1.734	1.697	1.665
98	2.757	2.358	2.141	2.003	1.907	1.835	1.779	1.734	1.696	1.665
99	2.757	2.357	2.140	2.003	1.906	1.835	1.778	1.733	1.696	1.664
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663

ν_2 \ ν_1	11	12	13	14	15	16	17	18	19	20
1	60.473	60.705	60.903	61.073	61.220	61.350	61.464	61.566	61.658	61.740
2	9.401	9.408	9.415	9.420	9.425	9.429	9.433	9.436	9.439	9.441
3	5.222	5.216	5.210	5.205	5.200	5.196	5.193	5.190	5.187	5.184
4	3.907	3.896	3.886	3.878	3.870	3.864	3.858	3.853	3.849	3.844
5	3.282	3.268	3.257	3.247	3.238	3.230	3.223	3.217	3.212	3.207
6	2.920	2.905	2.892	2.881	2.871	2.863	2.855	2.848	2.842	2.836
7	2.684	2.668	2.654	2.643	2.632	2.623	2.615	2.607	2.601	2.595
8	2.519	2.502	2.488	2.475	2.464	2.455	2.446	2.438	2.431	2.425
9	2.396	2.379	2.364	2.351	2.340	2.329	2.320	2.312	2.305	2.298
10	2.302	2.284	2.269	2.255	2.244	2.233	2.224	2.215	2.208	2.201
11	2.227	2.209	2.193	2.179	2.167	2.156	2.147	2.138	2.130	2.123
12	2.166	2.147	2.131	2.117	2.105	2.094	2.084	2.075	2.067	2.060
13	2.116	2.097	2.080	2.066	2.053	2.042	2.032	2.023	2.014	2.007
14	2.073	2.054	2.037	2.022	2.010	1.998	1.988	1.978	1.970	1.962
15	2.037	2.017	2.000	1.985	1.972	1.961	1.950	1.941	1.932	1.924
16	2.005	1.985	1.968	1.953	1.940	1.928	1.917	1.908	1.899	1.891
17	1.978	1.958	1.940	1.925	1.912	1.900	1.889	1.879	1.870	1.862
18	1.954	1.933	1.916	1.900	1.887	1.875	1.864	1.854	1.845	1.837
19	1.932	1.912	1.894	1.878	1.865	1.852	1.841	1.831	1.822	1.814
20	1.913	1.892	1.875	1.859	1.845	1.833	1.821	1.811	1.802	1.794

21	1.896	1.875	1.857	1.841	1.827	1.815	1.803	1.793	1.784	1.776
22	1.880	1.859	1.841	1.825	1.811	1.798	1.787	1.777	1.768	1.759
23	1.866	1.845	1.827	1.811	1.796	1.784	1.772	1.762	1.753	1.744
24	1.853	1.832	1.814	1.797	1.783	1.770	1.759	1.748	1.739	1.730
25	1.841	1.820	1.802	1.785	1.771	1.758	1.746	1.736	1.726	1.718
26	1.830	1.809	1.790	1.774	1.760	1.747	1.735	1.724	1.715	1.706
27	1.820	1.799	1.780	1.764	1.749	1.736	1.724	1.714	1.704	1.695
28	1.811	1.790	1.771	1.754	1.740	1.726	1.715	1.704	1.694	1.685
29	1.802	1.781	1.762	1.745	1.731	1.717	1.705	1.695	1.685	1.676
30	1.794	1.773	1.754	1.737	1.722	1.709	1.697	1.686	1.676	1.667
31	1.787	1.765	1.746	1.729	1.714	1.701	1.689	1.678	1.668	1.659
32	1.780	1.758	1.739	1.722	1.707	1.694	1.682	1.671	1.661	1.652
33	1.773	1.751	1.732	1.715	1.700	1.687	1.675	1.664	1.654	1.645
34	1.767	1.745	1.726	1.709	1.694	1.680	1.668	1.657	1.647	1.638
35	1.761	1.739	1.720	1.703	1.688	1.674	1.662	1.651	1.641	1.632
36	1.756	1.734	1.715	1.697	1.682	1.669	1.656	1.645	1.635	1.626
37	1.751	1.729	1.709	1.692	1.677	1.663	1.651	1.640	1.630	1.620
38	1.746	1.724	1.704	1.687	1.672	1.658	1.646	1.635	1.624	1.615
39	1.741	1.719	1.700	1.682	1.667	1.653	1.641	1.630	1.619	1.610
40	1.737	1.715	1.695	1.678	1.662	1.649	1.636	1.625	1.615	1.605
41	1.733	1.710	1.691	1.673	1.658	1.644	1.632	1.620	1.610	1.601
42	1.729	1.706	1.687	1.669	1.654	1.640	1.628	1.616	1.606	1.596
43	1.725	1.703	1.683	1.665	1.650	1.636	1.624	1.612	1.602	1.592

44	1.721	1.699	1.679	1.662	1.646	1.632	1.620	1.608	1.598	1.588
45	1.718	1.695	1.676	1.658	1.643	1.629	1.616	1.605	1.594	1.585
46	1.715	1.692	1.672	1.655	1.639	1.625	1.613	1.601	1.591	1.581
47	1.712	1.689	1.669	1.652	1.636	1.622	1.609	1.598	1.587	1.578
48	1.709	1.686	1.666	1.648	1.633	1.619	1.606	1.594	1.584	1.574
49	1.706	1.683	1.663	1.645	1.630	1.616	1.603	1.591	1.581	1.571
50	1.703	1.680	1.660	1.643	1.627	1.613	1.600	1.588	1.578	1.568
51	1.700	1.677	1.658	1.640	1.624	1.610	1.597	1.586	1.575	1.565
52	1.698	1.675	1.655	1.637	1.621	1.607	1.594	1.583	1.572	1.562
53	1.695	1.672	1.652	1.635	1.619	1.605	1.592	1.580	1.570	1.560
54	1.693	1.670	1.650	1.632	1.616	1.602	1.589	1.578	1.567	1.557
55	1.691	1.668	1.648	1.630	1.614	1.600	1.587	1.575	1.564	1.555
56	1.688	1.666	1.645	1.628	1.612	1.597	1.585	1.573	1.562	1.552
57	1.686	1.663	1.643	1.625	1.610	1.595	1.582	1.571	1.560	1.550
58	1.684	1.661	1.641	1.623	1.607	1.593	1.580	1.568	1.558	1.548
59	1.682	1.659	1.639	1.621	1.605	1.591	1.578	1.566	1.555	1.546
60	1.680	1.657	1.637	1.619	1.603	1.589	1.576	1.564	1.553	1.543
61	1.679	1.656	1.635	1.617	1.601	1.587	1.574	1.562	1.551	1.541
62	1.677	1.654	1.634	1.616	1.600	1.585	1.572	1.560	1.549	1.540
63	1.675	1.652	1.632	1.614	1.598	1.583	1.570	1.558	1.548	1.538
64	1.673	1.650	1.630	1.612	1.596	1.582	1.569	1.557	1.546	1.536
65	1.672	1.649	1.628	1.610	1.594	1.580	1.567	1.555	1.544	1.534
66	1.670	1.647	1.627	1.609	1.593	1.578	1.565	1.553	1.542	1.532

67	1.669	1.646	1.625	1.607	1.591	1.577	1.564	1.552	1.541	1.531
68	1.667	1.644	1.624	1.606	1.590	1.575	1.562	1.550	1.539	1.529
69	1.666	1.643	1.622	1.604	1.588	1.574	1.560	1.548	1.538	1.527
70	1.665	1.641	1.621	1.603	1.587	1.572	1.559	1.547	1.536	1.526
71	1.663	1.640	1.619	1.601	1.585	1.571	1.557	1.545	1.535	1.524
72	1.662	1.639	1.618	1.600	1.584	1.569	1.556	1.544	1.533	1.523
73	1.661	1.637	1.617	1.599	1.583	1.568	1.555	1.543	1.532	1.522
74	1.659	1.636	1.616	1.597	1.581	1.567	1.553	1.541	1.530	1.520
75	1.658	1.635	1.614	1.596	1.580	1.565	1.552	1.540	1.529	1.519
76	1.657	1.634	1.613	1.595	1.579	1.564	1.551	1.539	1.528	1.518
77	1.656	1.632	1.612	1.594	1.578	1.563	1.550	1.538	1.527	1.516
78	1.655	1.631	1.611	1.593	1.576	1.562	1.548	1.536	1.525	1.515
79	1.654	1.630	1.610	1.592	1.575	1.561	1.547	1.535	1.524	1.514
80	1.653	1.629	1.609	1.590	1.574	1.559	1.546	1.534	1.523	1.513
81	1.652	1.628	1.608	1.589	1.573	1.558	1.545	1.533	1.522	1.512
82	1.651	1.627	1.607	1.588	1.572	1.557	1.544	1.532	1.521	1.511
83	1.650	1.626	1.606	1.587	1.571	1.556	1.543	1.531	1.520	1.509
84	1.649	1.625	1.605	1.586	1.570	1.555	1.542	1.530	1.519	1.508
85	1.648	1.624	1.604	1.585	1.569	1.554	1.541	1.529	1.518	1.507
86	1.647	1.623	1.603	1.584	1.568	1.553	1.540	1.528	1.517	1.506
87	1.646	1.622	1.602	1.583	1.567	1.552	1.539	1.527	1.516	1.505
88	1.645	1.622	1.601	1.583	1.566	1.551	1.538	1.526	1.515	1.504
89	1.644	1.621	1.600	1.582	1.565	1.550	1.537	1.525	1.514	1.503

90	1.643	1.620	1.599	1.581	1.564	1.550	1.536	1.524	1.513	1.503
91	1.643	1.619	1.598	1.580	1.564	1.549	1.535	1.523	1.512	1.502
92	1.642	1.618	1.598	1.579	1.563	1.548	1.534	1.522	1.511	1.501
93	1.641	1.617	1.597	1.578	1.562	1.547	1.534	1.521	1.510	1.500
94	1.640	1.617	1.596	1.578	1.561	1.546	1.533	1.521	1.509	1.499
95	1.640	1.616	1.595	1.577	1.560	1.545	1.532	1.520	1.509	1.498
96	1.639	1.615	1.594	1.576	1.560	1.545	1.531	1.519	1.508	1.497
97	1.638	1.614	1.594	1.575	1.559	1.544	1.530	1.518	1.507	1.497
98	1.637	1.614	1.593	1.575	1.558	1.543	1.530	1.517	1.506	1.496
99	1.637	1.613	1.592	1.574	1.557	1.542	1.529	1.517	1.505	1.495
100	1.636	1.612	1.592	1.573	1.557	1.542	1.528	1.516	1.505	1.494

Upper critical values of the F distribution

for ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

1% significance level

$$F_{.01}(\nu_1, \nu_2)$$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10
1	4052.19	4999.52	5403.34	5624.62	5763.65	5858.97	5928.33	5981.10	6022.50	6055.85
2	98.502	99.000	99.166	99.249	99.300	99.333	99.356	99.374	99.388	99.399
3	34.116	30.816	29.457	28.710	28.237	27.911	27.672			

								27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421			

								3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979
31	7.530	5.362	4.484	3.993	3.675	3.449	3.281	3.149	3.043	2.955
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934
33	7.471	5.312	4.437	3.948	3.630	3.406	3.238	3.106	3.000	2.913
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894
35	7.419	5.268	4.396	3.908	3.592	3.368	3.200	3.069	2.963	2.876
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859
37	7.373	5.229	4.360	3.873	3.558	3.334	3.167	3.036	2.930	2.843
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828
39	7.333	5.194	4.327	3.843	3.528	3.305	3.137	3.006	2.901	2.814
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
41	7.296	5.163	4.299	3.815	3.501	3.278	3.111	2.980	2.875	2.788
42	7.280	5.149	4.285	3.802	3.488	3.266	3.099	2.968	2.863	2.776
43	7.264	5.136	4.273	3.790	3.476	3.254	3.087	2.957	2.851	2.764
44	7.248	5.123	4.261	3.778	3.465	3.243	3.076	2.946	2.840	2.754
45	7.234	5.110	4.249	3.767	3.454	3.232	3.066	2.935	2.830	2.743
46	7.220	5.099	4.238	3.757	3.444	3.222	3.056	2.925	2.820	2.733
47	7.207	5.087	4.228	3.747	3.434	3.213	3.046	2.916	2.811	2.724
48	7.194	5.077	4.218	3.737	3.425	3.204	3.037	2.907	2.802	2.715
49	7.182	5.066	4.208	3.728	3.416	3.195	3.028			

								2.898	2.793	2.706
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
51	7.159	5.047	4.191	3.711	3.400	3.178	3.012	2.882	2.777	2.690
52	7.149	5.038	4.182	3.703	3.392	3.171	3.005	2.874	2.769	2.683
53	7.139	5.030	4.174	3.695	3.384	3.163	2.997	2.867	2.762	2.675
54	7.129	5.021	4.167	3.688	3.377	3.156	2.990	2.860	2.755	2.668
55	7.119	5.013	4.159	3.681	3.370	3.149	2.983	2.853	2.748	2.662
56	7.110	5.006	4.152	3.674	3.363	3.143	2.977	2.847	2.742	2.655
57	7.102	4.998	4.145	3.667	3.357	3.136	2.971	2.841	2.736	2.649
58	7.093	4.991	4.138	3.661	3.351	3.130	2.965	2.835	2.730	2.643
59	7.085	4.984	4.132	3.655	3.345	3.124	2.959	2.829	2.724	2.637
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
61	7.070	4.971	4.120	3.643	3.333	3.113	2.948	2.818	2.713	2.626
62	7.062	4.965	4.114	3.638	3.328	3.108	2.942	2.813	2.708	2.621
63	7.055	4.959	4.109	3.632	3.323	3.103	2.937	2.808	2.703	2.616
64	7.048	4.953	4.103	3.627	3.318	3.098	2.932	2.803	2.698	2.611
65	7.042	4.947	4.098	3.622	3.313	3.093	2.928	2.798	2.693	2.607
66	7.035	4.942	4.093	3.618	3.308	3.088	2.923	2.793	2.689	2.602
67	7.029	4.937	4.088	3.613	3.304	3.084	2.919	2.789	2.684	2.598
68	7.023	4.932	4.083	3.608	3.299	3.080	2.914	2.785	2.680	2.593
69	7.017	4.927	4.079	3.604	3.295	3.075	2.910	2.781	2.676	2.589
70	7.011	4.922	4.074	3.600	3.291	3.071	2.906	2.777	2.672	2.585
71	7.006	4.917	4.070	3.596	3.287	3.067	2.902	2.773	2.668	2.581
72	7.001	4.913	4.066	3.591	3.283	3.063	2.898			

								2.769	2.664	2.578
73	6.995	4.908	4.062	3.588	3.279	3.060	2.895	2.765	2.660	2.574
74	6.990	4.904	4.058	3.584	3.275	3.056	2.891	2.762	2.657	2.570
75	6.985	4.900	4.054	3.580	3.272	3.052	2.887	2.758	2.653	2.567
76	6.981	4.896	4.050	3.577	3.268	3.049	2.884	2.755	2.650	2.563
77	6.976	4.892	4.047	3.573	3.265	3.046	2.881	2.751	2.647	2.560
78	6.971	4.888	4.043	3.570	3.261	3.042	2.877	2.748	2.644	2.557
79	6.967	4.884	4.040	3.566	3.258	3.039	2.874	2.745	2.640	2.554
80	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551
81	6.958	4.877	4.033	3.560	3.252	3.033	2.868	2.739	2.634	2.548
82	6.954	4.874	4.030	3.557	3.249	3.030	2.865	2.736	2.632	2.545
83	6.950	4.870	4.027	3.554	3.246	3.027	2.863	2.733	2.629	2.542
84	6.947	4.867	4.024	3.551	3.243	3.025	2.860	2.731	2.626	2.539
85	6.943	4.864	4.021	3.548	3.240	3.022	2.857	2.728	2.623	2.537
86	6.939	4.861	4.018	3.545	3.238	3.019	2.854	2.725	2.621	2.534
87	6.935	4.858	4.015	3.543	3.235	3.017	2.852	2.723	2.618	2.532
88	6.932	4.855	4.012	3.540	3.233	3.014	2.849	2.720	2.616	2.529
89	6.928	4.852	4.010	3.538	3.230	3.012	2.847	2.718	2.613	2.527
90	6.925	4.849	4.007	3.535	3.228	3.009	2.845	2.715	2.611	2.524
91	6.922	4.846	4.004	3.533	3.225	3.007	2.842	2.713	2.609	2.522
92	6.919	4.844	4.002	3.530	3.223	3.004	2.840	2.711	2.606	2.520
93	6.915	4.841	3.999	3.528	3.221	3.002	2.838	2.709	2.604	2.518
94	6.912	4.838	3.997	3.525	3.218	3.000	2.835	2.706	2.602	2.515
95	6.909	4.836	3.995	3.523	3.216	2.998	2.833			

								2.704	2.600	2.513
96	6.906	4.833	3.992	3.521	3.214	2.996	2.831	2.702	2.598	2.511
97	6.904	4.831	3.990	3.519	3.212	2.994	2.829	2.700	2.596	2.509
98	6.901	4.829	3.988	3.517	3.210	2.992	2.827	2.698	2.594	2.507
99	6.898	4.826	3.986	3.515	3.208	2.990	2.825	2.696	2.592	2.505
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503

ν_2 \ ν_1	11	12	13	14	15	16	17	18	19	20
1.	6083.35	6106.35	6125.86	6142.70	6157.28	6170.12	6181.42	6191.52	6200.58	6208.74
2.	99.408	99.416	99.422	99.428	99.432	99.437	99.440	99.444	99.447	99.449
3.	27.133	27.052	26.983	26.924	26.872	26.827	26.787	26.751	26.719	26.690
4.	14.452	14.374	14.307	14.249	14.198	14.154	14.115	14.080	14.048	14.020
5.	9.963	9.888	9.825	9.770	9.722	9.680	9.643	9.610	9.580	9.553
6.	7.790	7.718	7.657	7.605	7.559	7.519	7.483	7.451	7.422	7.396
7.	6.538	6.469	6.410	6.359	6.314	6.275	6.240	6.209	6.181	6.155
8.	5.734	5.667	5.609	5.559	5.515	5.477	5.442	5.412	5.384	5.359
9.	5.178	5.111	5.055	5.005	4.962	4.924	4.890	4.860	4.833	4.808
10.	4.772	4.706	4.650	4.601	4.558	4.520	4.487	4.457	4.430	4.405
11.	4.462	4.397	4.342	4.293	4.251	4.213	4.180	4.150	4.123	4.099
12.	4.220	4.155	4.100	4.052	4.010	3.972	3.939	3.909	3.883	3.858
13.	4.025	3.960	3.905	3.857	3.815	3.778	3.745	3.716	3.689	3.665
14.	3.864	3.800	3.745	3.698	3.656	3.619	3.586			

								3.556	3.529	3.505
15.	3.730	3.666	3.612	3.564	3.522	3.485	3.452	3.423	3.396	3.372
16.	3.616	3.553	3.498	3.451	3.409	3.372	3.339	3.310	3.283	3.259
17.	3.519	3.455	3.401	3.353	3.312	3.275	3.242	3.212	3.186	3.162
18.	3.434	3.371	3.316	3.269	3.227	3.190	3.158	3.128	3.101	3.077
19.	3.360	3.297	3.242	3.195	3.153	3.116	3.084	3.054	3.027	3.003
20.	3.294	3.231	3.177	3.130	3.088	3.051	3.018	2.989	2.962	2.938
21.	3.236	3.173	3.119	3.072	3.030	2.993	2.960	2.931	2.904	2.880
22.	3.184	3.121	3.067	3.019	2.978	2.941	2.908	2.879	2.852	2.827
23.	3.137	3.074	3.020	2.973	2.931	2.894	2.861	2.832	2.805	2.781
24.	3.094	3.032	2.977	2.930	2.889	2.852	2.819	2.789	2.762	2.738
25.	3.056	2.993	2.939	2.892	2.850	2.813	2.780	2.751	2.724	2.699
26.	3.021	2.958	2.904	2.857	2.815	2.778	2.745	2.715	2.688	2.664
27.	2.988	2.926	2.871	2.824	2.783	2.746	2.713	2.683	2.656	2.632
28.	2.959	2.896	2.842	2.795	2.753	2.716	2.683	2.653	2.626	2.602
29.	2.931	2.868	2.814	2.767	2.726	2.689	2.656	2.626	2.599	2.574
30.	2.906	2.843	2.789	2.742	2.700	2.663	2.630	2.600	2.573	2.549
31.	2.882	2.820	2.765	2.718	2.677	2.640	2.606	2.577	2.550	2.525
32.	2.860	2.798	2.744	2.696	2.655	2.618	2.584	2.555	2.527	2.503
33.	2.840	2.777	2.723	2.676	2.634	2.597	2.564	2.534	2.507	2.482
34.	2.821	2.758	2.704	2.657	2.615	2.578	2.545	2.515	2.488	2.463
35.	2.803	2.740	2.686	2.639	2.597	2.560	2.527	2.497	2.470	2.445
36.	2.786	2.723	2.669	2.622	2.580	2.543	2.510	2.480	2.453	2.428
37.	2.770	2.707	2.653	2.606	2.564	2.527	2.494			

								2.464	2.437	2.412
38.	2.755	2.692	2.638	2.591	2.549	2.512	2.479	2.449	2.421	2.397
39.	2.741	2.678	2.624	2.577	2.535	2.498	2.465	2.434	2.407	2.382
40.	2.727	2.665	2.611	2.563	2.522	2.484	2.451	2.421	2.394	2.369
41.	2.715	2.652	2.598	2.551	2.509	2.472	2.438	2.408	2.381	2.356
42.	2.703	2.640	2.586	2.539	2.497	2.460	2.426	2.396	2.369	2.344
43.	2.691	2.629	2.575	2.527	2.485	2.448	2.415	2.385	2.357	2.332
44.	2.680	2.618	2.564	2.516	2.475	2.437	2.404	2.374	2.346	2.321
45.	2.670	2.608	2.553	2.506	2.464	2.427	2.393	2.363	2.336	2.311
46.	2.660	2.598	2.544	2.496	2.454	2.417	2.384	2.353	2.326	2.301
47.	2.651	2.588	2.534	2.487	2.445	2.408	2.374	2.344	2.316	2.291
48.	2.642	2.579	2.525	2.478	2.436	2.399	2.365	2.335	2.307	2.282
49.	2.633	2.571	2.517	2.469	2.427	2.390	2.356	2.326	2.299	2.274
50.	2.625	2.562	2.508	2.461	2.419	2.382	2.348	2.318	2.290	2.265
51.	2.617	2.555	2.500	2.453	2.411	2.374	2.340	2.310	2.282	2.257
52.	2.610	2.547	2.493	2.445	2.403	2.366	2.333	2.302	2.275	2.250
53.	2.602	2.540	2.486	2.438	2.396	2.359	2.325	2.295	2.267	2.242
54.	2.595	2.533	2.479	2.431	2.389	2.352	2.318	2.288	2.260	2.235
55.	2.589	2.526	2.472	2.424	2.382	2.345	2.311	2.281	2.253	2.228
56.	2.582	2.520	2.465	2.418	2.376	2.339	2.305	2.275	2.247	2.222
57.	2.576	2.513	2.459	2.412	2.370	2.332	2.299	2.268	2.241	2.215
58.	2.570	2.507	2.453	2.406	2.364	2.326	2.293	2.262	2.235	2.209
59.	2.564	2.502	2.447	2.400	2.358	2.320	2.287	2.256	2.229	2.203
60.	2.559	2.496	2.442	2.394	2.352	2.315	2.281			

								2.251	2.223	2.198
61.	2.553	2.491	2.436	2.389	2.347	2.309	2.276	2.245	2.218	2.192
62.	2.548	2.486	2.431	2.384	2.342	2.304	2.270	2.240	2.212	2.187
63.	2.543	2.481	2.426	2.379	2.337	2.299	2.265	2.235	2.207	2.182
64.	2.538	2.476	2.421	2.374	2.332	2.294	2.260	2.230	2.202	2.177
65.	2.534	2.471	2.417	2.369	2.327	2.289	2.256	2.225	2.198	2.172
66.	2.529	2.466	2.412	2.365	2.322	2.285	2.251	2.221	2.193	2.168
67.	2.525	2.462	2.408	2.360	2.318	2.280	2.247	2.216	2.188	2.163
68.	2.520	2.458	2.403	2.356	2.314	2.276	2.242	2.212	2.184	2.159
69.	2.516	2.454	2.399	2.352	2.310	2.272	2.238	2.208	2.180	2.155
70.	2.512	2.450	2.395	2.348	2.306	2.268	2.234	2.204	2.176	2.150
71.	2.508	2.446	2.391	2.344	2.302	2.264	2.230	2.200	2.172	2.146
72.	2.504	2.442	2.388	2.340	2.298	2.260	2.226	2.196	2.168	2.143
73.	2.501	2.438	2.384	2.336	2.294	2.256	2.223	2.192	2.164	2.139
74.	2.497	2.435	2.380	2.333	2.290	2.253	2.219	2.188	2.161	2.135
75.	2.494	2.431	2.377	2.329	2.287	2.249	2.215	2.185	2.157	2.132
76.	2.490	2.428	2.373	2.326	2.284	2.246	2.212	2.181	2.154	2.128
77.	2.487	2.424	2.370	2.322	2.280	2.243	2.209	2.178	2.150	2.125
78.	2.484	2.421	2.367	2.319	2.277	2.239	2.206	2.175	2.147	2.122
79.	2.481	2.418	2.364	2.316	2.274	2.236	2.202	2.172	2.144	2.118
80.	2.478	2.415	2.361	2.313	2.271	2.233	2.199	2.169	2.141	2.115
81.	2.475	2.412	2.358	2.310	2.268	2.230	2.196	2.166	2.138	2.112
82.	2.472	2.409	2.355	2.307	2.265	2.227	2.193	2.163	2.135	2.109
83.	2.469	2.406	2.352	2.304	2.262	2.224	2.191			

								2.160	2.132	2.106
84.	2.466	2.404	2.349	2.302	2.259	2.222	2.188	2.157	2.129	2.104
85.	2.464	2.401	2.347	2.299	2.257	2.219	2.185	2.154	2.126	2.101
86.	2.461	2.398	2.344	2.296	2.254	2.216	2.182	2.152	2.124	2.098
87.	2.459	2.396	2.342	2.294	2.252	2.214	2.180	2.149	2.121	2.096
88.	2.456	2.393	2.339	2.291	2.249	2.211	2.177	2.147	2.119	2.093
89.	2.454	2.391	2.337	2.289	2.247	2.209	2.175	2.144	2.116	2.091
90.	2.451	2.389	2.334	2.286	2.244	2.206	2.172	2.142	2.114	2.088
91.	2.449	2.386	2.332	2.284	2.242	2.204	2.170	2.139	2.111	2.086
92.	2.447	2.384	2.330	2.282	2.240	2.202	2.168	2.137	2.109	2.083
93.	2.444	2.382	2.327	2.280	2.237	2.200	2.166	2.135	2.107	2.081
94.	2.442	2.380	2.325	2.277	2.235	2.197	2.163	2.133	2.105	2.079
95.	2.440	2.378	2.323	2.275	2.233	2.195	2.161	2.130	2.102	2.077
96.	2.438	2.375	2.321	2.273	2.231	2.193	2.159	2.128	2.100	2.075
97.	2.436	2.373	2.319	2.271	2.229	2.191	2.157	2.126	2.098	2.073
98.	2.434	2.371	2.317	2.269	2.227	2.189	2.155	2.124	2.096	2.071
99.	2.432	2.369	2.315	2.267	2.225	2.187	2.153	2.122	2.094	2.069
100.	2.430	2.368	2.313	2.265	2.223	2.185	2.151	2.120	2.092	2.067

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

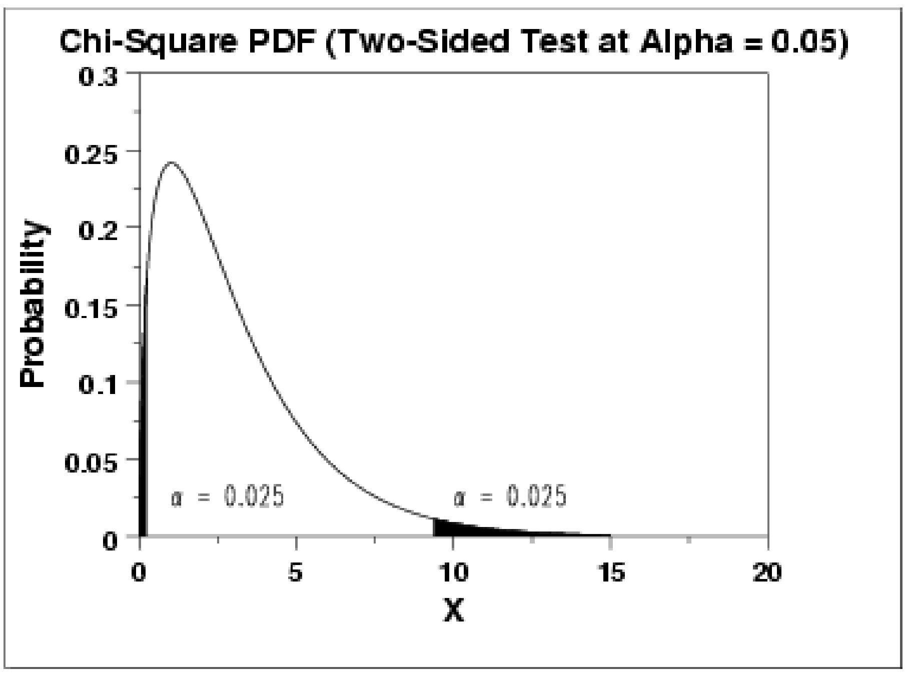
1.3.6.7.4. Critical Values of the Chi-Square Distribution

How to Use This Table

This table contains the critical values of the chi-square distribution. Because of the lack of symmetry of the chi-square distribution, separate tables are provided for the upper and lower tails of the distribution.

A test statistic with ν degrees of freedom is computed from the data. For upper one-sided tests, the test statistic is compared with a value from the table of upper critical values. For two-sided tests, the test statistic is compared with values from both the table for the upper critical value and the table for the lower critical value.

The significance level, α, is demonstrated with the graph below which shows a chi-square distribution with 3 degrees of freedom for a two-sided test at significance level α = 0.05. If the test statistic is greater than the upper critical value or less than the lower critical value, we reject the null hypothesis. Specific instructions are given below.



Given a specified value for α:

1. For a two-sided test, find the column corresponding to α/2 in the table for upper critical values and reject the null hypothesis if the test statistic is greater than the tabled value. Similarly, find the

column corresponding to **1 - α/2** in the table for lower critical values and reject the null hypothesis if the test statistic is less than the tabled value.

2. For an upper one-sided test, find the column corresponding to α in the upper critical values table and reject the null hypothesis if the test statistic is greater than the tabled value.
3. For a lower one-sided test, find the column corresponding to **1 - α** in the lower critical values table and reject the null hypothesis if the computed test statistic is less than the tabled value.

Upper critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264
12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797

22	30.813	33.924	36.781	40.289	48.268
23	32.007	35.172	38.076	41.638	49.728
24	33.196	36.415	39.364	42.980	51.179
25	34.382	37.652	40.646	44.314	52.620
26	35.563	38.885	41.923	45.642	54.052
27	36.741	40.113	43.195	46.963	55.476
28	37.916	41.337	44.461	48.278	56.892
29	39.087	42.557	45.722	49.588	58.301
30	40.256	43.773	46.979	50.892	59.703
31	41.422	44.985	48.232	52.191	61.098
32	42.585	46.194	49.480	53.486	62.487
33	43.745	47.400	50.725	54.776	63.870
34	44.903	48.602	51.966	56.061	65.247
35	46.059	49.802	53.203	57.342	66.619
36	47.212	50.998	54.437	58.619	67.985
37	48.363	52.192	55.668	59.893	69.347
38	49.513	53.384	56.896	61.162	70.703
39	50.660	54.572	58.120	62.428	72.055
40	51.805	55.758	59.342	63.691	73.402
41	52.949	56.942	60.561	64.950	74.745
42	54.090	58.124	61.777	66.206	76.084
43	55.230	59.304	62.990	67.459	77.419
44	56.369	60.481	64.201	68.710	78.750
45	57.505	61.656	65.410	69.957	80.077
46	58.641	62.830	66.617	71.201	81.400
47	59.774	64.001	67.821	72.443	82.720
48	60.907	65.171	69.023	73.683	84.037
49	62.038	66.339	70.222	74.919	85.351
50	63.167	67.505	71.420	76.154	86.661
51	64.295	68.669	72.616	77.386	87.968
52	65.422	69.832	73.810	78.616	89.272
53	66.548	70.993	75.002	79.843	90.573
54	67.673	72.153	76.192	81.069	91.872
55	68.796	73.311	77.380	82.292	93.168
56	69.919	74.468	78.567	83.513	94.461
57	71.040	75.624	79.752	84.733	95.751
58	72.160	76.778	80.936	85.950	97.039
59	73.279	77.931	82.117	87.166	98.324
60	74.397	79.082	83.298	88.379	99.607
61	75.514	80.232	84.476	89.591	100.888
62	76.630	81.381	85.654	90.802	102.166

63	77.745	82.529	86.830	92.010	103.442
64	78.860	83.675	88.004	93.217	104.716
65	79.973	84.821	89.177	94.422	105.988
66	81.085	85.965	90.349	95.626	107.258
67	82.197	87.108	91.519	96.828	108.526
68	83.308	88.250	92.689	98.028	109.791
69	84.418	89.391	93.856	99.228	111.055
70	85.527	90.531	95.023	100.425	112.317
71	86.635	91.670	96.189	101.621	113.577
72	87.743	92.808	97.353	102.816	114.835
73	88.850	93.945	98.516	104.010	116.092
74	89.956	95.081	99.678	105.202	117.346
75	91.061	96.217	100.839	106.393	118.599
76	92.166	97.351	101.999	107.583	119.850
77	93.270	98.484	103.158	108.771	121.100
78	94.374	99.617	104.316	109.958	122.348
79	95.476	100.749	105.473	111.144	123.594
80	96.578	101.879	106.629	112.329	124.839
81	97.680	103.010	107.783	113.512	126.083
82	98.780	104.139	108.937	114.695	127.324
83	99.880	105.267	110.090	115.876	128.565
84	100.980	106.395	111.242	117.057	129.804
85	102.079	107.522	112.393	118.236	131.041
86	103.177	108.648	113.544	119.414	132.277
87	104.275	109.773	114.693	120.591	133.512
88	105.372	110.898	115.841	121.767	134.746
89	106.469	112.022	116.989	122.942	135.978
90	107.565	113.145	118.136	124.116	137.208
91	108.661	114.268	119.282	125.289	138.438
92	109.756	115.390	120.427	126.462	139.666
93	110.850	116.511	121.571	127.633	140.893
94	111.944	117.632	122.715	128.803	142.119
95	113.038	118.752	123.858	129.973	143.344
96	114.131	119.871	125.000	131.141	144.567
97	115.223	120.990	126.141	132.309	145.789
98	116.315	122.108	127.282	133.476	147.010
99	117.407	123.225	128.422	134.642	148.230
100	118.498	124.342	129.561	135.807	149.449
100	118.498	124.342	129.561	135.807	149.449

Lower critical values of chi-square distribution with ν degrees of freedom

	Probability of exceeding the critical value				
ν	0.90	0.95	0.975	0.99	0.999
1.	.016	.004	.001	.000	.000
2.	.211	.103	.051	.020	.002
3.	.584	.352	.216	.115	.024
4.	1.064	.711	.484	.297	.091
5.	1.610	1.145	.831	.554	.210
6.	2.204	1.635	1.237	.872	.381
7.	2.833	2.167	1.690	1.239	.598
8.	3.490	2.733	2.180	1.646	.857
9.	4.168	3.325	2.700	2.088	1.152
10.	4.865	3.940	3.247	2.558	1.479
11.	5.578	4.575	3.816	3.053	1.834
12.	6.304	5.226	4.404	3.571	2.214
13.	7.042	5.892	5.009	4.107	2.617
14.	7.790	6.571	5.629	4.660	3.041
15.	8.547	7.261	6.262	5.229	3.483
16.	9.312	7.962	6.908	5.812	3.942
17.	10.085	8.672	7.564	6.408	4.416
18.	10.865	9.390	8.231	7.015	4.905
19.	11.651	10.117	8.907	7.633	5.407
20.	12.443	10.851	9.591	8.260	5.921
21.	13.240	11.591	10.283	8.897	6.447
22.	14.041	12.338	10.982	9.542	6.983
23.	14.848	13.091	11.689	10.196	7.529
24.	15.659	13.848	12.401	10.856	8.085
25.	16.473	14.611	13.120	11.524	8.649
26.	17.292	15.379	13.844	12.198	9.222
27.	18.114	16.151	14.573	12.879	9.803
28.	18.939	16.928	15.308	13.565	10.391
29.	19.768	17.708	16.047	14.256	10.986
30.	20.599	18.493	16.791	14.953	11.588

31.	21.434	19.281	17.539	15.655	12.196
32.	22.271	20.072	18.291	16.362	12.811
33.	23.110	20.867	19.047	17.074	13.431
34.	23.952	21.664	19.806	17.789	14.057
35.	24.797	22.465	20.569	18.509	14.688
36.	25.643	23.269	21.336	19.233	15.324
37.	26.492	24.075	22.106	19.960	15.965
38.	27.343	24.884	22.878	20.691	16.611
39.	28.196	25.695	23.654	21.426	17.262
40.	29.051	26.509	24.433	22.164	17.916
41.	29.907	27.326	25.215	22.906	18.575
42.	30.765	28.144	25.999	23.650	19.239
43.	31.625	28.965	26.785	24.398	19.906
44.	32.487	29.787	27.575	25.148	20.576
45.	33.350	30.612	28.366	25.901	21.251
46.	34.215	31.439	29.160	26.657	21.929
47.	35.081	32.268	29.956	27.416	22.610
48.	35.949	33.098	30.755	28.177	23.295
49.	36.818	33.930	31.555	28.941	23.983
50.	37.689	34.764	32.357	29.707	24.674
51.	38.560	35.600	33.162	30.475	25.368
52.	39.433	36.437	33.968	31.246	26.065
53.	40.308	37.276	34.776	32.018	26.765
54.	41.183	38.116	35.586	32.793	27.468
55.	42.060	38.958	36.398	33.570	28.173
56.	42.937	39.801	37.212	34.350	28.881
57.	43.816	40.646	38.027	35.131	29.592
58.	44.696	41.492	38.844	35.913	30.305
59.	45.577	42.339	39.662	36.698	31.020
60.	46.459	43.188	40.482	37.485	31.738
61.	47.342	44.038	41.303	38.273	32.459
62.	48.226	44.889	42.126	39.063	33.181
63.	49.111	45.741	42.950	39.855	33.906
64.	49.996	46.595	43.776	40.649	34.633
65.	50.883	47.450	44.603	41.444	35.362
66.	51.770	48.305	45.431	42.240	36.093
67.	52.659	49.162	46.261	43.038	36.826
68.	53.548	50.020	47.092	43.838	37.561
69.	54.438	50.879	47.924	44.639	38.298
70.	55.329	51.739	48.758	45.442	39.036
71.	56.221	52.600	49.592	46.246	39.777

72.	57.113	53.462	50.428	47.051	40.519
73.	58.006	54.325	51.265	47.858	41.264
74.	58.900	55.189	52.103	48.666	42.010
75.	59.795	56.054	52.942	49.475	42.757
76.	60.690	56.920	53.782	50.286	43.507
77.	61.586	57.786	54.623	51.097	44.258
78.	62.483	58.654	55.466	51.910	45.010
79.	63.380	59.522	56.309	52.725	45.764
80.	64.278	60.391	57.153	53.540	46.520
81.	65.176	61.261	57.998	54.357	47.277
82.	66.076	62.132	58.845	55.174	48.036
83.	66.976	63.004	59.692	55.993	48.796
84.	67.876	63.876	60.540	56.813	49.557
85.	68.777	64.749	61.389	57.634	50.320
86.	69.679	65.623	62.239	58.456	51.085
87.	70.581	66.498	63.089	59.279	51.850
88.	71.484	67.373	63.941	60.103	52.617
89.	72.387	68.249	64.793	60.928	53.386
90.	73.291	69.126	65.647	61.754	54.155
91.	74.196	70.003	66.501	62.581	54.926
92.	75.100	70.882	67.356	63.409	55.698
93.	76.006	71.760	68.211	64.238	56.472
94.	76.912	72.640	69.068	65.068	57.246
95.	77.818	73.520	69.925	65.898	58.022
96.	78.725	74.401	70.783	66.730	58.799
97.	79.633	75.282	71.642	67.562	59.577
98.	80.541	76.164	72.501	68.396	60.356
99.	81.449	77.046	73.361	69.230	61.137
100.	82.358	77.929	74.222	70.065	61.918

1. Exploratory Data Analysis
1.3. EDA Techniques
1.3.6. Probability Distributions
1.3.6.7. Tables for Probability Distributions

1.3.6.7.5. Critical Values of the t^* Distribution

How to Use This Table

This table contains upper critical values of the t* distribution that are appropriate for determining whether or not a calibration line is in a state of statistical control from measurements on a check standard at three points in the calibration interval. A test statistic with ν degrees of freedom is compared with the critical value. If the absolute value of the test statistic exceeds the tabled value, the calibration of the instrument is judged to be out of control.

Upper critical values of t* distribution at significance level 0.05 for testing the output of a linear calibration line at 3 points

ν	$t^*_{.05}(\nu)$	ν	$t^*_{.05}(\nu)$
1	37.544	61	2.455
2	7.582	62	2.454
3	4.826	63	2.453
4	3.941	64	2.452
5	3.518	65	2.451
6	3.274	66	2.450
7	3.115	67	2.449
8	3.004	68	2.448
9	2.923	69	2.447
10	2.860	70	2.446
11	2.811	71	2.445
12	2.770	72	2.445
13	2.737	73	2.444
14	2.709	74	2.443
15	2.685	75	2.442

16	2.665	76	2.441
17	2.647	77	2.441
18	2.631	78	2.440
19	2.617	79	2.439
20	2.605	80	2.439
21	2.594	81	2.438
22	2.584	82	2.437
23	2.574	83	2.437
24	2.566	84	2.436
25	2.558	85	2.436
26	2.551	86	2.435
27	2.545	87	2.435
28	2.539	88	2.434
29	2.534	89	2.434
30	2.528	90	2.433
31	2.524	91	2.432
32	2.519	92	2.432
33	2.515	93	2.431
34	2.511	94	2.431
35	2.507	95	2.431
36	2.504	96	2.430
37	2.501	97	2.430
38	2.498	98	2.429
39	2.495	99	2.429
40	2.492	100	2.428
41	2.489	101	2.428
42	2.487	102	2.428
43	2.484	103	2.427
44	2.482	104	2.427
45	2.480	105	2.426
46	2.478	106	2.426
47	2.476	107	2.426
48	2.474	108	2.425
49	2.472	109	2.425
50	2.470	110	2.425
51	2.469	111	2.424
52	2.467	112	2.424
53	2.466	113	2.424
54	2.464	114	2.423
55	2.463	115	2.423
56	2.461	116	2.423
57	2.460	117	2.422
58	2.459	118	2.422
59	2.457	119	2.422
60	2.456	120	2.422

1.3.6.7.6. Critical Values of the Normal PPCC Distribution

How to Use This Table

This table contains the critical values of the normal probability plot correlation coefficient (PPCC) distribution that are appropriate for determining whether or not a data set came from a population with approximately a normal distribution. It is used in conjuction with a normal probability plot. The test statistic is the correlation coefficient of the points that make up a normal probability plot. This test statistic is compared with the critical value below. If the test statistic is less than the tabulated value, the null hypothesis that the data came from a population with a normal distribution is rejected.

For example, suppose a set of 50 data points had a correlation coefficient of 0.985 from the normal probability plot. At the 5% significance level, the critical value is 0.965. Since 0.985 is greater than 0.965, we cannot reject the null hypothesis that the data came from a population with a normal distribution.

Since perferct normality implies perfect correlation (i.e., a correlation value of 1), we are only interested in rejecting normality for correlation values that are too low. That is, this is a lower one-tailed test.

The values in this table were determined from simulation studies by Filliben and Devaney.

Critical values of the normal PPCC for testing if data come from a normal distribution

N	0.01	0.05
3	0.8687	0.8790
4	0.8234	0.8666
5	0.8240	0.8786
6	0.8351	0.8880
7	0.8474	0.8970
8	0.8590	0.9043
9	0.8689	0.9115
10	0.8765	0.9173
11	0.8838	0.9223
12	0.8918	0.9267
13	0.8974	0.9310
14	0.9029	0.9343
15	0.9080	0.9376
16	0.9121	0.9405
17	0.9160	0.9433
18	0.9196	0.9452
19	0.9230	0.9479
20	0.9256	0.9498
21	0.9285	0.9515
22	0.9308	0.9535
23	0.9334	0.9548
24	0.9356	0.9564
25	0.9370	0.9575
26	0.9393	0.9590
27	0.9413	0.9600
28	0.9428	0.9615
29	0.9441	0.9622
30	0.9462	0.9634
31	0.9476	0.9644
32	0.9490	0.9652
33	0.9505	0.9661
34	0.9521	0.9671
35	0.9530	0.9678
36	0.9540	0.9686
37	0.9551	0.9693
38	0.9555	0.9700
39	0.9568	0.9704

40	0.9576	0.9712
41	0.9589	0.9719
42	0.9593	0.9723
43	0.9609	0.9730
44	0.9611	0.9734
45	0.9620	0.9739
46	0.9629	0.9744
47	0.9637	0.9748
48	0.9640	0.9753
49	0.9643	0.9758
50	0.9654	0.9761
55	0.9683	0.9781
60	0.9706	0.9797
65	0.9723	0.9809
70	0.9742	0.9822
75	0.9758	0.9831
80	0.9771	0.9841
85	0.9784	0.9850
90	0.9797	0.9857
95	0.9804	0.9864
100	0.9814	0.9869
110	0.9830	0.9881
120	0.9841	0.9889
130	0.9854	0.9897
140	0.9865	0.9904
150	0.9871	0.9909
160	0.9879	0.9915
170	0.9887	0.9919
180	0.9891	0.9923
190	0.9897	0.9927
200	0.9903	0.9930
210	0.9907	0.9933
220	0.9910	0.9936
230	0.9914	0.9939
240	0.9917	0.9941
250	0.9921	0.9943
260	0.9924	0.9945
270	0.9926	0.9947
280	0.9929	0.9949
290	0.9931	0.9951
300	0.9933	0.9952
310	0.9936	0.9954
320	0.9937	0.9955
330	0.9939	0.9956
340	0.9941	0.9957
350	0.9942	0.9958

360	0.9944	0.9959
370	0.9945	0.9960
380	0.9947	0.9961
390	0.9948	0.9962
400	0.9949	0.9963
410	0.9950	0.9964
420	0.9951	0.9965
430	0.9953	0.9966
440	0.9954	0.9966
450	0.9954	0.9967
460	0.9955	0.9968
470	0.9956	0.9968
480	0.9957	0.9969
490	0.9958	0.9969
500	0.9959	0.9970
525	0.9961	0.9972
550	0.9963	0.9973
575	0.9964	0.9974
600	0.9965	0.9975
625	0.9967	0.9976
650	0.9968	0.9977
675	0.9969	0.9977
700	0.9970	0.9978
725	0.9971	0.9979
750	0.9972	0.9980
775	0.9973	0.9980
800	0.9974	0.9981
825	0.9975	0.9981
850	0.9975	0.9982
875	0.9976	0.9982
900	0.9977	0.9983
925	0.9977	0.9983
950	0.9978	0.9984
975	0.9978	0.9984
1000	0.9979	0.9984

1. Exploratory Data Analysis

1.4. EDA Case Studies

Summary

This section presents a series of case studies that demonstrate the application of EDA methods to specific problems. In some cases, we have focused on just one EDA technique that uncovers virtually all there is to know about the data. For other case studies, we need several EDA techniques, the selection of which is dictated by the outcome of the previous step in the analaysis sequence. Note in these case studies how the flow of the analysis is motivated by the focus on underlying assumptions and general EDA principles.

Table of Contents for Section 4

1.4.1. Case Studies Introduction

Purpose

The purpose of the first eight case studies is to show how EDA graphics and quantitative measures and tests are applied to data from scientific processes and to critique those data with regard to the following assumptions that typically underlie a measurement process; namely, that the data behave like:

- random drawings
- from a fixed distribution
- with a fixed location
- with a fixed standard deviation

Case studies 9 and 10 show the use of EDA techniques in distributional modeling and the analysis of a designed experiment, respectively.

$Y_i = C + E_i$

If the above assumptions are satisfied, the process is said to be statistically "in control" with the core characteristic of having "predictability". That is, probability statements can be made about the process, not only in the past, but also in the future.

An appropriate model for an "in control" process is

$$\boldsymbol{Y_i = C + E_i}$$

where $\boldsymbol{C}$ is a constant (the "deterministic" or "structural" component), and where $\boldsymbol{E_i}$ is the error term (or "random" component).

The constant $\boldsymbol{C}$ is the average value of the process--it is the primary summary number which shows up on any report. Although $\boldsymbol{C}$ is (assumed) fixed, it is unknown, and so a primary analysis objective of the engineer is to arrive at an estimate of $\boldsymbol{C}$.

This goal partitions into 4 sub-goals:

1. Is the most common estimator of $\boldsymbol{C}$, $\bar{Y}$, the best estimator for $\boldsymbol{C}$? What does "best" mean?
2. If $\bar{Y}$ is best, what is the uncertainty $s_{\bar{Y}}$ for $\bar{Y}$. In particular, is

the usual formula for the uncertainty of $\bar{Y}$:

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

valid? Here, ***s*** is the standard deviation of the data and ***N*** is the sample size.

3. If $\bar{Y}$ is **not** the best estimator for ***C***, what is a better estimator for ***C*** (for example, median, midrange, midmean)?
4. If there is a better estimator, $\hat{C}$, what is its uncertainty? That is, what is $s_{\hat{C}}$?

EDA and the routine checking of underlying assumptions provides insight into all of the above.

1. Location and variation checks provide information as to whether ***C*** is really constant.
2. Distributional checks indicate whether $\bar{Y}$ is the best estimator. Techniques for distributional checking include histograms, normal probability plots, and probability plot correlation coefficient plots.
3. Randomness checks ascertain whether the usual

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

is valid.
4. Distributional tests assist in determining a better estimator, if needed.
5. Simulator tools (namely bootstrapping) provide values for the uncertainty of alternative estimators.

Assumptions not satisfied

If one or more of the above assumptions is not satisfied, then we use EDA techniques, or some mix of EDA and classical techniques, to find a more appropriate model for the data. That is,

$$Y_i = D + E_i$$

where ***D*** is the deterministic part and ***E*** is an error component.

If the data are not random, then we may investigate fitting some simple time series models to the data. If the constant location and scale assumptions are violated, we may need to investigate the measurement process to see if there is an explanation.

The assumptions on the error term are still quite relevant in the sense that for an appropriate model the error component should follow the assumptions. The criterion for validating the model, or comparing competing models, is framed in terms of these assumptions.

Multivariable data

Although the case studies in this chapter utilize univariate data, the assumptions above are relevant for multivariable data as well.

If the data are not univariate, then we are trying to find a model

$$Y_i = F(X_1, ..., X_k) + E_i$$

where F is some function based on one or more variables. The error component, which is a univariate data set, of a good model should satisfy the assumptions given above. The criterion for validating and comparing models is based on how well the error component follows these assumptions.

The load cell calibration case study in the process modeling chapter shows an example of this in the regression context.

First three case studies utilize data with known characteristics

The first three case studies utilize data that are randomly generated from the following distributions:

- normal distribution with mean 0 and standard deviation 1
- uniform distribution with mean 0 and standard deviation $\sqrt{1/12}$ (uniform over the interval (0,1))
- random walk

The other univariate case studies utilize data from scientific processes. The goal is to determine if

$$Y_i = C + E_i$$

is a reasonable model. This is done by testing the underlying assumptions. If the assumptions are satisfied, then an estimate of C and an estimate of the uncertainty of C are computed. If the assumptions are not satisfied, we attempt to find a model where the error component does satisfy the underlying assumptions.

Graphical methods that are applied to the data

To test the underlying assumptions, each data set is analyzed using four graphical methods that are particularly suited for this purpose:

1. run sequence plot which is useful for detecting shifts of location or scale
2. lag plot which is useful for detecting non-randomness in the data
3. histogram which is useful for trying to determine the underlying distribution
4. normal probability plot for deciding whether the data follow the normal distribution

There are a number of other techniques for addressing the underlying assumptions. However, the four plots listed above provide an excellent opportunity for addressing all of the assumptions on a single page of graphics.

Additional graphical techniques are used in certain case studies to develop models that do have error components that satisfy the underlying assumptions.

Quantitative methods that are applied to the data

The normal and uniform random number data sets are also analyzed with the following quantitative techniques, which are explained in more detail in an earlier section:

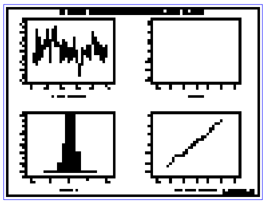
1. Summary statistics which include:
   - mean
   - standard deviation
   - autocorrelation coefficient to test for randomness
   - normal and uniform probability plot correlation coefficients (ppcc) to test for a normal or uniform distribution, respectively
   - Wilk-Shapiro test for a normal distribution
2. Linear fit of the data as a function of time to assess drift (test for fixed location)
3. Bartlett test for fixed variance
4. Autocorrelation plot and coefficient to test for randomness
5. Runs test to test for lack of randomness
6. Anderson-Darling test for a normal distribution
7. Grubbs test for outliers
8. Summary report

Although the graphical methods applied to the normal and uniform random numbers are sufficient to assess the validity of the underlying assumptions, the quantitative techniques are used to show the different flavor of the graphical and quantitative approaches.

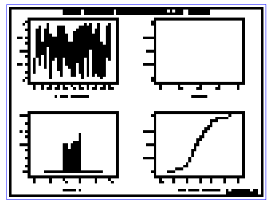
The remaining case studies intermix one or more of these quantitative techniques into the analysis where appropriate.

1.4.2. Case Studies

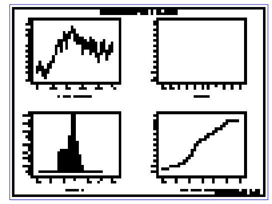
Univariate
$Y_i = C + E_i$



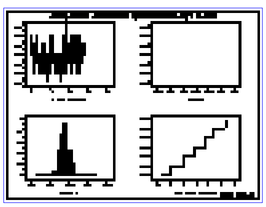
Normal Random Numbers



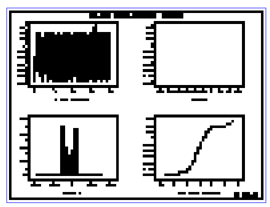
Uniform Random Numbers



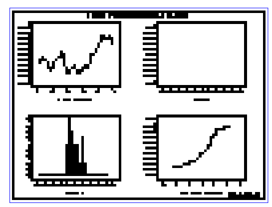
Random Walk



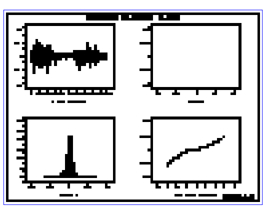
Josephson Junction Cryothermometry



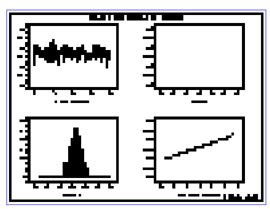
Beam Deflections



Filter Transmittance

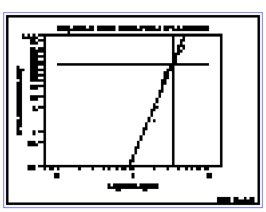


Standard Resistor



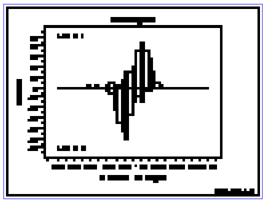
Heat Flow Meter 1

Reliability



Airplane Glass Failure Time

Multi-Factor



Ceramic Strength

1.4.2.1. Normal Random Numbers

Normal Random Numbers

This example illustrates the univariate analysis of a set of normal random numbers.

1.4.2.1.1. Background and Data

Generation

The normal random numbers used in this case study are from a Rand Corporation publication.

The motivation for studying a set of normal random numbers is to illustrate the ideal case where all four underlying assumptions hold.

Software

Most general purpose statistical software programs, including Dataplot, can generate normal random numbers.

Resulting Data

The following is the set of normal random numbers used for this case study.

```
 -1.2760 -1.2180 -0.4530 -0.3500  0.7230
  0.6760 -1.0990 -0.3140 -0.3940 -0.6330
 -0.3180 -0.7990 -1.6640  1.3910  0.3820
  0.7330  0.6530  0.2190 -0.6810  1.1290
 -1.3770 -1.2570  0.4950 -0.1390 -0.8540
  0.4280 -1.3220 -0.3150 -0.7320 -1.3480
  2.3340 -0.3370 -1.9550 -0.6360 -1.3180
 -0.4330  0.5450  0.4280 -0.2970  0.2760
 -1.1360  0.6420  3.4360 -1.6670  0.8470
 -1.1730 -0.3550  0.0350  0.3590  0.9300
  0.4140 -0.0110  0.6660 -1.1320 -0.4100
 -1.0770  0.7340  1.4840 -0.3400  0.7890
 -0.4940  0.3640 -1.2370 -0.0440 -0.1110
 -0.2100  0.9310  0.6160 -0.3770 -0.4330
  1.0480  0.0370  0.7590  0.6090 -2.0430
 -0.2900  0.4040 -0.5430  0.4860  0.8690
  0.3470  2.8160 -0.4640 -0.6320 -1.6140
  0.3720 -0.0740 -0.9160  1.3140 -0.0380
  0.6370  0.5630 -0.1070  0.1310 -1.8080
 -1.1260  0.3790  0.6100 -0.3640 -2.6260
```

2.1760	0.3930	-0.9240	1.9110	-1.0400
-1.1680	0.4850	0.0760	-0.7690	1.6070
-1.1850	-0.9440	-1.6040	0.1850	-0.2580
-0.3000	-0.5910	-0.5450	0.0180	-0.4850
0.9720	1.7100	2.6820	2.8130	-1.5310
-0.4900	2.0710	1.4440	-1.0920	0.4780
1.2100	0.2940	-0.2480	0.7190	1.1030
1.0900	0.2120	-1.1850	-0.3380	-1.1340
2.6470	0.7770	0.4500	2.2470	1.1510
-1.6760	0.3840	1.1330	1.3930	0.8140
0.3980	0.3180	-0.9280	2.4160	-0.9360
1.0360	0.0240	-0.5600	0.2030	-0.8710
0.8460	-0.6990	-0.3680	0.3440	-0.9260
-0.7970	-1.4040	-1.4720	-0.1180	1.4560
0.6540	-0.9550	2.9070	1.6880	0.7520
-0.4340	0.7460	0.1490	-0.1700	-0.4790
0.5220	0.2310	-0.6190	-0.2650	0.4190
0.5580	-0.5490	0.1920	-0.3340	1.3730
-1.2880	-0.5390	-0.8240	0.2440	-1.0700
0.0100	0.4820	-0.4690	-0.0900	1.1710
1.3720	1.7690	-1.0570	1.6460	0.4810
-0.6000	-0.5920	0.6100	-0.0960	-1.3750
0.8540	-0.5350	1.6070	0.4280	-0.6150
0.3310	-0.3360	-1.1520	0.5330	-0.8330
-0.1480	-1.1440	0.9130	0.6840	1.0430
0.5540	-0.0510	-0.9440	-0.4400	-0.2120
-1.1480	-1.0560	0.6350	-0.3280	-1.2210
0.1180	-2.0450	-1.9770	-1.1330	0.3380
0.3480	0.9700	-0.0170	1.2170	-0.9740
-1.2910	-0.3990	-1.2090	-0.2480	0.4800
0.2840	0.4580	1.3070	-1.6250	-0.6290
-0.5040	-0.0560	-0.1310	0.0480	1.8790
-1.0160	0.3600	-0.1190	2.3310	1.6720
-1.0530	0.8400	-0.2460	0.2370	-1.3120
1.6030	-0.9520	-0.5660	1.6000	0.4650
1.9510	0.1100	0.2510	0.1160	-0.9570
-0.1900	1.4790	-0.9860	1.2490	1.9340
0.0700	-1.3580	-1.2460	-0.9590	-1.2970
-0.7220	0.9250	0.7830	-0.4020	0.6190
1.8260	1.2720	-0.9450	0.4940	0.0500
-1.6960	1.8790	0.0630	0.1320	0.6820
0.5440	-0.4170	-0.6660	-0.1040	-0.2530
-2.5430	-1.3330	1.9870	0.6680	0.3600
1.9270	1.1830	1.2110	1.7650	0.3500
-0.3590	0.1930	-1.0230	-0.2220	-0.6160
-0.0600	-1.3190	0.7850	-0.4300	-0.2980

```
 0.2480 -0.0880 -1.3790  0.2950 -0.1150
-0.6210 -0.6180  0.2090  0.9790  0.9060
-0.0990 -1.3760  1.0470 -0.8720 -2.2000
-1.3840  1.4250 -0.8120  0.7480 -1.0930
-0.4630 -1.2810 -2.5140  0.6750  1.1450
 1.0830 -0.6670 -0.2230 -1.5920 -1.2780
 0.5030  1.4340  0.2900  0.3970 -0.8370
-0.9730 -0.1200 -1.5940 -0.9960 -1.2440
-0.8570 -0.3710 -0.2160  0.1480 -2.1060
-1.4530  0.6860 -0.0750 -0.2430 -0.1700
-0.1220  1.1070 -1.0390 -0.6360 -0.8600
-0.8950 -1.4580 -0.5390 -0.1590 -0.4200
 1.6320  0.5860 -0.4680 -0.3860 -0.3540
 0.2030 -1.2340  2.3810 -0.3880 -0.0630
 2.0720 -1.4450 -0.6800  0.2240 -0.1200
 1.7530 -0.5710  1.2230 -0.1260  0.0340
-0.4350 -0.3750 -0.9850 -0.5850 -0.2030
-0.5560  0.0240  0.1260  1.2500 -0.6150
 0.8760 -1.2270 -2.6470 -0.7450  1.7970
-1.2310  0.5470 -0.6340 -0.8360 -0.7190
 0.8330  1.2890 -0.0220 -0.4310  0.5820
 0.7660 -0.5740 -1.1530  0.5200 -1.0180
-0.8910  0.3320 -0.4530 -1.1270  2.0850
-0.7220 -1.5080  0.4890 -0.4960 -0.0250
 0.6440 -0.2330 -0.1530  1.0980  0.7570
-0.0390 -0.4600  0.3930  2.0120  1.3560
 0.1050 -0.1710 -0.1100 -1.1450  0.8780
-0.9090 -0.3280  1.0210 -1.6130  1.5600
-1.1920  1.7700 -0.0030  0.3690  0.0520
 0.6470  1.0290  1.5260  0.2370 -1.3280
-0.0420  0.5530  0.7700  0.3240 -0.4890
-0.3670  0.3780  0.6010 -1.9960 -0.7380
 0.4980  1.0720  1.5670  0.3020  1.1570
-0.7200  1.4030  0.6980 -0.3700 -0.5510
```

1.4.2.1.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

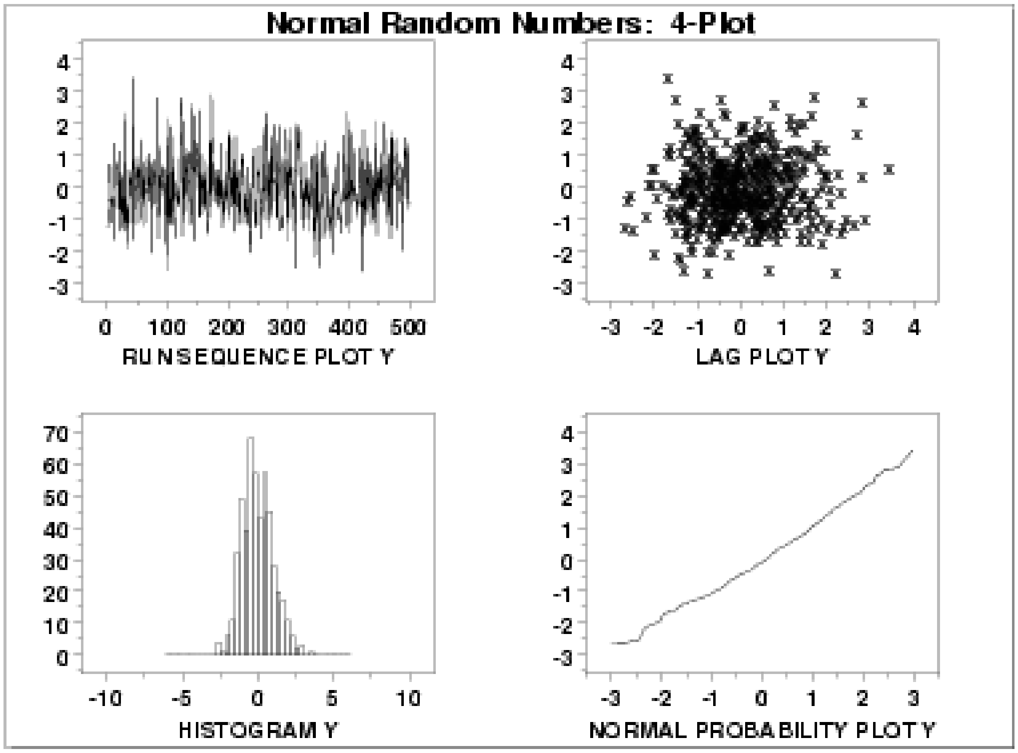
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

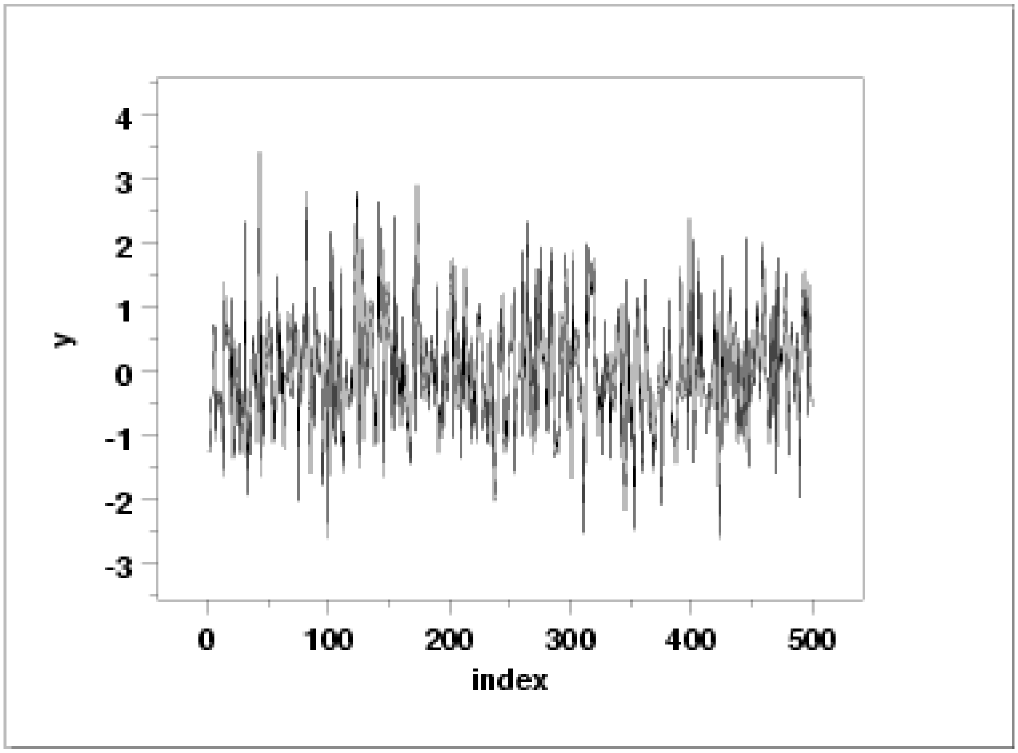
1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time. The run sequence plot does not show any obvious outliers.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram (lower left) shows that the data are reasonably symmetric, there do not appear to be significant outliers in the tails, and that it is reasonable to assume that the data are from approximately a normal distribution.
4. The normal probability plot (lower right) verifies that an assumption of normality is in fact reasonable.

From the above plots, we conclude that the underlying assumptions are valid and the data follow approximately a normal distribution. Therefore, the confidence interval form given previously is appropriate for quantifying the uncertainty of the population mean. The numerical values for this model are given in the Quantitative Output and Interpretation section.

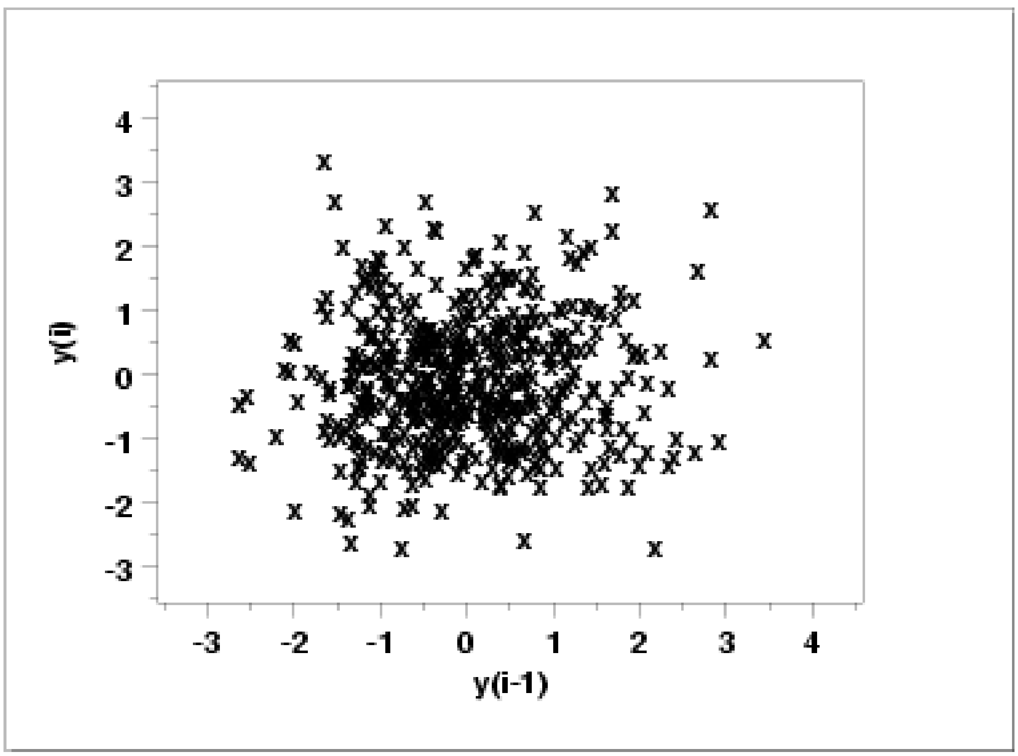
Individual Plots

Although it is usually not necessary, the plots can be generated individually to give more detail.

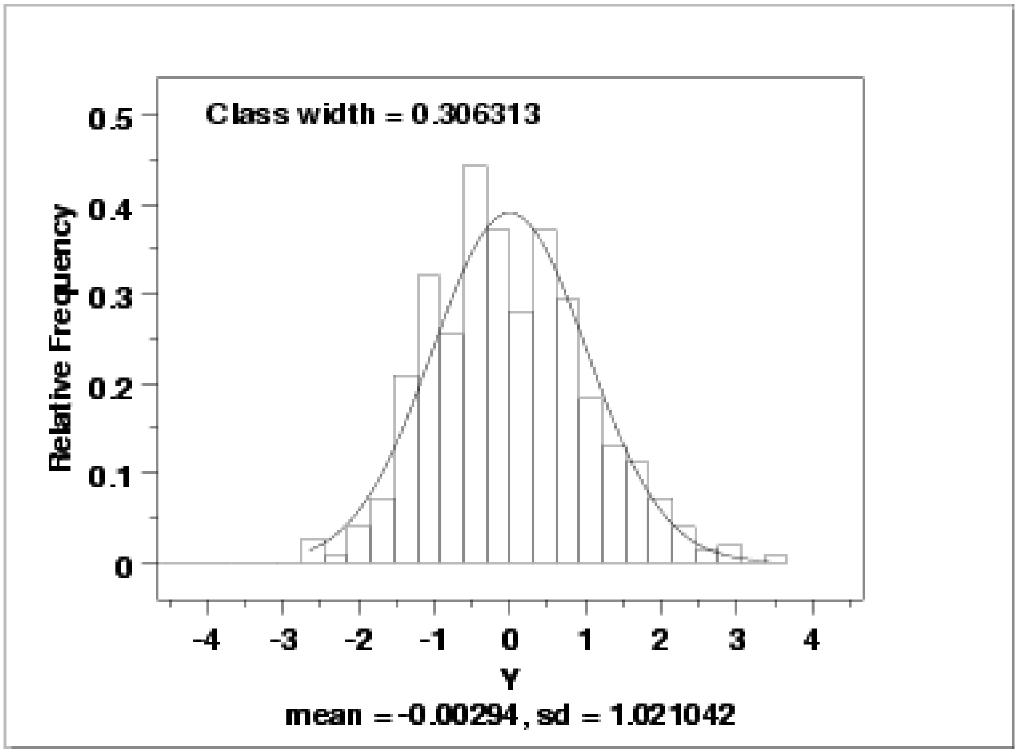
Run Sequence Plot



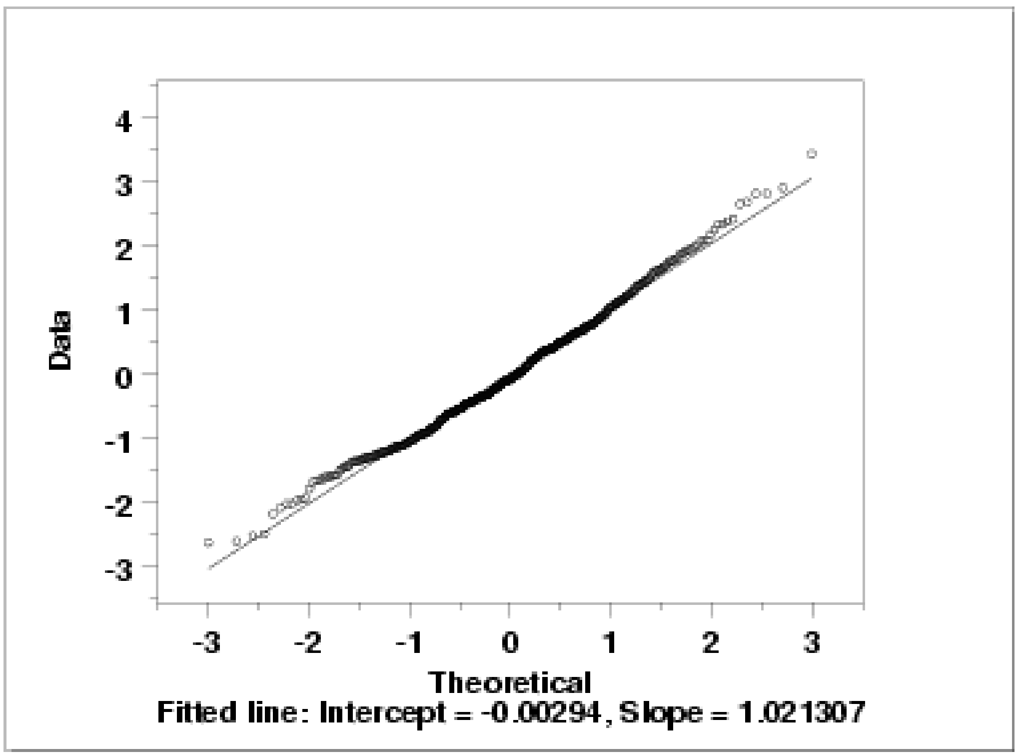
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot



1.4.2.1.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                        SUMMARY

             NUMBER OF OBSERVATIONS =      500


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.3945000E+00  *  RANGE        =   0.6083000E+01  *
*  MEAN         =  -0.2935997E-02  *  STAND. DEV.  =   0.1021041E+01  *
*  MIDMEAN      =   0.1623600E-01  *  AV. AB. DEV. =   0.8174360E+00  *
*  MEDIAN       =  -0.9300000E-01  *  MINIMUM      =  -0.2647000E+01  *
*               =                  *  LOWER QUART. =  -0.7204999E+00  *
*               =                  *  LOWER HINGE  =  -0.7210000E+00  *
*               =                  *  UPPER HINGE  =   0.6455001E+00  *
*               =                  *  UPPER QUART. =   0.6447501E+00  *
*               =                  *  MAXIMUM      =   0.3436000E+01  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.4505888E-01  *  ST. 3RD MOM. =   0.3072273E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.2990314E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =   0.7515639E+01  *
*               =                  *  UNIFORM PPCC =   0.9756625E+00  *
*               =                  *  NORMAL  PPCC =   0.9961721E+00  *
*               =                  *  TUK -.5 PPCC =   0.8366451E+00  *
*               =                  *  CAUCHY  PPCC =   0.4922674E+00  *
***********************************************************************
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set, using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generated the following output:

```
LEAST SQUARES MULTILINEAR FIT
SAMPLE SIZE N         =      500
NUMBER OF VARIABLES =          1
NO REPLICATION CASE


        PARAMETER ESTIMATES            (APPROX. ST. DEV.)    T VALUE
 1  A0                  0.699127E-02   (0.9155E-01)        0.7636E-01
 2  A1        X        -0.396298E-04   (0.3167E-03)        -0.1251

RESIDUAL    STANDARD DEVIATION =         1.02205
RESIDUAL    DEGREES OF FREEDOM =         498
```

The slope parameter, A1, has a t value of -0.13 which is statistically not significant. This indicates that the slope can in fact be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test, after dividing the data set into several equal-sized intervals. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Bartlett test.

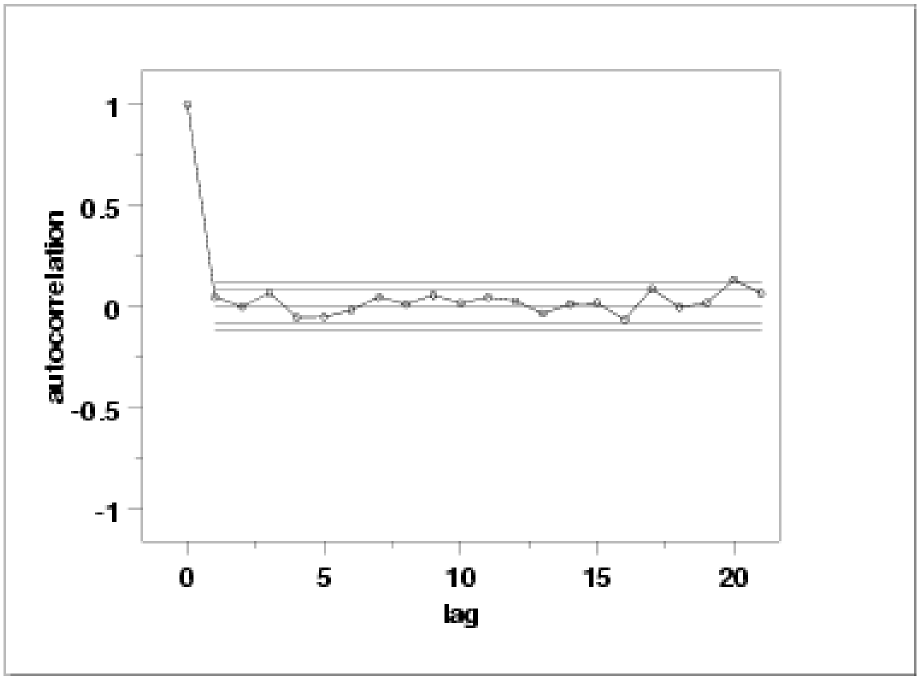
```
               BARTLETT TEST
           (STANDARD DEFINITION)
 NULL HYPOTHESIS UNDER TEST--ALL SIGMA(I) ARE EQUAL

 TEST:
    DEGREES OF FREEDOM          =    3.000000

    TEST STATISTIC VALUE        =    2.373660
    CUTOFF: 95% PERCENT POINT   =    7.814727
    CUTOFF: 99% PERCENT POINT   =    11.34487

    CHI-SQUARE CDF VALUE        =    0.501443

   NULL           NULL HYPOTHESIS       NULL HYPOTHESIS
   HYPOTHESIS    ACCEPTANCE INTERVAL    CONCLUSION
 ALL SIGMA EQUAL    (0.000,0.950)         ACCEPT
```

In this case, the Bartlett test indicates that the standard deviations are not significantly different in the 4 intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot above is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of most interest, is 0.045. The critical values at the 5% significance level are -0.087 and 0.087. Thus, since 0.045 is in the interval, the lag 1 autocorrelation is not statistically significant, so there is no evidence of non-randomness.

A common test for randomness is the runs test.

```
                     RUNS UP
          STATISTIC = NUMBER OF RUNS UP
              OF LENGTH EXACTLY I
 I          STAT      EXP(STAT)    SD(STAT)        Z

 1         98.0    104.2083     10.2792       -0.60
 2         43.0     45.7167      5.2996       -0.51
 3         13.0     13.1292      3.2297       -0.04
 4          6.0      2.8563      1.6351        1.92
 5          1.0      0.5037      0.7045        0.70
 6          0.0      0.0749      0.2733       -0.27
 7          0.0      0.0097      0.0982       -0.10
 8          0.0      0.0011      0.0331       -0.03
 9          0.0      0.0001      0.0106       -0.01
10          0.0      0.0000      0.0032        0.00
          STATISTIC = NUMBER OF RUNS UP
              OF LENGTH I OR MORE
 I          STAT      EXP(STAT)    SD(STAT)        Z
```

1	161.0	166.5000	6.6546	-0.83
2	63.0	62.2917	4.4454	0.16
3	20.0	16.5750	3.4338	1.00
4	7.0	3.4458	1.7786	2.00
5	1.0	0.5895	0.7609	0.54
6	0.0	0.0858	0.2924	-0.29
7	0.0	0.0109	0.1042	-0.10
8	0.0	0.0012	0.0349	-0.03
9	0.0	0.0001	0.0111	-0.01
10	0.0	0.0000	0.0034	0.00

RUNS DOWN

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	91.0	104.2083	10.2792	-1.28
2	55.0	45.7167	5.2996	1.75
3	14.0	13.1292	3.2297	0.27
4	1.0	2.8563	1.6351	-1.14
5	0.0	0.5037	0.7045	-0.71
6	0.0	0.0749	0.2733	-0.27
7	0.0	0.0097	0.0982	-0.10
8	0.0	0.0011	0.0331	-0.03
9	0.0	0.0001	0.0106	-0.01
10	0.0	0.0000	0.0032	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	161.0	166.5000	6.6546	-0.83
2	70.0	62.2917	4.4454	1.73
3	15.0	16.5750	3.4338	-0.46
4	1.0	3.4458	1.7786	-1.38
5	0.0	0.5895	0.7609	-0.77
6	0.0	0.0858	0.2924	-0.29
7	0.0	0.0109	0.1042	-0.10
8	0.0	0.0012	0.0349	-0.03
9	0.0	0.0001	0.0111	-0.01
10	0.0	0.0000	0.0034	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	189.0	208.4167	14.5370	-1.34
2	98.0	91.4333	7.4947	0.88
3	27.0	26.2583	4.5674	0.16
4	7.0	5.7127	2.3123	0.56
5	1.0	1.0074	0.9963	-0.01
6	0.0	0.1498	0.3866	-0.39
7	0.0	0.0193	0.1389	-0.14
8	0.0	0.0022	0.0468	-0.05
9	0.0	0.0002	0.0150	-0.01
10	0.0	0.0000	0.0045	0.00

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	322.0	333.0000	9.4110	-1.17
2	133.0	124.5833	6.2868	1.34
3	35.0	33.1500	4.8561	0.38
4	8.0	6.8917	2.5154	0.44

```
 5          1.0      1.1790      1.0761       -0.17
 6          0.0      0.1716      0.4136       -0.41
 7          0.0      0.0217      0.1474       -0.15
 8          0.0      0.0024      0.0494       -0.05
 9          0.0      0.0002      0.0157       -0.02
10          0.0      0.0000      0.0047        0.00
        LENGTH OF THE LONGEST RUN UP           =     5
        LENGTH OF THE LONGEST RUN DOWN         =     4
        LENGTH OF THE LONGEST RUN UP OR DOWN =     5

        NUMBER OF POSITIVE DIFFERENCES =   252
        NUMBER OF NEGATIVE DIFFERENCES =   247
        NUMBER OF ZERO     DIFFERENCES =     0
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. The runs test does not indicate any significant non-randomness.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot. For this data set the correlation coefficient is 0.996. Since this is greater than the critical value of 0.987 (this is a tabulated value), the normality assumption is not rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. Dataplot generates the following output for the Anderson-Darling normality test.

```
               ANDERSON-DARLING 1-SAMPLE TEST
               THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

 1. STATISTICS:
       NUMBER OF OBSERVATIONS                =      500
       MEAN                                  =   -0.2935997E-02
       STANDARD DEVIATION                    =    1.021041

       ANDERSON-DARLING TEST STATISTIC VALUE =    1.061249
       ADJUSTED TEST STATISTIC VALUE         =    1.069633

 2. CRITICAL VALUES:
       90         % POINT    =   0.6560000
       95         % POINT    =   0.7870000
       97.5       % POINT    =   0.9180000
       99         % POINT    =    1.092000

 3. CONCLUSION (AT THE 5% LEVEL):
       THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.
```

The Anderson-Darling test rejects the normality assumption at the 5% level but accepts it at the 1% level.

Outlier Analysis

A test for outliers is the Grubbs test. Dataplot generated the following output for Grubbs' test.

```
                GRUBBS TEST FOR OUTLIERS
                (ASSUMPTION: NORMALITY)

 1. STATISTICS:
       NUMBER OF OBSERVATIONS      =      500
       MINIMUM                     =   -2.647000
       MEAN                        =  -0.2935997E-02
       MAXIMUM                     =    3.436000
    STANDARD DEVIATION          =    1.021041

    GRUBBS TEST STATISTIC       =    3.368068

 2. PERCENT POINTS OF THE REFERENCE DISTRIBUTION
    FOR GRUBBS TEST STATISTIC
       0          % POINT    =   0.0000000E+00
       50         % POINT    =    3.274338
       75         % POINT    =    3.461431
       90         % POINT    =    3.695134
       95         % POINT    =    3.863087
       99         % POINT    =    4.228033

 3. CONCLUSION (AT THE 5% LEVEL):
       THERE ARE NO OUTLIERS.
```

For this data set, Grubbs' test does not detect any outliers at the 25%, 10%, 5%, and 1% significance levels.

Model

Since the underlying assumptions were validated both graphically and analytically, we conclude that a reasonable model for the data is:

$$\boldsymbol{Y_i = -0.00294 + E_i}$$

We can express the uncertainty for ***C*** as the 95% confidence interval (-0.09266,0.086779).

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report. The report for the 500 normal random numbers follows.

```
 Analysis for 500 normal random numbers

 1: Sample Size                           = 500

 2: Location
    Mean                                  = -0.00294
    Standard Deviation of Mean            = 0.045663
    95% Confidence Interval for Mean      = (-0.09266,0.086779)
    Drift with respect to location?       = NO

 3: Variation
    Standard Deviation                    = 1.021042
    95% Confidence Interval for SD        = (0.961437,1.088585)
    Drift with respect to variation?
    (based on Bartletts test on quarters
    of the data)                          = NO

 4: Distribution
    Normal PPCC                           = 0.996173
```

```
   Data are Normal?
     (as measured by Normal PPCC)        = YES

5: Randomness
   Autocorrelation                       = 0.045059
   Data are Random?
     (as measured by autocorrelation)    = YES

6: Statistical Control
   (i.e., no drift in location or scale,
   data are random, distribution is
   fixed, here we are testing only for
   fixed normal)
   Data Set is in Statistical Control?   = YES

7: Outliers?
   (as determined by Grubbs' test)       = NO
```

1.4.2.1.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale, and the data seem to follow a normal distribution.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a

4. Generate a normal probability plot.	normal distribution is a good distribution for these data. 4. The normal probability plot verifies that the normal distribution is a reasonable distribution for these data.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics.	1. The summary statistics table displays 25+ statistics.
2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location.	2. The mean is -0.00294 and a 95% confidence interval is (-0.093,0.087). The linear fit indicates no drift in location since the slope parameter is statistically not significant.
3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Barltett's test for equal standard deviations.	3. The standard deviation is 1.02 with a 95% confidence interval of (0.96,1.09). Bartlett's test indicates no significant change in variation.
4. Check for randomness by generating an autocorrelation plot and a runs test.	4. The lag 1 autocorrelation is 0.04. From the autocorrelation plot, this is within the 95% confidence interval bands.
5. Check for normality by computing the normal probability plot correlation coefficient.	5. The normal probability plot correlation coefficient is 0.996. At the 5% level, we cannot reject the normality assumption.
6. Check for outliers using Grubbs' test.	6. Grubbs' test detects no outliers at the 5% level.
7. Print a univariate report (this assumes steps 2 thru 6 have already been run).	7. The results are summarized in a convenient report.

1.4.2.2. Uniform Random Numbers

Uniform Random Numbers

This example illustrates the univariate analysis of a set of uniform random numbers.

1.4.2.2.1. Background and Data

Generation

The uniform random numbers used in this case study are from a Rand Corporation publication.

The motivation for studying a set of uniform random numbers is to illustrate the effects of a known underlying non-normal distribution.

Software

Most general purpose statistical software programs, including Dataplot, can generate uniform random numbers.

Resulting Data

The following is the set of uniform random numbers used for this case study.

```
.100973   .253376   .520135   .863467   .354876
.809590   .911739   .292749   .375420   .480564
.894742   .962480   .524037   .206361   .040200
.822916   .084226   .895319   .645093   .032320
.902560   .159533   .476435   .080336   .990190
.252909   .376707   .153831   .131165   .886767
.439704   .436276   .128079   .997080   .157361
.476403   .236653   .989511   .687712   .171768
.660657   .471734   .072768   .503669   .736170
.658133   .988511   .199291   .310601   .080545
.571824   .063530   .342614   .867990   .743923
.403097   .852697   .760202   .051656   .926866
.574818   .730538   .524718   .623885   .635733
.213505   .325470   .489055   .357548   .284682
.870983   .491256   .737964   .575303   .529647
.783580   .834282   .609352   .034435   .273884
.985201   .776714   .905686   .072210   .940558
.609709   .343350   .500739   .118050   .543139
.808277   .325072   .568248   .294052   .420152
.775678   .834529   .963406   .288980   .831374
```

.670078	.184754	.061068	.711778	.886854
.020086	.507584	.013676	.667951	.903647
.649329	.609110	.995946	.734887	.517649
.699182	.608928	.937856	.136823	.478341
.654811	.767417	.468509	.505804	.776974
.730395	.718640	.218165	.801243	.563517
.727080	.154531	.822374	.211157	.825314
.385537	.743509	.981777	.402772	.144323
.600210	.455216	.423796	.286026	.699162
.680366	.252291	.483693	.687203	.766211
.399094	.400564	.098932	.050514	.225685
.144642	.756788	.962977	.882254	.382145
.914991	.452368	.479276	.864616	.283554
.947508	.992337	.089200	.803369	.459826
.940368	.587029	.734135	.531403	.334042
.050823	.441048	.194985	.157479	.543297
.926575	.576004	.088122	.222064	.125507
.374211	.100020	.401286	.074697	.966448
.943928	.707258	.636064	.932916	.505344
.844021	.952563	.436517	.708207	.207317
.611969	.044626	.457477	.745192	.433729
.653945	.959342	.582605	.154744	.526695
.270799	.535936	.783848	.823961	.011833
.211594	.945572	.857367	.897543	.875462
.244431	.911904	.259292	.927459	.424811
.621397	.344087	.211686	.848767	.030711
.205925	.701466	.235237	.831773	.208898
.376893	.591416	.262522	.966305	.522825
.044935	.249475	.246338	.244586	.251025
.619627	.933565	.337124	.005499	.765464
.051881	.599611	.963896	.546928	.239123
.287295	.359631	.530726	.898093	.543335
.135462	.779745	.002490	.103393	.598080
.839145	.427268	.428360	.949700	.130212
.489278	.565201	.460588	.523601	.390922
.867728	.144077	.939108	.364770	.617429
.321790	.059787	.379252	.410556	.707007
.867431	.715785	.394118	.692346	.140620
.117452	.041595	.660000	.187439	.242397
.118963	.195654	.143001	.758753	.794041
.921585	.666743	.680684	.962852	.451551
.493819	.476072	.464366	.794543	.590479
.003320	.826695	.948643	.199436	.168108
.513488	.881553	.015403	.545605	.014511
.980862	.482645	.240284	.044499	.908896
.390947	.340735	.441318	.331851	.623241

.941509	.498943	.548581	.886954	.199437
.548730	.809510	.040696	.382707	.742015
.123387	.250162	.529894	.624611	.797524
.914071	.961282	.966986	.102591	.748522
.053900	.387595	.186333	.253798	.145065
.713101	.024674	.054556	.142777	.938919
.740294	.390277	.557322	.709779	.017119
.525275	.802180	.814517	.541784	.561180
.993371	.430533	.512969	.561271	.925536
.040903	.116644	.988352	.079848	.275938
.171539	.099733	.344088	.461233	.483247
.792831	.249647	.100229	.536870	.323075
.754615	.020099	.690749	.413887	.637919
.763558	.404401	.105182	.161501	.848769
.091882	.009732	.825395	.270422	.086304
.833898	.737464	.278580	.900458	.549751
.981506	.549493	.881997	.918707	.615068
.476646	.731895	.020747	.677262	.696229
.064464	.271246	.701841	.361827	.757687
.649020	.971877	.499042	.912272	.953750
.587193	.823431	.540164	.405666	.281310
.030068	.227398	.207145	.329507	.706178
.083586	.991078	.542427	.851366	.158873
.046189	.755331	.223084	.283060	.326481
.333105	.914051	.007893	.326046	.047594
.119018	.538408	.623381	.594136	.285121
.590290	.284666	.879577	.762207	.917575
.374161	.613622	.695026	.390212	.557817
.651483	.483470	.894159	.269400	.397583
.911260	.717646	.489497	.230694	.541374
.775130	.382086	.864299	.016841	.482774
.519081	.398072	.893555	.195023	.717469
.979202	.885521	.029773	.742877	.525165
.344674	.218185	.931393	.278817	.570568

1.4.2.2.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

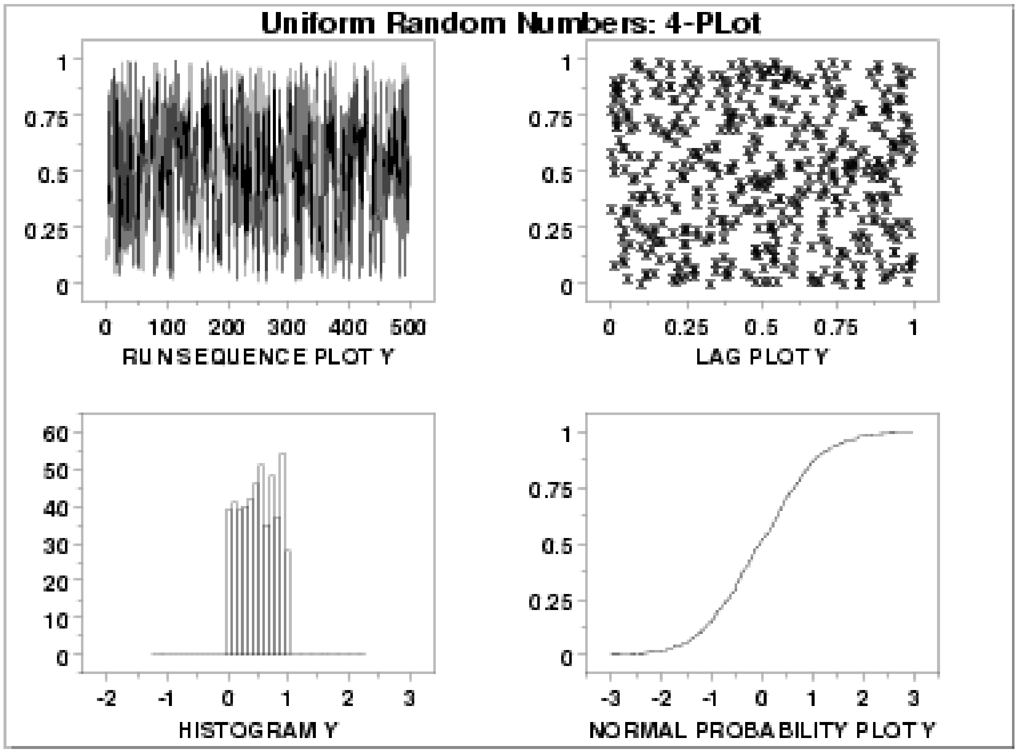
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

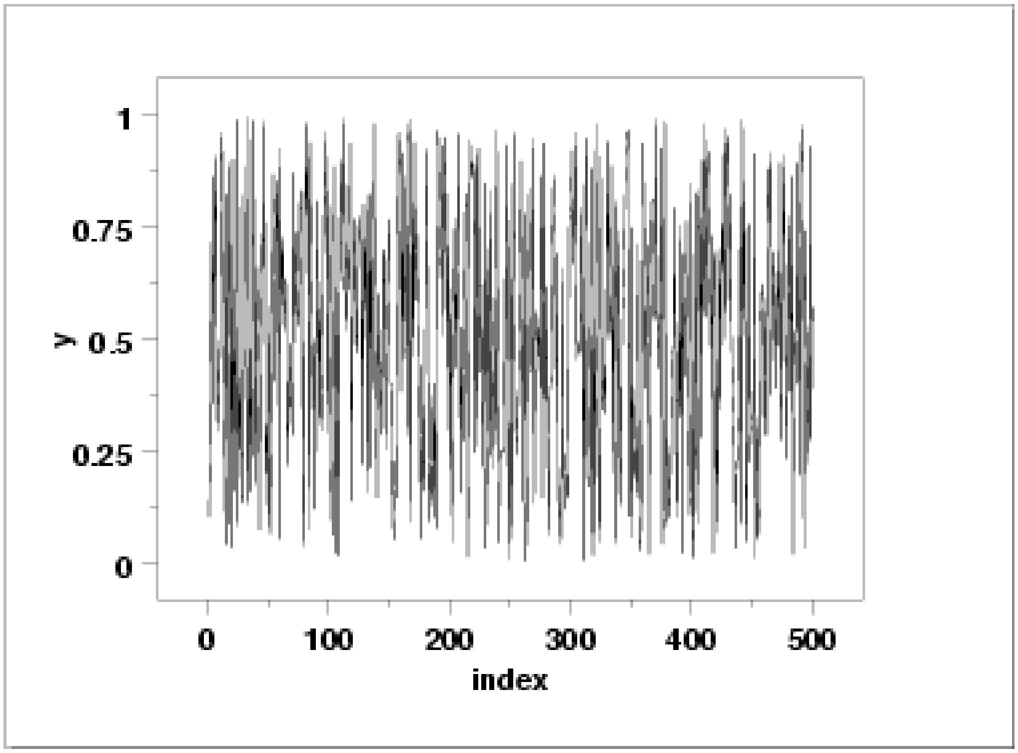
1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram shows that the frequencies are relatively flat across the range of the data. This suggests that the uniform distribution might provide a better distributional fit than the normal distribution.
4. The normal probability plot verifies that an assumption of normality is not reasonable. In this case, the 4-plot should be followed up by a uniform probability plot to determine if it provides a better fit to the data. This is shown below.

From the above plots, we conclude that the underlying assumptions are valid. Therefore, the model $Y_i = C + E_i$ is valid. However, since the data are not normally distributed, using the mean as an estimate of C and the confidence interval cited above for quantifying its uncertainty are not valid or appropriate.

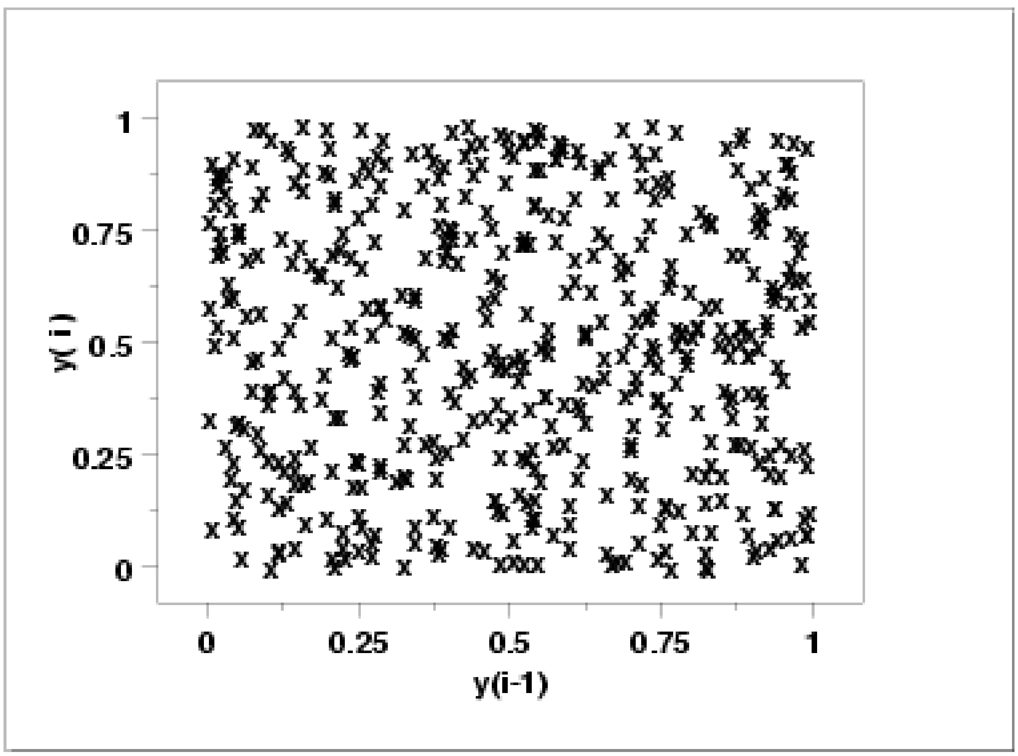
Individual Plots

Although it is usually not necessary, the plots can be generated individually to give more detail.

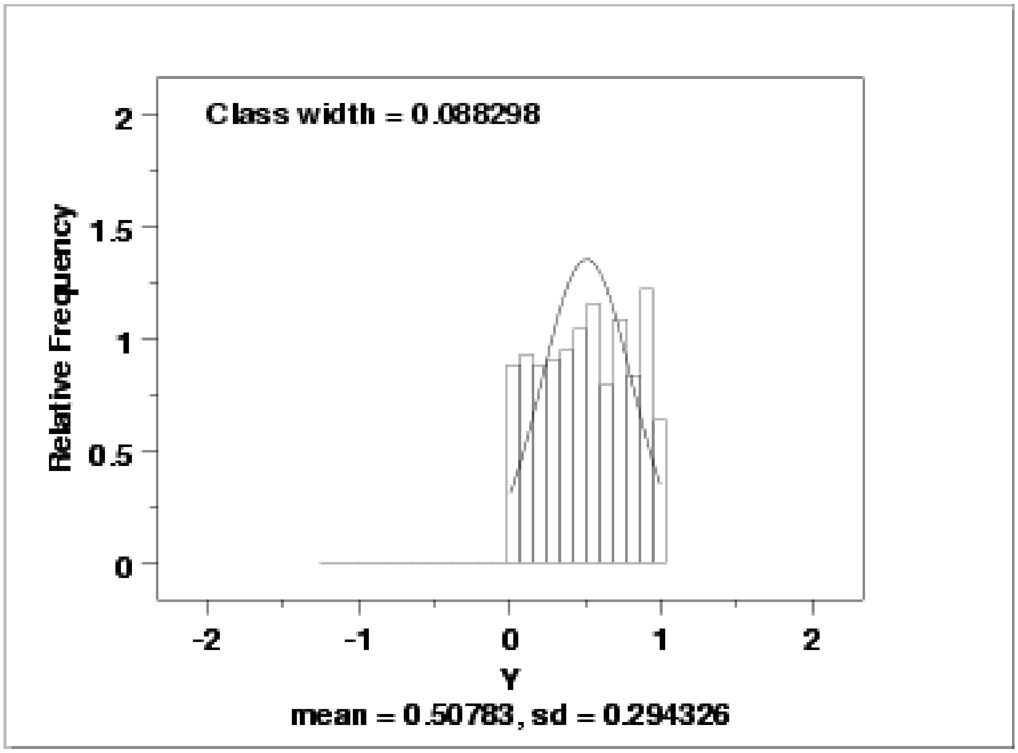
Run Sequence Plot



Lag Plot

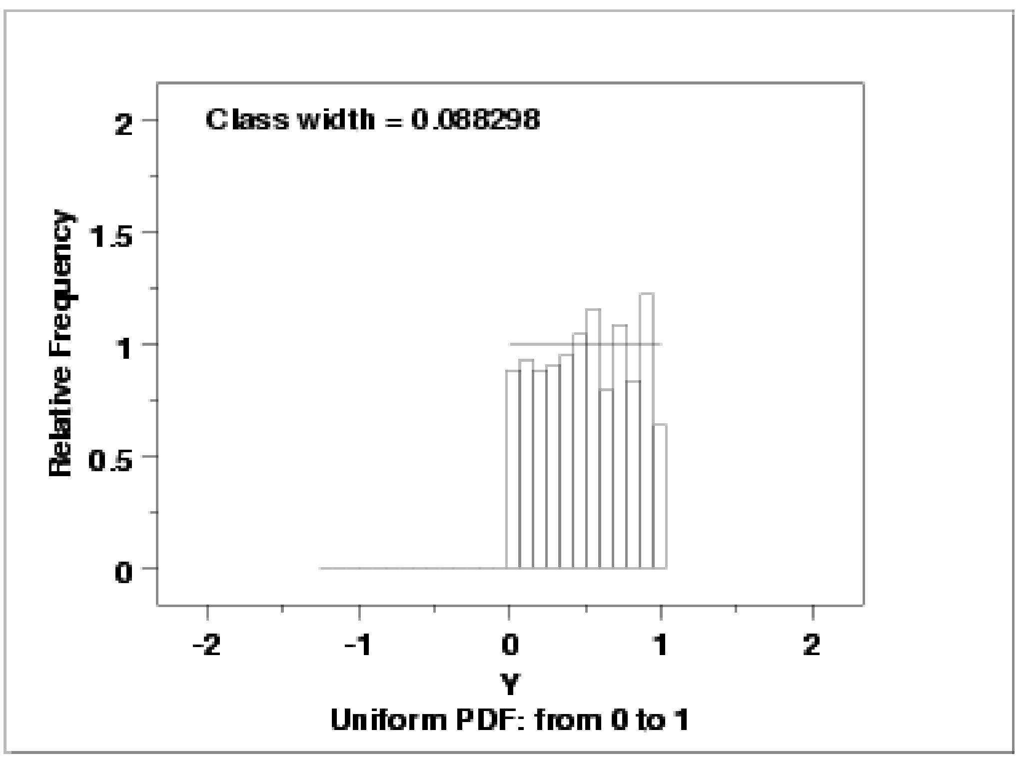


Histogram (with overlaid Normal PDF)



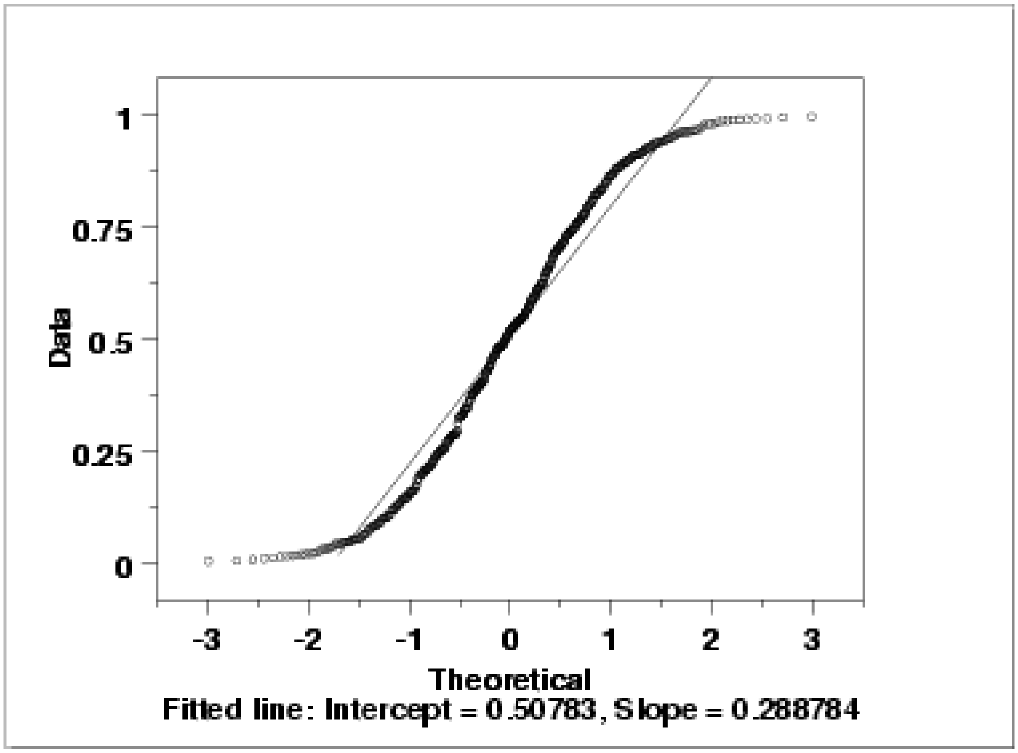
This plot shows that a normal distribution is a poor fit. The flatness of the histogram suggests that a uniform distribution might be a better fit.

Histogram (with overlaid Uniform PDF)



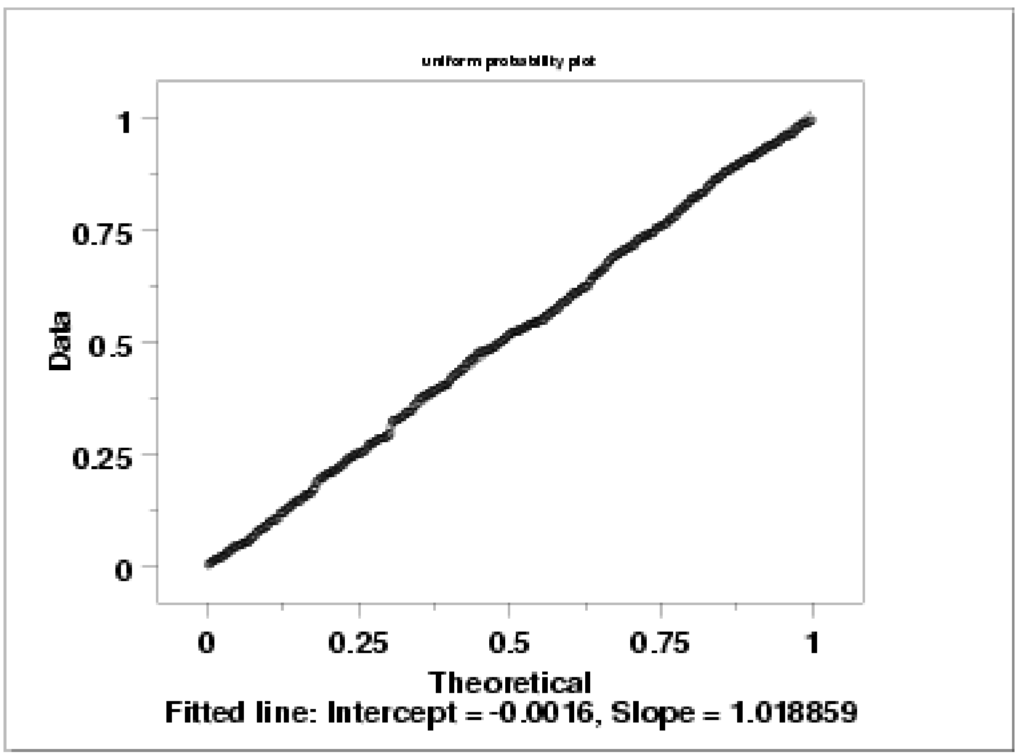
Since the histogram from the 4-plot suggested that the uniform distribution might be a good fit, we overlay a uniform distribution on top of the histogram. This indicates a much better fit than a normal distribution.

Normal Probability Plot



As with the histogram, the normal probability plot shows that the normal distribution does not fit these data well.

Uniform Probability Plot



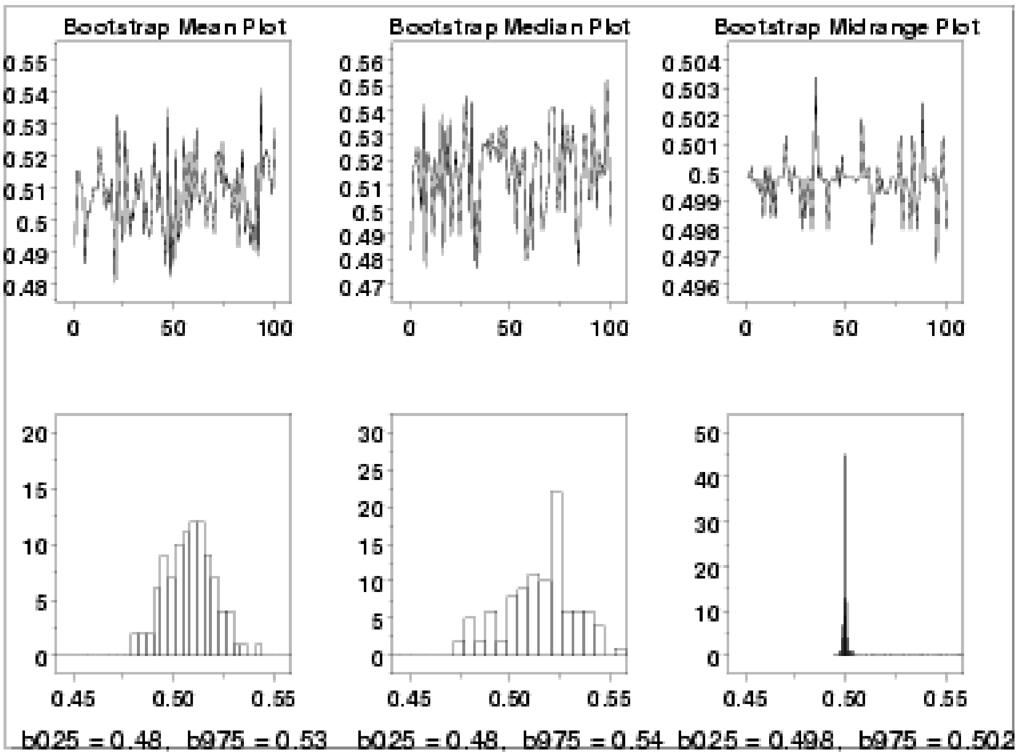
Since the above plots suggested that a uniform distribution might be appropriate, we generate a uniform probability plot. This plot shows that the uniform distribution provides an excellent fit to the data.

Better Model

Since the data follow the underlying assumptions, but with a uniform distribution rather than a normal distribution, we would still like to characterize ***C*** by a typical value plus or minus a confidence interval. In this case, we would like to find a location estimator with the smallest variability.

The bootstrap plot is an ideal tool for this purpose. The following plots show the bootstrap plot, with the corresponding histogram, for the mean, median, mid-range, and median absolute deviation.

Bootstrap Plots



Mid-Range is Best

From the above histograms, it is obvious that for these data, the mid-range is far superior to the mean or median as an estimate for location.

Using the mean, the location estimate is 0.507 and a 95% confidence interval for the mean is (0.482,0.534). Using the mid-range, the location estimate is 0.499 and the 95% confidence interval for the mid-range is (0.497,0.503).

Although the values for the location are similar, the difference in the uncertainty intervals is quite large.

Note that in the case of a uniform distribution it is known theoretically that the mid-range is the best linear unbiased estimator for location. However, in many applications, the most appropriate estimator will not be known or it will be mathematically intractable to determine a valid condfidence interval. The bootstrap provides a method for determining (and comparing) confidence intervals in these cases.

1.4.2.2.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                                  SUMMARY

                       NUMBER OF OBSERVATIONS =      500


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.4997850E+00  *  RANGE        =   0.9945900E+00  *
*  MEAN         =   0.5078304E+00  *  STAND. DEV.  =   0.2943252E+00  *
*  MIDMEAN      =   0.5045621E+00  *  AV. AB. DEV. =   0.2526468E+00  *
*  MEDIAN       =   0.5183650E+00  *  MINIMUM      =   0.2490000E-02  *
*               =                  *  LOWER QUART. =   0.2508093E+00  *
*               =                  *  LOWER HINGE  =   0.2505935E+00  *
*               =                  *  UPPER HINGE  =   0.7594775E+00  *
*               =                  *  UPPER QUART. =   0.7591152E+00  *
*               =                  *  MAXIMUM      =   0.9970800E+00  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =  -0.3098569E-01  *  ST. 3RD MOM. =  -0.3443941E-01  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.1796969E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =  -0.2004886E+02  *
*               =                  *  UNIFORM PPCC =   0.9995682E+00  *
*               =                  *  NORMAL  PPCC =   0.9771602E+00  *
*               =                  *  TUK -.5 PPCC =   0.7229201E+00  *
*               =                  *  CAUCHY  PPCC =   0.3591767E+00  *
***********************************************************************
```

Note that under the distributional measures the uniform probability plot correlation coefficient (PPCC) value is significantly larger than the normal PPCC value. This is evidence that the uniform distribution fits these data better than does a normal distribution.

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generated the following output:

```
LEAST SQUARES MULTILINEAR FIT
SAMPLE SIZE N        =      500
NUMBER OF VARIABLES =         1
NO REPLICATION CASE


        PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
 1  A0                  0.522923        (0.2638E-01)        19.82
 2  A1        X        -0.602478E-04   (0.9125E-04)        -0.6603

RESIDUAL    STANDARD DEVIATION =         0.2944917
RESIDUAL    DEGREES OF FREEDOM =         498
```

The slope parameter, A1, has a t value of -0.66 which is statistically not significant. This indicates that the slope can in fact be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since we know this data set is not approximated well by the normal distribution, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

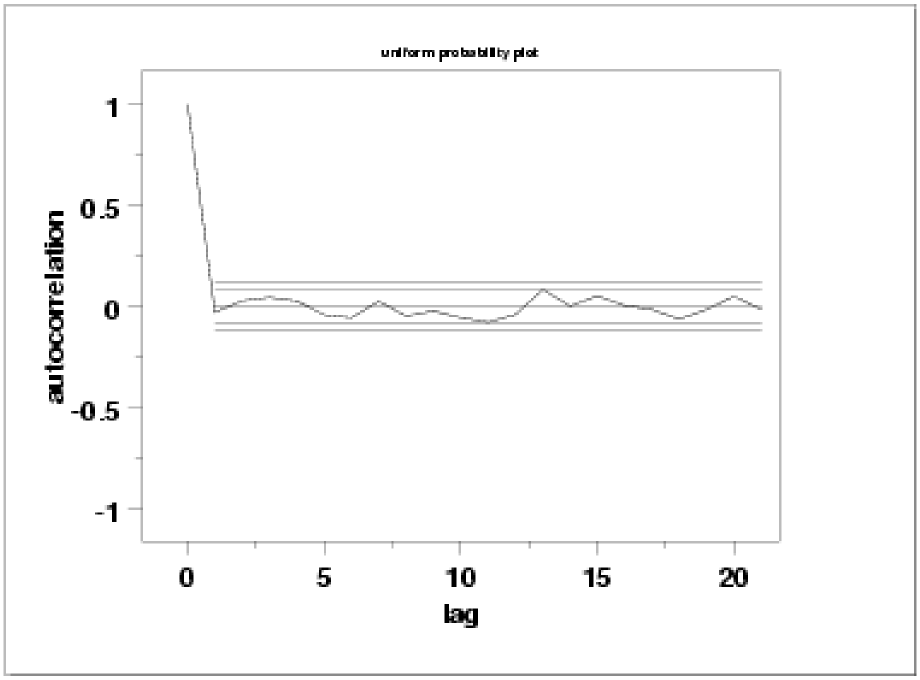
```
               LEVENE F-TEST FOR SHIFT IN VARIATION
                      (ASSUMPTION: NORMALITY)

 1. STATISTICS
       NUMBER OF OBSERVATIONS    =      500
       NUMBER OF GROUPS          =        4
       LEVENE F TEST STATISTIC   =   0.7983007E-01


    FOR LEVENE TEST STATISTIC
       0           % POINT    =   0.0000000E+00
       50          % POINT    =   0.7897459
       75          % POINT    =    1.373753
       90          % POINT    =    2.094885
       95          % POINT    =    2.622929
       99          % POINT    =    3.821479
       99.9        % POINT    =    5.506884


          2.905608       % Point:    0.7983007E-01

 3. CONCLUSION (AT THE 5% LEVEL):
       THERE IS NO SHIFT IN VARIATION.
       THUS: HOMOGENEOUS WITH RESPECT TO VARIATION.
```

In this case, the Levene test indicates that the standard deviations are not significantly different in the 4 intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot in the previous section is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted using 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of most interest, is 0.03. The critical values at the 5% significance level are -0.087 and 0.087. This indicates that the lag 1 autocorrelation is not statistically significant, so there is no evidence of non-randomness.

A common test for randomness is the runs test.

```
                    RUNS UP
         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH EXACTLY I
 I         STAT      EXP(STAT)    SD(STAT)        Z

 1       103.0    104.2083     10.2792       -0.12
 2        48.0     45.7167      5.2996        0.43
 3        11.0     13.1292      3.2297       -0.66
 4         6.0      2.8563      1.6351        1.92
 5         0.0      0.5037      0.7045       -0.71
 6         0.0      0.0749      0.2733       -0.27
 7         1.0      0.0097      0.0982       10.08
 8         0.0      0.0011      0.0331       -0.03
 9         0.0      0.0001      0.0106       -0.01
10         0.0      0.0000      0.0032        0.00
         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH I OR MORE
 I         STAT      EXP(STAT)    SD(STAT)        Z
```

```
  1       169.0    166.5000      6.6546        0.38
  2        66.0     62.2917      4.4454        0.83
  3        18.0     16.5750      3.4338        0.41
  4         7.0      3.4458      1.7786        2.00
  5         1.0      0.5895      0.7609        0.54
  6         1.0      0.0858      0.2924        3.13
  7         1.0      0.0109      0.1042        9.49
  8         0.0      0.0012      0.0349       -0.03
  9         0.0      0.0001      0.0111       -0.01
 10         0.0      0.0000      0.0034        0.00
                  RUNS DOWN
          STATISTIC = NUMBER OF RUNS DOWN
              OF LENGTH EXACTLY I
  I         STAT     EXP(STAT)    SD(STAT)        Z

  1       113.0    104.2083     10.2792        0.86
  2        43.0     45.7167      5.2996       -0.51
  3        11.0     13.1292      3.2297       -0.66
  4         1.0      2.8563      1.6351       -1.14
  5         0.0      0.5037      0.7045       -0.71
  6         0.0      0.0749      0.2733       -0.27
  7         0.0      0.0097      0.0982       -0.10
  8         0.0      0.0011      0.0331       -0.03
  9         0.0      0.0001      0.0106       -0.01
 10         0.0      0.0000      0.0032        0.00
          STATISTIC = NUMBER OF RUNS DOWN
              OF LENGTH I OR MORE
  I         STAT     EXP(STAT)    SD(STAT)        Z

  1       168.0    166.5000      6.6546        0.23
  2        55.0     62.2917      4.4454       -1.64
  3        12.0     16.5750      3.4338       -1.33
  4         1.0      3.4458      1.7786       -1.38
  5         0.0      0.5895      0.7609       -0.77
  6         0.0      0.0858      0.2924       -0.29
  7         0.0      0.0109      0.1042       -0.10
  8         0.0      0.0012      0.0349       -0.03
  9         0.0      0.0001      0.0111       -0.01
 10         0.0      0.0000      0.0034        0.00
          RUNS TOTAL = RUNS UP + RUNS DOWN
        STATISTIC = NUMBER OF RUNS TOTAL
             OF LENGTH EXACTLY I
  I         STAT     EXP(STAT)    SD(STAT)        Z

  1       216.0    208.4167     14.5370        0.52
  2        91.0     91.4333      7.4947       -0.06
  3        22.0     26.2583      4.5674       -0.93
  4         7.0      5.7127      2.3123        0.56
  5         0.0      1.0074      0.9963       -1.01
  6         0.0      0.1498      0.3866       -0.39
  7         1.0      0.0193      0.1389        7.06
  8         0.0      0.0022      0.0468       -0.05
  9         0.0      0.0002      0.0150       -0.01
 10         0.0      0.0000      0.0045        0.00
        STATISTIC = NUMBER OF RUNS TOTAL
              OF LENGTH I OR MORE
  I         STAT     EXP(STAT)    SD(STAT)        Z

  1       337.0    333.0000      9.4110        0.43
  2       121.0    124.5833      6.2868       -0.57
  3        30.0     33.1500      4.8561       -0.65
```

```
 4          8.0       6.8917      2.5154         0.44
 5          1.0       1.1790      1.0761        -0.17
 6          1.0       0.1716      0.4136         2.00
 7          1.0       0.0217      0.1474         6.64
 8          0.0       0.0024      0.0494        -0.05
 9          0.0       0.0002      0.0157        -0.02
10          0.0       0.0000      0.0047         0.00
        LENGTH OF THE LONGEST RUN UP          =     7
        LENGTH OF THE LONGEST RUN DOWN        =     4
        LENGTH OF THE LONGEST RUN UP OR DOWN =     7

        NUMBER OF POSITIVE DIFFERENCES =   263
        NUMBER OF NEGATIVE DIFFERENCES =   236
        NUMBER OF ZERO     DIFFERENCES =     0
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. This runs test does not indicate any significant non-randomness. There is a statistically significant value for runs of length 7. However, further examination of the table shows that there is in fact a single run of length 7 when near 0 are expected. This is not sufficient evidence to conclude that the data are non-random.

Distributional Analysis

Probability plots are a graphical test of assessing whether a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot. For this data set the correlation coefficient, from the summary table above, is 0.977. Since this is less than the critical value of 0.987 (this is a tabulated value), the normality assumption is rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. Dataplot generates the following output for the Anderson-Darling normality test.

```
               ANDERSON-DARLING 1-SAMPLE TEST
               THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

 1. STATISTICS:
       NUMBER OF OBSERVATIONS                =      500
       MEAN                                  =   0.5078304
       STANDARD DEVIATION                    =   0.2943252

       ANDERSON-DARLING TEST STATISTIC VALUE =    5.719849
       ADJUSTED TEST STATISTIC VALUE         =    5.765036

 2. CRITICAL VALUES:
       90          % POINT    =   0.6560000
       95          % POINT    =   0.7870000
       97.5        % POINT    =   0.9180000
       99          % POINT    =    1.092000

 3. CONCLUSION (AT THE 5% LEVEL):
       THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.
```

The Anderson-Darling test rejects the normality assumption because the value of the test statistic, 5.72, is larger than the critical value of 1.092 at the 1% significance level.

Model

Based on the graphical and quantitative analysis, we use the model

$$\boldsymbol{Y_i = C + E_i}$$

where $\boldsymbol{C}$ is estimated by the mid-range and the uncertainty interval for $\boldsymbol{C}$ is based on a bootstrap analysis. Specifically,

$\boldsymbol{C}$ = 0.499
95% confidence limit for $\boldsymbol{C}$ = (0.497,0.503)

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report. The report for the 500 uniform random numbers follows.

```
Analysis for 500 uniform random numbers

1: Sample Size                              = 500

2: Location
   Mean                                     = 0.50783
   Standard Deviation of Mean               = 0.013163
   95% Confidence Interval for Mean         = (0.48197,0.533692)
   Drift with respect to location?          = NO

3: Variation
   Standard Deviation                       = 0.294326
   95% Confidence Interval for SD           = (0.277144,0.313796)
   Drift with respect to variation?
   (based on Levene's test on quarters
   of the data)                             = NO

4: Distribution
   Normal PPCC                              = 0.999569
   Data are Normal?
     (as measured by Normal PPCC)           = NO

   Uniform PPCC                             = 0.9995
   Data are Uniform?
     (as measured by Uniform PPCC)          = YES

5: Randomness
   Autocorrelation                          = -0.03099
   Data are Random?
     (as measured by autocorrelation)       = YES

6: Statistical Control
   (i.e., no drift in location or scale,
   data is random, distribution is
   fixed, here we are testing only for
   fixed uniform)
   Data Set is in Statistical Control?      = YES
```

1.4.2.2.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	 1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	 1. Based on the 4-plot, there are no shifts in location or scale, and the data do not seem to follow a normal distribution.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf.	 1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is not a good

4. Generate a histogram with an overlaid uniform pdf. 5. Generate a normal probability plot. 6. Generate a uniform probability plot.	distribution for these data. 4. The histogram indicates that a uniform distribution is a good distribution for these data. 5. The normal probability plot verifies that the normal distribution is not a reasonable distribution for these data. 6. The uniform probability plot verifies that the uniform distribution is a reasonable distribution for these data.
4. Generate the bootstrap plot. 1. Generate a bootstrap plot.	1. The bootstrap plot clearly shows the superiority of the mid-range over the mean and median as the location estimator of choice for this problem.
5. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Barltetts test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Check for normality by computing the normal probability plot correlation coefficient.	1. The summary statistics table displays 25+ statistics. 2. The mean is 0.5078 and a 95% confidence interval is (0.482,0.534). The linear fit indicates no drift in location since the slope parameter is statistically not significant. 3. The standard deviation is 0.29 with a 95% confidence interval of (0.277,0.314). Levene's test indicates no significant drift in variation. 4. The lag 1 autocorrelation is -0.03. From the autocorrelation plot, this is within the 95% confidence interval bands. 5. The uniform probability plot correlation coefficient is 0.9995. This indicates that the uniform distribution is a good fit.

6. Print a univariate report (this assumes steps 2 thru 6 have already been run).	6. The results are summarized in a convenient report.

1.4.2.3. Random Walk

Random Walk

This example illustrates the univariate analysis of a set of numbers derived from a random walk.

1.4.2.3.1. Background and Data

Generation

A random walk can be generated from a set of uniform random numbers by the formula:

$$R_i = \sum_{j=1}^{i} (U_j - 0.5)$$

where U is a set of uniform random numbers.

The motivation for studying a set of random walk data is to illustrate the effects of a known underlying autocorrelation structure (i.e., non-randomness) in the data.

Software

Most general purpose statistical software programs, including Dataplot, can generate data for a random walk.

Resulting Data

The following is the set of random walk numbers used for this case study.

```
-0.399027
-0.645651
-0.625516
-0.262049
-0.407173
-0.097583
 0.314156
 0.106905
-0.017675
-0.037111
 0.357631
 0.820111
 0.844148
 0.550509
 0.090709
```

```
 0.413625
-0.002149
 0.393170
 0.538263
 0.070583
 0.473143
 0.132676
 0.109111
-0.310553
 0.179637
-0.067454
-0.190747
-0.536916
-0.905751
-0.518984
-0.579280
-0.643004
-1.014925
-0.517845
-0.860484
-0.884081
-1.147428
-0.657917
-0.470205
-0.798437
-0.637780
-0.666046
-1.093278
-1.089609
-0.853439
-0.695306
-0.206795
-0.507504
-0.696903
-1.116358
-1.044534
-1.481004
-1.638390
-1.270400
-1.026477
-1.123380
-0.770683
-0.510481
-0.958825
-0.531959
-0.457141
```

```
-0.226603
-0.201885
-0.078000
 0.057733
-0.228762
-0.403292
-0.414237
-0.556689
-0.772007
-0.401024
-0.409768
-0.171804
-0.096501
-0.066854
 0.216726
 0.551008
 0.660360
 0.194795
-0.031321
 0.453880
 0.730594
 1.136280
 0.708490
 1.149048
 1.258757
 1.102107
 1.102846
 0.720896
 0.764035
 1.072312
 0.897384
 0.965632
 0.759684
 0.679836
 0.955514
 1.290043
 1.753449
 1.542429
 1.873803
 2.043881
 1.728635
 1.289703
 1.501481
 1.888335
 1.408421
 1.416005
```

```
0.929681
1.097632
1.501279
1.650608
1.759718
2.255664
2.490551
2.508200
2.707382
2.816310
3.254166
2.890989
2.869330
3.024141
3.291558
3.260067
3.265871
3.542845
3.773240
3.991880
3.710045
4.011288
4.074805
4.301885
3.956416
4.278790
3.989947
4.315261
4.200798
4.444307
4.926084
4.828856
4.473179
4.573389
4.528605
4.452401
4.238427
4.437589
4.617955
4.370246
4.353939
4.541142
4.807353
4.706447
4.607011
4.205943
```

```
3.756457
3.482142
3.126784
3.383572
3.846550
4.228803
4.110948
4.525939
4.478307
4.457582
4.822199
4.605752
5.053262
5.545598
5.134798
5.438168
5.397993
5.838361
5.925389
6.159525
6.190928
6.024970
5.575793
5.516840
5.211826
4.869306
4.912601
5.339177
5.415182
5.003303
4.725367
4.350873
4.225085
3.825104
3.726391
3.301088
3.767535
4.211463
4.418722
4.554786
4.987701
4.993045
5.337067
5.789629
5.726147
5.934353
```

```
5.641670
5.753639
5.298265
5.255743
5.500935
5.434664
5.588610
6.047952
6.130557
5.785299
5.811995
5.582793
5.618730
5.902576
6.226537
5.738371
5.449965
5.895537
6.252904
6.650447
7.025909
6.770340
7.182244
6.941536
7.368996
7.293807
7.415205
7.259291
6.970976
7.319743
6.850454
6.556378
6.757845
6.493083
6.824855
6.533753
6.410646
6.502063
6.264585
6.730889
6.753715
6.298649
6.048126
5.794463
5.539049
5.290072
```

```
5.409699
5.843266
5.680389
5.185889
5.451353
5.003233
5.102844
5.566741
5.613668
5.352791
5.140087
4.999718
5.030444
5.428537
5.471872
5.107334
5.387078
4.889569
4.492962
4.591042
4.930187
4.857455
4.785815
5.235515
4.865727
4.855005
4.920206
4.880794
4.904395
4.795317
5.163044
4.807122
5.246230
5.111000
5.228429
5.050220
4.610006
4.489258
4.399814
4.606821
4.974252
5.190037
5.084155
5.276501
4.917121
4.534573
```

```
4.076168
4.236168
3.923607
3.666004
3.284967
2.980621
2.623622
2.882375
3.176416
3.598001
3.764744
3.945428
4.408280
4.359831
4.353650
4.329722
4.294088
4.588631
4.679111
4.182430
4.509125
4.957768
4.657204
4.325313
4.338800
4.720353
4.235756
4.281361
3.795872
4.276734
4.259379
3.999663
3.544163
3.953058
3.844006
3.684740
3.626058
3.457909
3.581150
4.022659
4.021602
4.070183
4.457137
4.156574
4.205304
4.514814
```

```
4.055510
3.938217
4.180232
3.803619
3.553781
3.583675
3.708286
4.005810
4.419880
4.881163
5.348149
4.950740
5.199262
4.753162
4.640757
4.327090
4.080888
3.725953
3.939054
3.463728
3.018284
2.661061
3.099980
3.340274
3.230551
3.287873
3.497652
3.014771
3.040046
3.342226
3.656743
3.698527
3.759707
4.253078
4.183611
4.196580
4.257851
4.683387
4.224290
3.840934
4.329286
3.909134
3.685072
3.356611
2.956344
2.800432
```

```
2.761665
2.744913
3.037743
2.787390
2.387619
2.424489
2.247564
2.502179
2.022278
2.213027
2.126914
2.264833
2.528391
2.432792
2.037974
1.699475
2.048244
1.640126
1.149858
1.475253
1.245675
0.831979
1.165877
1.403341
1.181921
1.582379
1.632130
2.113636
2.163129
2.545126
2.963833
3.078901
3.055547
3.287442
2.808189
2.985451
3.181679
2.746144
2.517390
2.719231
2.581058
2.838745
2.987765
3.459642
3.458684
3.870956
```

```
4.324706
4.411899
4.735330
4.775494
4.681160
4.462470
3.992538
3.719936
3.427081
3.256588
3.462766
3.046353
3.537430
3.579857
3.931223
3.590096
3.136285
3.391616
3.114700
2.897760
2.724241
2.557346
2.971397
2.479290
2.305336
1.852930
1.471948
1.510356
1.633737
1.727873
1.512994
1.603284
1.387950
1.767527
2.029734
2.447309
2.321470
2.435092
2.630118
2.520330
2.578147
2.729630
2.713100
3.107260
2.876659
2.774242
```

```
3.185503
3.403148
3.392646
3.123339
3.164713
3.439843
3.321929
3.686229
3.203069
3.185843
3.204924
3.102996
3.496552
3.191575
3.409044
3.888246
4.273767
3.803540
4.046417
4.071581
3.916256
3.634441
4.065834
3.844651
3.915219
```

1.4.2.3.2. Test Underlying Assumptions

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

$$Y_i = C + E_i$$

   is appropriate and valid.

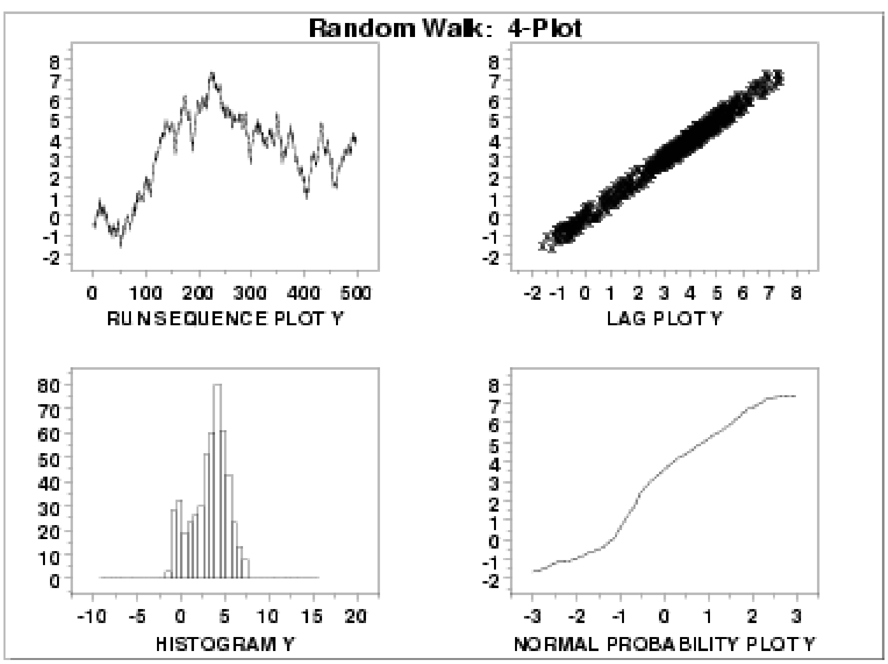
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.

3. Determine if the confidence interval

$$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid, with ***s*** denoting the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

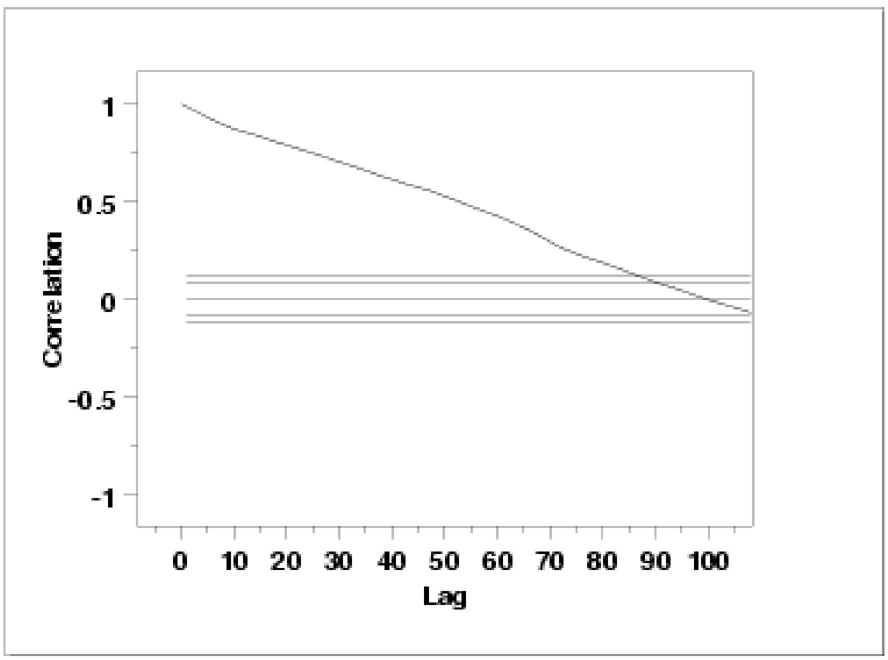
1. The run sequence plot (upper left) indicates significant shifts in location over time.
2. The lag plot (upper right) indicates significant non-randomness in the data.
3. When the assumptions of randomness and constant location and scale are not satisfied, the distributional assumptions are not meaningful. Therefore we do not attempt to make any interpretation of the histogram (lower left) or the normal probability plot (lower right).

From the above plots, we conclude that the underlying assumptions are seriously violated. Therefore the $Y_i = C + E_i$ model is not valid.

When the randomness assumption is seriously violated, a time series model may be appropriate. The lag plot often suggests a reasonable model. For example, in this case the strongly linear appearance of the lag plot suggests a model fitting Y_i versus Y_{i-1} might be appropriate. When the data are non-random, it is helpful to supplement the lag plot with an autocorrelation plot and a spectral plot. Although in this case the lag plot is enough to suggest an appropriate model, we provide the autocorrelation and spectral plots for comparison.

Autocorrelation Plot

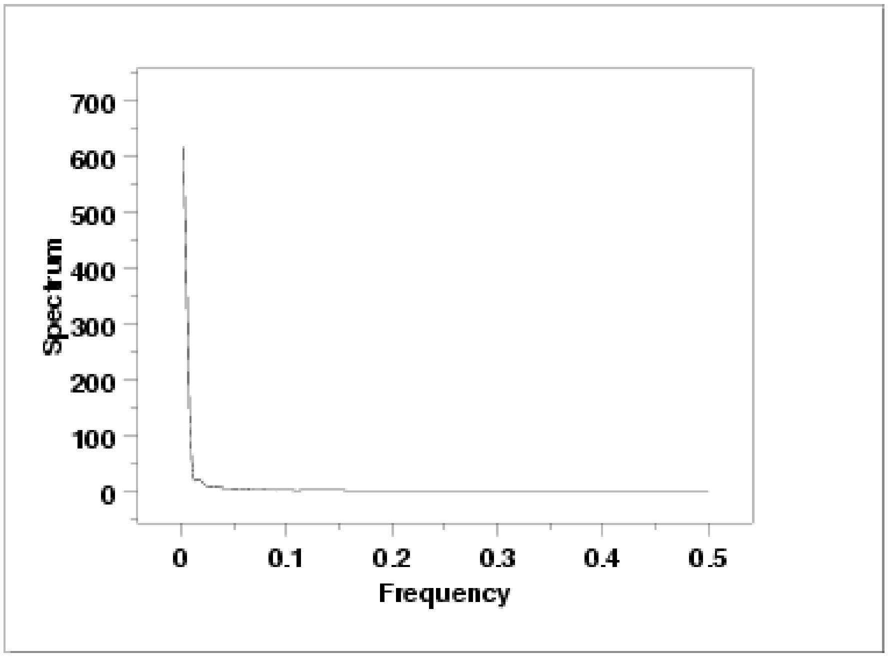
When the lag plot indicates significant non-randomness, it can be helpful to follow up with a an autocorrelation plot.



This autocorrelation plot shows significant autocorrelation at lags 1 through 100 in a linearly decreasing fashion.

Spectral Plot

Another useful plot for non-random data is the spectral plot.



This spectral plot shows a single dominant low frequency peak.

Quantitative Output

Although the 4-plot above clearly shows the violation of the assumptions, we supplement the graphical output with some quantitative measures.

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                                SUMMARY

                     NUMBER OF OBSERVATIONS =      500


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.2888407E+01  *  RANGE        =   0.9053595E+01  *
*  MEAN         =   0.3216681E+01  *  STAND. DEV.  =   0.2078675E+01  *
*  MIDMEAN      =   0.4791331E+01  *  AV. AB. DEV. =   0.1660585E+01  *
*  MEDIAN       =   0.3612030E+01  *  MINIMUM      =  -0.1638390E+01  *
*               =                  *  LOWER QUART. =   0.1747245E+01  *
*               =                  *  LOWER HINGE  =   0.1741042E+01  *
*               =                  *  UPPER HINGE  =   0.4682273E+01  *
*               =                  *  UPPER QUART. =   0.4681717E+01  *
*               =                  *  MAXIMUM      =   0.7415205E+01  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.9868608E+00  *  ST. 3RD MOM. =  -0.4448926E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.2397789E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =  -0.1279870E+02  *
*               =                  *  UNIFORM PPCC =   0.9765666E+00  *
*               =                  *  NORMAL  PPCC =   0.9811183E+00  *
*               =                  *  TUK -.5 PPCC =   0.7754489E+00  *
*               =                  *  CAUCHY  PPCC =   0.4165502E+00  *
***********************************************************************
```

The value of the autocorrelation statistic, 0.987, is evidence of a very strong autocorrelation.

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generates the following output:

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N         =      500
       NUMBER OF VARIABLES =        1
       NO REPLICATION CASE


               PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
        1  A0                      1.83351       (0.1721     )        10.65
        2  A1         X          0.552164E-02   (0.5953E-03)        9.275

       RESIDUAL    STANDARD DEVIATION =         1.921416
       RESIDUAL    DEGREES OF FREEDOM =         498

       COEF AND SD(COEF) WRITTEN OUT TO FILE DPST1F.DAT
       SD(PRED),95LOWER,95UPPER,99LOWER,99UPPER
                         WRITTEN OUT TO FILE DPST2F.DAT
       REGRESSION DIAGNOSTICS WRITTEN OUT TO FILE DPST3F.DAT
       PARAMETER VARIANCE-COVARIANCE MATRIX AND
       INVERSE OF X-TRANSPOSE X MATRIX
       WRITTEN OUT TO FILE DPST4F.DAT
```

The slope parameter, A1, has a t value of 9.3 which is statistically significant. This indicates that the slope cannot in fact be considered zero and so the conclusion is that we do not have constant location.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since we know this data set is not approximated well by the normal distribution, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

```
               LEVENE F-TEST FOR SHIFT IN VARIATION
                      (ASSUMPTION: NORMALITY)

 1. STATISTICS
       NUMBER OF OBSERVATIONS    =      500
       NUMBER OF GROUPS          =        4
       LEVENE F TEST STATISTIC   =    10.45940


    FOR LEVENE TEST STATISTIC
       0          % POINT    =   0.0000000E+00
       50         % POINT    =   0.7897459
       75         % POINT    =    1.373753
       90         % POINT    =    2.094885
       95         % POINT    =    2.622929
       99         % POINT    =    3.821479
       99.9       % POINT    =    5.506884
```

```
                  99.99989       % Point:     10.45940

         3. CONCLUSION (AT THE 5% LEVEL):
               THERE IS A SHIFT IN VARIATION.
               THUS: NOT HOMOGENEOUS WITH RESPECT TO VARIATION.
```

In this case, the Levene test indicates that the standard deviations are significantly different in the 4 intervals since the test statistic of 10.46 is greater than the 95% critical value of 2.62. Therefore we conclude that the scale is not constant.

Randomness

Although the lag 1 autocorrelation coefficient above clearly shows the non-randomness, we show the output from a runs test as well.

```
                    RUNS UP

         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH EXACTLY I

  I         STAT     EXP(STAT)    SD(STAT)        Z

  1         63.0    104.2083     10.2792       -4.01
  2         34.0     45.7167      5.2996       -2.21
  3         17.0     13.1292      3.2297        1.20
  4          4.0      2.8563      1.6351        0.70
  5          1.0      0.5037      0.7045        0.70
  6          5.0      0.0749      0.2733       18.02
  7          1.0      0.0097      0.0982       10.08
  8          1.0      0.0011      0.0331       30.15
  9          0.0      0.0001      0.0106       -0.01
 10          1.0      0.0000      0.0032      311.40


         STATISTIC = NUMBER OF RUNS UP
             OF LENGTH I OR MORE

  I         STAT     EXP(STAT)    SD(STAT)        Z

  1        127.0    166.5000      6.6546       -5.94
  2         64.0     62.2917      4.4454        0.38
  3         30.0     16.5750      3.4338        3.91
  4         13.0      3.4458      1.7786        5.37
  5          9.0      0.5895      0.7609       11.05
  6          8.0      0.0858      0.2924       27.06
  7          3.0      0.0109      0.1042       28.67
  8          2.0      0.0012      0.0349       57.21
  9          1.0      0.0001      0.0111       90.14
 10          1.0      0.0000      0.0034      298.08


                   RUNS DOWN

         STATISTIC = NUMBER OF RUNS DOWN
             OF LENGTH EXACTLY I

  I         STAT     EXP(STAT)    SD(STAT)        Z

  1         69.0    104.2083     10.2792       -3.43
  2         32.0     45.7167      5.2996       -2.59
  3         11.0     13.1292      3.2297       -0.66
  4          6.0      2.8563      1.6351        1.92
  5          5.0      0.5037      0.7045        6.38
  6          2.0      0.0749      0.2733        7.04
```

7	2.0	0.0097	0.0982	20.26
8	0.0	0.0011	0.0331	-0.03
9	0.0	0.0001	0.0106	-0.01
10	0.0	0.0000	0.0032	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	127.0	166.5000	6.6546	-5.94
2	58.0	62.2917	4.4454	-0.97
3	26.0	16.5750	3.4338	2.74
4	15.0	3.4458	1.7786	6.50
5	9.0	0.5895	0.7609	11.05
6	4.0	0.0858	0.2924	13.38
7	2.0	0.0109	0.1042	19.08
8	0.0	0.0012	0.0349	-0.03
9	0.0	0.0001	0.0111	-0.01
10	0.0	0.0000	0.0034	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	132.0	208.4167	14.5370	-5.26
2	66.0	91.4333	7.4947	-3.39
3	28.0	26.2583	4.5674	0.38
4	10.0	5.7127	2.3123	1.85
5	6.0	1.0074	0.9963	5.01
6	7.0	0.1498	0.3866	17.72
7	3.0	0.0193	0.1389	21.46
8	1.0	0.0022	0.0468	21.30
9	0.0	0.0002	0.0150	-0.01
10	1.0	0.0000	0.0045	220.19

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	254.0	333.0000	9.4110	-8.39
2	122.0	124.5833	6.2868	-0.41
3	56.0	33.1500	4.8561	4.71
4	28.0	6.8917	2.5154	8.39
5	18.0	1.1790	1.0761	15.63
6	12.0	0.1716	0.4136	28.60
7	5.0	0.0217	0.1474	33.77
8	2.0	0.0024	0.0494	40.43
9	1.0	0.0002	0.0157	63.73
10	1.0	0.0000	0.0047	210.77

```
LENGTH OF THE LONGEST RUN UP         =    10
LENGTH OF THE LONGEST RUN DOWN       =     7
LENGTH OF THE LONGEST RUN UP OR DOWN =    10

NUMBER OF POSITIVE DIFFERENCES =   258
```

```
NUMBER OF NEGATIVE DIFFERENCES =   241
NUMBER OF ZERO     DIFFERENCES =     0
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. Numerous values in this column are much larger than +/-1.96, so we conclude that the data are not random.

Distributional Assumptions

Since the quantitative tests show that the assumptions of randomness and constant location and scale are not met, the distributional measures will not be meaningful. Therefore these quantitative tests are omitted.

1.4.2.3.3. Develop A Better Model

Lag Plot Suggests Better Model

Since the underlying assumptions did not hold, we need to develop a better model.

The lag plot showed a distinct linear pattern. Given the definition of the lag plot, Y_i versus Y_{i-1}, a good candidate model is a model of the form

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

Fit Output

A linear fit of this model generated the following output.

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N         =      499
       NUMBER OF VARIABLES =          1
       NO REPLICATION CASE


               PARAMETER ESTIMATES            (APPROX. ST. DEV.)     T VALUE
        1   A0                   0.501650E-01   (0.2417E-01)         2.075
        2   A1         YIM1      0.987087       (0.6313E-02)         156.4

       RESIDUAL    STANDARD DEVIATION =        0.2931194
       RESIDUAL    DEGREES OF FREEDOM =         497
```

The slope parameter, A1, has a t value of 156.4 which is statistically significant. Also, the residual standard deviation is 0.29. This can be compared to the standard deviation shown in the summary table, which is 2.08. That is, the fit to the autoregressive model has reduced the variability by a factor of 7.

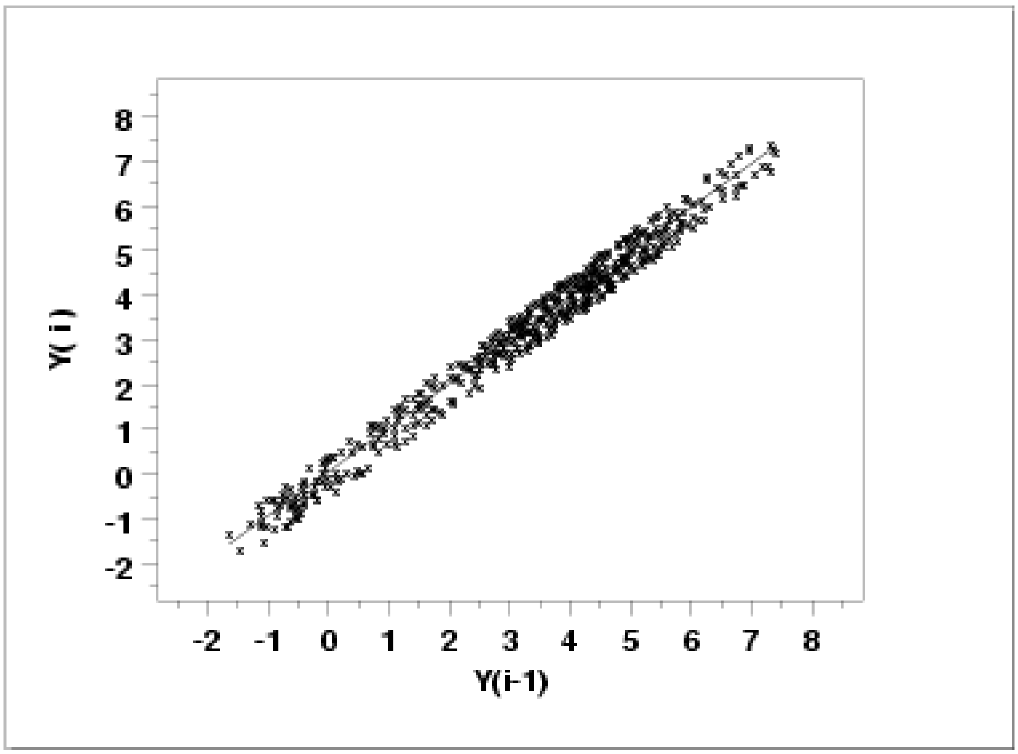
Time Series Model

This model is an example of a time series model. More extensive discussion of time series is given in the Process Monitoring chapter.

1.4.2.3.4. Validate New Model

Plot Predicted with Original Data

The first step in verifying the model is to plot the predicted values from the fit with the original data.

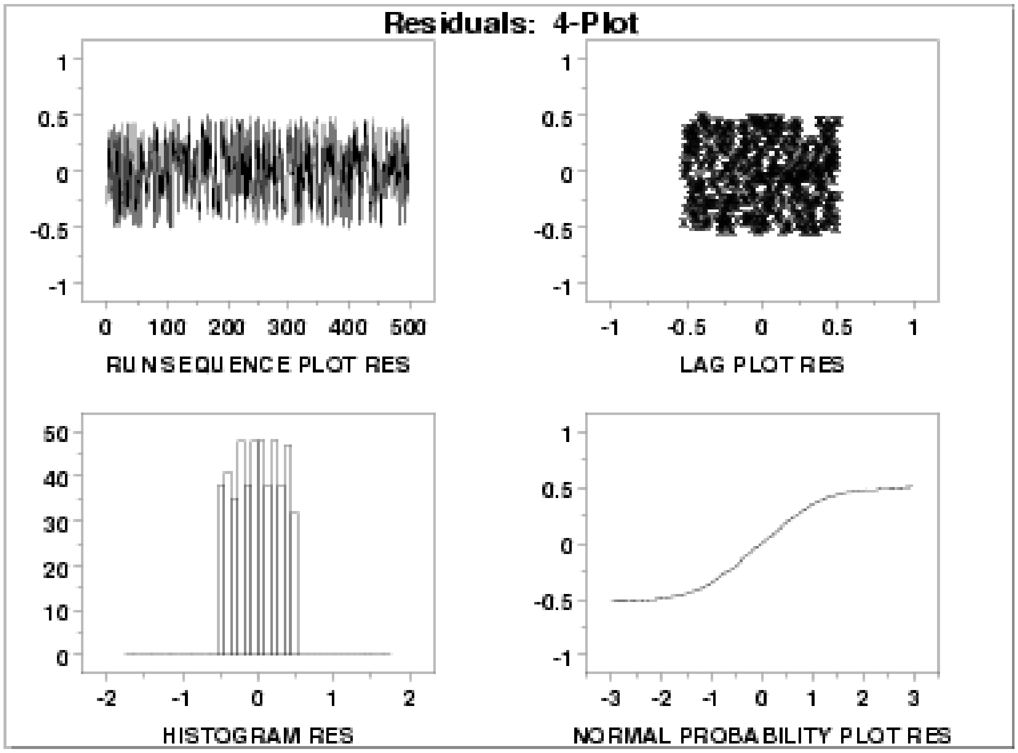


This plot indicates a reasonably good fit.

Test Underlying Assumptions on the Residuals

In addition to the plot of the predicted values, the residual standard deviation from the fit also indicates a significant improvement for the model. The next step is to validate the underlying assumptions for the error component, or residuals, from this model.

4-Plot of Residuals



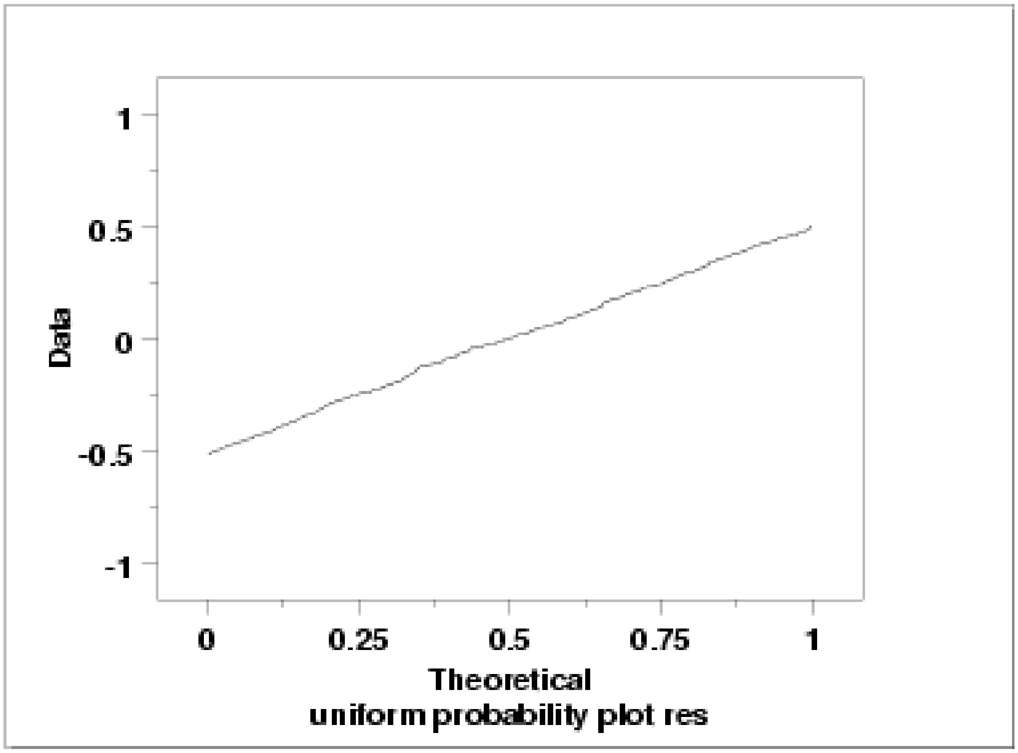
Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates no significant shifts in location or scale over time.
2. The lag plot (upper right) exhibits a random appearance.
3. The histogram shows a relatively flat appearance. This indicates that a uniform probability distribution may be an appropriate model for the error component (or residuals).
4. The normal probability plot clearly shows that the normal distribution is not an appropriate model for the error component.

A uniform probability plot can be used to further test the suggestion that a uniform distribution might be a good model for the error component.

Uniform Probability Plot of Residuals



Since the uniform probability plot is nearly linear, this verifies that a uniform distribution is a good model for the error component.

Conclusions

Since the residuals from our model satisfy the underlying assumptions, we conlude that

$$Y_i = 0.0502 + 0.987 * Y_{i-1} + E_i$$

where the E_i follow a uniform distribution is a good model for this data set. We could simplify this model to

$$Y_i = 1.0 * Y_{i-1} + E_i$$

This has the advantage of simplicity (the current point is simply the previous point plus a uniformly distributed error term).

Using Scientific and Engineering Knowledge

In this case, the above model makes sense based on our definition of the random walk. That is, a random walk is the cumulative sum of uniformly distributed data points. It makes sense that modeling the current point as the previous point plus a uniformly distributed error term is about as good as we can do. Although this case is a bit artificial in that we knew how the data were constructed, it is common and desirable to use scientific and engineering knowledge of the process that generated the data in formulating and testing models for the data. Quite often, several competing models will produce nearly equivalent mathematical results. In this case, selecting the model that best approximates the scientific understanding of the process is a reasonable choice.

Time Series Model

This model is an example of a time series model. More extensive discussion of time series is given in the Process Monitoring chapter.

1.4.2.3.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. Validate assumptions. 1. 4-plot of Y. 2. Generate a table of summary statistics. 3. Generate a linear fit to detect drift in location. 4. Detect drift in variation by dividing the data into quarters and computing Levene's test for equal	1. Based on the 4-plot, there are shifts in location and scale and the data are not random. 2. The summary statistics table displays 25+ statistics. 3. The linear fit indicates drift in location since the slope parameter is statistically significant. 4. Levene's test indicates significant drift in variation.

standard deviations. 5. Check for randomness by generating a runs test.	5. The runs test indicates significant non-randomness.
3. Generate the randomness plots. 1. Generate an autocorrelation plot. 2. Generate a spectral plot.	1. The autocorrelation plot shows significant autocorrelation at lag 1. 2. The spectral plot shows a single dominant low frequency peak.
4. Fit $Y_i = A0 + A1*Y_{i-1} + E_i$ and validate. 1. Generate the fit. 2. Plot fitted line with original data. 3. Generate a 4-plot of the residuals from the fit. 4. Generate a uniform probability plot of the residuals.	1. The residual standard deviation from the fit is 0.29 (compared to the standard deviation of 2.08 from the original data). 2. The plot of the predicted values with the original data indicates a good fit. 3. The 4-plot indicates that the assumptions of constant location and scale are valid. The lag plot indicates that the data are random. However, the histogram and normal probability plot indicate that the uniform disribution might be a better model for the residuals than the normal distribution. 4. The uniform probability plot verifies that the residuals can be fit by a uniform distribution.

1.4.2.4. Josephson Junction Cryothermometry

Josephson Junction Cryothermometry

This example illustrates the univariate analysis of Josephson junction cyrothermometry.

1. Exploratory Data Analysis
1.4. EDA Case Studies
1.4.2. Case Studies
1.4.2.4. Josephson Junction Cryothermometry

1.4.2.4.1. Background and Data

Generation

This data set was collected by Bob Soulen of NIST in October, 1971 as a sequence of observations collected equi-spaced in time from a volt meter to ascertain the process temperature in a Josephson junction cryothermometry (low temperature) experiment. The response variable is voltage counts.

Motivation

The motivation for studying this data set is to illustrate the case where there is discreteness in the measurements, but the underlying assumptions hold. In this case, the discreteness is due to the data being integers.

This file can be read by Dataplot with the following commands:

```
SKIP 25
SET READ FORMAT 5F5.0
READ SOULEN.DAT Y
SET READ FORMAT
```

Resulting Data

The following are the data used for this case study.

```
2899 2898 2898 2900 2898
2901 2899 2901 2900 2898
2898 2898 2898 2900 2898
2897 2899 2897 2899 2899
2900 2897 2900 2900 2899
2898 2898 2899 2899 2899
2899 2899 2898 2899 2899
2899 2902 2899 2900 2898
2899 2899 2899 2899 2899
2899 2900 2899 2900 2898
2901 2900 2899 2899 2899
2899 2899 2900 2899 2898
2898 2898 2900 2896 2897
```

2899	2899	2900	2898	2900
2901	2898	2899	2901	2900
2898	2900	2899	2899	2897
2899	2898	2899	2899	2898
2899	2897	2899	2899	2897
2899	2897	2899	2897	2897
2899	2897	2898	2898	2899
2897	2898	2897	2899	2899
2898	2898	2897	2898	2895
2897	2898	2898	2896	2898
2898	2897	2896	2898	2898
2897	2897	2898	2898	2896
2898	2898	2896	2899	2898
2898	2898	2899	2899	2898
2898	2899	2899	2899	2900
2900	2901	2899	2898	2898
2900	2899	2898	2901	2897
2898	2898	2900	2899	2899
2898	2898	2899	2898	2901
2900	2897	2897	2898	2898
2900	2898	2899	2898	2898
2898	2896	2895	2898	2898
2898	2898	2897	2897	2895
2897	2897	2900	2898	2896
2897	2898	2898	2899	2898
2897	2898	2898	2896	2900
2899	2898	2896	2898	2896
2896	2896	2897	2897	2896
2897	2897	2896	2898	2896
2898	2896	2897	2896	2897
2897	2898	2897	2896	2895
2898	2896	2896	2898	2896
2898	2898	2897	2897	2898
2897	2899	2896	2897	2899
2900	2898	2898	2897	2898
2899	2899	2900	2900	2900
2900	2899	2899	2899	2898
2900	2901	2899	2898	2900
2901	2901	2900	2899	2898
2901	2899	2901	2900	2901
2898	2900	2900	2898	2900
2900	2898	2899	2901	2900
2899	2899	2900	2900	2899
2900	2901	2899	2898	2898
2899	2896	2898	2897	2898
2898	2897	2897	2897	2898

```
2897 2899 2900 2899 2897
2898 2900 2900 2898 2898
2899 2900 2898 2900 2900
2898 2900 2898 2898 2898
2898 2898 2899 2898 2900
2897 2899 2898 2899 2898
2897 2900 2901 2899 2898
2898 2901 2898 2899 2897
2899 2897 2896 2898 2898
2899 2900 2896 2897 2897
2898 2899 2899 2898 2898
2897 2897 2898 2897 2897
2898 2898 2898 2896 2895
2898 2898 2898 2896 2898
2898 2898 2897 2897 2899
2896 2900 2897 2897 2898
2896 2897 2898 2898 2898
2897 2897 2898 2899 2897
2898 2899 2897 2900 2896
2899 2897 2898 2897 2900
2899 2900 2897 2897 2898
2897 2899 2899 2898 2897
2901 2900 2898 2901 2899
2900 2899 2898 2900 2900
2899 2898 2897 2900 2898
2898 2897 2899 2898 2900
2899 2898 2899 2897 2900
2898 2902 2897 2898 2899
2899 2899 2898 2897 2898
2897 2898 2899 2900 2900
2899 2898 2899 2900 2899
2900 2899 2899 2899 2899
2899 2898 2899 2899 2900
2902 2899 2900 2900 2901
2899 2901 2899 2899 2902
2898 2898 2898 2898 2899
2899 2900 2900 2900 2898
2899 2899 2900 2899 2900
2899 2900 2898 2898 2898
2900 2898 2899 2900 2899
2899 2900 2898 2898 2899
2899 2899 2899 2898 2898
2897 2898 2899 2897 2897
2901 2898 2897 2898 2899
2898 2897 2899 2898 2897
2898 2898 2897 2898 2899
```

2899	2899	2899	2900	2899
2899	2897	2898	2899	2900
2898	2897	2901	2899	2901
2898	2899	2901	2900	2900
2899	2900	2900	2900	2900
2901	2900	2901	2899	2897
2900	2900	2901	2899	2898
2900	2899	2899	2900	2899
2900	2899	2900	2899	2901
2900	2900	2899	2899	2898
2899	2900	2898	2899	2899
2901	2898	2898	2900	2899
2899	2898	2897	2898	2897
2899	2899	2899	2898	2898
2897	2898	2899	2897	2897
2899	2898	2898	2899	2899
2901	2899	2899	2899	2897
2900	2896	2898	2898	2900
2897	2899	2897	2896	2898
2897	2898	2899	2896	2899
2901	2898	2898	2896	2897
2899	2897	2898	2899	2898
2898	2898	2898	2898	2898
2899	2900	2899	2901	2898
2899	2899	2898	2900	2898
2899	2899	2901	2900	2901
2899	2901	2899	2901	2899
2900	2902	2899	2898	2899
2900	2899	2900	2900	2901
2900	2899	2901	2901	2899
2898	2901	2897	2898	2901
2900	2902	2899	2900	2898
2900	2899	2900	2899	2899
2899	2898	2900	2898	2899
2899	2899	2899	2898	2900

1.4.2.4.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

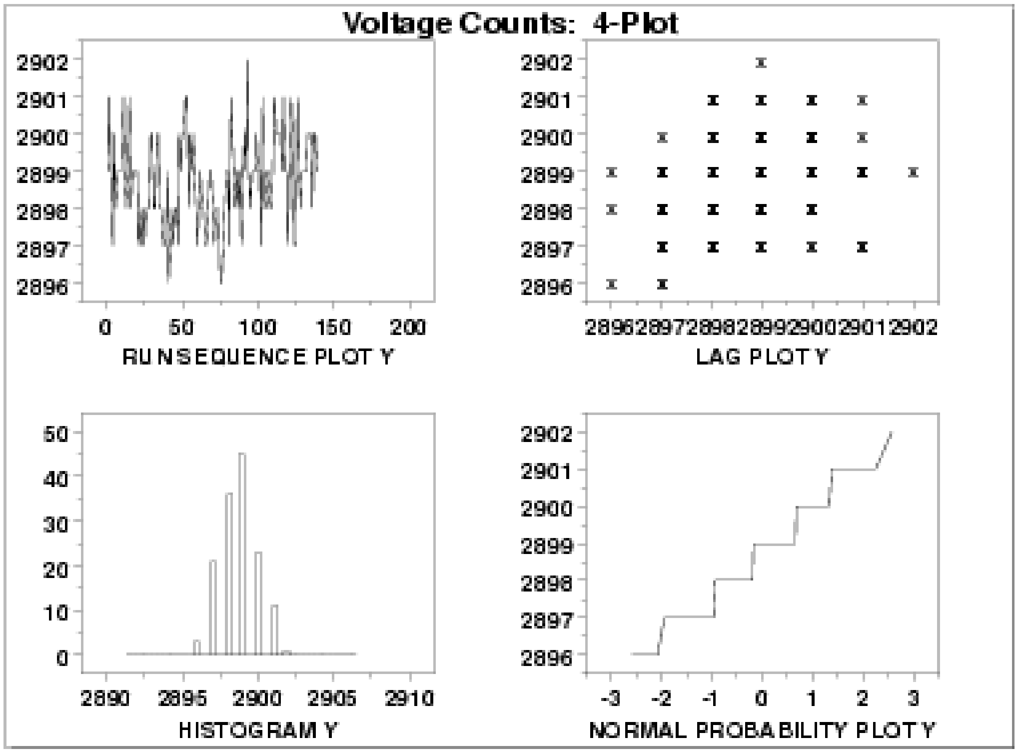
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram (lower left) shows that the data are reasonably symmetric, there does not appear to be significant outliers in the tails, and that it is reasonable to assume that the data can be fit with a normal distribution.
4. The normal probability plot (lower right) is difficult to interpret due to the fact that there are only a few distinct values with many repeats.

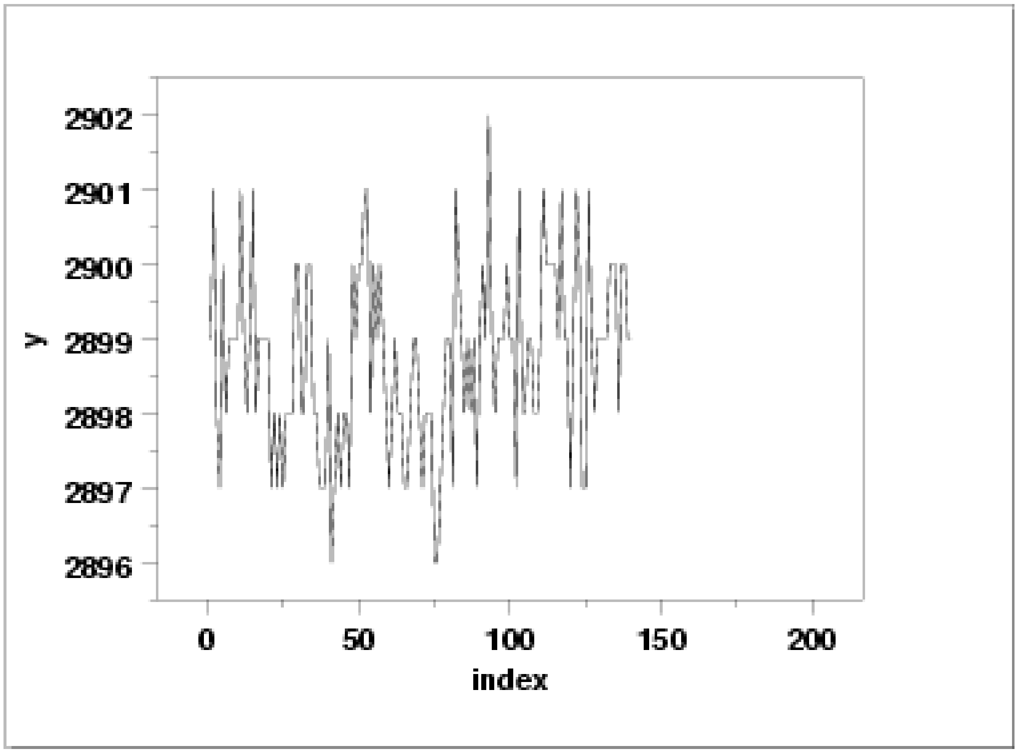
The integer data with only a few distinct values and many repeats accounts for the discrete appearance of several of the plots (e.g., the lag plot and the normal probability plot). In this case, the nature of the data makes the normal probability plot difficult to interpret, especially since each number is repeated many times. However, the histogram indicates that a normal distribution should provide an adequate model for the data.

From the above plots, we conclude that the underlying assumptions are valid and the data can be reasonably approximated with a normal distribution. Therefore, the commonly used uncertainty standard is valid and appropriate. The numerical values for this model are given in the Quantitative Output and Interpretation section.

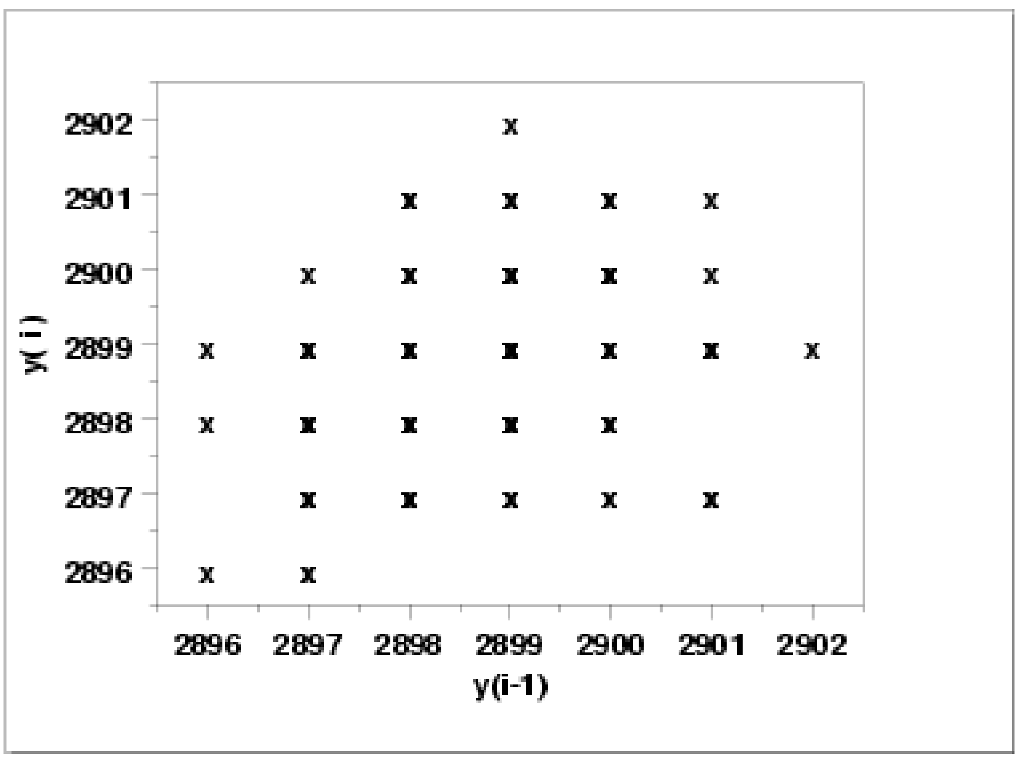
Individual Plots

Although it is normally not necessary, the plots can be generated individually to give more detail.

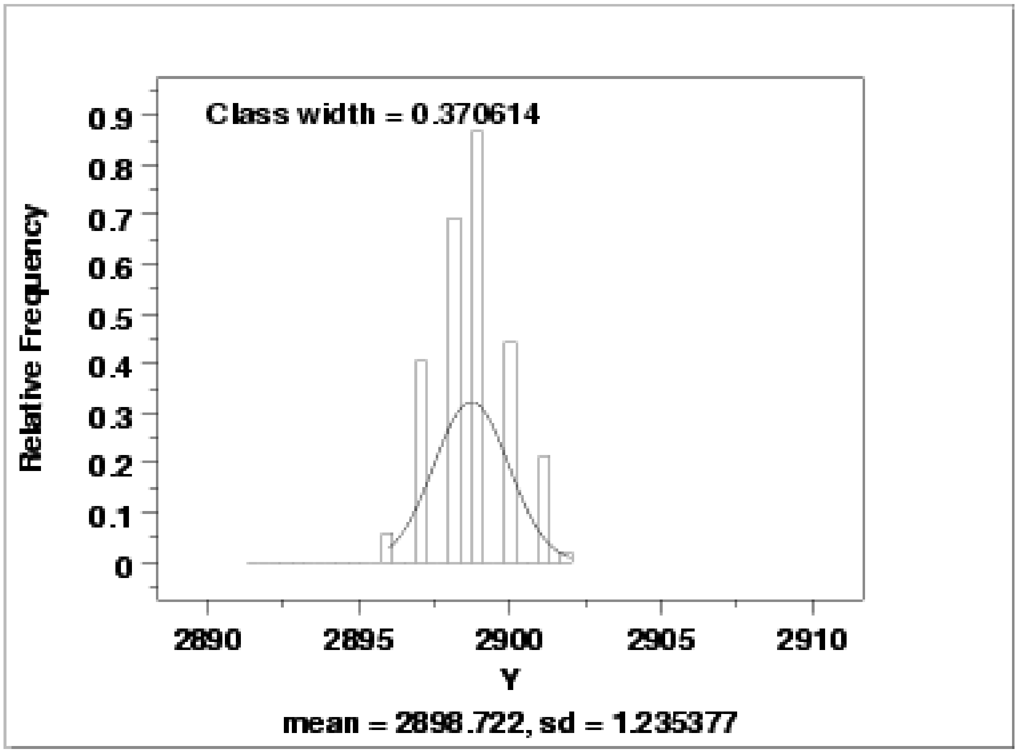
Run Sequence Plot



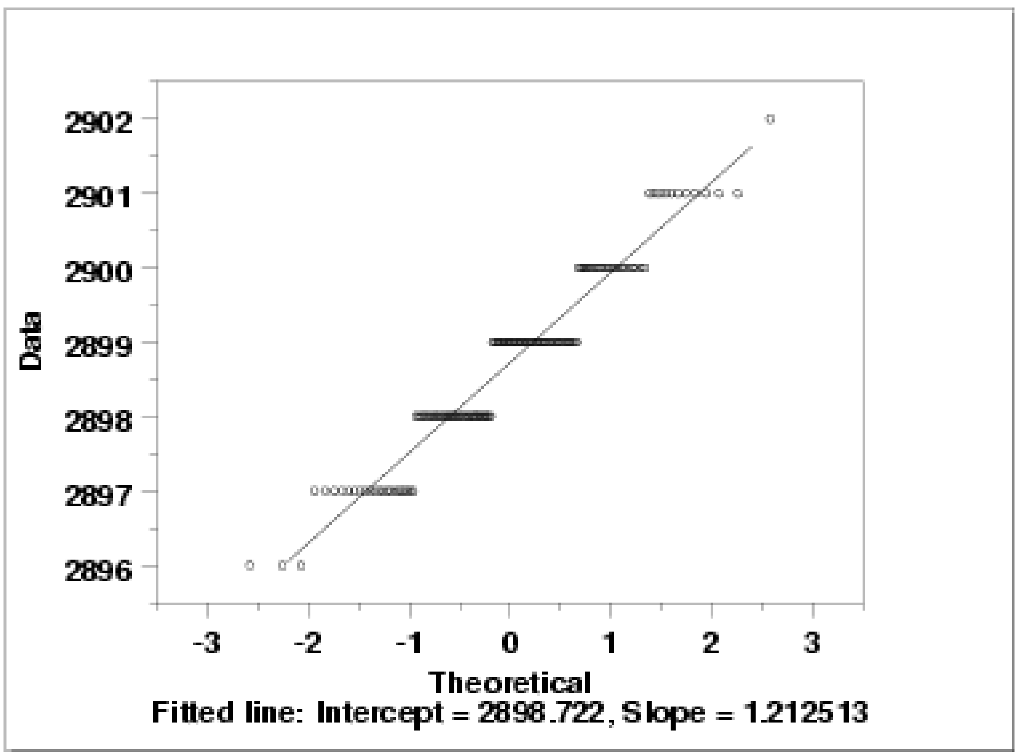
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot



1.4.2.4.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                          SUMMARY

               NUMBER OF OBSERVATIONS =      140


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.2899000E+04  *  RANGE        =   0.6000000E+01  *
*  MEAN         =   0.2898721E+04  *  STAND. DEV.  =   0.1235377E+01  *
*  MIDMEAN      =   0.2898457E+04  *  AV. AB. DEV. =   0.9642857E+00  *
*  MEDIAN       =   0.2899000E+04  *  MINIMUM      =   0.2896000E+04  *
*               =                  *  LOWER QUART. =   0.2898000E+04  *
*               =                  *  LOWER HINGE  =   0.2898000E+04  *
*               =                  *  UPPER HINGE  =   0.2899500E+04  *
*               =                  *  UPPER QUART. =   0.2899250E+04  *
*               =                  *  MAXIMUM      =   0.2902000E+04  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.2925397E+00  *  ST. 3RD MOM. =   0.1271097E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.2571418E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =  -0.3911592E+01  *
*               =                  *  UNIFORM PPCC =   0.9580541E+00  *
*               =                  *  NORMAL  PPCC =   0.9701443E+00  *
*               =                  *  TUK -.5 PPCC =   0.8550686E+00  *
*               =                  *  CAUCHY  PPCC =   0.6239791E+00  *
***********************************************************************
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generates the following output:

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N          =      140
       NUMBER OF VARIABLES =        1
       NO REPLICATION CASE


               PARAMETER ESTIMATES           (APPROX. ST. DEV.)      T VALUE
        1   A0                    2898.34     (0.2074     )       0.1398E+05
        2   A1        X         0.539896E-02  (0.2552E-02)         2.116
```

```
RESIDUAL    STANDARD DEVIATION =         1.220212
RESIDUAL    DEGREES OF FREEDOM =         138

COEF AND SD(COEF) WRITTEN OUT TO FILE DPST1F.DAT
SD(PRED),95LOWER,95UPPER,99LOWER,99UPPER
                  WRITTEN OUT TO FILE DPST2F.DAT
REGRESSION DIAGNOSTICS WRITTEN OUT TO FILE DPST3F.DAT
PARAMETER VARIANCE-COVARIANCE MATRIX AND
INVERSE OF X-TRANSPOSE X MATRIX
WRITTEN OUT TO FILE DPST4F.DAT
```

The slope parameter, A1, has a t value of 2.1 which is statistically significant (the critical value is 1.98). However, the value of the slope is 0.0054. Given that the slope is nearly zero, the assumption of constant location is not seriously violated even though it is (just barely) statistically significant.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since the nature of the data (a few distinct points repeated many times) makes the normality assumption questionable, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

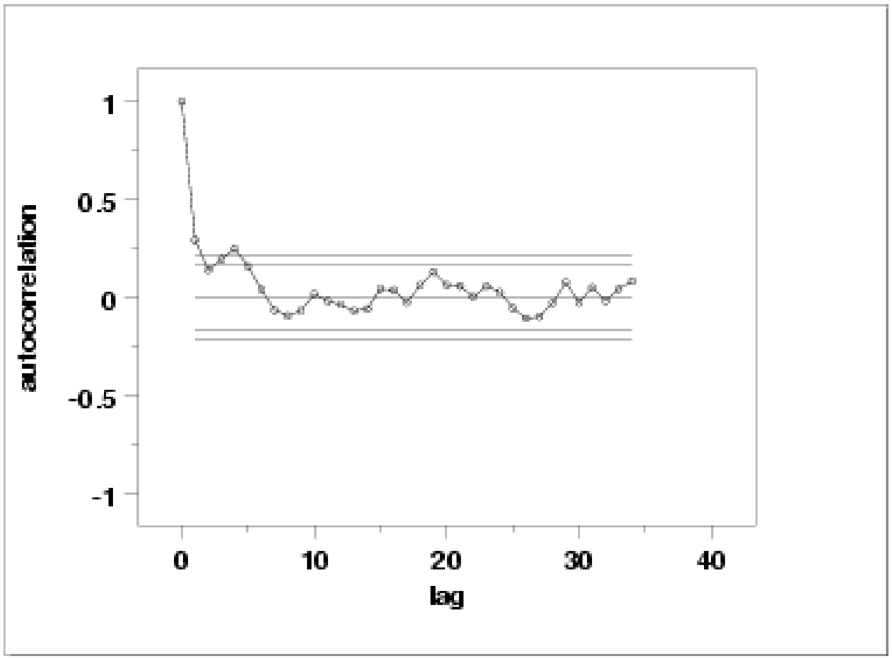
```
              LEVENE F-TEST FOR SHIFT IN VARIATION
                     (ASSUMPTION: NORMALITY)

 1. STATISTICS
       NUMBER OF OBSERVATIONS    =      140
       NUMBER OF GROUPS          =        4
       LEVENE F TEST STATISTIC   =   0.4128718


    FOR LEVENE TEST STATISTIC
       0          % POINT    =   0.0000000E+00
       50         % POINT    =   0.7926317
       75         % POINT    =    1.385201
       90         % POINT    =    2.124494
       95         % POINT    =    2.671178
       99         % POINT    =    3.928924
       99.9       % POINT    =    5.737571


          25.59809       % Point:    0.4128718

 3. CONCLUSION (AT THE 5% LEVEL):
       THERE IS NO SHIFT IN VARIATION.
       THUS: HOMOGENEOUS WITH RESPECT TO VARIATION.
```

Since the Levene test statistic value of 0.41 is less than the 95% critical value of 2.67, we conclude that the standard deviations are not significantly different in the 4 intervals.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the previous section is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of most interest, is 0.29. The critical values at the 5% level of significance are -0.087 and 0.087. This indicates that the lag 1 autocorrelation is statistically significant, so there is some evidence for non-randomness.

A common test for randomness is the runs test.

```
                         RUNS UP

              STATISTIC = NUMBER OF RUNS UP
                  OF LENGTH EXACTLY I

      I         STAT      EXP(STAT)    SD(STAT)        Z

      1         15.0      29.2083      5.4233       -2.62
      2         10.0      12.7167      2.7938       -0.97
      3          2.0       3.6292      1.6987       -0.96
      4          4.0       0.7849      0.8573        3.75
      5          2.0       0.1376      0.3683        5.06
      6          0.0       0.0204      0.1425       -0.14
      7          1.0       0.0026      0.0511       19.54
      8          0.0       0.0003      0.0172       -0.02
      9          0.0       0.0000      0.0055       -0.01
     10          0.0       0.0000      0.0017        0.00


              STATISTIC = NUMBER OF RUNS UP
                  OF LENGTH I OR MORE
```

I	STAT	EXP(STAT)	SD(STAT)	Z
1	34.0	46.5000	3.5048	-3.57
2	19.0	17.2917	2.3477	0.73
3	9.0	4.5750	1.8058	2.45
4	7.0	0.9458	0.9321	6.49
5	3.0	0.1609	0.3976	7.14
6	1.0	0.0233	0.1524	6.41
7	1.0	0.0029	0.0542	18.41
8	0.0	0.0003	0.0181	-0.02
9	0.0	0.0000	0.0057	-0.01
10	0.0	0.0000	0.0017	0.00

RUNS DOWN

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	16.0	29.2083	5.4233	-2.44
2	10.0	12.7167	2.7938	-0.97
3	5.0	3.6292	1.6987	0.81
4	1.0	0.7849	0.8573	0.25
5	2.0	0.1376	0.3683	5.06
6	0.0	0.0204	0.1425	-0.14
7	0.0	0.0026	0.0511	-0.05
8	0.0	0.0003	0.0172	-0.02
9	0.0	0.0000	0.0055	-0.01
10	0.0	0.0000	0.0017	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	34.0	46.5000	3.5048	-3.57
2	18.0	17.2917	2.3477	0.30
3	8.0	4.5750	1.8058	1.90
4	3.0	0.9458	0.9321	2.20
5	2.0	0.1609	0.3976	4.63
6	0.0	0.0233	0.1524	-0.15
7	0.0	0.0029	0.0542	-0.05
8	0.0	0.0003	0.0181	-0.02
9	0.0	0.0000	0.0057	-0.01
10	0.0	0.0000	0.0017	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	31.0	58.4167	7.6697	-3.57
2	20.0	25.4333	3.9510	-1.38
3	7.0	7.2583	2.4024	-0.11
4	5.0	1.5698	1.2124	2.83
5	4.0	0.2752	0.5208	7.15
6	0.0	0.0407	0.2015	-0.20

```
      7          1.0       0.0052       0.0722       13.78
      8          0.0       0.0006       0.0243       -0.02
      9          0.0       0.0001       0.0077       -0.01
     10          0.0       0.0000       0.0023        0.00


            STATISTIC = NUMBER OF RUNS TOTAL
                  OF LENGTH I OR MORE

      I          STAT      EXP(STAT)    SD(STAT)        Z

      1         68.0      93.0000       4.9565       -5.04
      2         37.0      34.5833       3.3202        0.73
      3         17.0       9.1500       2.5537        3.07
      4         10.0       1.8917       1.3182        6.15
      5          5.0       0.3218       0.5623        8.32
      6          1.0       0.0466       0.2155        4.42
      7          1.0       0.0059       0.0766       12.98
      8          0.0       0.0007       0.0256       -0.03
      9          0.0       0.0001       0.0081       -0.01
     10          0.0       0.0000       0.0024        0.00


             LENGTH OF THE LONGEST RUN UP          =     7
             LENGTH OF THE LONGEST RUN DOWN        =     5
             LENGTH OF THE LONGEST RUN UP OR DOWN =     7

             NUMBER OF POSITIVE DIFFERENCES =    48
             NUMBER OF NEGATIVE DIFFERENCES =    49
             NUMBER OF ZERO     DIFFERENCES =    42
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. The runs test indicates some mild non-randomness.

Although the runs test and lag 1 autocorrelation indicate some mild non-randomness, it is not sufficient to reject the $Y_i = C + E_i$ model. At least part of the non-randomness can be explained by the discrete nature of the data.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot. For this data set the correlation coefficient is 0.970. Since this is less than the critical value of 0.987 (this is a tabulated value), the normality assumption is rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. Dataplot generates the following output for the Anderson-Darling normality test.

```
                  ANDERSON-DARLING 1-SAMPLE TEST
                  THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

  1. STATISTICS:
        NUMBER OF OBSERVATIONS                =      140
        MEAN                                  =   2898.721
        STANDARD DEVIATION                    =   1.235377

        ANDERSON-DARLING TEST STATISTIC VALUE =   3.839233
        ADJUSTED TEST STATISTIC VALUE         =   3.944029
```

```
 2. CRITICAL VALUES:
       90          % POINT    =   0.6560000
       95          % POINT    =   0.7870000
       97.5        % POINT    =   0.9180000
       99          % POINT    =    1.092000

 3. CONCLUSION (AT THE 5% LEVEL):
       THE DATA DO NOT COME FROM A NORMAL DISTRIBUTION.
```

The Anderson-Darling test rejects the normality assumption because the test statistic, 3.84, is greater than the 99% critical value 1.092.

Although the data are not strictly normal, the violation of the normality assumption is not severe enough to conclude that the $Y_i = C + E_i$ model is unreasonable. At least part of the non-normality can be explained by the discrete nature of the data.

Outlier Analysis

A test for outliers is the Grubbs test. Dataplot generated the following output for Grubbs' test.

```
                GRUBBS TEST FOR OUTLIERS
                (ASSUMPTION: NORMALITY)

 1. STATISTICS:
       NUMBER OF OBSERVATIONS      =      140
       MINIMUM                     =    2896.000
       MEAN                        =    2898.721
       MAXIMUM                     =    2902.000
    STANDARD DEVIATION          =    1.235377

    GRUBBS TEST STATISTIC       =    2.653898

 2. PERCENT POINTS OF THE REFERENCE DISTRIBUTION
    FOR GRUBBS TEST STATISTIC
       0           % POINT    =   0.0000000E+00
       50          % POINT    =    2.874578
       75          % POINT    =    3.074733
       90          % POINT    =    3.320834
       95          % POINT    =    3.495103
       99          % POINT    =    3.867364

 3. CONCLUSION (AT THE 5% LEVEL):
       THERE ARE NO OUTLIERS.
```

For this data set, Grubbs' test does not detect any outliers at the 10%, 5%, and 1% significance levels.

Model

Although the randomness and normality assumptions were mildly violated, we conclude that a reasonable model for the data is:

$$Y_i = 2898.7 + E_i$$

In addition, a 95% confidence interval for the mean value is (2898.515,2898.928).

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
Analysis for Josephson Junction Cryothermometry Data

1: Sample Size                            = 140

2: Location
   Mean                                   = 2898.722
   Standard Deviation of Mean             = 0.104409
   95% Confidence Interval for Mean       = (2898.515,2898.928)
   Drift with respect to location?        = YES
   (Further analysis indicates that
   the drift, while statistically
   significant, is not practically
   significant)

3: Variation
   Standard Deviation                     = 1.235377
   95% Confidence Interval for SD         = (1.105655,1.399859)
   Drift with respect to variation?
   (based on Levene's test on quarters
   of the data)                           = NO

4: Distribution
   Normal PPCC                            = 0.970145
   Data are Normal?
     (as measured by Normal PPCC)         = NO

5: Randomness
   Autocorrelation                        = 0.29254
   Data are Random?
     (as measured by autocorrelation)     = NO

6: Statistical Control
   (i.e., no drift in location or scale,
   data are random, distribution is
   fixed, here we are testing only for
   fixed normal)
   Data Set is in Statistical Control?    = NO

   Note: Although we have violations of
   the assumptions, they are mild enough,
   and at least partially explained by the
   discrete nature of the data, so we may model
   the data as if it were in statistical
   control

7: Outliers?
   (as determined by Grubbs test)         = NO
```

1. Exploratory Data Analysis
1.4. EDA Case Studies
1.4.2. Case Studies
1.4.2.4. Josephson Junction Cryothermometry

1.4.2.4.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale. Due to the nature of the data (a few distinct points with many repeats), the normality assumption is questionable.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random).

overlaid normal pdf.	3. The histogram indicates that a normal distribution is a good distribution for these data.
4. Generate a normal probability plot.	4. The discrete nature of the data masks the normality or non-normality of the data somewhat. The plot indicates that a normal distribution provides a rough approximation for the data.
4. Generate summary statistics, quantitative analysis, and print a univariate report.	
1. Generate a table of summary statistics.	1. The summary statistics table displays 25+ statistics.
2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location.	2. The mean is 2898.27 and a 95% confidence interval is (2898.52,2898.93). The linear fit indicates no meaningful drift in location since the value of the slope parameter is near zero.
3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Levene's test for equal standard deviations.	3. The standard devaition is 1.24 with a 95% confidence interval of (1.11,1.40). Levene's test indicates no significant drift in variation.
4. Check for randomness by generating an autocorrelation plot and a runs test.	4. The lag 1 autocorrelation is 0.29. This indicates some mild non-randomness.
5. Check for normality by computing the normal probability plot correlation coefficient.	5. The normal probability plot correlation coefficient is 0.970. At the 5% level, we reject the normality assumption.
6. Check for outliers using Grubbs' test.	6. Grubbs' test detects no outliers at the 5% level.
7. Print a univariate report (this assumes steps 2 thru 6 have already been run).	7. The results are summarized in a convenient report.

1.4.2.5. Beam Deflections

Beam Deflection

This example illustrates the univariate analysis of beam deflection data.

1.4.2.5.1. Background and Data

Generation

This data set was collected by H. S. Lew of NIST in 1969 to measure steel-concrete beam deflections. The response variable is the deflection of a beam from the center point.

The motivation for studying this data set is to show how the underlying assumptions are affected by periodic data.

This file can be read by Dataplot with the following commands:

```
SKIP 25
READ LEW.DAT Y
```

Resulting Data

The following are the data used for this case study.

```
 -213
 -564
  -35
  -15
  141
  115
 -420
 -360
  203
 -338
 -431
  194
 -220
 -513
  154
 -125
 -559
   92
  -21
 -579
```

```
 -52
  99
-543
-175
 162
-457
-346
 204
-300
-474
 164
-107
-572
  -8
  83
-541
-224
 180
-420
-374
 201
-236
-531
  83
  27
-564
-112
 131
-507
-254
 199
-311
-495
 143
 -46
-579
 -90
 136
-472
-338
 202
-287
-477
 169
-124
-568
```

```
  17
  48
-568
-135
 162
-430
-422
 172
 -74
-577
 -13
  92
-534
-243
 194
-355
-465
 156
 -81
-578
 -64
 139
-449
-384
 193
-198
-538
 110
 -44
-577
  -6
  66
-552
-164
 161
-460
-344
 205
-281
-504
 134
 -28
-576
-118
 156
-437
```

```
-381
 200
-220
-540
  83
  11
-568
-160
 172
-414
-408
 188
-125
-572
 -32
 139
-492
-321
 205
-262
-504
 142
 -83
-574
   0
  48
-571
-106
 137
-501
-266
 190
-391
-406
 194
-186
-553
  83
 -13
-577
 -49
 103
-515
-280
 201
 300
```

```
-506
 131
 -45
-578
 -80
 138
-462
-361
 201
-211
-554
  32
  74
-533
-235
 187
-372
-442
 182
-147
-566
  25
  68
-535
-244
 194
-351
-463
 174
-125
-570
  15
  72
-550
-190
 172
-424
-385
 198
-218
-536
  96
```

1.4.2.5.2. Test Underlying Assumptions

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

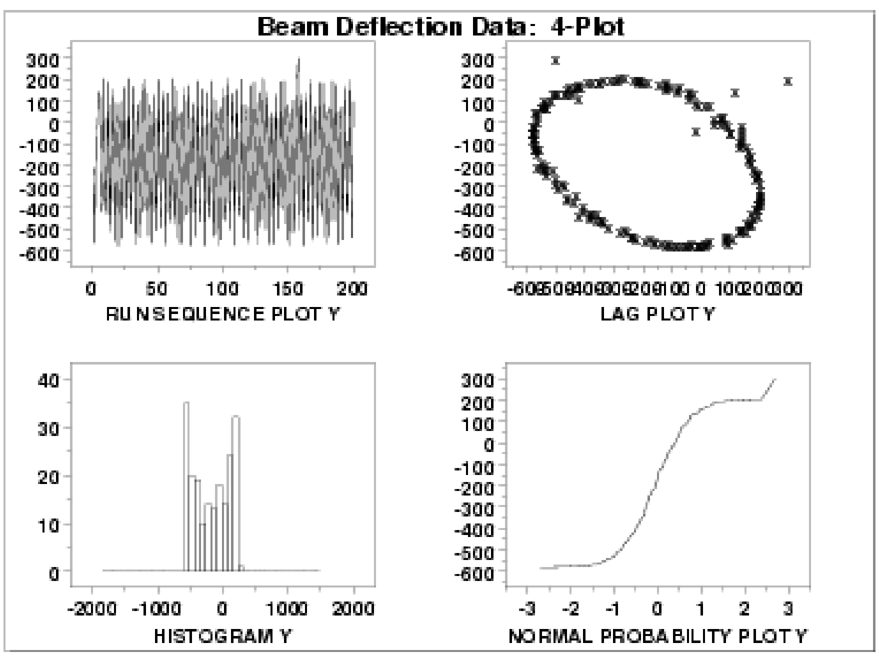
$$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

$$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) shows that the data are not random. The lag plot further indicates the presence of a few outliers.
3. When the randomness assumption is thus seriously violated, the histogram (lower left) and normal probability plot (lower right) are ignored since determining the distribution of the data is only meaningful when the data are random.

From the above plots we conclude that the underlying randomness assumption is not valid. Therefore, the model

$$Y_i = C + E_i$$

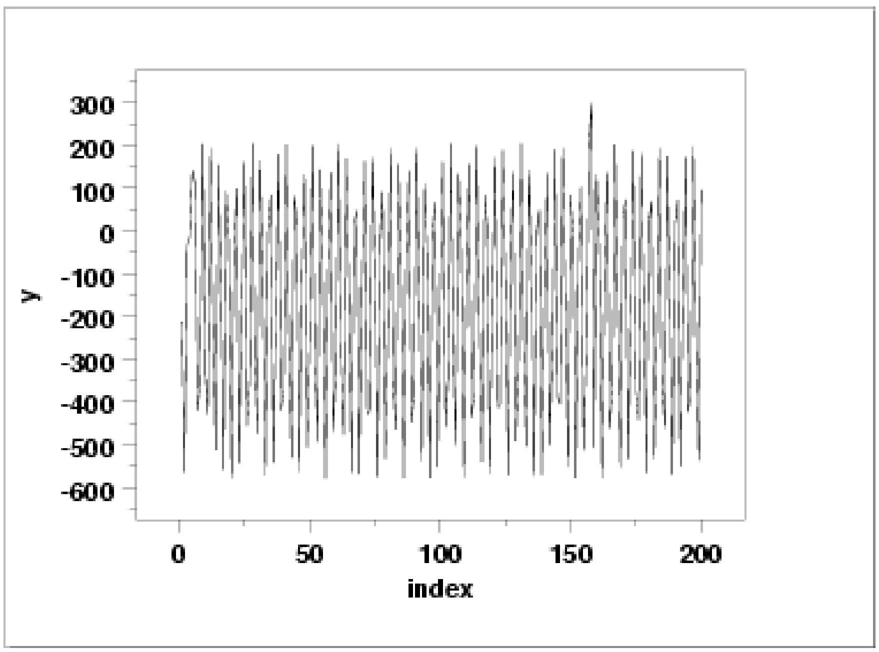
is not appropriate.

We need to develop a better model. Non-random data can frequently be modeled using time series mehtodology. Specifically, the circular pattern in the lag plot indicates that a sinusoidal model might be appropriate. The sinusoidal model will be developed in the next section.

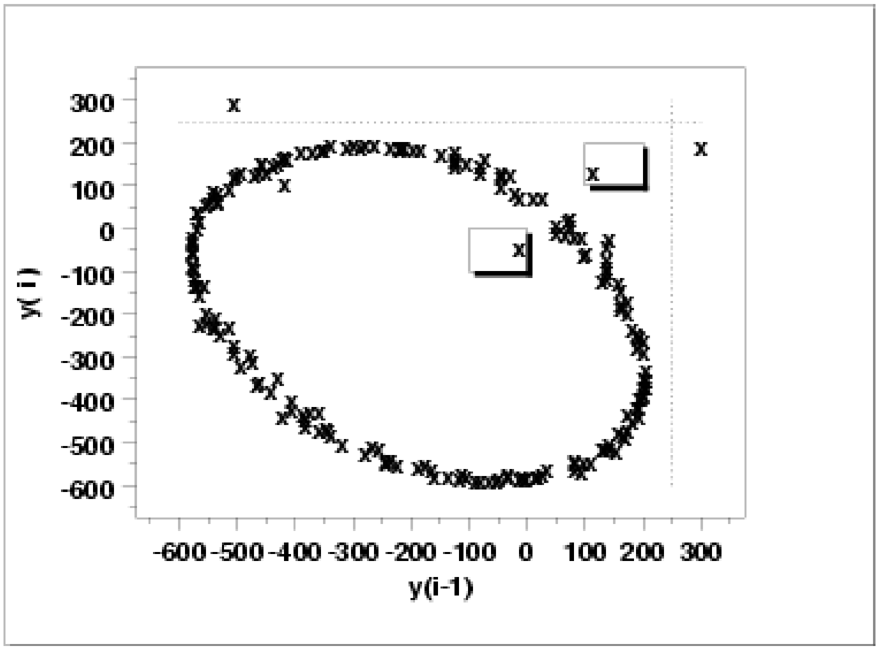
Individual Plots

The plots can be generated individually for more detail. In this case, only the run sequence plot and the lag plot are drawn since the distributional plots are not meaningful.

Run Sequence Plot



Lag Plot



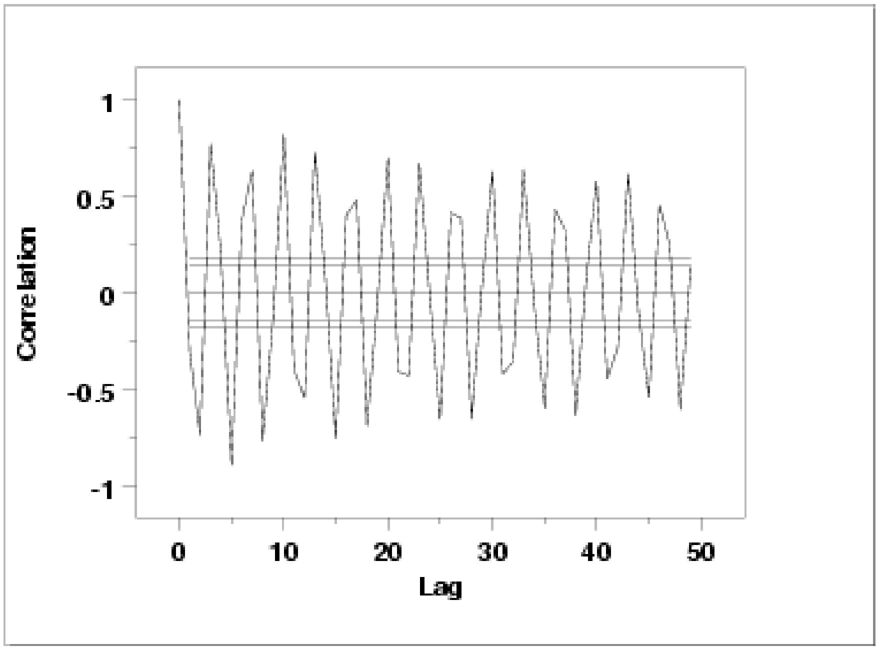
We have drawn some lines and boxes on the plot to better isolate the outliers. The following output helps identify the points that are generating the outliers on the lag plot.

```
     *****************************************************
     **  print y index xplot yplot subset yplot > 250  **
     *****************************************************


VARIABLES--Y               INDEX          XPLOT          YPLOT

        300.00         158.00         -506.00         300.00

     *****************************************************
     **  print y index xplot yplot subset xplot > 250  **
     *****************************************************


VARIABLES--Y               INDEX          XPLOT          YPLOT

        201.00         157.00          300.00         201.00

     *********************************************************
     **  print y index xplot yplot subset yplot -100 to 0
                                  subset xplot -100 to 0  **
     *********************************************************


VARIABLES--Y               INDEX          XPLOT          YPLOT

        -35.00           3.00          -15.00         -35.00

     **********************************************************
     **  print y index xplot yplot subset yplot 100 to 200
                                  subset xplot 100 to 200  **
     **********************************************************


VARIABLES--Y               INDEX          XPLOT          YPLOT

        141.00           5.00          115.00         141.00
```

That is, the third, fifth, and 158th points appear to be outliers.

Autocorrelation Plot

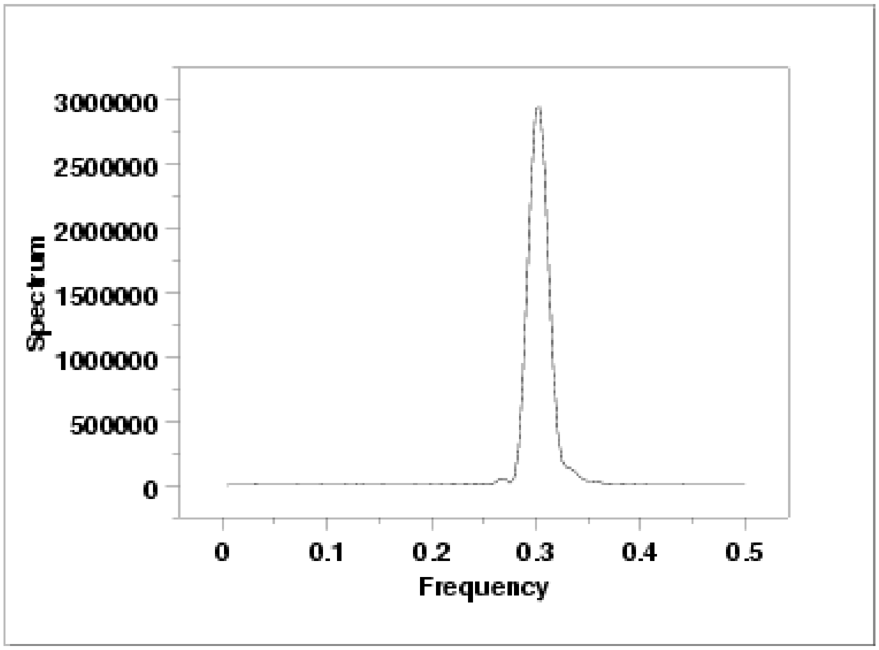
When the lag plot indicates significant non-randomness, it can be helpful to follow up with a an autocorrelation plot.



This autocorrelation plot shows a distinct cyclic pattern. As with the lag plot, this suggests a sinusoidal model.

Spectral Plot

Another useful plot for non-random data is the spectral plot.



This spectral plot shows a single dominant peak at a frequency of 0.3. This frequency of 0.3 will be used in fitting the sinusoidal model in the next section.

Quantitative Output

Although the lag plot, autocorrelation plot, and spectral plot clearly show the violation of the randomness assumption, we supplement the graphical output with some quantitative measures.

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                              SUMMARY

                   NUMBER OF OBSERVATIONS =      200


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =  -0.1395000E+03  *  RANGE         =   0.8790000E+03  *
*  MEAN         =  -0.1774350E+03  *  STAND. DEV.   =   0.2773322E+03  *
*  MIDMEAN      =  -0.1797600E+03  *  AV. AB. DEV.  =   0.2492250E+03  *
*  MEDIAN       =  -0.1620000E+03  *  MINIMUM       =  -0.5790000E+03  *
*               =                  *  LOWER QUART.  =  -0.4510000E+03  *
*               =                  *  LOWER HINGE   =  -0.4530000E+03  *
*               =                  *  UPPER HINGE   =   0.9400000E+02  *
*               =                  *  UPPER QUART.  =   0.9300000E+02  *
*               =                  *  MAXIMUM       =   0.3000000E+03  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =  -0.3073048E+00  *  ST. 3RD MOM.  =  -0.5010057E-01  *
*               =   0.0000000E+00  *  ST. 4TH MOM.  =   0.1503684E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA  =  -0.1883372E+02  *
*               =                  *  UNIFORM PPCC  =   0.9925535E+00  *
*               =                  *  NORMAL  PPCC  =   0.9540811E+00  *
*               =                  *  TUK -.5 PPCC  =   0.7313794E+00  *
*               =                  *  CAUCHY  PPCC  =   0.4408355E+00  *
***********************************************************************
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generates the following output:

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N         =      200
       NUMBER OF VARIABLES =          1
       NO REPLICATION CASE


               PARAMETER ESTIMATES           (APPROX. ST. DEV.)     T VALUE
        1  A0                    -178.175         ( 39.47     )       -4.514
        2  A1         X          0.736593E-02     (0.3405     )       0.2163E-01

       RESIDUAL    STANDARD DEVIATION =         278.0313
       RESIDUAL    DEGREES OF FREEDOM =         198
```

The slope parameter, A1, has a *t* value of 0.022 which is statistically not significant. This indicates that the slope can in fact be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett the non-randomness of this data does not allows us to assume normality, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

```
              LEVENE F-TEST FOR SHIFT IN VARIATION
                     (ASSUMPTION: NORMALITY)

 1. STATISTICS
       NUMBER OF OBSERVATIONS    =      200
       NUMBER OF GROUPS          =        4
       LEVENE F TEST STATISTIC   =   0.9378599E-01


    FOR LEVENE TEST STATISTIC
       0          % POINT    =   0.0000000E+00
       50         % POINT    =   0.7914120
       75         % POINT    =    1.380357
       90         % POINT    =    2.111936
       95         % POINT    =    2.650676
       99         % POINT    =    3.883083
       99.9       % POINT    =    5.638597


          3.659895       % Point:    0.9378599E-01

 3. CONCLUSION (AT THE 5% LEVEL):
       THERE IS NO SHIFT IN VARIATION.
       THUS: HOMOGENEOUS WITH RESPECT TO VARIATION.
```

In this case, the Levene test indicates that the standard deviations are significantly different in the 4 intervals since the test statistic of 13.2 is greater than the 95% critical value of 2.6. Therefore we conclude that the scale is not constant.

Randomness

A runs test is used to check for randomness

```
                     RUNS UP

             STATISTIC = NUMBER OF RUNS UP
                 OF LENGTH EXACTLY I

  I          STAT      EXP(STAT)     SD(STAT)        Z

  1          63.0     104.2083     10.2792       -4.01
  2          34.0      45.7167      5.2996       -2.21
  3          17.0      13.1292      3.2297        1.20
  4           4.0       2.8563      1.6351        0.70
  5           1.0       0.5037      0.7045        0.70
  6           5.0       0.0749      0.2733       18.02
  7           1.0       0.0097      0.0982       10.08
  8           1.0       0.0011      0.0331       30.15
  9           0.0       0.0001      0.0106       -0.01
 10           1.0       0.0000      0.0032      311.40


             STATISTIC = NUMBER OF RUNS UP
                 OF LENGTH I OR MORE
```

I	STAT	EXP(STAT)	SD(STAT)	Z
1	127.0	166.5000	6.6546	-5.94
2	64.0	62.2917	4.4454	0.38
3	30.0	16.5750	3.4338	3.91
4	13.0	3.4458	1.7786	5.37
5	9.0	0.5895	0.7609	11.05
6	8.0	0.0858	0.2924	27.06
7	3.0	0.0109	0.1042	28.67
8	2.0	0.0012	0.0349	57.21
9	1.0	0.0001	0.0111	90.14
10	1.0	0.0000	0.0034	298.08

RUNS DOWN

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	69.0	104.2083	10.2792	-3.43
2	32.0	45.7167	5.2996	-2.59
3	11.0	13.1292	3.2297	-0.66
4	6.0	2.8563	1.6351	1.92
5	5.0	0.5037	0.7045	6.38
6	2.0	0.0749	0.2733	7.04
7	2.0	0.0097	0.0982	20.26
8	0.0	0.0011	0.0331	-0.03
9	0.0	0.0001	0.0106	-0.01
10	0.0	0.0000	0.0032	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	127.0	166.5000	6.6546	-5.94
2	58.0	62.2917	4.4454	-0.97
3	26.0	16.5750	3.4338	2.74
4	15.0	3.4458	1.7786	6.50
5	9.0	0.5895	0.7609	11.05
6	4.0	0.0858	0.2924	13.38
7	2.0	0.0109	0.1042	19.08
8	0.0	0.0012	0.0349	-0.03
9	0.0	0.0001	0.0111	-0.01
10	0.0	0.0000	0.0034	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	132.0	208.4167	14.5370	-5.26
2	66.0	91.4333	7.4947	-3.39
3	28.0	26.2583	4.5674	0.38
4	10.0	5.7127	2.3123	1.85
5	6.0	1.0074	0.9963	5.01
6	7.0	0.1498	0.3866	17.72
7	3.0	0.0193	0.1389	21.46

```
  8          1.0      0.0022      0.0468       21.30
  9          0.0      0.0002      0.0150       -0.01
 10          1.0      0.0000      0.0045      220.19


              STATISTIC = NUMBER OF RUNS TOTAL
                    OF LENGTH I OR MORE

  I           STAT     EXP(STAT)    SD(STAT)       Z

  1        254.0    333.0000      9.4110       -8.39
  2        122.0    124.5833      6.2868       -0.41
  3         56.0     33.1500      4.8561        4.71
  4         28.0      6.8917      2.5154        8.39
  5         18.0      1.1790      1.0761       15.63
  6         12.0      0.1716      0.4136       28.60
  7          5.0      0.0217      0.1474       33.77
  8          2.0      0.0024      0.0494       40.43
  9          1.0      0.0002      0.0157       63.73
 10          1.0      0.0000      0.0047      210.77


               LENGTH OF THE LONGEST RUN UP         =    10
               LENGTH OF THE LONGEST RUN DOWN       =     7
               LENGTH OF THE LONGEST RUN UP OR DOWN =    10

               NUMBER OF POSITIVE DIFFERENCES =   258
               NUMBER OF NEGATIVE DIFFERENCES =   241
               NUMBER OF ZERO     DIFFERENCES =     0
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. Numerous values in this column are much larger than +/-1.96, so we conclude that the data are not random.

Distributional Assumptions

Since the quantitative tests show that the assumptions of constant scale and non-randomness are not met, the distributional measures will not be meaningful. Therefore these quantitative tests are omitted.

1.4.2.5.3. Develop a Better Model

Sinusoidal Model

The lag plot and autocorrelation plot in the previous section strongly suggested a sinusoidal model might be appropriate. The basic sinusoidal model is:

$$Y_i = C + \alpha \sin(2\pi\omega T_i + \phi) + E_i$$

where C is constant defining a mean level, α is an amplitude for the sine function, ω is the frequency, T_i is a time variable, and ϕ is the phase. This sinusoidal model can be fit using non-linear least squares.

To obtain a good fit, sinusoidal models require good starting values for C, the amplitude, and the frequency.

Good Starting Value for C

A good starting value for C can be obtained by calculating the mean of the data. If the data show a trend, i.e., the assumption of constant location is violated, we can replace C with a linear or quadratic least squares fit. That is, the model becomes

$$Y_i = (B_0 + B_1 * T_i) + \alpha \sin(2\pi\omega T_i + \phi) + E_i$$

or

$$Y_i = (B_0 + B_1 * T_i + B2 * T_i^2) + \alpha \sin(2\pi\omega T_i + \phi) + E_i$$

Since our data did not have any meaningful change of location, we can fit the simpler model with C equal to the mean. From the summary output in the previous page, the mean is -177.44.

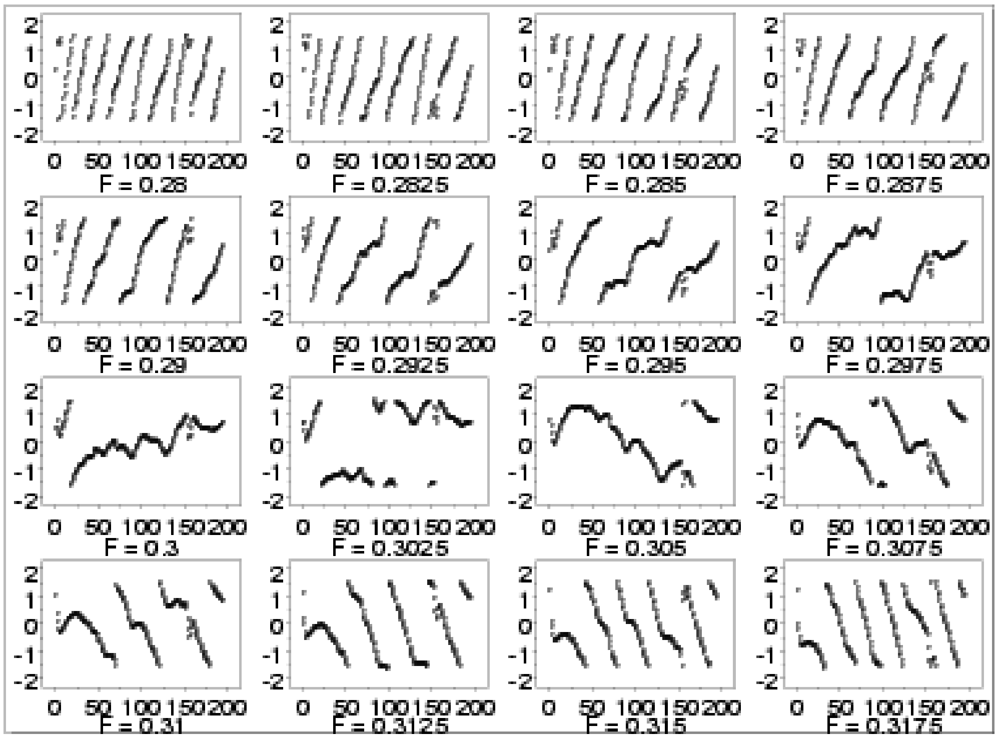
Good Starting Value for Frequency

The starting value for the frequency can be obtained from the spectral plot, which shows the dominant frequency is about 0.3.

Complex Demodulation Phase Plot

The complex demodulation phase plot can be used to refine this initial estimate for the frequency.

For the complex demodulation plot, if the lines slope from left to right, the frequency should be increased. If the lines slope from right to left, it should be decreased. A relatively flat (i.e., horizontal) slope indicates a good frequency. We could generate the demodulation phase plot for 0.3 and then use trial and error to obtain a better estimate for the frequency. To simplify this, we generate 16 of these plots on a single page starting with a frequency of 0.28, increasing in increments of 0.0025, and stopping at 0.3175.



Interpretation

The plots start with lines sloping from left to right but gradually change to a right to left slope. The relatively flat slope occurs for frequency 0.3025 (third row, second column). The complex demodulation phase plot restricts the range from $\pi/2$ to $-\pi/2$. This is why the plot appears to show some breaks.

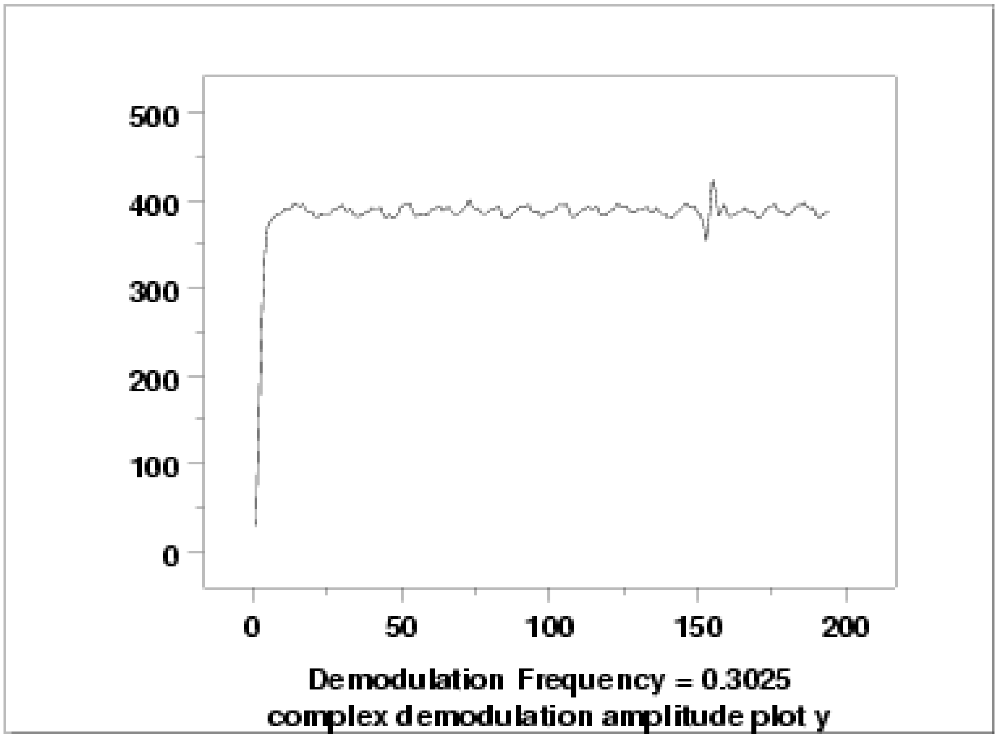
Good Starting Values for Amplitude

The complex demodulation amplitude plot is used to find a good starting value for the amplitude. In addition, this plot indicates whether or not the amplitude is constant over the entire range of the data or if it varies. If the plot is essentially flat, i.e., zero slope, then it is reasonable to assume a constant amplitude in the non-linear model. However, if the slope varies over the range of the plot, we may need to adjust the model to be:

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

That is, we replace α with a function of time. A linear fit is specified in the model above, but this can be replaced with a more elaborate function if needed.

Complex Demodulation Amplitude Plot



The complex demodulation amplitude plot for this data shows that:

1. The amplitude is fixed at approximately 390.
2. There is a short start-up effect.
3. There is a change in amplitude at around x=160 that should be investigated for an outlier.

In terms of a non-linear model, the plot indicates that fitting a single constant for α should be adequate for this data set.

Fit Output

Using starting estimates of 0.3025 for the frequency, 390 for the amplitude, and -177.44 for C, Dataplot generated the following output for the fit.

```
LEAST SQUARES NON-LINEAR FIT
      SAMPLE SIZE N =      200
      MODEL--Y =C + AMP*SIN(2*3.14159*FREQ*T + PHASE)
      NO REPLICATION CASE

ITERATION  CONVERGENCE  RESIDUAL   *  PARAMETER
 NUMBER      MEASURE    STANDARD   *  ESTIMATES
                        DEVIATION *
----------------------------------*-----------
    1--   0.10000E-01  0.52903E+03 *-0.17743E+03 0.39000E+03 0.30250E+00 0.10000E+01
    2--   0.50000E-02  0.22218E+03 *-0.17876E+03-0.33137E+03 0.30238E+00 0.71471E+00
    3--   0.25000E-02  0.15634E+03 *-0.17886E+03-0.24523E+03 0.30233E+00 0.14022E+01
    4--   0.96108E-01  0.15585E+03 *-0.17879E+03-0.36177E+03 0.30260E+00 0.14654E+01

        FINAL PARAMETER ESTIMATES          (APPROX. ST. DEV.)    T VALUE
       1  C                      -178.786       ( 11.02    )       -16.22
       2  AMP                    -361.766       ( 26.19    )       -13.81
```

```
 3  FREQ                    0.302596       (0.1510E-03)        2005.
 4  PHASE                    1.46536       (0.4909E-01)        29.85

RESIDUAL    STANDARD DEVIATION =          155.8484
RESIDUAL    DEGREES OF FREEDOM =          196
```

Model From the fit output, our proposed model is:

$$Y_i = -178.79 - 361.77 * (2\pi * 0.302596 T_i + 1.465) + E_i$$

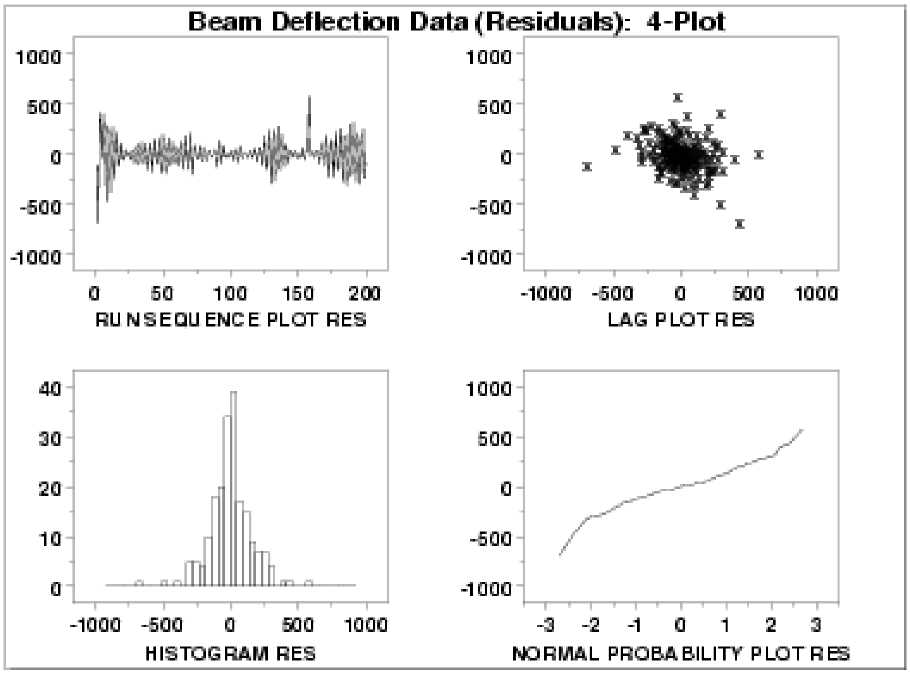
We will evaluate the adequacy of this model in the next section.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

1.4.2.5.4. Validate New Model

4-Plot of Residuals

The first step in evaluating the fit is to generate a 4-plot of the residuals.



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location. There does seem to be some shifts in scale. A start-up effect was detected previously by the complex demodulation amplitude plot. There does appear to be a few outliers.
2. The lag plot (upper right) shows that the data are random. The outliers also appear in the lag plot.
3. The histogram (lower left) and the normal probability plot (lower right) do not show any serious non-normality in the residuals. However, the bend in the left portion of the normal probability plot shows some cause for concern.

The 4-plot indicates that this fit is reasonably good. However, we will attempt to improve the fit by removing the outliers.

Fit Output with Outliers Removed

Dataplot generated the following fit output after removing 3 outliers.

```
    LEAST SQUARES NON-LINEAR FIT
          SAMPLE SIZE N =      197
          MODEL--Y =C + AMP*SIN(2*3.14159*FREQ*T + PHASE)
          NO REPLICATION CASE

    ITERATION  CONVERGENCE  RESIDUAL  *  PARAMETER
     NUMBER      MEASURE    STANDARD  *  ESTIMATES
                            DEVIATION *
    ----------------------------------*-----------
        1--  0.10000E-01  0.14834E+03 *-0.17879E+03-0.36177E+03 0.30260E+00 0.14654E+01

        2--  0.37409E+02  0.14834E+03 *-0.17879E+03-0.36176E+03 0.30260E+00 0.14653E+01

            FINAL PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
           1  C                      -178.788       ( 10.57     )       -16.91
           2  AMP                    -361.759       ( 25.45     )       -14.22
           3  FREQ                   0.302597       (0.1457E-03)        2077.
           4  PHASE                   1.46533       (0.4715E-01)        31.08

          RESIDUAL    STANDARD DEVIATION =         148.3398
          RESIDUAL    DEGREES OF FREEDOM =         193
```

New Fit to Edited Data

The original fit, with a residual standard deviation of 155.84, was:

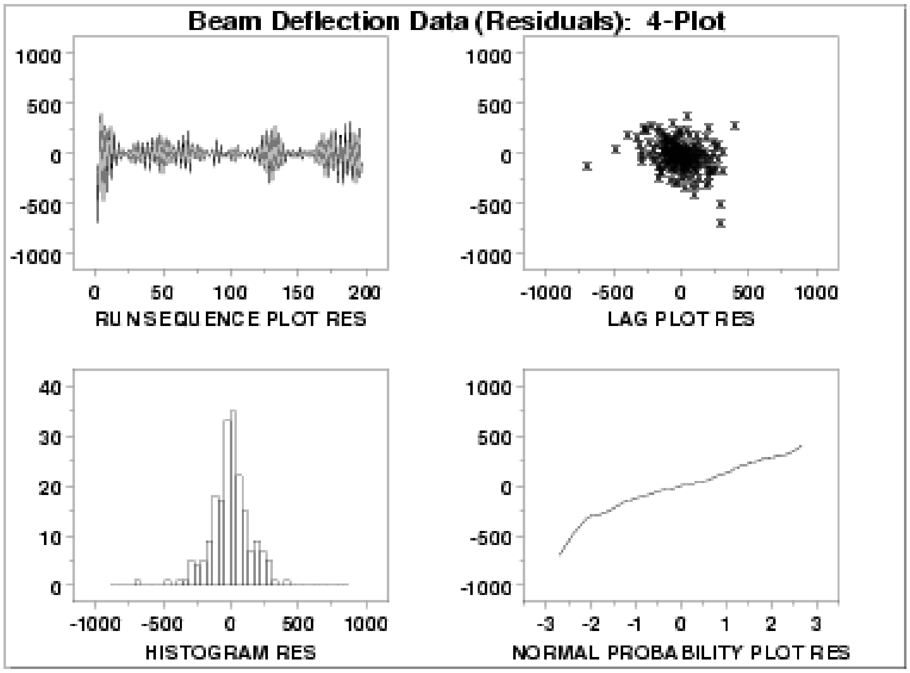
$$Y_i = -178.79 - 361.77 * (2\pi * 0.302596 T_i + 1.465) + E_i$$

The new fit, with a residual standard deviation of 148.34, is:

$$Y_i = -178.79 - 361.76 * (2\pi * 0.302597 T_i + 1.465) + E_i$$

There is minimal change in the parameter estimates and about a 5% reduction in the residual standard deviation. In this case, removing the residuals has a modest benefit in terms of reducing the variability of the model.

4-Plot for New Fit



This plot shows that the underlying assumptions are satisfied and therefore the new fit is a good descriptor of the data.

In this case, it is a judgment call whether to use the fit with or without the outliers removed.

1.4.2.5.5. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. Validate assumptions. 1. 4-plot of Y. 2. Generate a run sequence plot. 3. Generate a lag plot. 4. Generate an autocorrelation plot.	1. Based on the 4-plot, there are no obvious shifts in location and scale, but the data are not random. 2. Based on the run sequence plot, there are no obvious shifts in location and scale. 3. Based on the lag plot, the data are not random. 4. The autocorrelation plot shows significant autocorrelation at lag 1. 5. The spectral plot shows a single dominant

	low frequency peak.
5. Generate a spectral plot.	
6. Generate a table of summary statistics.	6. The summary statistics table displays 25+ statistics.
7. Generate a linear fit to detect drift in location.	7. The linear fit indicates no drift in location since the slope parameter is not statistically significant.
8. Detect drift in variation by dividing the data into quarters and computing Levene's test statistic for equal standard deviations.	8. Levene's test indicates no significant drift in variation.
9. Check for randomness by generating a runs test.	9. The runs test indicates significant non-randomness.
3. Fit $Y_i = C + A*SIN(2*PI*omega*t_i+phi)$.	
1. Generate a complex demodulation phase plot.	1. Complex demodulation phase plot indicates a starting frequency of 0.3025.
2. Generate a complex demodulation amplitude plot.	2. Complex demodulation amplitude plot indicates an amplitude of 390 (but there is a short start-up effect).
3. Fit the non-linear model.	3. Non-linear fit generates final parameter estimates. The residual standard deviation from the fit is 155.85 (compared to the standard deviation of 277.73 from the original data).

4. Validate fit.	
1. Generate a 4-plot of the residuals from the fit.	1. The 4-plot indicates that the assumptions of constant location and scale are valid. The lag plot indicates that the data are random. The histogram and normal probability plot indicate that the residuals that the normality assumption for the residuals are not seriously violated, although there is a bend on the probablity plot that warrants attention.
2. Generate a nonlinear fit with outliers removed.	2. The fit after removing 3 outliers shows some marginal improvement in the model (a 5% reduction in the residual standard deviation).
3. Generate a 4-plot of the residuals from the fit with the outliers removed.	3. The 4-plot of the model fit after 3 outliers removed shows marginal improvement in satisfying model assumptions.

1.4.2.6. Filter Transmittance

Filter Transmittance

This example illustrates the univariate analysis of filter transmittance data.

1.4.2.6.1. Background and Data

Generation

This data set was collected by NIST chemist Radu Mavrodineaunu in the 1970's from an automatic data acquisition system for a filter transmittance experiment. The response variable is transmittance.

The motivation for studying this data set is to show how the underlying autocorrelation structure in a relatively small data set helped the scientist detect problems with his automatic data acquisition system.

This file can be read by Dataplot with the following commands:

```
SKIP 25
READ MAVRO.DAT Y
```

Resulting Data

The following are the data used for this case study.

```
2.00180
2.00170
2.00180
2.00190
2.00180
2.00170
2.00150
2.00140
2.00150
2.00150
2.00170
2.00180
2.00180
2.00190
2.00190
2.00210
2.00200
2.00160
2.00140
```

```
2.00130
2.00130
2.00150
2.00150
2.00160
2.00150
2.00140
2.00130
2.00140
2.00150
2.00140
2.00150
2.00160
2.00150
2.00160
2.00190
2.00200
2.00200
2.00210
2.00220
2.00230
2.00240
2.00250
2.00270
2.00260
2.00260
2.00260
2.00270
2.00260
2.00250
2.00240
```

1.4.2.6.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

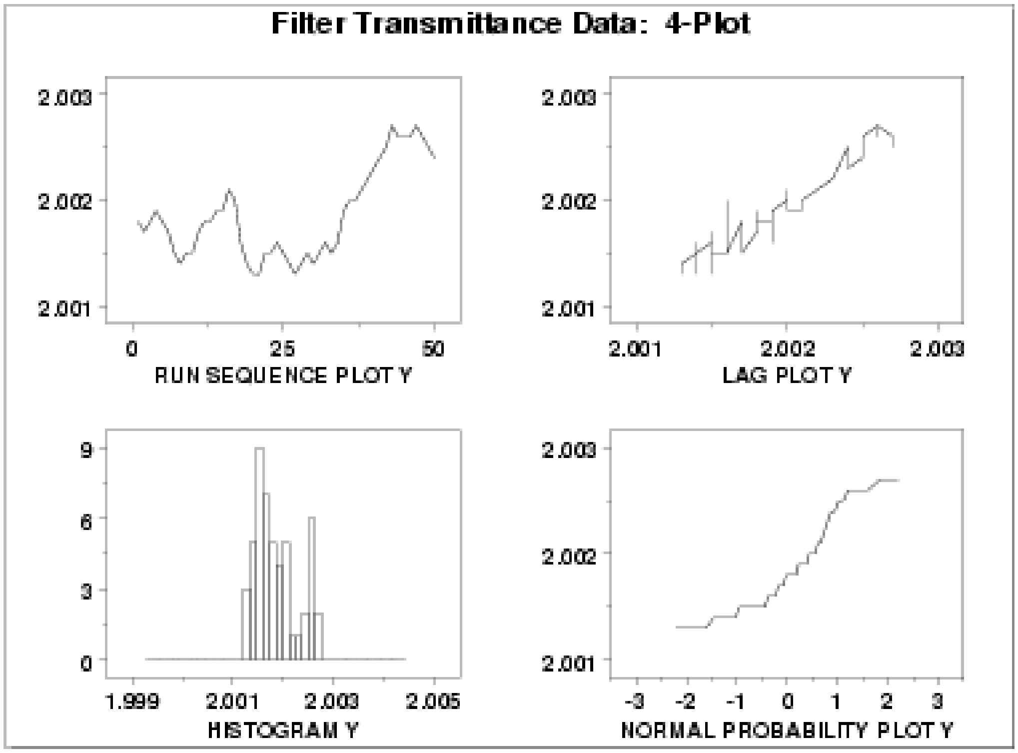
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates a significant shift in location around x=35.
2. The linear appearance in the lag plot (upper right) indicates a non-random pattern in the data.
3. Since the lag plot indicates significant non-randomness, we do not make any interpretation of either the histogram (lower left) or the normal probability plot (lower right).

The serious violation of the non-randomness assumption means that the univariate model

$$Y_i = C + E_i$$

is not valid. Given the linear appearance of the lag plot, the first step might be to consider a model of the type

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

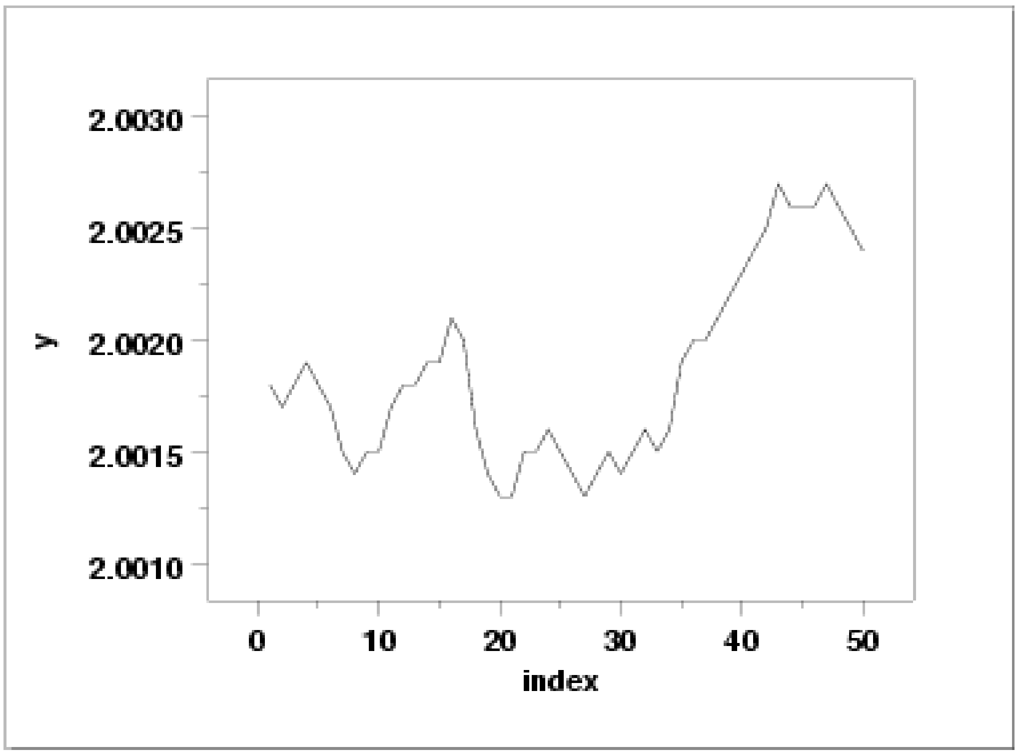
However, in this case discussions with the scientist revealed that non-randomness was entirely unexpected. An examination of the experimental process revealed that the sampling rate for the automatic data acquisition system was too fast. That is, the equipment did not have sufficient time to reset before the next sample started, resulting in the current measurement being contaminated by the previous measurement. The solution was to rerun the experiment allowing more time between samples.

Simple graphical techniques can be quite effective in revealing unexpected results in the data. When this occurs, it is important to investigate whether the unexpected result is due to problems in the experiment and data collection or is indicative of unexpected underlying structure in the data. This determination cannot be made on the basis of statistics alone. The role of the graphical and statistical analysis is to detect problems or unexpected results in the data. Resolving the issues requires the knowledge of the scientist or engineer.

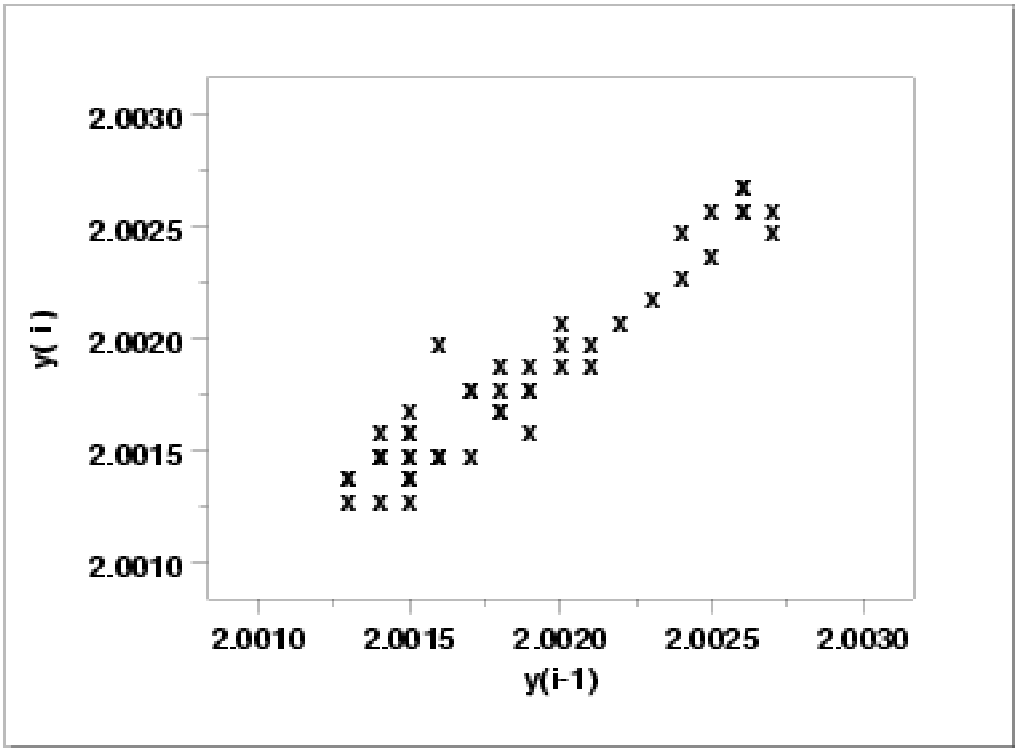
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail. Since the lag plot indicates significant non-randomness, we omit the distributional plots.

Run Sequence Plot



Lag Plot


2.0030
2.0025
2.0020
2.0015
2.0010
y(i)
2.0010
2.0015
2.0020
2.0025
2.0030
y(i-1)

1.4.2.6.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                         SUMMARY

               NUMBER OF OBSERVATIONS =       50


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.2002000E+01  *  RANGE        =   0.1399994E-02  *
*  MEAN         =   0.2001856E+01  *  STAND. DEV.  =   0.4291329E-03  *
*  MIDMEAN      =   0.2001638E+01  *  AV. AB. DEV. =   0.3480196E-03  *
*  MEDIAN       =   0.2001800E+01  *  MINIMUM      =   0.2001300E+01  *
*               =                  *  LOWER QUART. =   0.2001500E+01  *
*               =                  *  LOWER HINGE  =   0.2001500E+01  *
*               =                  *  UPPER HINGE  =   0.2002100E+01  *
*               =                  *  UPPER QUART. =   0.2002175E+01  *
*               =                  *  MAXIMUM      =   0.2002700E+01  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.9379919E+00  *  ST. 3RD MOM. =   0.6191616E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.2098746E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =  -0.4995516E+01  *
*               =                  *  UNIFORM PPCC =   0.9666610E+00  *
*               =                  *  NORMAL  PPCC =   0.9558001E+00  *
*               =                  *  TUK -.5 PPCC =   0.8462552E+00  *
*               =                  *  CAUCHY  PPCC =   0.6822084E+00  *
***********************************************************************
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generates the following output:

```
LEAST SQUARES MULTILINEAR FIT
      SAMPLE SIZE N         =       50
      NUMBER OF VARIABLES =         1
      NO REPLICATION CASE


              PARAMETER ESTIMATES           (APPROX. ST. DEV.)    T VALUE
       1  A0                    2.00138      (0.9695E-04)        0.2064E+05
       2  A1        X         0.184685E-04   (0.3309E-05)         5.582

      RESIDUAL    STANDARD DEVIATION =        0.3376404E-03
      RESIDUAL    DEGREES OF FREEDOM =          48
```

The slope parameter, A1, has a *t* value of 5.6, which is statistically significant. The value of the slope parameter is 0.0000185. Although this number is nearly zero, we need to take into account that the original scale of the data is from about 2.0012 to 2.0028. In this case, we conclude that there is a drift in location, although by a relatively minor amount.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal sized intervals. However, the Bartlett test is not robust for non-normality. Since the normality assumption is questionable for these data, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

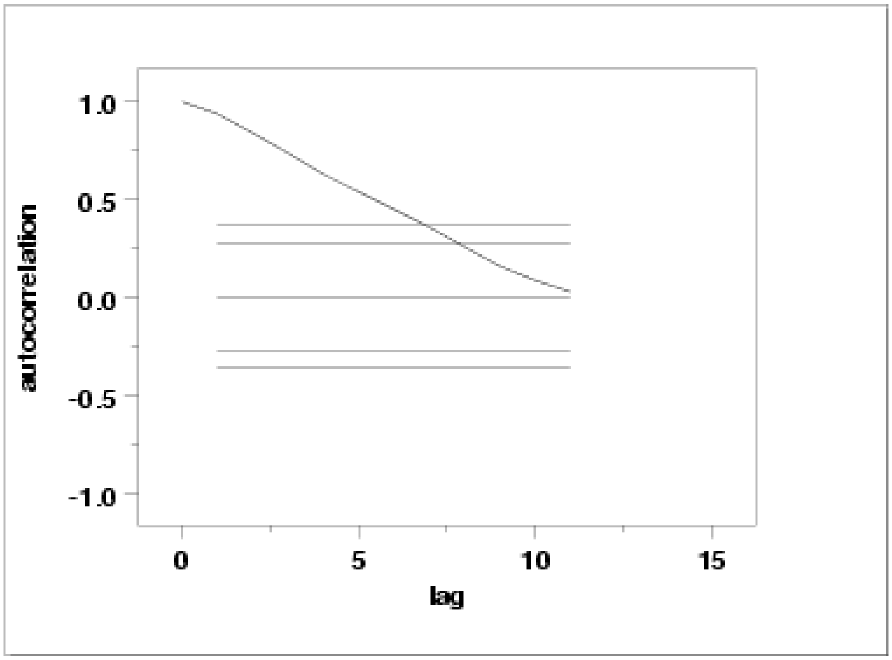
```
               LEVENE F-TEST FOR SHIFT IN VARIATION
                      (ASSUMPTION: NORMALITY)

1. STATISTICS
      NUMBER OF OBSERVATIONS     =       50
      NUMBER OF GROUPS           =        4
      LEVENE F TEST STATISTIC    =   0.9714893


   FOR LEVENE TEST STATISTIC
      0           % POINT    =   0.0000000E+00
      50          % POINT    =   0.8004835
      75          % POINT    =    1.416631
      90          % POINT    =    2.206890
      95          % POINT    =    2.806845
      99          % POINT    =    4.238307
      99.9        % POINT    =    6.424733


         58.56597       % Point:    0.9714893

3. CONCLUSION (AT THE 5% LEVEL):
      THERE IS NO SHIFT IN VARIATION.
      THUS: HOMOGENEOUS WITH RESPECT TO VARIATION.
```

In this case, since the Levene test statistic value of 0.971 is less than the critical value of 2.806 at the 5% level, we conclude that there is no evidence of a change in variation.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot in the previous seciton is a simple graphical technique.

One check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of most interest, is 0.93. The critical values at the 5% level are -0.277 and 0.277. This indicates that the lag 1 autocorrelation is statistically significant, so there is strong evidence of non-randomness.

A common test for randomness is the runs test.

```
                  RUNS UP

       STATISTIC = NUMBER OF RUNS UP
           OF LENGTH EXACTLY I

   I         STAT     EXP(STAT)    SD(STAT)        Z

   1         1.0     10.4583      3.2170       -2.94
   2         3.0      4.4667      1.6539       -0.89
   3         1.0      1.2542      0.9997       -0.25
   4         0.0      0.2671      0.5003       -0.53
   5         0.0      0.0461      0.2132       -0.22
   6         0.0      0.0067      0.0818       -0.08
   7         0.0      0.0008      0.0291       -0.03
   8         1.0      0.0001      0.0097      103.06
   9         0.0      0.0000      0.0031        0.00
  10         1.0      0.0000      0.0009     1087.63


       STATISTIC = NUMBER OF RUNS UP
           OF LENGTH I OR MORE
```

I	STAT	EXP(STAT)	SD(STAT)	Z
1	7.0	16.5000	2.0696	-4.59
2	6.0	6.0417	1.3962	-0.03
3	3.0	1.5750	1.0622	1.34
4	2.0	0.3208	0.5433	3.09
5	2.0	0.0538	0.2299	8.47
6	2.0	0.0077	0.0874	22.79
7	2.0	0.0010	0.0308	64.85
8	2.0	0.0001	0.0102	195.70
9	1.0	0.0000	0.0032	311.64
10	1.0	0.0000	0.0010	1042.19

RUNS DOWN

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	3.0	10.4583	3.2170	-2.32
2	0.0	4.4667	1.6539	-2.70
3	3.0	1.2542	0.9997	1.75
4	1.0	0.2671	0.5003	1.46
5	1.0	0.0461	0.2132	4.47
6	0.0	0.0067	0.0818	-0.08
7	0.0	0.0008	0.0291	-0.03
8	0.0	0.0001	0.0097	-0.01
9	0.0	0.0000	0.0031	0.00
10	0.0	0.0000	0.0009	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	8.0	16.5000	2.0696	-4.11
2	5.0	6.0417	1.3962	-0.75
3	5.0	1.5750	1.0622	3.22
4	2.0	0.3208	0.5433	3.09
5	1.0	0.0538	0.2299	4.12
6	0.0	0.0077	0.0874	-0.09
7	0.0	0.0010	0.0308	-0.03
8	0.0	0.0001	0.0102	-0.01
9	0.0	0.0000	0.0032	0.00
10	0.0	0.0000	0.0010	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	4.0	20.9167	4.5496	-3.72
2	3.0	8.9333	2.3389	-2.54
3	4.0	2.5083	1.4138	1.06
4	1.0	0.5341	0.7076	0.66
5	1.0	0.0922	0.3015	3.01

```
       6          0.0       0.0134      0.1157        -0.12
       7          0.0       0.0017      0.0411        -0.04
       8          1.0       0.0002      0.0137        72.86
       9          0.0       0.0000      0.0043         0.00
      10          1.0       0.0000      0.0013       769.07


              STATISTIC = NUMBER OF RUNS TOTAL
                    OF LENGTH I OR MORE

       I          STAT     EXP(STAT)    SD(STAT)        Z

       1         15.0     33.0000      2.9269        -6.15
       2         11.0     12.0833      1.9745        -0.55
       3          8.0      3.1500      1.5022         3.23
       4          4.0      0.6417      0.7684         4.37
       5          3.0      0.1075      0.3251         8.90
       6          2.0      0.0153      0.1236        16.05
       7          2.0      0.0019      0.0436        45.83
       8          2.0      0.0002      0.0145       138.37
       9          1.0      0.0000      0.0045       220.36
      10          1.0      0.0000      0.0014       736.94


               LENGTH OF THE LONGEST RUN UP         =    10
               LENGTH OF THE LONGEST RUN DOWN       =     5
               LENGTH OF THE LONGEST RUN UP OR DOWN =    10

               NUMBER OF POSITIVE DIFFERENCES =    23
               NUMBER OF NEGATIVE DIFFERENCES =    18
               NUMBER OF ZERO     DIFFERENCES =     8
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. Due to the number of values that are much larger than the 1.96 cut-off, we conclude that the data are not random.

Distributional Analysis

Since we rejected the randomness assumption, the distributional tests are not meaningful. Therefore, these quantitative tests are omitted. We also omit Grubbs' outlier test since it also assumes the data are approximately normally distributed.

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
Analysis for filter transmittance data

1: Sample Size                              = 50

2: Location
   Mean                                     = 2.001857
   Standard Deviation of Mean               = 0.00006
   95% Confidence Interval for Mean         = (2.001735,2.001979)
   Drift with respect to location?          = NO

3: Variation
   Standard Deviation                       = 0.00043
   95% Confidence Interval for SD           = (0.000359,0.000535)
   Change in variation?
   (based on Levene's test on quarters
   of the data)                             = NO
```

```
4: Distribution
   Distributional tests omitted due to
   non-randomness of the data

5: Randomness
   Lag One Autocorrelation               = 0.937998
   Data are Random?
     (as measured by autocorrelation)    = NO

6: Statistical Control
   (i.e., no drift in location or scale,
   data are random, distribution is
   fixed, here we are testing only for
   normal)
   Data Set is in Statistical Control?   = NO

7: Outliers?
   (Grubbs' test omitted)                = NO
```

1.4.2.6.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*

`1. Invoke Dataplot and read data.`
 	
`1. Read in the data.`
	`1. You have read 1 column of numbers into Dataplot, variable Y.`

`2. 4-plot of the data.`
 	
`1. 4-plot of Y.`
	`1. Based on the 4-plot, there is a shift in location and the data are not random.`

`3. Generate the individual plots.`
 	
`1. Generate a run sequence plot.`
 	
`2. Generate a lag plot.`
	`1. The run sequence plot indicates that there is a shift in location.`
	
	`2. The strong linear pattern of the lag plot indicates significant non-randomness.`

4. Generate summary statistics, quantitative analysis, and print a univariate report.	
1. Generate a table of summary statistics.	1. The summary statistics table displays 25+ statistics.
2. Compute a linear fit based on quarters of the data to detect drift in location.	2. The linear fit indicates a slight drift in location since the slope parameter is statistically significant, but small.
3. Compute Levene's test based on quarters of the data to detect changes in variation.	3. Levene's test indicates no significant drift in variation.
4. Check for randomness by generating an autocorrelation plot and a runs test.	4. The lag 1 autocorrelation is 0.94. This is outside the 95% confidence interval bands which indicates significant non-randomness.
5. Print a univariate report (this assumes steps 2 thru 4 have already been run).	5. The results are summarized in a convenient report.

1. Exploratory Data Analysis
1.4. EDA Case Studies
1.4.2. Case Studies

1.4.2.7. Standard Resistor

Standard Resistor

This example illustrates the univariate analysis of standard resistor data.

1.4.2.7.1. Background and Data

Generation

This data set was collected by Ron Dziuba of NIST over a 5-year period from 1980 to 1985. The response variable is resistor values.

The motivation for studying this data set is to illustrate data that violate the assumptions of constant location and scale.

This file can be read by Dataplot with the following commands:

SKIP 25
COLUMN LIMITS 10 80
READ DZIUBA1.DAT Y
COLUMN LIMITS

Resulting Data

The following are the data used for this case study.

```
27.8680
27.8929
27.8773
27.8530
27.8876
27.8725
27.8743
27.8879
27.8728
27.8746
27.8863
27.8716
27.8818
27.8872
27.8885
27.8945
27.8797
27.8627
27.8870
```

```
27.8895
27.9138
27.8931
27.8852
27.8788
27.8827
27.8939
27.8558
27.8814
27.8479
27.8479
27.8848
27.8809
27.8479
27.8611
27.8630
27.8679
27.8637
27.8985
27.8900
27.8577
27.8848
27.8869
27.8976
27.8610
27.8567
27.8417
27.8280
27.8555
27.8639
27.8702
27.8582
27.8605
27.8900
27.8758
27.8774
27.9008
27.8988
27.8897
27.8990
27.8958
27.8830
27.8967
27.9105
27.9028
27.8977
```

```
27.8953
27.8970
27.9190
27.9180
27.8997
27.9204
27.9234
27.9072
27.9152
27.9091
27.8882
27.9035
27.9267
27.9138
27.8955
27.9203
27.9239
27.9199
27.9646
27.9411
27.9345
27.8712
27.9145
27.9259
27.9317
27.9239
27.9247
27.9150
27.9444
27.9457
27.9166
27.9066
27.9088
27.9255
27.9312
27.9439
27.9210
27.9102
27.9083
27.9121
27.9113
27.9091
27.9235
27.9291
27.9253
27.9092
```

```
27.9117
27.9194
27.9039
27.9515
27.9143
27.9124
27.9128
27.9260
27.9339
27.9500
27.9530
27.9430
27.9400
27.8850
27.9350
27.9120
27.9260
27.9660
27.9280
27.9450
27.9390
27.9429
27.9207
27.9205
27.9204
27.9198
27.9246
27.9366
27.9234
27.9125
27.9032
27.9285
27.9561
27.9616
27.9530
27.9280
27.9060
27.9380
27.9310
27.9347
27.9339
27.9410
27.9397
27.9472
27.9235
27.9315
```

```
27.9368
27.9403
27.9529
27.9263
27.9347
27.9371
27.9129
27.9549
27.9422
27.9423
27.9750
27.9339
27.9629
27.9587
27.9503
27.9573
27.9518
27.9527
27.9589
27.9300
27.9629
27.9630
27.9660
27.9730
27.9660
27.9630
27.9570
27.9650
27.9520
27.9820
27.9560
27.9670
27.9520
27.9470
27.9720
27.9610
27.9437
27.9660
27.9580
27.9660
27.9700
27.9600
27.9660
27.9770
27.9110
27.9690
```

```
27.9698
27.9616
27.9371
27.9700
27.9265
27.9964
27.9842
27.9667
27.9610
27.9943
27.9616
27.9397
27.9799
28.0086
27.9709
27.9741
27.9675
27.9826
27.9676
27.9703
27.9789
27.9786
27.9722
27.9831
28.0043
27.9548
27.9875
27.9495
27.9549
27.9469
27.9744
27.9744
27.9449
27.9837
27.9585
28.0096
27.9762
27.9641
27.9854
27.9877
27.9839
27.9817
27.9845
27.9877
27.9880
27.9822
```

```
27.9836
28.0030
27.9678
28.0146
27.9945
27.9805
27.9785
27.9791
27.9817
27.9805
27.9782
27.9753
27.9792
27.9704
27.9794
27.9814
27.9794
27.9795
27.9881
27.9772
27.9796
27.9736
27.9772
27.9960
27.9795
27.9779
27.9829
27.9829
27.9815
27.9811
27.9773
27.9778
27.9724
27.9756
27.9699
27.9724
27.9666
27.9666
27.9739
27.9684
27.9861
27.9901
27.9879
27.9865
27.9876
27.9814
```

```
27.9842
27.9868
27.9834
27.9892
27.9864
27.9843
27.9838
27.9847
27.9860
27.9872
27.9869
27.9602
27.9852
27.9860
27.9836
27.9813
27.9623
27.9843
27.9802
27.9863
27.9813
27.9881
27.9850
27.9850
27.9830
27.9866
27.9888
27.9841
27.9863
27.9903
27.9961
27.9905
27.9945
27.9878
27.9929
27.9914
27.9914
27.9997
28.0006
27.9999
28.0004
28.0020
28.0029
28.0008
28.0040
28.0078
```

```
28.0065
27.9959
28.0073
28.0017
28.0042
28.0036
28.0055
28.0007
28.0066
28.0011
27.9960
28.0083
27.9978
28.0108
28.0088
28.0088
28.0139
28.0092
28.0092
28.0049
28.0111
28.0120
28.0093
28.0116
28.0102
28.0139
28.0113
28.0158
28.0156
28.0137
28.0236
28.0171
28.0224
28.0184
28.0199
28.0190
28.0204
28.0170
28.0183
28.0201
28.0182
28.0183
28.0175
28.0127
28.0211
28.0057
```

```
28.0180
28.0183
28.0149
28.0185
28.0182
28.0192
28.0213
28.0216
28.0169
28.0162
28.0167
28.0167
28.0169
28.0169
28.0161
28.0152
28.0179
28.0215
28.0194
28.0115
28.0174
28.0178
28.0202
28.0240
28.0198
28.0194
28.0171
28.0134
28.0121
28.0121
28.0141
28.0101
28.0114
28.0122
28.0124
28.0171
28.0165
28.0166
28.0159
28.0181
28.0200
28.0116
28.0144
28.0141
28.0116
28.0107
```

```
28.0169
28.0105
28.0136
28.0138
28.0114
28.0122
28.0122
28.0116
28.0025
28.0097
28.0066
28.0072
28.0066
28.0068
28.0067
28.0130
28.0091
28.0088
28.0091
28.0091
28.0115
28.0087
28.0128
28.0139
28.0095
28.0115
28.0101
28.0121
28.0114
28.0121
28.0122
28.0121
28.0168
28.0212
28.0219
28.0221
28.0204
28.0169
28.0141
28.0142
28.0147
28.0159
28.0165
28.0144
28.0182
28.0155
```

28.0155
28.0192
28.0204
28.0185
28.0248
28.0185
28.0226
28.0271
28.0290
28.0240
28.0302
28.0243
28.0288
28.0287
28.0301
28.0273
28.0313
28.0293
28.0300
28.0344
28.0308
28.0291
28.0287
28.0358
28.0309
28.0286
28.0308
28.0291
28.0380
28.0411
28.0420
28.0359
28.0368
28.0327
28.0361
28.0334
28.0300
28.0347
28.0359
28.0344
28.0370
28.0355
28.0371
28.0318
28.0390
28.0390

```
28.0390
28.0376
28.0376
28.0377
28.0345
28.0333
28.0429
28.0379
28.0401
28.0401
28.0423
28.0393
28.0382
28.0424
28.0386
28.0386
28.0373
28.0397
28.0412
28.0565
28.0419
28.0456
28.0426
28.0423
28.0391
28.0403
28.0388
28.0408
28.0457
28.0455
28.0460
28.0456
28.0464
28.0442
28.0416
28.0451
28.0432
28.0434
28.0448
28.0448
28.0373
28.0429
28.0392
28.0469
28.0443
28.0356
```

```
28.0474
28.0446
28.0348
28.0368
28.0418
28.0445
28.0533
28.0439
28.0474
28.0435
28.0419
28.0538
28.0538
28.0463
28.0491
28.0441
28.0411
28.0507
28.0459
28.0519
28.0554
28.0512
28.0507
28.0582
28.0471
28.0539
28.0530
28.0502
28.0422
28.0431
28.0395
28.0177
28.0425
28.0484
28.0693
28.0490
28.0453
28.0494
28.0522
28.0393
28.0443
28.0465
28.0450
28.0539
28.0566
28.0585
```

```
28.0486
28.0427
28.0548
28.0616
28.0298
28.0726
28.0695
28.0629
28.0503
28.0493
28.0537
28.0613
28.0643
28.0678
28.0564
28.0703
28.0647
28.0579
28.0630
28.0716
28.0586
28.0607
28.0601
28.0611
28.0606
28.0611
28.0066
28.0412
28.0558
28.0590
28.0750
28.0483
28.0599
28.0490
28.0499
28.0565
28.0612
28.0634
28.0627
28.0519
28.0551
28.0696
28.0581
28.0568
28.0572
28.0529
```

```
28.0421
28.0432
28.0211
28.0363
28.0436
28.0619
28.0573
28.0499
28.0340
28.0474
28.0534
28.0589
28.0466
28.0448
28.0576
28.0558
28.0522
28.0480
28.0444
28.0429
28.0624
28.0610
28.0461
28.0564
28.0734
28.0565
28.0503
28.0581
28.0519
28.0625
28.0583
28.0645
28.0642
28.0535
28.0510
28.0542
28.0677
28.0416
28.0676
28.0596
28.0635
28.0558
28.0623
28.0718
28.0585
28.0552
```

28.0684
28.0646
28.0590
28.0465
28.0594
28.0303
28.0533
28.0561
28.0585
28.0497
28.0582
28.0507
28.0562
28.0715
28.0468
28.0411
28.0587
28.0456
28.0705
28.0534
28.0558
28.0536
28.0552
28.0461
28.0598
28.0598
28.0650
28.0423
28.0442
28.0449
28.0660
28.0506
28.0655
28.0512
28.0407
28.0475
28.0411
28.0512
28.1036
28.0641
28.0572
28.0700
28.0577
28.0637
28.0534
28.0461

```
28.0701
28.0631
28.0575
28.0444
28.0592
28.0684
28.0593
28.0677
28.0512
28.0644
28.0660
28.0542
28.0768
28.0515
28.0579
28.0538
28.0526
28.0833
28.0637
28.0529
28.0535
28.0561
28.0736
28.0635
28.0600
28.0520
28.0695
28.0608
28.0608
28.0590
28.0290
28.0939
28.0618
28.0551
28.0757
28.0698
28.0717
28.0529
28.0644
28.0613
28.0759
28.0745
28.0736
28.0611
28.0732
28.0782
```

```
28.0682
28.0756
28.0857
28.0739
28.0840
28.0862
28.0724
28.0727
28.0752
28.0732
28.0703
28.0849
28.0795
28.0902
28.0874
28.0971
28.0638
28.0877
28.0751
28.0904
28.0971
28.0661
28.0711
28.0754
28.0516
28.0961
28.0689
28.1110
28.1062
28.0726
28.1141
28.0913
28.0982
28.0703
28.0654
28.0760
28.0727
28.0850
28.0877
28.0967
28.1185
28.0945
28.0834
28.0764
28.1129
28.0797
```

```
28.0707
28.1008
28.0971
28.0826
28.0857
28.0984
28.0869
28.0795
28.0875
28.1184
28.0746
28.0816
28.0879
28.0888
28.0924
28.0979
28.0702
28.0847
28.0917
28.0834
28.0823
28.0917
28.0779
28.0852
28.0863
28.0942
28.0801
28.0817
28.0922
28.0914
28.0868
28.0832
28.0881
28.0910
28.0886
28.0961
28.0857
28.0859
28.1086
28.0838
28.0921
28.0945
28.0839
28.0877
28.0803
28.0928
```

```
28.0885
28.0940
28.0856
28.0849
28.0955
28.0955
28.0846
28.0871
28.0872
28.0917
28.0931
28.0865
28.0900
28.0915
28.0963
28.0917
28.0950
28.0898
28.0902
28.0867
28.0843
28.0939
28.0902
28.0911
28.0909
28.0949
28.0867
28.0932
28.0891
28.0932
28.0887
28.0925
28.0928
28.0883
28.0946
28.0977
28.0914
28.0959
28.0926
28.0923
28.0950
28.1006
28.0924
28.0963
28.0893
28.0956
```

```
28.0980
28.0928
28.0951
28.0958
28.0912
28.0990
28.0915
28.0957
28.0976
28.0888
28.0928
28.0910
28.0902
28.0950
28.0995
28.0965
28.0972
28.0963
28.0946
28.0942
28.0998
28.0911
28.1043
28.1002
28.0991
28.0959
28.0996
28.0926
28.1002
28.0961
28.0983
28.0997
28.0959
28.0988
28.1029
28.0989
28.1000
28.0944
28.0979
28.1005
28.1012
28.1013
28.0999
28.0991
28.1059
28.0961
```

```
28.0981
28.1045
28.1047
28.1042
28.1146
28.1113
28.1051
28.1065
28.1065
28.0985
28.1000
28.1066
28.1041
28.0954
28.1090
```

1.4.2.7.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

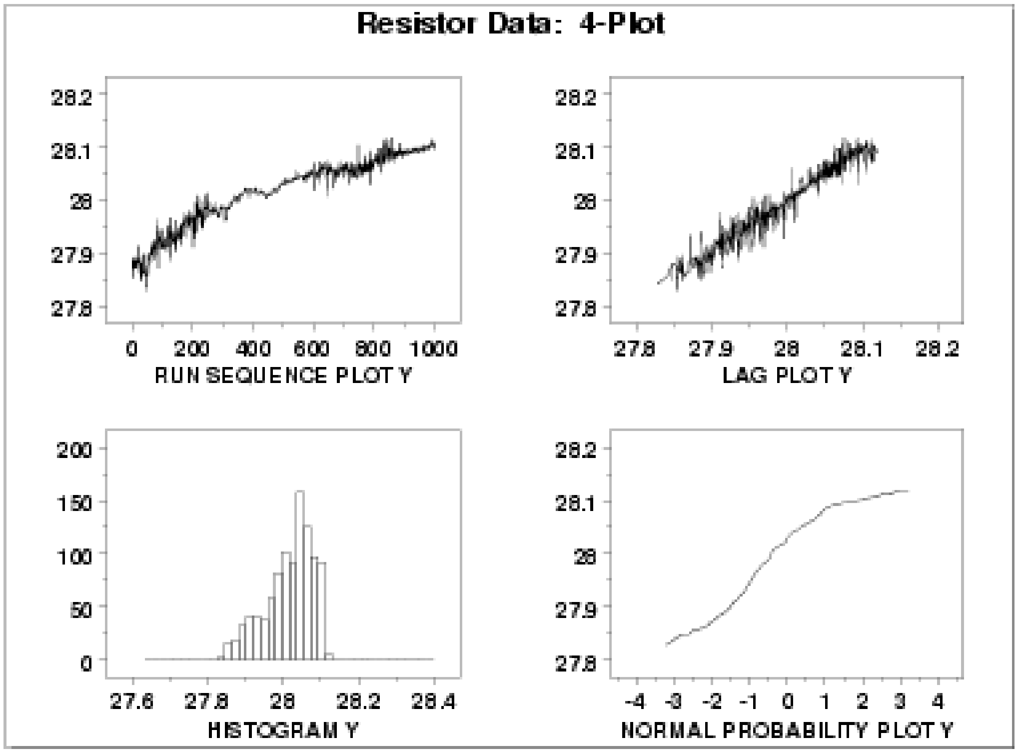
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates significant shifts in both location and variation. Specifically, the location is increasing with time. The variability seems greater in the first and last third of the data than it does in the middle third.
2. The lag plot (upper right) shows a significant non-random pattern in the data. Specifically, the strong linear appearance of this plot is indicative of a model that relates Y_t to Y_{t-1}.
3. The distributional plots, the histogram (lower left) and the normal probability plot (lower right), are not interpreted since the randomness assumption is so clearly violated.

The serious violation of the non-randomness assumption means that the univariate model

$$Y_i = C + E_i$$

is not valid. Given the linear appearance of the lag plot, the first step might be to consider a model of the type

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

However, discussions with the scientist revealed the following:

1. the drift with respect to location was expected.
2. the non-constant variability was not expected.

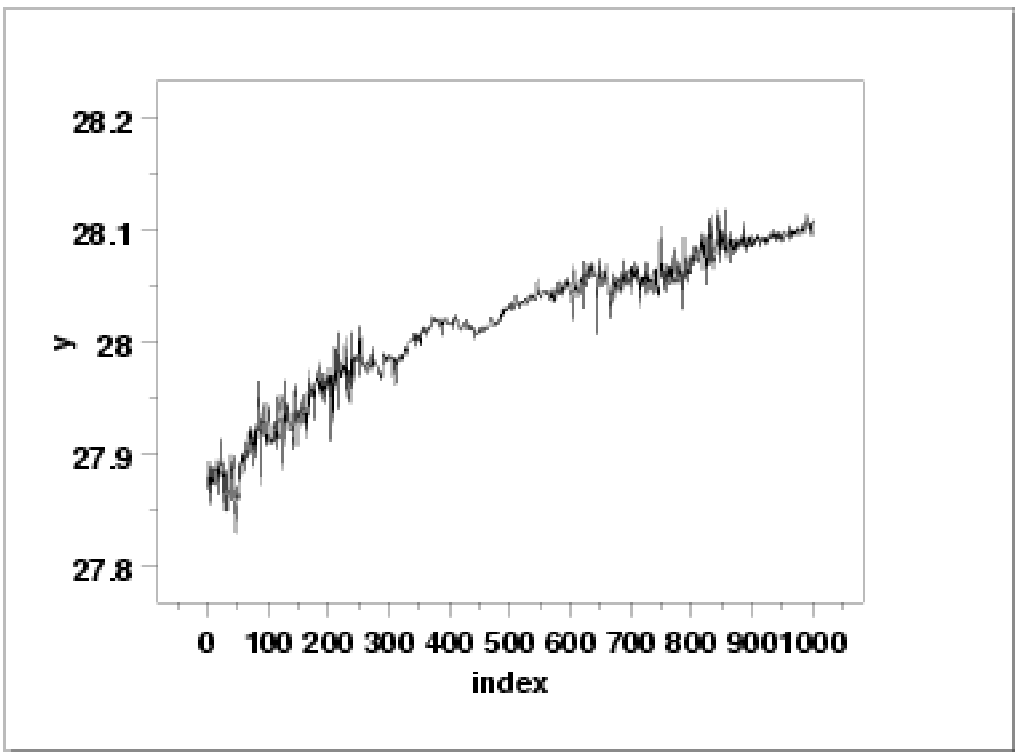
The scientist examined the data collection device and determined that the non-constant variation was a seasonal effect. The high variability data in the first and last thirds was collected in winter while the more stable middle third was collected in the summer. The seasonal effect was determined to be caused by the amount of humidity affecting the measurement equipment. In this case, the solution was to modify the test equipment to be less sensitive to enviromental factors.

Simple graphical techniques can be quite effective in revealing unexpected results in the data. When this occurs, it is important to investigate whether the unexpected result is due to problems in the experiment and data collection, or is it in fact indicative of an unexpected underlying structure in the data. This determination cannot be made on the basis of statistics alone. The role of the graphical and statistical analysis is to detect problems or unexpected results in the data. Resolving the issues requires the knowledge of the scientist or engineer.

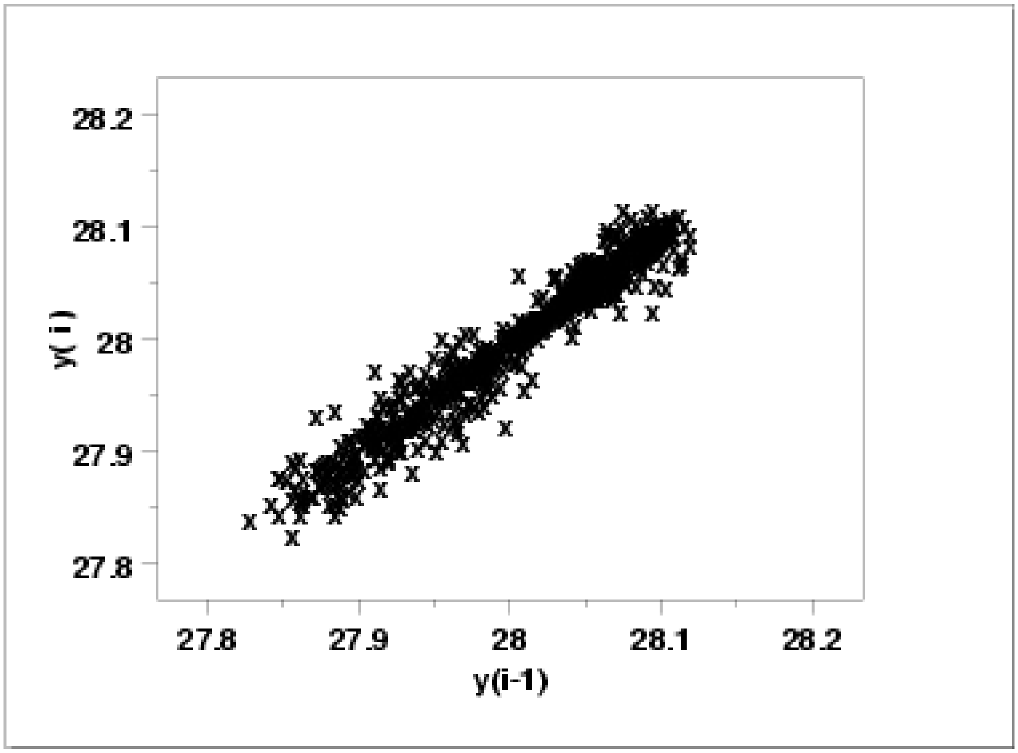
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail. Since the lag plot indicates significant non-randomness, we omit the distributional plots.

Run Sequence Plot



Lag Plot



1.4.2.7.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                          SUMMARY

               NUMBER OF OBSERVATIONS =     1000

***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.2797325E+02  *  RANGE        =   0.2905006E+00  *
*  MEAN         =   0.2801634E+02  *  STAND. DEV.  =   0.6349404E-01  *
*  MIDMEAN      =   0.2802659E+02  *  AV. AB. DEV. =   0.5101655E-01  *
*  MEDIAN       =   0.2802910E+02  *  MINIMUM      =   0.2782800E+02  *
*               =                  *  LOWER QUART. =   0.2797905E+02  *
*               =                  *  LOWER HINGE  =   0.2797900E+02  *
*               =                  *  UPPER HINGE  =   0.2806295E+02  *
*               =                  *  UPPER QUART. =   0.2806293E+02  *
*               =                  *  MAXIMUM      =   0.2811850E+02  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.9721591E+00  *  ST. 3RD MOM. =  -0.6936395E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.2689681E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =  -0.4216419E+02  *
*               =                  *  UNIFORM PPCC =   0.9689648E+00  *
*               =                  *  NORMAL  PPCC =   0.9718416E+00  *
*               =                  *  TUK -.5 PPCC =   0.7334843E+00  *
*               =                  *  CAUCHY  PPCC =   0.3347875E+00  *
***********************************************************************
```

The autocorrelation coefficient of 0.972 is evidence of significant non-randomness.

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter estimate should be zero. For this data set, Dataplot generates the following output:

```
 LEAST SQUARES MULTILINEAR FIT
       SAMPLE SIZE N          =     1000
       NUMBER OF VARIABLES =        1
       NO REPLICATION CASE


               PARAMETER ESTIMATES          (APPROX. ST. DEV.)    T VALUE
        1  A0                     27.9114       (0.1209E-02)        0.2309E+05
        2  A1       X            0.209670E-03   (0.2092E-05)         100.2

       RESIDUAL    STANDARD DEVIATION =        0.1909796E-01
       RESIDUAL    DEGREES OF FREEDOM =         998

       COEF AND SD(COEF) WRITTEN OUT TO FILE DPST1F.DAT
       SD(PRED),95LOWER,95UPPER,99LOWER,99UPPER
                         WRITTEN OUT TO FILE DPST2F.DAT
       REGRESSION DIAGNOSTICS WRITTEN OUT TO FILE DPST3F.DAT
       PARAMETER VARIANCE-COVARIANCE MATRIX AND
       INVERSE OF X-TRANSPOSE X MATRIX
       WRITTEN OUT TO FILE DPST4F.DAT
```

The slope parameter, A1, has a *t* value of 100 which is statistically significant. The value of the slope parameter estimate is 0.00021. Although this number is nearly zero, we need to take into account that the original scale of the data is from about 27.8 to 28.2. In this case, we conclude that there is a drift in location.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. However, the Bartlett test is not robust for non-normality. Since the normality assumption is questionable for these data, we use the alternative Levene test. In partiuclar, we use the Levene test based on the median rather the mean. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Levene test.

```
               LEVENE F-TEST FOR SHIFT IN VARIATION
                      (ASSUMPTION: NORMALITY)

 1. STATISTICS
       NUMBER OF OBSERVATIONS    =     1000
       NUMBER OF GROUPS          =        4
       LEVENE F TEST STATISTIC   =    140.8509


    FOR LEVENE TEST STATISTIC
       0          % POINT    =   0.0000000E+00
       50         % POINT    =   0.7891988
       75         % POINT    =    1.371589
       90         % POINT    =    2.089303
       95         % POINT    =    2.613852
       99         % POINT    =    3.801369
       99.9       % POINT    =    5.463994


          100.0000       % Point:     140.8509
```

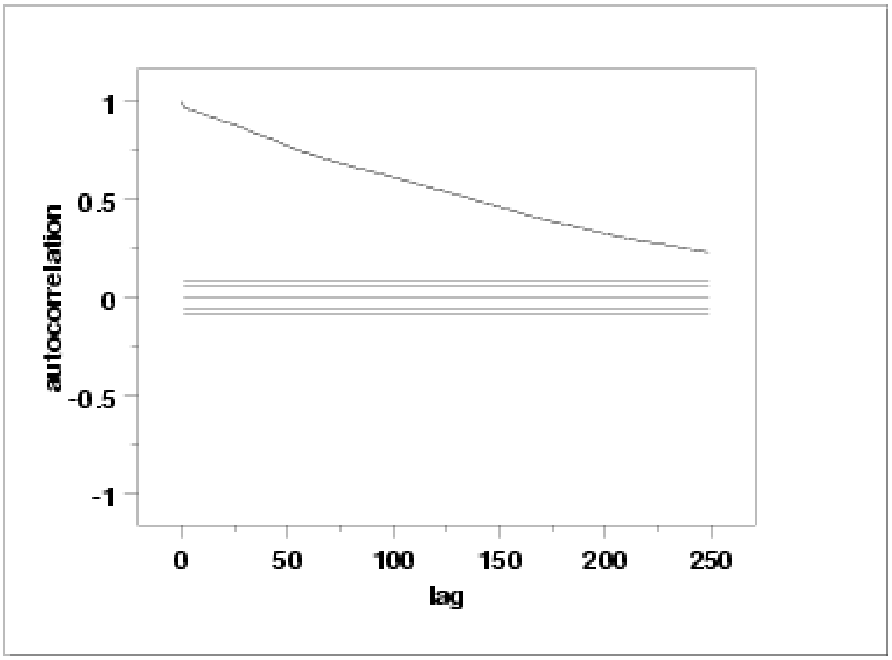
```
3. CONCLUSION (AT THE 5% LEVEL):
      THERE IS A SHIFT IN VARIATION.
      THUS: NOT HOMOGENEOUS WITH RESPECT TO VARIATION.
```

In this case, since the Levene test statistic value of 140.9 is greater than the 5% significance level critical value of 2.6, we conclude that there is significant evidence of nonconstant variation.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the 4-plot in the previous section is a simple graphical technique.

One check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of greatest interest, is 0.97. The critical values at the 5% significance level are -0.062 and 0.062. This indicates that the lag 1 autocorrelation is statistically significant, so there is strong evidence of non-randomness.

A common test for randomness is the runs test.

```
                   RUNS UP

          STATISTIC = NUMBER OF RUNS UP
              OF LENGTH EXACTLY I

I         STAT      EXP(STAT)    SD(STAT)        Z

1        178.0    208.3750     14.5453       -2.09
2         90.0     91.5500      7.5002       -0.21
3         29.0     26.3236      4.5727        0.59
4         16.0      5.7333      2.3164        4.43
5          2.0      1.0121      0.9987        0.99
6          0.0      0.1507      0.3877       -0.39
```

```
         7         0.0      0.0194      0.1394        -0.14
         8         0.0      0.0022      0.0470        -0.05
         9         0.0      0.0002      0.0150        -0.02
        10         0.0      0.0000      0.0046         0.00


                STATISTIC = NUMBER OF RUNS UP
                    OF LENGTH I OR MORE

         I          STAT     EXP(STAT)    SD(STAT)        Z

         1       315.0    333.1667      9.4195        -1.93
         2       137.0    124.7917      6.2892         1.94
         3        47.0     33.2417      4.8619         2.83
         4        18.0      6.9181      2.5200         4.40
         5         2.0      1.1847      1.0787         0.76
         6         0.0      0.1726      0.4148        -0.42
         7         0.0      0.0219      0.1479        -0.15
         8         0.0      0.0025      0.0496        -0.05
         9         0.0      0.0002      0.0158        -0.02
        10         0.0      0.0000      0.0048         0.00


                          RUNS DOWN

                STATISTIC = NUMBER OF RUNS DOWN
                    OF LENGTH EXACTLY I

         I          STAT     EXP(STAT)    SD(STAT)        Z

         1       195.0    208.3750     14.5453        -0.92
         2        81.0     91.5500      7.5002        -1.41
         3        32.0     26.3236      4.5727         1.24
         4         4.0      5.7333      2.3164        -0.75
         5         1.0      1.0121      0.9987        -0.01
         6         1.0      0.1507      0.3877         2.19
         7         0.0      0.0194      0.1394        -0.14
         8         0.0      0.0022      0.0470        -0.05
         9         0.0      0.0002      0.0150        -0.02
        10         0.0      0.0000      0.0046         0.00


                STATISTIC = NUMBER OF RUNS DOWN
                    OF LENGTH I OR MORE


         I          STAT     EXP(STAT)    SD(STAT)        Z

         1       314.0    333.1667      9.4195        -2.03
         2       119.0    124.7917      6.2892        -0.92
         3        38.0     33.2417      4.8619         0.98
         4         6.0      6.9181      2.5200        -0.36
         5         2.0      1.1847      1.0787         0.76
         6         1.0      0.1726      0.4148         1.99
         7         0.0      0.0219      0.1479        -0.15
         8         0.0      0.0025      0.0496        -0.05
         9         0.0      0.0002      0.0158        -0.02
        10         0.0      0.0000      0.0048         0.00


                RUNS TOTAL = RUNS UP + RUNS DOWN

              STATISTIC = NUMBER OF RUNS TOTAL
                   OF LENGTH EXACTLY I
```

```
 I         STAT     EXP(STAT)    SD(STAT)       Z

 1       373.0    416.7500     20.5701       -2.13
 2       171.0    183.1000     10.6068       -1.14
 3        61.0     52.6472      6.4668        1.29
 4        20.0     11.4667      3.2759        2.60
 5         3.0      2.0243      1.4123        0.69
 6         1.0      0.3014      0.5483        1.27
 7         0.0      0.0389      0.1971       -0.20
 8         0.0      0.0044      0.0665       -0.07
 9         0.0      0.0005      0.0212       -0.02
10         0.0      0.0000      0.0065       -0.01


       STATISTIC = NUMBER OF RUNS TOTAL
             OF LENGTH I OR MORE

 I         STAT     EXP(STAT)    SD(STAT)       Z

 1       629.0    666.3333     13.3212       -2.80
 2       256.0    249.5833      8.8942        0.72
 3        85.0     66.4833      6.8758        2.69
 4        24.0     13.8361      3.5639        2.85
 5         4.0      2.3694      1.5256        1.07
 6         1.0      0.3452      0.5866        1.12
 7         0.0      0.0438      0.2092       -0.21
 8         0.0      0.0049      0.0701       -0.07
 9         0.0      0.0005      0.0223       -0.02
10         0.0      0.0000      0.0067       -0.01


        LENGTH OF THE LONGEST RUN UP         =     5
        LENGTH OF THE LONGEST RUN DOWN       =     6
        LENGTH OF THE LONGEST RUN UP OR DOWN =     6

        NUMBER OF POSITIVE DIFFERENCES =   505
        NUMBER OF NEGATIVE DIFFERENCES =   469
        NUMBER OF ZERO     DIFFERENCES =    25
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. Due to the number of values that are larger than the 1.96 cut-off, we conclude that the data are not random. However, in this case the evidence from the runs test is not nearly as strong as it is from the autocorrelation plot.

Distributional Analysis

Since we rejected the randomness assumption, the distributional tests are not meaningful. Therefore, these quantitative tests are omitted. Since the Grubbs' test for outliers also assumes the approximate normality of the data, we omit Grubbs' test as well.

Univariate Report

It is sometimes useful and convenient to summarize the above results in a report.

```
Analysis for resistor case study

1: Sample Size                              = 1000

2: Location
   Mean                                     = 28.01635
   Standard Deviation of Mean               = 0.002008
   95% Confidence Interval for Mean         = (28.0124,28.02029)
   Drift with respect to location?          = NO

3: Variation
   Standard Deviation                       = 0.063495
   95% Confidence Interval for SD           = (0.060829,0.066407)
   Change in variation?
   (based on Levene's test on quarters
   of the data)                             = YES

4: Randomness
   Autocorrelation                          = 0.972158
   Data Are Random?
     (as measured by autocorrelation)       = NO

5: Distribution
   Distributional test omitted due to
   non-randomness of the data

6: Statistical Control
   (i.e., no drift in location or scale,
   data are random, distribution is
   fixed)
   Data Set is in Statistical Control?      = NO

7: Outliers?
   (Grubbs' test omitted due to
   non-randomness of the data
```

1.4.2.7.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step. *NOTE: This case study has 1,000 points. For better performance, it is highly recommended that you check the "No Update" box on the Spreadsheet window for this case study. This will suppress subsequent updating of the Spreadsheet window as the data are created or modified.*	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*

`1. Invoke Dataplot and read data.`
 	
`1. Read in the data.`
	`1. You have read 1 column of numbers into Dataplot, variable Y.`

`2. 4-plot of the data.`
 	
`1. 4-plot of Y.`
	`1. Based on the 4-plot, there are shifts in location and variation and the data are not random.`

`3. Generate the individual plots.`
 	
`1. Generate a run sequence plot.`
 	
`2. Generate a lag plot.`
	`1. The run sequence plot indicates that there are shifts of location and variation.`
	
	`2. The lag plot shows a strong linear pattern, which indicates significant non-randomness.`

4. Generate summary statistics, quantitative analysis, and print a univariate report.	
1. Generate a table of summary statistics.	1. The summary statistics table displays 25+ statistics.
2. Generate the sample mean, a confidence interval for the population mean, and compute a linear fit to detect drift in location.	2. The mean is 28.0163 and a 95% confidence interval is (28.0124,28.02029). The linear fit indicates drift in location since the slope parameter estimate is statistically significant.
3. Generate the sample standard deviation, a confidence interval for the population standard deviation, and detect drift in variation by dividing the data into quarters and computing Levene's test for equal standard deviations.	3. The standard deviation is 0.0635 with a 95% confidence interval of (0.060829,0.066407). Levene's test indicates significant change in variation.
4. Check for randomness by generating an autocorrelation plot and a runs test.	4. The lag 1 autocorrelation is 0.97. From the autocorrelation plot, this is outside the 95% confidence interval bands, indicating significant non-randomness.
5. Print a univariate report (this assumes steps 2 thru 5 have already been run).	5. The results are summarized in a convenient report.

1.4.2.8. Heat Flow Meter 1

Heat Flow Meter Calibration and Stability

This example illustrates the univariate analysis of standard resistor data.

1.4.2.8.1. Background and Data

Generation

This data set was collected by Bob Zarr of NIST in January, 1990 from a heat flow meter calibration and stability analysis. The response variable is a calibration factor.

The motivation for studying this data set is to illustrate a well-behaved process where the underlying assumptions hold and the process is in statistical control.

This file can be read by Dataplot with the following commands:

```
SKIP 25
READ ZARR13.DAT Y
```

Resulting Data

The following are the data used for this case study.

```
9.206343
9.299992
9.277895
9.305795
9.275351
9.288729
9.287239
9.260973
9.303111
9.275674
9.272561
9.288454
9.255672
9.252141
9.297670
9.266534
9.256689
9.277542
9.248205
```

```
9.252107
9.276345
9.278694
9.267144
9.246132
9.238479
9.269058
9.248239
9.257439
9.268481
9.288454
9.258452
9.286130
9.251479
9.257405
9.268343
9.291302
9.219460
9.270386
9.218808
9.241185
9.269989
9.226585
9.258556
9.286184
9.320067
9.327973
9.262963
9.248181
9.238644
9.225073
9.220878
9.271318
9.252072
9.281186
9.270624
9.294771
9.301821
9.278849
9.236680
9.233988
9.244687
9.221601
9.207325
9.258776
9.275708
```

```
9.268955
9.257269
9.264979
9.295500
9.292883
9.264188
9.280731
9.267336
9.300566
9.253089
9.261376
9.238409
9.225073
9.235526
9.239510
9.264487
9.244242
9.277542
9.310506
9.261594
9.259791
9.253089
9.245735
9.284058
9.251122
9.275385
9.254619
9.279526
9.275065
9.261952
9.275351
9.252433
9.230263
9.255150
9.268780
9.290389
9.274161
9.255707
9.261663
9.250455
9.261952
9.264041
9.264509
9.242114
9.239674
9.221553
```

```
9.241935
9.215265
9.285930
9.271559
9.266046
9.285299
9.268989
9.267987
9.246166
9.231304
9.240768
9.260506
9.274355
9.292376
9.271170
9.267018
9.308838
9.264153
9.278822
9.255244
9.229221
9.253158
9.256292
9.262602
9.219793
9.258452
9.267987
9.267987
9.248903
9.235153
9.242933
9.253453
9.262671
9.242536
9.260803
9.259825
9.253123
9.240803
9.238712
9.263676
9.243002
9.246826
9.252107
9.261663
9.247311
9.306055
```

```
9.237646
9.248937
9.256689
9.265777
9.299047
9.244814
9.287205
9.300566
9.256621
9.271318
9.275154
9.281834
9.253158
9.269024
9.282077
9.277507
9.284910
9.239840
9.268344
9.247778
9.225039
9.230750
9.270024
9.265095
9.284308
9.280697
9.263032
9.291851
9.252072
9.244031
9.283269
9.196848
9.231372
9.232963
9.234956
9.216746
9.274107
9.273776
```

1.4.2.8.2. Graphical Output and Interpretation

Goal

The goal of this analysis is threefold:

1. Determine if the univariate model:

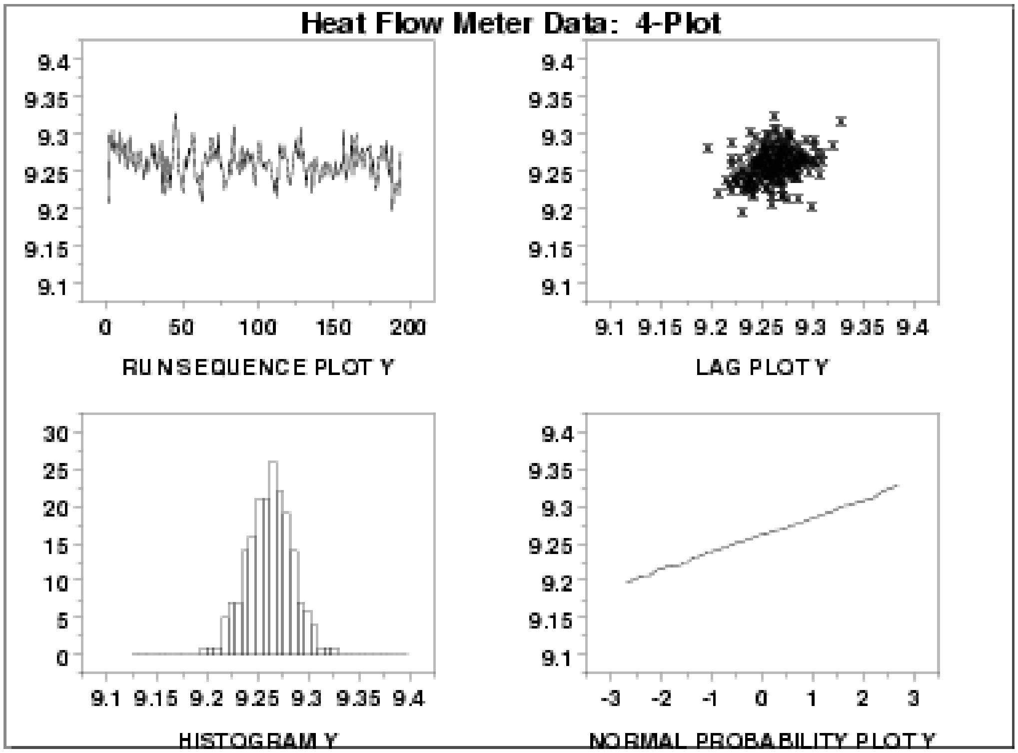
   $$Y_i = C + E_i$$

   is appropriate and valid.
2. Determine if the typical underlying assumptions for an "in control" measurement process are valid. These assumptions are:
   1. random drawings;
   2. from a fixed distribution;
   3. with the distribution having a fixed location; and
   4. the distribution having a fixed scale.
3. Determine if the confidence interval

   $$\bar{Y} \pm 2s/\sqrt{N}$$

   is appropriate and valid where ***s*** is the standard deviation of the original data.

4-Plot of Data



Interpretation

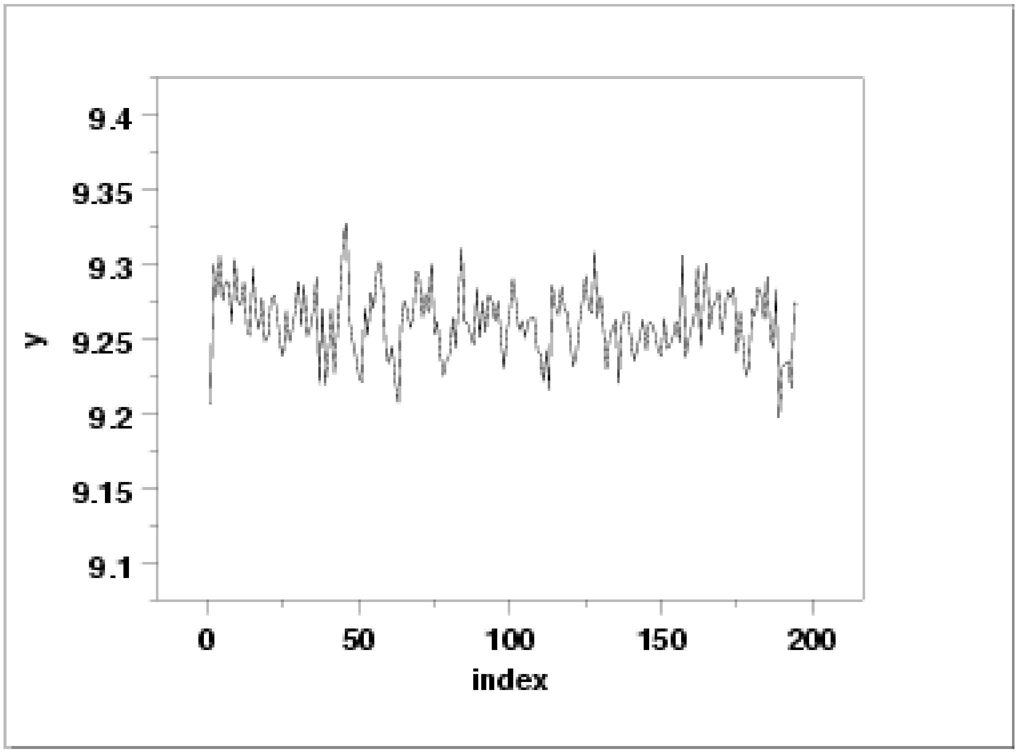
The assumptions are addressed by the graphics shown above:

1. The run sequence plot (upper left) indicates that the data do not have any significant shifts in location or scale over time.
2. The lag plot (upper right) does not indicate any non-random pattern in the data.
3. The histogram (lower left) shows that the data are reasonably symmetric, there does not appear to be significant outliers in the tails, and it seems reasonable to assume that the data are from approximately a normal distribution.
4. The normal probability plot (lower right) verifies that an assumption of normality is in fact reasonable.

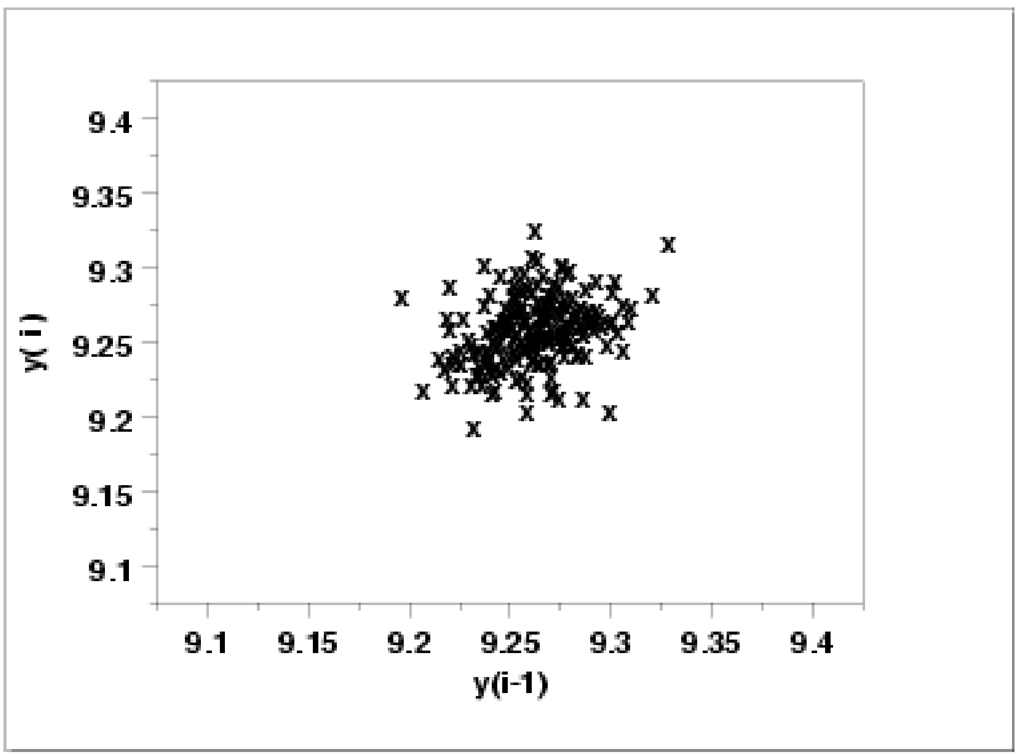
Individual Plots

Although it is generally unnecessary, the plots can be generated individually to give more detail.

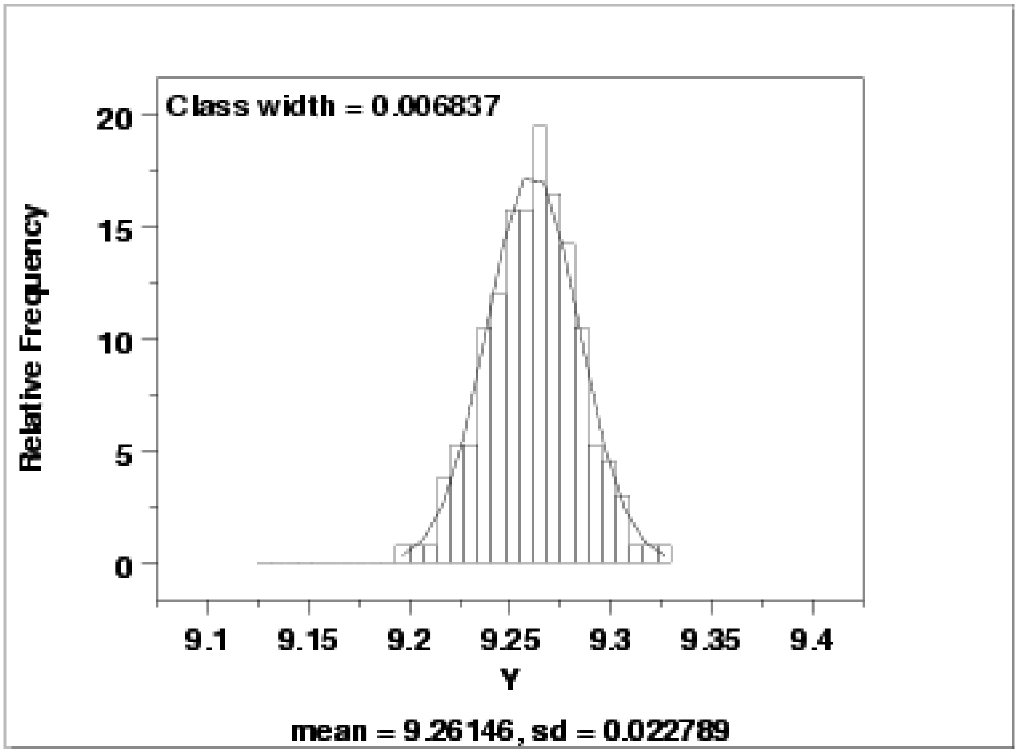
Run Sequence Plot



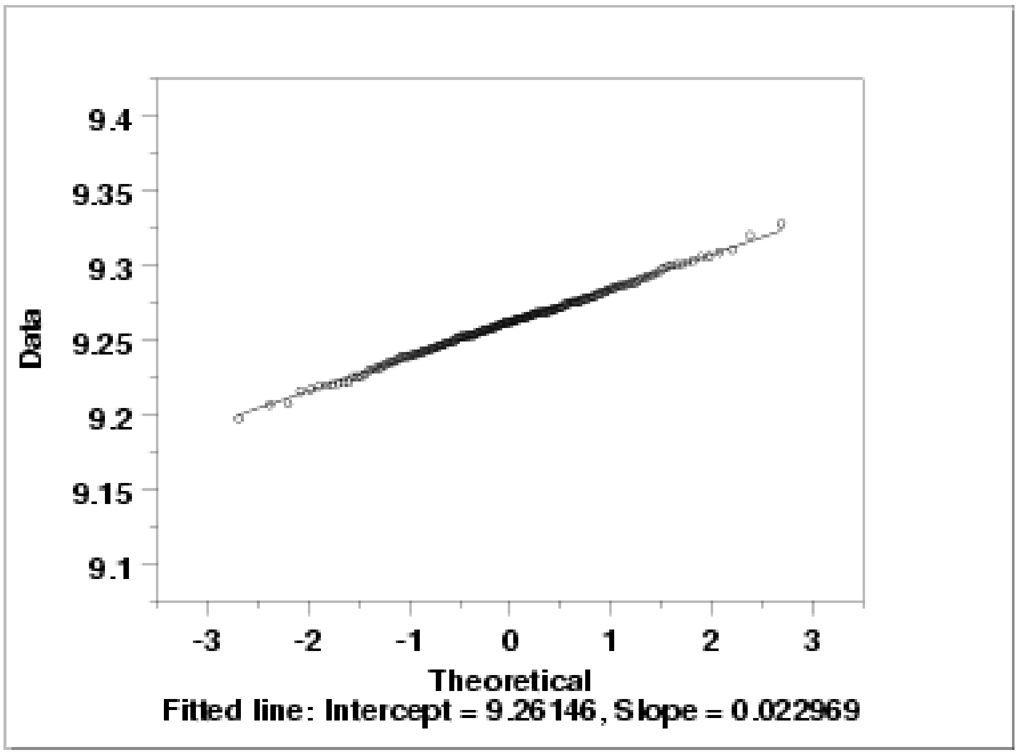
Lag Plot



Histogram (with overlaid Normal PDF)



Normal Probability Plot



1.4.2.8.3. Quantitative Output and Interpretation

Summary Statistics

As a first step in the analysis, a table of summary statistics is computed from the data. The following table, generated by Dataplot, shows a typical set of statistics.

```
                        SUMMARY

                NUMBER OF OBSERVATIONS =      195


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.9262411E+01  *  RANGE        =   0.1311255E+00  *
*  MEAN         =   0.9261460E+01  *  STAND. DEV.  =   0.2278881E-01  *
*  MIDMEAN      =   0.9259412E+01  *  AV. AB. DEV. =   0.1788945E-01  *
*  MEDIAN       =   0.9261952E+01  *  MINIMUM      =   0.9196848E+01  *
*               =                  *  LOWER QUART. =   0.9246826E+01  *
*               =                  *  LOWER HINGE  =   0.9246496E+01  *
*               =                  *  UPPER HINGE  =   0.9275530E+01  *
*               =                  *  UPPER QUART. =   0.9275708E+01  *
*               =                  *  MAXIMUM      =   0.9327973E+01  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =   0.2805789E+00  *  ST. 3RD MOM. =  -0.8537455E-02  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.3049067E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =   0.9458605E+01  *
*               =                  *  UNIFORM PPCC =   0.9735289E+00  *
*               =                  *  NORMAL  PPCC =   0.9989640E+00  *
*               =                  *  TUK -.5 PPCC =   0.8927904E+00  *
*               =                  *  CAUCHY  PPCC =   0.6360204E+00  *
***********************************************************************
```

Location

One way to quantify a change in location over time is to fit a straight line to the data set using the index variable X = 1, 2, ..., N, with N denoting the number of observations. If there is no significant drift in the location, the slope parameter should be zero. For this data set, Dataplot generates the following output:

```
LEAST SQUARES MULTILINEAR FIT
      SAMPLE SIZE N          =       195
      NUMBER OF VARIABLES =            1
      NO REPLICATION CASE


              PARAMETER ESTIMATES            (APPROX. ST. DEV.)    T VALUE
       1  A0                       9.26699       (0.3253E-02)        2849.
       2  A1       X          -0.564115E-04    (0.2878E-04)        -1.960

      RESIDUAL    STANDARD DEVIATION =        0.2262372E-01
      RESIDUAL    DEGREES OF FREEDOM =         193
```

The slope parameter, A1, has a *t* value of -1.96 which is (barely) statistically significant since it is essentially equal to the 95% level cutoff of -1.96. However, notice that the value of the slope parameter estimate is -0.00056. This slope, even though statistically significant, can essentially be considered zero.

Variation

One simple way to detect a change in variation is with a Bartlett test after dividing the data set into several equal-sized intervals. The choice of the number of intervals is somewhat arbitrary, although values of 4 or 8 are reasonable. Dataplot generated the following output for the Bartlett test.

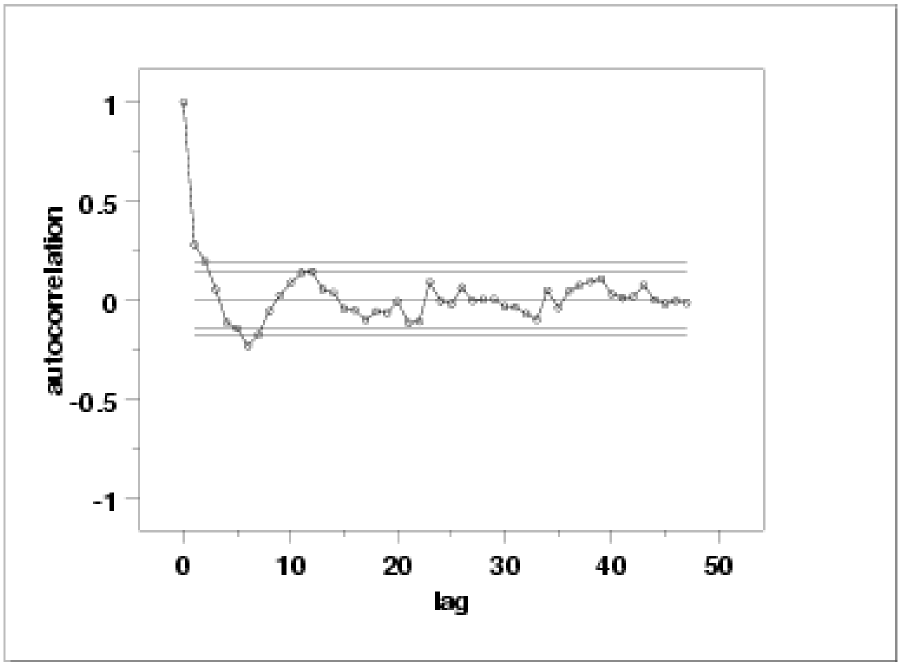
```
              BARTLETT TEST
          (STANDARD DEFINITION)
NULL HYPOTHESIS UNDER TEST--ALL SIGMA(I) ARE EQUAL

TEST:
   DEGREES OF FREEDOM          =    3.000000

   TEST STATISTIC VALUE        =    3.147338
   CUTOFF: 95% PERCENT POINT   =    7.814727
   CUTOFF: 99% PERCENT POINT   =    11.34487

   CHI-SQUARE CDF VALUE        =    0.630538

  NULL          NULL HYPOTHESIS       NULL HYPOTHESIS
  HYPOTHESIS    ACCEPTANCE INTERVAL   CONCLUSION
ALL SIGMA EQUAL    (0.000,0.950)        ACCEPT
```

In this case, since the Bartlett test statistic of 3.14 is less than the critical value at the 5% significance level of 7.81, we conclude that the standard deviations are not significantly different in the 4 intervals. That is, the assumption of constant scale is valid.

Randomness

There are many ways in which data can be non-random. However, most common forms of non-randomness can be detected with a few simple tests. The lag plot in the previous section is a simple graphical technique.

Another check is an autocorrelation plot that shows the autocorrelations for various lags. Confidence bands can be plotted at the 95% and 99% confidence levels. Points outside this band indicate statistically significant values (lag 0 is always 1). Dataplot generated the following autocorrelation plot.



The lag 1 autocorrelation, which is generally the one of greatest interest, is 0.281. The critical values at the 5% significance level are -0.087 and 0.087. This indicates that the lag 1 autocorrelation is statistically significant, so there is evidence of non-randomness.

A common test for randomness is the runs test.

```
                         RUNS UP

          STATISTIC = NUMBER OF RUNS UP
              OF LENGTH EXACTLY I

   I         STAT      EXP(STAT)    SD(STAT)       Z

   1         35.0      40.6667      6.4079       -0.88
   2          8.0      17.7583      3.3021       -2.96
   3         12.0       5.0806      2.0096        3.44
   4          3.0       1.1014      1.0154        1.87
   5          0.0       0.1936      0.4367       -0.44
   6          0.0       0.0287      0.1692       -0.17
   7          0.0       0.0037      0.0607       -0.06
   8          0.0       0.0004      0.0204       -0.02
   9          0.0       0.0000      0.0065       -0.01
  10          0.0       0.0000      0.0020        0.00


          STATISTIC = NUMBER OF RUNS UP
              OF LENGTH I OR MORE
```

I	STAT	EXP(STAT)	SD(STAT)	Z
1	58.0	64.8333	4.1439	-1.65
2	23.0	24.1667	2.7729	-0.42
3	15.0	6.4083	2.1363	4.02
4	3.0	1.3278	1.1043	1.51
5	0.0	0.2264	0.4716	-0.48
6	0.0	0.0328	0.1809	-0.18
7	0.0	0.0041	0.0644	-0.06
8	0.0	0.0005	0.0215	-0.02
9	0.0	0.0000	0.0068	-0.01
10	0.0	0.0000	0.0021	0.00

RUNS DOWN

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	33.0	40.6667	6.4079	-1.20
2	18.0	17.7583	3.3021	0.07
3	3.0	5.0806	2.0096	-1.04
4	3.0	1.1014	1.0154	1.87
5	1.0	0.1936	0.4367	1.85
6	0.0	0.0287	0.1692	-0.17
7	0.0	0.0037	0.0607	-0.06
8	0.0	0.0004	0.0204	-0.02
9	0.0	0.0000	0.0065	-0.01
10	0.0	0.0000	0.0020	0.00

STATISTIC = NUMBER OF RUNS DOWN
OF LENGTH I OR MORE

I	STAT	EXP(STAT)	SD(STAT)	Z
1	58.0	64.8333	4.1439	-1.65
2	25.0	24.1667	2.7729	0.30
3	7.0	6.4083	2.1363	0.28
4	4.0	1.3278	1.1043	2.42
5	1.0	0.2264	0.4716	1.64
6	0.0	0.0328	0.1809	-0.18
7	0.0	0.0041	0.0644	-0.06
8	0.0	0.0005	0.0215	-0.02
9	0.0	0.0000	0.0068	-0.01
10	0.0	0.0000	0.0021	0.00

RUNS TOTAL = RUNS UP + RUNS DOWN

STATISTIC = NUMBER OF RUNS TOTAL
OF LENGTH EXACTLY I

I	STAT	EXP(STAT)	SD(STAT)	Z
1	68.0	81.3333	9.0621	-1.47
2	26.0	35.5167	4.6698	-2.04
3	15.0	10.1611	2.8420	1.70
4	6.0	2.2028	1.4360	2.64
5	1.0	0.3871	0.6176	0.99

```
     6          0.0       0.0574      0.2392        -0.24
     7          0.0       0.0074      0.0858        -0.09
     8          0.0       0.0008      0.0289        -0.03
     9          0.0       0.0001      0.0092        -0.01
    10          0.0       0.0000      0.0028         0.00


               STATISTIC = NUMBER OF RUNS TOTAL
                     OF LENGTH I OR MORE

     I          STAT      EXP(STAT)    SD(STAT)        Z

     1        116.0     129.6667      5.8604        -2.33
     2         48.0      48.3333      3.9215        -0.09
     3         22.0      12.8167      3.0213         3.04
     4          7.0       2.6556      1.5617         2.78
     5          1.0       0.4528      0.6669         0.82
     6          0.0       0.0657      0.2559        -0.26
     7          0.0       0.0083      0.0911        -0.09
     8          0.0       0.0009      0.0305        -0.03
     9          0.0       0.0001      0.0097        -0.01
    10          0.0       0.0000      0.0029         0.00


                LENGTH OF THE LONGEST RUN UP         =     4
                LENGTH OF THE LONGEST RUN DOWN       =     5
                LENGTH OF THE LONGEST RUN UP OR DOWN =     5

                NUMBER OF POSITIVE DIFFERENCES =    98
                NUMBER OF NEGATIVE DIFFERENCES =    95
                NUMBER OF ZERO     DIFFERENCES =     1
```

Values in the column labeled "Z" greater than 1.96 or less than -1.96 are statistically significant at the 5% level. The runs test does indicate some non-randomness.

Although the autocorrelation plot and the runs test indicate some mild non-randomness, the violation of the randomness assumption is not serious enough to warrant developing a more sophisticated model. It is common in practice that some of the assumptions are mildly violated and it is a judgement call as to whether or not the violations are serious enough to warrant developing a more sophisticated model for the data.

Distributional Analysis

Probability plots are a graphical test for assessing if a particular distribution provides an adequate fit to a data set.

A quantitative enhancement to the probability plot is the correlation coefficient of the points on the probability plot. For this data set the correlation coefficient is 0.996. Since this is greater than the critical value of 0.987 (this is a tabulated value), the normality assumption is not rejected.

Chi-square and Kolmogorov-Smirnov goodness-of-fit tests are alternative methods for assessing distributional adequacy. The Wilk-Shapiro and Anderson-Darling tests can be used to test for normality. Dataplot generates the following output for the Anderson-Darling normality test.

```
                     ANDERSON-DARLING 1-SAMPLE TEST
                     THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

  1. STATISTICS:
```

```
        NUMBER OF OBSERVATIONS                =      195
        MEAN                                  =    9.261460
        STANDARD DEVIATION                    =   0.2278881E-01

        ANDERSON-DARLING TEST STATISTIC VALUE =   0.1264954
        ADJUSTED TEST STATISTIC VALUE         =   0.1290070

  2. CRITICAL VALUES:
        90         % POINT    =   0.6560000
        95         % POINT    =   0.7870000
        97.5       % POINT    =   0.9180000
        99         % POINT    =    1.092000

  3. CONCLUSION (AT THE 5% LEVEL):
        THE DATA DO COME FROM A NORMAL DISTRIBUTION.
```

The Anderson-Darling test also does not reject the normality assumption because the test statistic, 0.129, is less than the critical value at the 5% significance level of 0.918.

Outlier Analysis

A test for outliers is the Grubbs' test. Dataplot generated the following output for Grubbs' test.

```
                GRUBBS TEST FOR OUTLIERS
                (ASSUMPTION: NORMALITY)

  1. STATISTICS:
        NUMBER OF OBSERVATIONS      =      195
        MINIMUM                     =    9.196848
        MEAN                        =    9.261460
        MAXIMUM                     =    9.327973
     STANDARD DEVIATION          =   0.2278881E-01

     GRUBBS TEST STATISTIC       =    2.918673

  2. PERCENT POINTS OF THE REFERENCE DISTRIBUTION
     FOR GRUBBS TEST STATISTIC
        0          % POINT    =   0.0000000E+00
        50         % POINT    =    2.984294
        75         % POINT    =    3.181226
        90         % POINT    =    3.424672
        95         % POINT    =    3.597898
        99         % POINT    =    3.970215

  3. CONCLUSION (AT THE 5% LEVEL):
        THERE ARE NO OUTLIERS.
```

For this data set, Grubbs' test does not detect any outliers at the 25%, 10%, 5%, and 1% significance levels.

Model

Since the underlying assumptions were validated both graphically and analytically, with a mild violation of the randomness assumption, we conclude that a reasonable model for the data is:

$$Y_i = 9.26146 + E_i$$

We can express the uncertainty for ***C***, here estimated by 9.26146, as the 95% confidence interval (9.258242,9.26479).

Univariate Report	It is sometimes useful and convenient to summarize the above results in a report. The report for the heat flow meter data follows.

```
Analysis for heat flow meter data

1: Sample Size                              = 195

2: Location
   Mean                                     = 9.26146
   Standard Deviation of Mean               = 0.001632
   95% Confidence Interval for Mean         = (9.258242,9.264679)
   Drift with respect to location?          = NO

3: Variation
   Standard Deviation                       = 0.022789
   95% Confidence Interval for SD           = (0.02073,0.025307)
   Drift with respect to variation?
   (based on Bartlett's test on quarters
   of the data)                             = NO

4: Randomness
   Autocorrelation                          = 0.280579
   Data are Random?
     (as measured by autocorrelation)       = NO

5: Distribution
   Normal PPCC                              = 0.998965
   Data are Normal?
     (as measured by Normal PPCC)           = YES

6: Statistical Control
   (i.e., no drift in location or scale,
   data are random, distribution is
   fixed, here we are testing only for
   fixed normal)
   Data Set is in Statistical Control?      = YES

7: Outliers?
   (as determined by Grubbs' test)           = NO
```

1.4.2.8.4. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. Based on the 4-plot, there are no shifts in location or scale, and the data seem to follow a normal distribution.
3. Generate the individual plots. 1. Generate a run sequence plot. 2. Generate a lag plot. 3. Generate a histogram with an overlaid normal pdf.	1. The run sequence plot indicates that there are no shifts of location or scale. 2. The lag plot does not indicate any significant patterns (which would show the data were not random). 3. The histogram indicates that a normal distribution is a good distribution for these data.

4. Generate a normal probability plot.	4. The normal probability plot verifies that the normal distribution is a reasonable distribution for these data.
4. Generate summary statistics, quantitative analysis, and print a univariate report. 1. Generate a table of summary statistics. 2. Generate the mean, a confidence interval for the mean, and compute a linear fit to detect drift in location. 3. Generate the standard deviation, a confidence interval for the standard deviation, and detect drift in variation by dividing the data into quarters and computing Bartlett's test for equal standard deviations. 4. Check for randomness by generating an autocorrelation plot and a runs test. 5. Check for normality by computing the normal probability plot correlation coefficient. 6. Check for outliers using Grubbs' test. 7. Print a univariate report (this assumes steps 2 thru 6 have already been run).	1. The summary statistics table displays 25+ statistics. 2. The mean is 9.261 and a 95% confidence interval is (9.258,9.265). The linear fit indicates no drift in location since the slope parameter estimate is essentially zero. 3. The standard deviation is 0.023 with a 95% confidence interval of (0.0207,0.0253). Bartlett's test indicates no significant change in variation. 4. The lag 1 autocorrelation is 0.28. From the autocorrelation plot, this is statistically significant at the 95% level. 5. The normal probability plot correlation coefficient is 0.999. At the 5% level, we cannot reject the normality assumption. 6. Grubbs' test detects no outliers at the 5% level. 7. The results are summarized in a convenient report.

1.4.2.9. Airplane Glass Failure Time

Airplane Glass Failure Time

This example illustrates the univariate analysis of airplane glass failure time data.

1.4.2.9.1. Background and Data

Generation

This data set was collected by Ed Fuller of NIST in December, 1993. The response variable is time to failure for airplane glass under test.

Purpose of Analysis

The goal of this case study is to find a good distributional model for the data. Once a good distributional model has been determined, various percent points for glass failure will be computed.

Since the data are failure times, this case study is a form of reliability analysis. The assessing product reliability chapter contains a more complete discussion of reliabilty methods. This case study is meant to complement that chapter by showing the use of graphical techniques in one aspect of reliability modeling.

Failure times are basically extreme values that do not follow a normal distribution; non-parametric methods (techniques that do not rely on a specific distribution) are frequently recommended for developing confidence intervals for failure data. One problem with this approach is that sample sizes are often small due to the expense involved in collecting the data, and non-parametric methods do not work well for small sample sizes. For this reason, a parametric method based on a specific distributional model of the data is preferred if the data can be shown to follow a specific distribution. Parametric models typically have greater efficiency at the cost of more specific assumptions about the data, but, it is important to verify that the distributional assumption is indeed valid. If the distributional assumption is not justified, then the conclusions drawn from the model may not be valid.

This file can be read by Dataplot with the following commands:

```
SKIP 25
READ FULLER2.DAT Y
```

Resulting Data

The following are the data used for this case study.

```
18.830
20.800
21.657
23.030
23.230
24.050
24.321
25.500
25.520
25.800
26.690
26.770
26.780
27.050
27.670
29.900
31.110
33.200
33.730
33.760
33.890
34.760
35.750
35.910
36.980
37.080
37.090
39.580
44.045
45.290
45.381
```

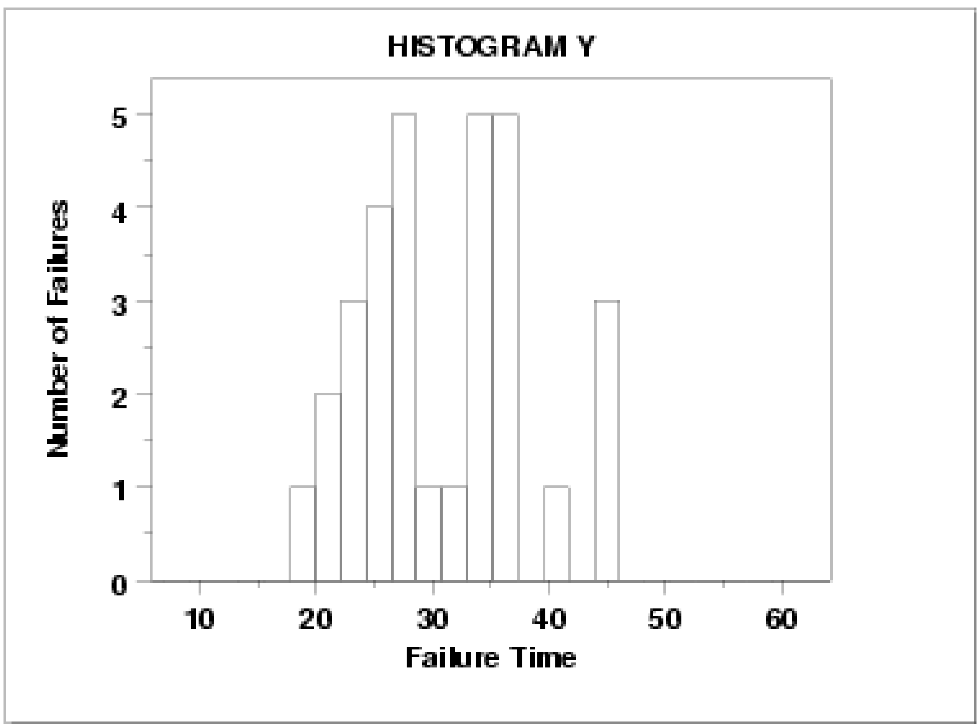
1.4.2.9.2. Graphical Output and Interpretation

Goal

The goal of this analysis is to determine a good distributional model for these failure time data. A secondary goal is to provide estimates for various percent points of the data. Percent points provide an answer to questions of the type "At what time do we expect 5% of the airplane glass to have failed?".

Initial Plots of the Data

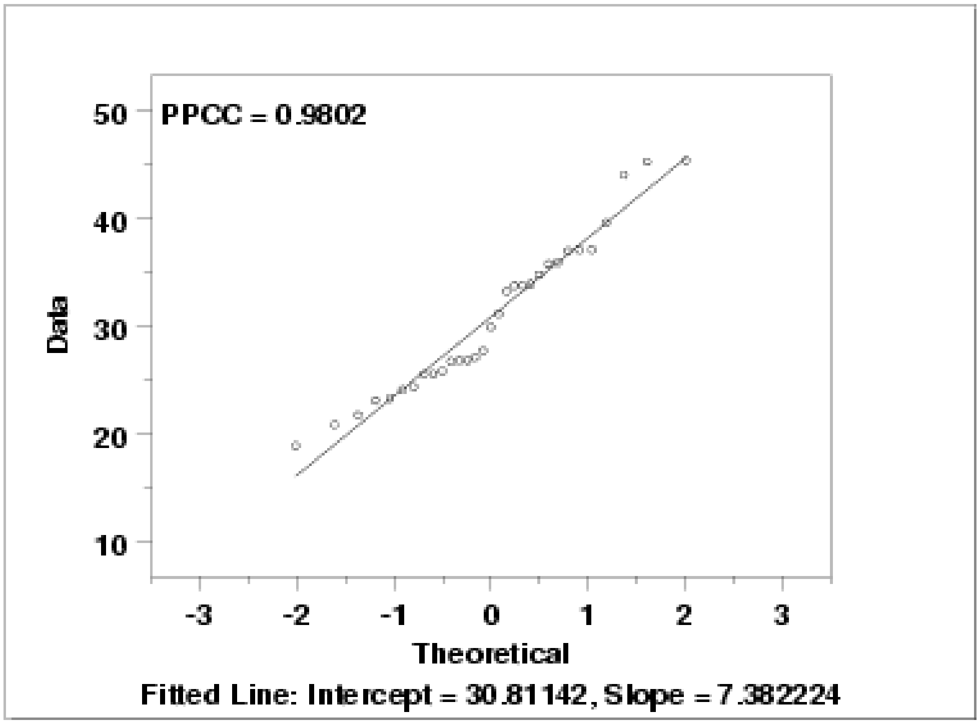
The first step is to generate a histogram to get an overall feel for the data.



The histogram shows the following:

- The failure times range between slightly greater than 15 to slightly less than 50.
- There are modes at approximately 28 and 38 with a gap in-between.
- The data are somewhat symmetric, but with a gap in the middle.

We next generate a normal probability plot.



The normal probability plot has a correlation coefficient of 0.980. We can use this number as a reference baseline when comparing the performance of other distributional fits.

Other Potential Distributions

There is a large number of distributions that would be distributional model candidates for the data. However, we will restrict ourselves to consideration of the following distributional models because these have proven to be useful in reliability studies.

1. Normal distribution
2. Exponential distribution
3. Weibull distribution
4. Lognormal distribution
5. Gamma distribution
6. Power normal distribution
7. Power lognormal distribution

Approach

There are two basic questions that need to be addressed.

1. Does a given distributional model provide an adequate fit to the data?
2. Of the candidate distributional models, is there one distribution that fits the data better than the other candidate distributional models?

The use of probability plots and probability plot correlation coefficient (PPCC) plots provide answers to both of these questions.

If the distribution does not have a shape parameter, we simply generate a probability plot.

1. If we fit a straight line to the points on the probability plot, the intercept and slope of that line provide estimates of the location and scale parameters, respectively.
2. Our critierion for the "best fit" distribution is the one with the most linear probability plot. The correlation coefficient of the fitted line of the points on the probability plot, referred to as the PPCC value, provides a measure of the linearity of the probability plot, and thus a measure of how well the distribution fits the data. The PPCC values for multiple distributions can be compared to address the second question above.

If the distribution does have a shape parameter, then we are actually addressing a family of distributions rather than a single distribution. We first need to find the optimal value of the shape parameter. The PPCC plot can be used to determine the optimal parameter. We will use the PPCC plots in two stages. The first stage will be over a broad range of parameter values while the second stage will be in the neighborhood of the largest values. Although we could go further than two stages, for practical purposes two stages is sufficient. After determining an optimal value for the shape parameter, we use the probability plot as above to obtain estimates of the location and scale parameters and to determine the PPCC value. This PPCC value can be compared to the PPCC values obtained from other distributional models.

Analyses for Specific Distributions

We analyzed the data using the approach described above for the following distributional models:

1. Normal distribution - from the 4-plot above, the PPCC value was 0.980.
2. Exponential distribution - the exponential distribution is a special case of the Weibull with shape parameter equal to 1. If the Weibull analysis yields a shape parameter close to 1, then we would consider using the simpler exponential model.
3. Weibull distribution
4. Lognormal distribution
5. Gamma distribution
6. Power normal distribution
7. Power lognormal distribution

Summary of Results

The results are summarized below.

Normal Distribution

Max PPCC = 0.980
Estimate of location = 30.81
Estimate of scale = 7.38

Weibull Distribution

Max PPCC = 0.988
Estimate of shape = 2.13
Estimate of location = 15.9
Estimate of scale = 16.92

Lognormal Distribution

Max PPCC = 0.986
Estimate of shape = 0.18
Estimate of location = -9.96
Estimate of scale = 40.17

Gamma Distribution

Max PPCC = 0.987
Estimate of shape = 11.8
Estimate of location = 5.19
Estimate of scale = 2.17

Power Normal Distribution

Max PPCC = 0.987
Estimate of shape = 0.11
Estimate of location = 20.9
Estimate of scale = 3.3

Power Lognormal Distribution

Max PPCC = 0.988
Estimate of shape = 50
Estimate of location = 13.5
Estimate of scale = 150.8

These results indicate that several of these distributions provide an adequate distributional model for the data. We choose the 3-parameter Weibull distribution as the most appropriate model because it provides the best balance between simplicity and best fit.

Percent Point Estimates

The final step in this analysis is to compute percent point estimates for the 1%, 2.5%, 5%, 95%, 97.5%, and 99% percent points. A percent point estimate is an estimate of the time by which a given percentage of the units will have failed. For example, the 5% point is the time at which we estimate 5% of the units will have failed.

To calculate these values, we use the Weibull percent point function with the appropriate estimates of the shape, location, and scale parameters. The Weibull percent point function can be computed in many general purpose statistical software programs, including Dataplot.

Dataplot generated the following estimates for the percent points:

```
Estimated percent points using Weibull Distribution

PERCENT POINT          FAILURE TIME
0.01                   17.86
0.02                   18.92
0.05                   20.10
0.95                   44.21
0.97                   47.11
0.99                   50.53
```

Quantitative Measures of Goodness of Fit

Although it is generally unnecessary, we can include quantitative measures of distributional goodness-of-fit. Three of the commonly used measures are:

1. Chi-square goodness-of-fit.
2. Kolmogorov-Smirnov goodness-of-fit.
3. Anderson-Darling goodness-of-fit.

In this case, the sample size of 31 precludes the use of the chi-square test since the chi-square approximation is not valid for small sample sizes. Specifically, the smallest expected frequency should be at least 5. Although we could combine classes, we will instead use one of the other tests. The Kolmogorov-Smirnov test requires a fully specified distribution. Since we need to use the data to estimate the shape, location, and scale parameters, we do not use this test here. The Anderson-Darling test is a refinement of the Kolmogorov-Smirnov test. We run this test for the normal, lognormal, and Weibull distributions.

Normal Anderson-Darling Output

```
**************************************
**  Anderson-Darling normal test y  **
**************************************


        ANDERSON-DARLING 1-SAMPLE TEST
        THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

1. STATISTICS:
      NUMBER OF OBSERVATIONS                =       31
      MEAN                                  =    30.81142
      STANDARD DEVIATION                    =    7.253381

      ANDERSON-DARLING TEST STATISTIC VALUE =   0.5321903
      ADJUSTED TEST STATISTIC VALUE         =   0.5870153

2. CRITICAL VALUES:
      90          % POINT    =   0.6160000
      95          % POINT    =   0.7350000
      97.5        % POINT    =   0.8610000
      99          % POINT    =    1.021000

3. CONCLUSION (AT THE 5% LEVEL):
      THE DATA DO COME FROM A NORMAL DISTRIBUTION.
```

Lognormal Anderson-Darling Output

```
*****************************************
**  Anderson-Darling lognormal test y  **
*****************************************


        ANDERSON-DARLING 1-SAMPLE TEST
        THAT THE DATA CAME FROM A LOGNORMAL DISTRIBUTION

1. STATISTICS:
      NUMBER OF OBSERVATIONS                =       31
      MEAN                                  =    3.401242
      STANDARD DEVIATION                    =   0.2349026

      ANDERSON-DARLING TEST STATISTIC VALUE =   0.3888340
      ADJUSTED TEST STATISTIC VALUE         =   0.4288908

2. CRITICAL VALUES:
      90          % POINT    =   0.6160000
      95          % POINT    =   0.7350000
      97.5        % POINT    =   0.8610000
      99          % POINT    =    1.021000

3. CONCLUSION (AT THE 5% LEVEL):
      THE DATA DO COME FROM A LOGNORMAL DISTRIBUTION.
```

Weibull Anderson-Darling Output

```
**************************************
**  Anderson-Darling Weibull test y  **
**************************************


               ANDERSON-DARLING 1-SAMPLE TEST
               THAT THE DATA CAME FROM A WEIBULL DISTRIBUTION

 1. STATISTICS:
       NUMBER OF OBSERVATIONS                =       31
       MEAN                                  =    30.81142
       STANDARD DEVIATION                    =    7.253381
       SHAPE PARAMETER                       =    4.635379
       SCALE PARAMETER                       =    33.67423

       ANDERSON-DARLING TEST STATISTIC VALUE =   0.5973396
       ADJUSTED TEST STATISTIC VALUE         =   0.6187967

 2. CRITICAL VALUES:
       90          % POINT    =   0.6370000
       95          % POINT    =   0.7570000
       97.5        % POINT    =   0.8770000
       99          % POINT    =    1.038000

 3. CONCLUSION (AT THE 5% LEVEL):
       THE DATA DO COME FROM A WEIBULL DISTRIBUTION.
```

Note that for the Weibull distribution, the Anderson-Darling test is actually testing the 2-parameter Weibull distribution (based on maximum likelihood estimates), not the 3-parameter Weibull distribution. However, passing the 2-parameter Weibull distribution does give evidence that the Weibull is an appropriate distributional model even though we used a different parameter estimation method.

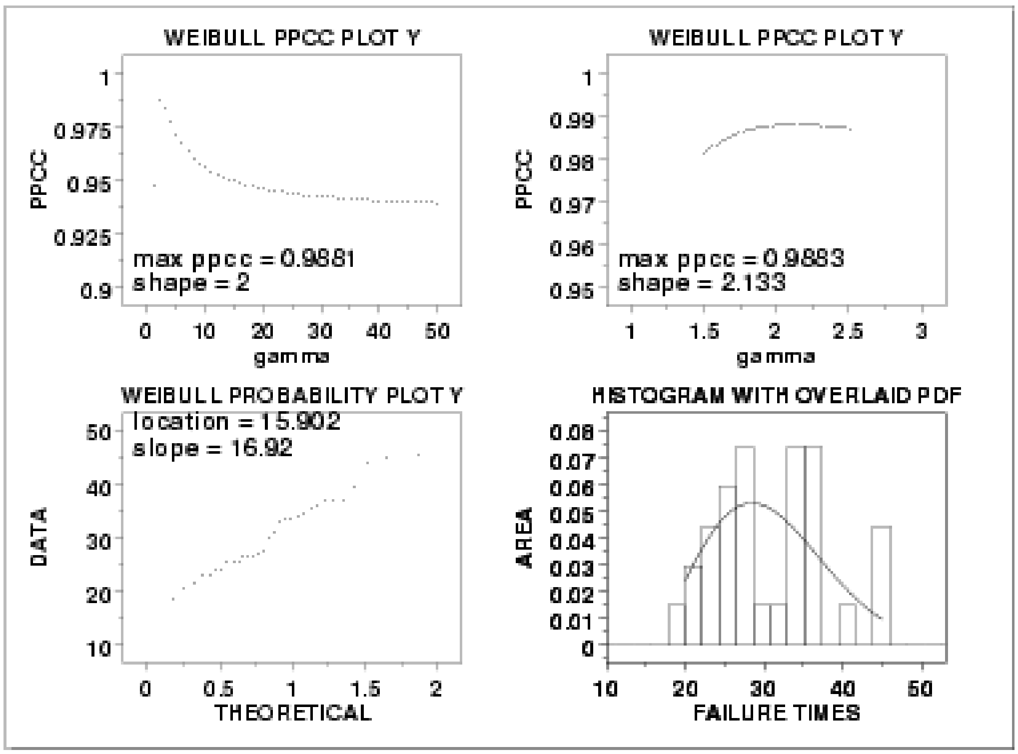
Conclusions

The Anderson-Darling test passes all three of these distributions.

1.4.2.9.3. Weibull Analysis

Plots for Weibull Distribution

The following plots were generated for a Weibull distribution.



Conclusions

We can make the following conclusions from these plots.

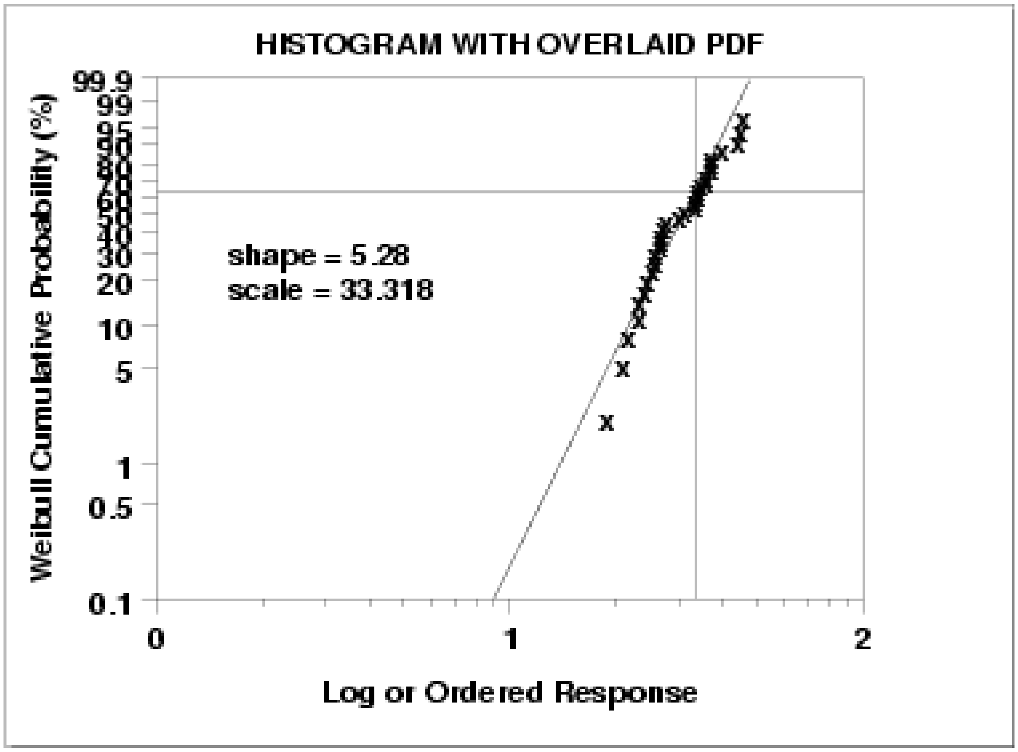
1. The optimal value, in the sense of having the most linear probability plot, of the shape parameter gamma is 2.13.
2. At the optimal value of the shape parameter, the PPCC value is 0.988.
3. At the optimal value of the shape parameter, the estimate of the location parameter is 15.90 and the estimate of the scale parameter is 16.92.
4. Fine tuning the estimate of gamma (from 2 to 2.13) has minimal impact on the PPCC value.

Alternative Plots

The Weibull plot and the Weibull hazard plot are alternative graphical analysis procedures to the PPCC plots and probability plots.

These two procedures, especially the Weibull plot, are very commonly employed. That not withstanding, the disadvantage of these two procedures is that they both assume that the location parameter (i.e., the lower bound) is zero and that we are fitting a 2-parameter Weibull instead of a 3-parameter Weibull. The advantage is that there is an extensive literature on these methods and they have been designed to work with either censored or uncensored data.

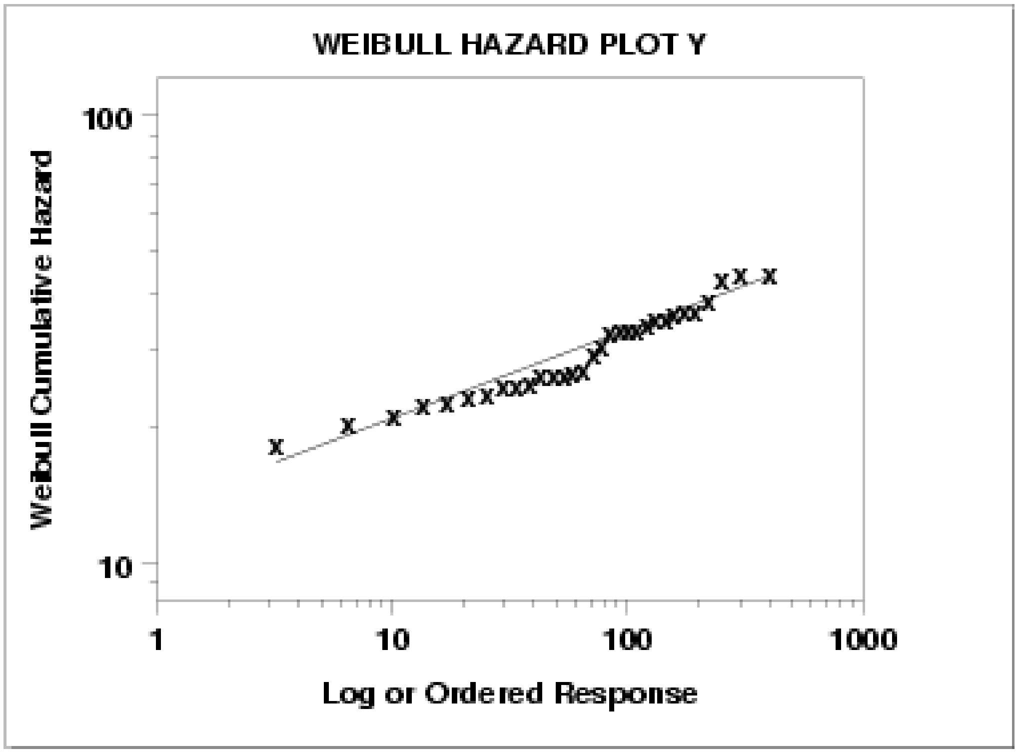
Weibull Plot



This Weibull plot shows the following

1. The Weibull plot is approximately linear indicating that the 2-parameter Weibull provides an adequate fit to the data.
2. The estimate of the shape parameter is 5.28 and the estimate of the scale parameter is 33.32.

Weibull Hazard Plot

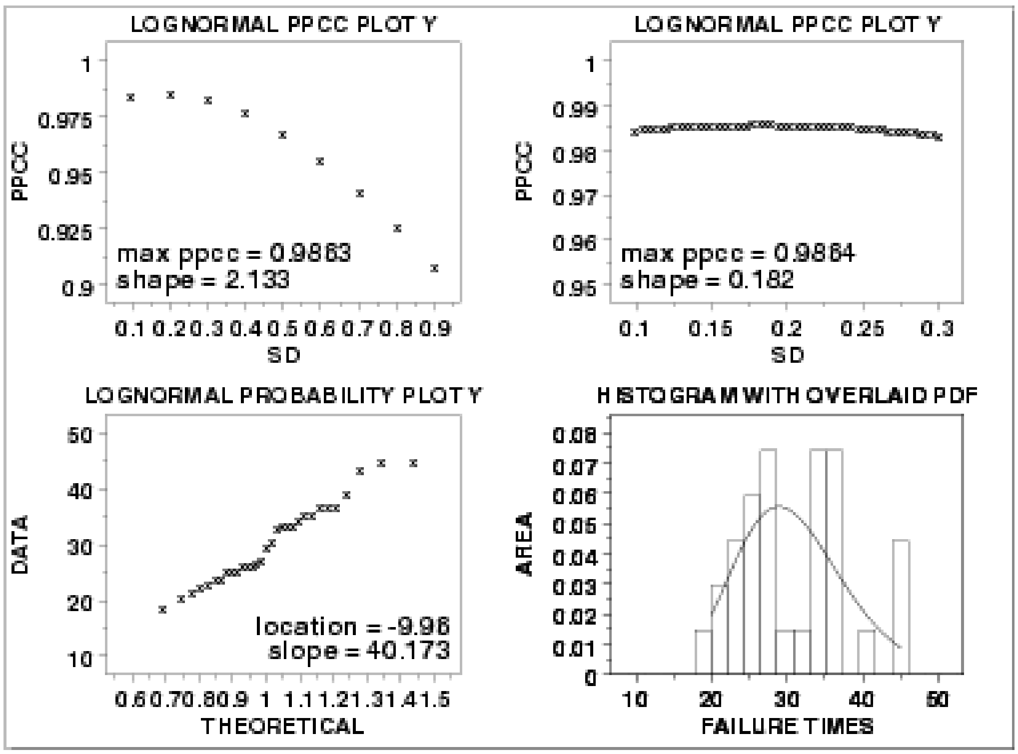


The construction and interpretation of the Weibull hazard plot is discussed in the [Assessing Product Reliability](#) chapter.

1.4.2.9.4. Lognormal Analysis

Plots for Lognormal Distribution

The following plots were generated for a lognormal distribution.



Conclusions

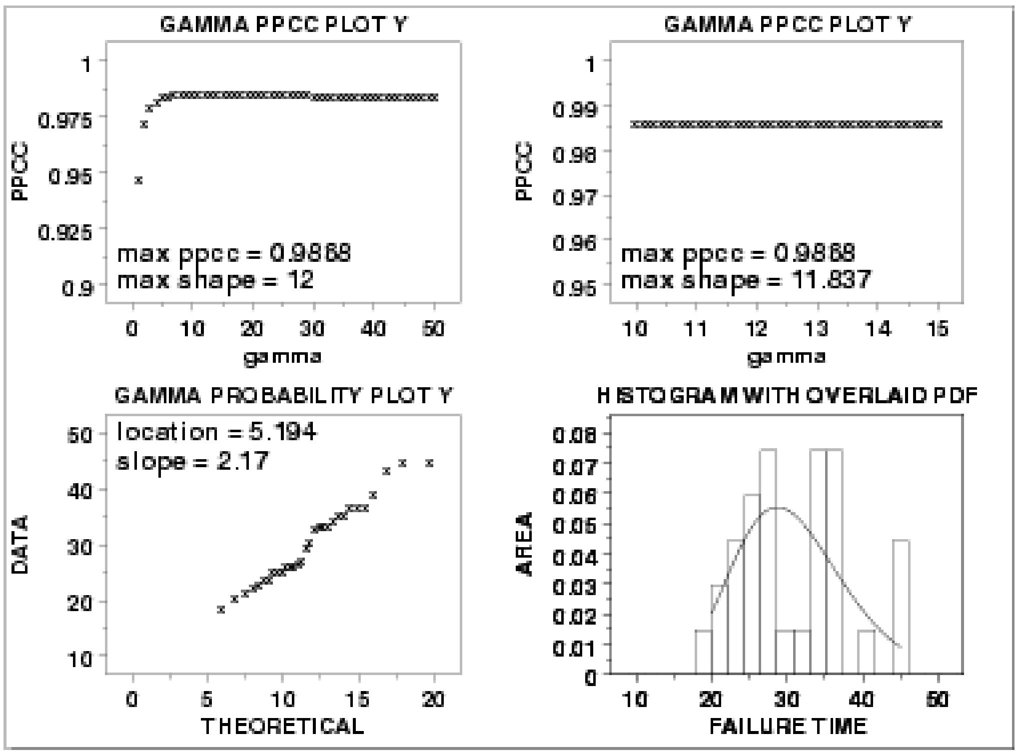
We can make the following conclusions from these plots.

1. The optimal value, in the sense of having the most linear probability plot, of the shape parameter σ is 0.18.
2. At the optimal value of the shape parameter, the PPCC value is 0.986.
3. At the optimal value of the shape parameter, the estimate of the location parameter is -9.96 and the estimate of the scale parameter is 40.17.
4. Fine tuning the estimate of the shape parameter (from 0.2 to 0.18) has minimal impact on the PPCC value.

1.4.2.9.5. Gamma Analysis

Plots for Gamma Distribution

The following plots were generated for a gamma distribution.



Conclusions

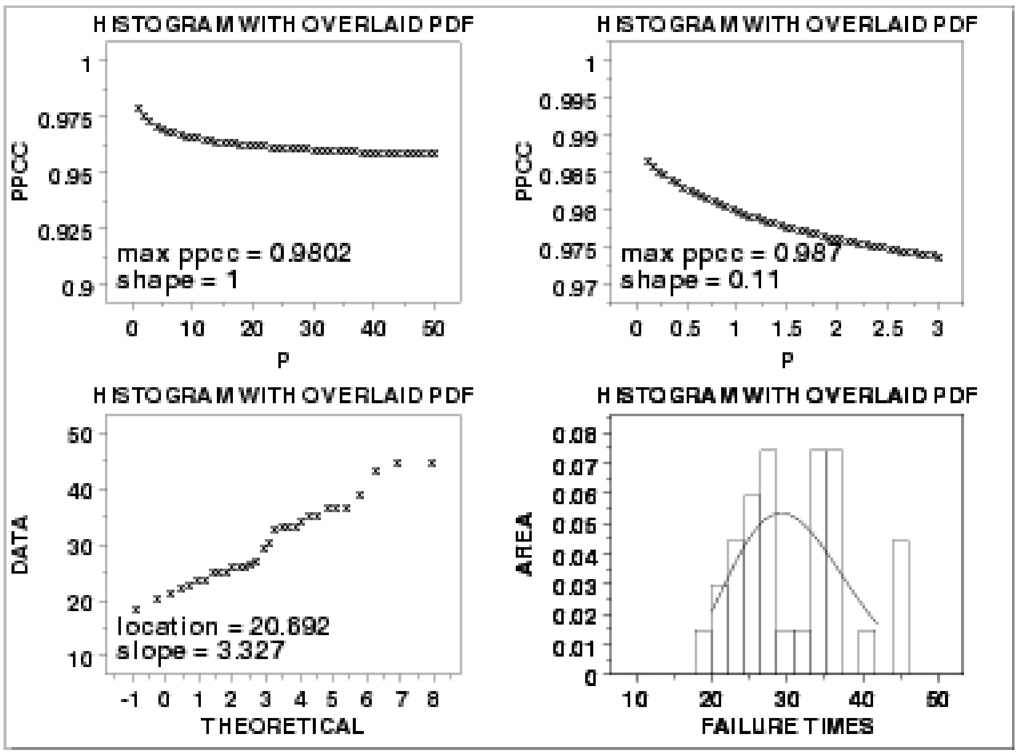
We can make the following conclusions from these plots.

1. The optimal value, in the sense of having the most linear probability plot, of the shape parameter γ is 11.8.
2. At the optimal value of the shape parameter, the PPCC value is 0.987.
3. At the optimal value of the shape parameter, the estimate of the location parameter is 5.19 and the estimate of the scale parameter is 2.17.
4. Fine tuning the estimate of γ (from 12 to 11.8) has some impact on the PPCC value (from 0.978 to 0.987).

1.4.2.9.6. Power Normal Analysis

Plots for Power Normal Distribution

The following plots were generated for a power normal distribution.



Conclusions

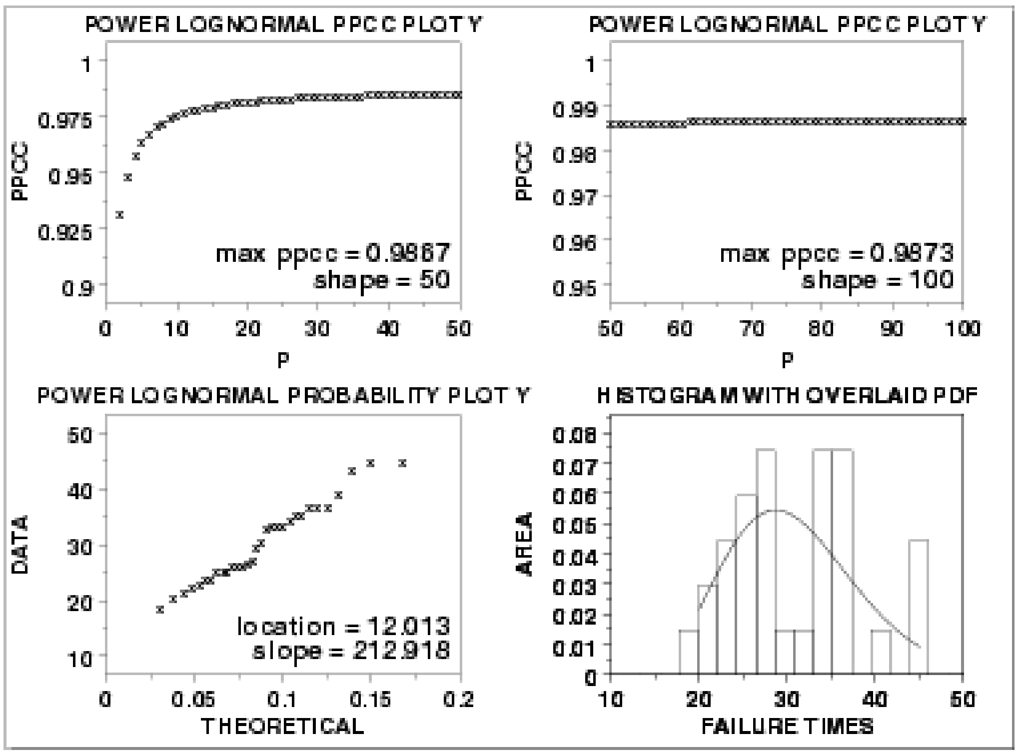
We can make the following conclusions from these plots.

1. A reasonable value, in the sense of having the most linear probability plot, of the shape parameter p is 0.11.
2. At the this value of the shape parameter, the PPCC value is 0.987.
3. At the optimal value of the shape parameter, the estimate of the location parameter is 20.9 and the estimate of the scale parameter is 3.3.
4. Fine tuning the estimate of p (from 1 to 0.11) results in a slight improvement of the the computed PPCC value (from 0.980 to 0.987).

1.4.2.9.7. Power Lognormal Analysis

Plots for Power Lognormal Distribution

The following plots were generated for a power lognormal distribution.



Conclusions

We can make the following conclusions from these plots.

1. A reasonable value, in the sense of having the most linear probability plot, of the shape parameter p is 100 (i.e., p is asymptotically increasing).
2. At this value of the shape parameter, the PPCC value is 0.987.
3. At this value of the shape parameter, the estimate of the location parameter is 12.01 and the estimate of the scale parameter is 212.92.
4. Fine tuning the estimate of p (from 50 to 100) has minimal impact on the PPCC value.

1.4.2.9.8. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to repeat the analysis outlined in the case study description on the previous page using Dataplot . It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*
1. Invoke Dataplot and read data. 1. Read in the data.	1. You have read 1 column of numbers into Dataplot, variable Y.
2. 4-plot of the data. 1. 4-plot of Y.	1. The failure times are in the range 15 to 50. The histogram and normal probability plot indicate a normal distribution fits the data reasonably well, but we can probably do better.
3. Generate the Weibull analysis. 1. Generate 2 iterations of the Weibull PPCC plot, a Weibull probability plot, and estimate some percent points. 2. Generate a Weibull plot.	1. The Weibull analysis results in a maximum PPCC value of 0.988. 2. The Weibull plot permits the estimation of a 2-parameter Weibull model.

3. Generate a Weibull hazard plot.	3. The Weibull hazard plot is approximately linear, indicating that the Weibull provides a good distributional model for these data.
4. Generate the lognormal analysis. 1. Generate 2 iterations of the lognormal PPCC plot and a lognormal probability plot.	 1. The lognormal analysis results in a maximum PPCC value of 0.986.
5. Generate the gamma analysis. 1. Generate 2 iterations of the gamma PPCC plot and a gamma probability plot.	 1. The gamma analysis results in a maximum PPCC value of 0.987.
6. Generate the power normal analysis. 1. Generate 2 iterations of the power normal PPCC plot and a power normal probability plot.	 1. The power normal analysis results in a maximum PPCC value of 0.988.
7. Generate the power lognormal analysis. 1. Generate 2 iterations of the power lognormal PPCC plot and a power lognormal probability plot.	 1. The power lognormal analysis results in a maximum PPCC value of 0.987.
8. Generate quantitative goodness of fit tests 1. Generate Anderson-Darling test for normality. 2. Generate Anderson-Darling test for lognormal distribution. 3. Generate Anderson-Darling test for Weibull distribution.	 1. The Anderson-Darling normality test indicates the normal distribution provides an adequate fit to the data. 2. The Anderson-Darling lognormal test indicates the lognormal distribution provides an adequate fit to the data. 3. The Anderson-Darling Weibull test indicates the lognormal distribution provides an adequate fit to the data.

1.4.2.10. Ceramic Strength

Ceramic Strength

This case study analyzes the effect of machining factors on the strength of ceramics.

1.4.2.10.1. Background and Data

Generation

The data for this case study were collected by Said Jahanmir of the NIST Ceramics Division in 1996 in connection with a NIST/industry ceramics consortium for strength optimization of ceramic strength

The motivation for studying this data set is to illustrate the analysis of multiple factors from a designed experiment

This case study will utilize only a subset of a full study that was conducted by Lisa Gill and James Filliben of the NIST Statistical Engineering Division

The response variable is a measure of the strength of the ceramic material (bonded S_i nitrate). The complete data set contains the following variables:

1. Factor 1 = Observation ID, i.e., run number (1 to 960)
2. Factor 2 = Lab (1 to 8)
3. Factor 3 = Bar ID within lab (1 to 30)
4. Factor 4 = Test number (1 to 4)
5. Response Variable = Strength of Ceramic
6. Factor 5 = Table speed (2 levels: 0.025 and 0.125)
7. Factor 6 = Down feed rate (2 levels: 0.050 and 0.125)
8. Factor 7 = Wheel grit size (2 levels: 150 and 80)
9. Factor 8 = Direction (2 levels: longitudinal and transverse)
10. Factor 9 = Treatment (1 to 16)
11. Factor 10 = Set of 15 within lab (2 levels: 1 and 2)
12. Factor 11 = Replication (2 levels: 1 and 2)
13. Factor 12 = Bar Batch (1 and 2)

The four primary factors of interest are:

1. Table speed (X1)
2. Down feed rate (X2)
3. Wheel grit size (X3)

4. Direction (X4)

For this case study, we are using only half the data. Specifically, we are using the data with the direction longitudinal. Therefore, we have only three primary factors

In addtion, we are interested in the nuisance factors

1. Lab
2. Batch

The complete file can be read into Dataplot with the following commands:

```
DIMENSION 20 VARIABLES
SKIP 50
READ JAHANMI2.DAT RUN RUN LAB BAR SET Y X1 TO X8 BATCH
```

Purpose of Analysis

The goals of this case study are:

1. Determine which of the four primary factors has the strongest effect on the strength of the ceramic material
2. Estimate the magnitude of the effects
3. Determine the optimal settings for the primary factors
4. Determine if the nuisance factors (lab and batch) have an effect on the ceramic strength

This case study is an example of a designed experiment. The Process Improvement chapter contains a detailed discussion of the construction and analysis of designed experiments. This case study is meant to complement the material in that chapter by showing how an EDA approach (emphasizing the use of graphical techniques) can be used in the analysis of designed experiments

Resulting Data

The following are the data used for this case study

Run	Lab	Batch	Y	X1	X2	X3
1	1	1	608.781	-1	-1	-1
2	1	2	569.670	-1	-1	-1
3	1	1	689.556	-1	-1	-1
4	1	2	747.541	-1	-1	-1
5	1	1	618.134	-1	-1	-1
6	1	2	612.182	-1	-1	-1
7	1	1	680.203	-1	-1	-1
8	1	2	607.766	-1	-1	-1
9	1	1	726.232	-1	-1	-1
10	1	2	605.380	-1	-1	-1
11	1	1	518.655	-1	-1	-1
12	1	2	589.226	-1	-1	-1

13	1	1	740.447	-1	-1	-1
14	1	2	588.375	-1	-1	-1
15	1	1	666.830	-1	-1	-1
16	1	2	531.384	-1	-1	-1
17	1	1	710.272	-1	-1	-1
18	1	2	633.417	-1	-1	-1
19	1	1	751.669	-1	-1	-1
20	1	2	619.060	-1	-1	-1
21	1	1	697.979	-1	-1	-1
22	1	2	632.447	-1	-1	-1
23	1	1	708.583	-1	-1	-1
24	1	2	624.256	-1	-1	-1
25	1	1	624.972	-1	-1	-1
26	1	2	575.143	-1	-1	-1
27	1	1	695.070	-1	-1	-1
28	1	2	549.278	-1	-1	-1
29	1	1	769.391	-1	-1	-1
30	1	2	624.972	-1	-1	-1
61	1	1	720.186	-1	1	1
62	1	2	587.695	-1	1	1
63	1	1	723.657	-1	1	1
64	1	2	569.207	-1	1	1
65	1	1	703.700	-1	1	1
66	1	2	613.257	-1	1	1
67	1	1	697.626	-1	1	1
68	1	2	565.737	-1	1	1
69	1	1	714.980	-1	1	1
70	1	2	662.131	-1	1	1
71	1	1	657.712	-1	1	1
72	1	2	543.177	-1	1	1
73	1	1	609.989	-1	1	1
74	1	2	512.394	-1	1	1
75	1	1	650.771	-1	1	1
76	1	2	611.190	-1	1	1
77	1	1	707.977	-1	1	1
78	1	2	659.982	-1	1	1
79	1	1	712.199	-1	1	1
80	1	2	569.245	-1	1	1
81	1	1	709.631	-1	1	1
82	1	2	725.792	-1	1	1
83	1	1	703.160	-1	1	1
84	1	2	608.960	-1	1	1
85	1	1	744.822	-1	1	1
86	1	2	586.060	-1	1	1
87	1	1	719.217	-1	1	1
88	1	2	617.441	-1	1	1

89	1	1	619.137	-1	1	1
90	1	2	592.845	-1	1	1
151	2	1	753.333	1	1	1
152	2	2	631.754	1	1	1
153	2	1	677.933	1	1	1
154	2	2	588.113	1	1	1
155	2	1	735.919	1	1	1
156	2	2	555.724	1	1	1
157	2	1	695.274	1	1	1
158	2	2	702.411	1	1	1
159	2	1	504.167	1	1	1
160	2	2	631.754	1	1	1
161	2	1	693.333	1	1	1
162	2	2	698.254	1	1	1
163	2	1	625.000	1	1	1
164	2	2	616.791	1	1	1
165	2	1	596.667	1	1	1
166	2	2	551.953	1	1	1
167	2	1	640.898	1	1	1
168	2	2	636.738	1	1	1
169	2	1	720.506	1	1	1
170	2	2	571.551	1	1	1
171	2	1	700.748	1	1	1
172	2	2	521.667	1	1	1
173	2	1	691.604	1	1	1
174	2	2	587.451	1	1	1
175	2	1	636.738	1	1	1
176	2	2	700.422	1	1	1
177	2	1	731.667	1	1	1
178	2	2	595.819	1	1	1
179	2	1	635.079	1	1	1
180	2	2	534.236	1	1	1
181	2	1	716.926	1	-1	-1
182	2	2	606.188	1	-1	-1
183	2	1	759.581	1	-1	-1
184	2	2	575.303	1	-1	-1
185	2	1	673.903	1	-1	-1
186	2	2	590.628	1	-1	-1
187	2	1	736.648	1	-1	-1
188	2	2	729.314	1	-1	-1
189	2	1	675.957	1	-1	-1
190	2	2	619.313	1	-1	-1
191	2	1	729.230	1	-1	-1
192	2	2	624.234	1	-1	-1
193	2	1	697.239	1	-1	-1
194	2	2	651.304	1	-1	-1

195	2	1	728.499	1	-1	-1
196	2	2	724.175	1	-1	-1
197	2	1	797.662	1	-1	-1
198	2	2	583.034	1	-1	-1
199	2	1	668.530	1	-1	-1
200	2	2	620.227	1	-1	-1
201	2	1	815.754	1	-1	-1
202	2	2	584.861	1	-1	-1
203	2	1	777.392	1	-1	-1
204	2	2	565.391	1	-1	-1
205	2	1	712.140	1	-1	-1
206	2	2	622.506	1	-1	-1
207	2	1	663.622	1	-1	-1
208	2	2	628.336	1	-1	-1
209	2	1	684.181	1	-1	-1
210	2	2	587.145	1	-1	-1
271	3	1	629.012	1	-1	1
272	3	2	584.319	1	-1	1
273	3	1	640.193	1	-1	1
274	3	2	538.239	1	-1	1
275	3	1	644.156	1	-1	1
276	3	2	538.097	1	-1	1
277	3	1	642.469	1	-1	1
278	3	2	595.686	1	-1	1
279	3	1	639.090	1	-1	1
280	3	2	648.935	1	-1	1
281	3	1	439.418	1	-1	1
282	3	2	583.827	1	-1	1
283	3	1	614.664	1	-1	1
284	3	2	534.905	1	-1	1
285	3	1	537.161	1	-1	1
286	3	2	569.858	1	-1	1
287	3	1	656.773	1	-1	1
288	3	2	617.246	1	-1	1
289	3	1	659.534	1	-1	1
290	3	2	610.337	1	-1	1
291	3	1	695.278	1	-1	1
292	3	2	584.192	1	-1	1
293	3	1	734.040	1	-1	1
294	3	2	598.853	1	-1	1
295	3	1	687.665	1	-1	1
296	3	2	554.774	1	-1	1
297	3	1	710.858	1	-1	1
298	3	2	605.694	1	-1	1
299	3	1	701.716	1	-1	1
300	3	2	627.516	1	-1	1

301	3	1	382.133	1	1	-1
302	3	2	574.522	1	1	-1
303	3	1	719.744	1	1	-1
304	3	2	582.682	1	1	-1
305	3	1	756.820	1	1	-1
306	3	2	563.872	1	1	-1
307	3	1	690.978	1	1	-1
308	3	2	715.962	1	1	-1
309	3	1	670.864	1	1	-1
310	3	2	616.430	1	1	-1
311	3	1	670.308	1	1	-1
312	3	2	778.011	1	1	-1
313	3	1	660.062	1	1	-1
314	3	2	604.255	1	1	-1
315	3	1	790.382	1	1	-1
316	3	2	571.906	1	1	-1
317	3	1	714.750	1	1	-1
318	3	2	625.925	1	1	-1
319	3	1	716.959	1	1	-1
320	3	2	682.426	1	1	-1
321	3	1	603.363	1	1	-1
322	3	2	707.604	1	1	-1
323	3	1	713.796	1	1	-1
324	3	2	617.400	1	1	-1
325	3	1	444.963	1	1	-1
326	3	2	689.576	1	1	-1
327	3	1	723.276	1	1	-1
328	3	2	676.678	1	1	-1
329	3	1	745.527	1	1	-1
330	3	2	563.290	1	1	-1
361	4	1	778.333	-1	-1	1
362	4	2	581.879	-1	-1	1
363	4	1	723.349	-1	-1	1
364	4	2	447.701	-1	-1	1
365	4	1	708.229	-1	-1	1
366	4	2	557.772	-1	-1	1
367	4	1	681.667	-1	-1	1
368	4	2	593.537	-1	-1	1
369	4	1	566.085	-1	-1	1
370	4	2	632.585	-1	-1	1
371	4	1	687.448	-1	-1	1
372	4	2	671.350	-1	-1	1
373	4	1	597.500	-1	-1	1
374	4	2	569.530	-1	-1	1
375	4	1	637.410	-1	-1	1
376	4	2	581.667	-1	-1	1

377	4	1	755.864	-1	-1	1
378	4	2	643.449	-1	-1	1
379	4	1	692.945	-1	-1	1
380	4	2	581.593	-1	-1	1
381	4	1	766.532	-1	-1	1
382	4	2	494.122	-1	-1	1
383	4	1	725.663	-1	-1	1
384	4	2	620.948	-1	-1	1
385	4	1	698.818	-1	-1	1
386	4	2	615.903	-1	-1	1
387	4	1	760.000	-1	-1	1
388	4	2	606.667	-1	-1	1
389	4	1	775.272	-1	-1	1
390	4	2	579.167	-1	-1	1
421	4	1	708.885	-1	1	-1
422	4	2	662.510	-1	1	-1
423	4	1	727.201	-1	1	-1
424	4	2	436.237	-1	1	-1
425	4	1	642.560	-1	1	-1
426	4	2	644.223	-1	1	-1
427	4	1	690.773	-1	1	-1
428	4	2	586.035	-1	1	-1
429	4	1	688.333	-1	1	-1
430	4	2	620.833	-1	1	-1
431	4	1	743.973	-1	1	-1
432	4	2	652.535	-1	1	-1
433	4	1	682.461	-1	1	-1
434	4	2	593.516	-1	1	-1
435	4	1	761.430	-1	1	-1
436	4	2	587.451	-1	1	-1
437	4	1	691.542	-1	1	-1
438	4	2	570.964	-1	1	-1
439	4	1	643.392	-1	1	-1
440	4	2	645.192	-1	1	-1
441	4	1	697.075	-1	1	-1
442	4	2	540.079	-1	1	-1
443	4	1	708.229	-1	1	-1
444	4	2	707.117	-1	1	-1
445	4	1	746.467	-1	1	-1
446	4	2	621.779	-1	1	-1
447	4	1	744.819	-1	1	-1
448	4	2	585.777	-1	1	-1
449	4	1	655.029	-1	1	-1
450	4	2	703.980	-1	1	-1
541	5	1	715.224	-1	-1	-1
542	5	2	698.237	-1	-1	-1

543	5	1	614.417	-1	-1	-1
544	5	2	757.120	-1	-1	-1
545	5	1	761.363	-1	-1	-1
546	5	2	621.751	-1	-1	-1
547	5	1	716.106	-1	-1	-1
548	5	2	472.125	-1	-1	-1
549	5	1	659.502	-1	-1	-1
550	5	2	612.700	-1	-1	-1
551	5	1	730.781	-1	-1	-1
552	5	2	583.170	-1	-1	-1
553	5	1	546.928	-1	-1	-1
554	5	2	599.771	-1	-1	-1
555	5	1	734.203	-1	-1	-1
556	5	2	549.227	-1	-1	-1
557	5	1	682.051	-1	-1	-1
558	5	2	605.453	-1	-1	-1
559	5	1	701.341	-1	-1	-1
560	5	2	569.599	-1	-1	-1
561	5	1	759.729	-1	-1	-1
562	5	2	637.233	-1	-1	-1
563	5	1	689.942	-1	-1	-1
564	5	2	621.774	-1	-1	-1
565	5	1	769.424	-1	-1	-1
566	5	2	558.041	-1	-1	-1
567	5	1	715.286	-1	-1	-1
568	5	2	583.170	-1	-1	-1
569	5	1	776.197	-1	-1	-1
570	5	2	345.294	-1	-1	-1
571	5	1	547.099	1	-1	1
572	5	2	570.999	1	-1	1
573	5	1	619.942	1	-1	1
574	5	2	603.232	1	-1	1
575	5	1	696.046	1	-1	1
576	5	2	595.335	1	-1	1
577	5	1	573.109	1	-1	1
578	5	2	581.047	1	-1	1
579	5	1	638.794	1	-1	1
580	5	2	455.878	1	-1	1
581	5	1	708.193	1	-1	1
582	5	2	627.880	1	-1	1
583	5	1	502.825	1	-1	1
584	5	2	464.085	1	-1	1
585	5	1	632.633	1	-1	1
586	5	2	596.129	1	-1	1
587	5	1	683.382	1	-1	1
588	5	2	640.371	1	-1	1

589	5	1	684.812	1	-1	1
590	5	2	621.471	1	-1	1
591	5	1	738.161	1	-1	1
592	5	2	612.727	1	-1	1
593	5	1	671.492	1	-1	1
594	5	2	606.460	1	-1	1
595	5	1	709.771	1	-1	1
596	5	2	571.760	1	-1	1
597	5	1	685.199	1	-1	1
598	5	2	599.304	1	-1	1
599	5	1	624.973	1	-1	1
600	5	2	579.459	1	-1	1
601	6	1	757.363	1	1	1
602	6	2	761.511	1	1	1
603	6	1	633.417	1	1	1
604	6	2	566.969	1	1	1
605	6	1	658.754	1	1	1
606	6	2	654.397	1	1	1
607	6	1	664.666	1	1	1
608	6	2	611.719	1	1	1
609	6	1	663.009	1	1	1
610	6	2	577.409	1	1	1
611	6	1	773.226	1	1	1
612	6	2	576.731	1	1	1
613	6	1	708.261	1	1	1
614	6	2	617.441	1	1	1
615	6	1	739.086	1	1	1
616	6	2	577.409	1	1	1
617	6	1	667.786	1	1	1
618	6	2	548.957	1	1	1
619	6	1	674.481	1	1	1
620	6	2	623.315	1	1	1
621	6	1	695.688	1	1	1
622	6	2	621.761	1	1	1
623	6	1	588.288	1	1	1
624	6	2	553.978	1	1	1
625	6	1	545.610	1	1	1
626	6	2	657.157	1	1	1
627	6	1	752.305	1	1	1
628	6	2	610.882	1	1	1
629	6	1	684.523	1	1	1
630	6	2	552.304	1	1	1
631	6	1	717.159	-1	1	-1
632	6	2	545.303	-1	1	-1
633	6	1	721.343	-1	1	-1
634	6	2	651.934	-1	1	-1

635	6	1	750.623	-1	1	-1
636	6	2	635.240	-1	1	-1
637	6	1	776.488	-1	1	-1
638	6	2	641.083	-1	1	-1
639	6	1	750.623	-1	1	-1
640	6	2	645.321	-1	1	-1
641	6	1	600.840	-1	1	-1
642	6	2	566.127	-1	1	-1
643	6	1	686.196	-1	1	-1
644	6	2	647.844	-1	1	-1
645	6	1	687.870	-1	1	-1
646	6	2	554.815	-1	1	-1
647	6	1	725.527	-1	1	-1
648	6	2	620.087	-1	1	-1
649	6	1	658.796	-1	1	-1
650	6	2	711.301	-1	1	-1
651	6	1	690.380	-1	1	-1
652	6	2	644.355	-1	1	-1
653	6	1	737.144	-1	1	-1
654	6	2	713.812	-1	1	-1
655	6	1	663.851	-1	1	-1
656	6	2	696.707	-1	1	-1
657	6	1	766.630	-1	1	-1
658	6	2	589.453	-1	1	-1
659	6	1	625.922	-1	1	-1
660	6	2	634.468	-1	1	-1
721	7	1	694.430	1	1	-1
722	7	2	599.751	1	1	-1
723	7	1	730.217	1	1	-1
724	7	2	624.542	1	1	-1
725	7	1	700.770	1	1	-1
726	7	2	723.505	1	1	-1
727	7	1	722.242	1	1	-1
728	7	2	674.717	1	1	-1
729	7	1	763.828	1	1	-1
730	7	2	608.539	1	1	-1
731	7	1	695.668	1	1	-1
732	7	2	612.135	1	1	-1
733	7	1	688.887	1	1	-1
734	7	2	591.935	1	1	-1
735	7	1	531.021	1	1	-1
736	7	2	676.656	1	1	-1
737	7	1	698.915	1	1	-1
738	7	2	647.323	1	1	-1
739	7	1	735.905	1	1	-1
740	7	2	811.970	1	1	-1

741	7	1	732.039	1	1	-1
742	7	2	603.883	1	1	-1
743	7	1	751.832	1	1	-1
744	7	2	608.643	1	1	-1
745	7	1	618.663	1	1	-1
746	7	2	630.778	1	1	-1
747	7	1	744.845	1	1	-1
748	7	2	623.063	1	1	-1
749	7	1	690.826	1	1	-1
750	7	2	472.463	1	1	-1
811	7	1	666.893	-1	1	1
812	7	2	645.932	-1	1	1
813	7	1	759.860	-1	1	1
814	7	2	577.176	-1	1	1
815	7	1	683.752	-1	1	1
816	7	2	567.530	-1	1	1
817	7	1	729.591	-1	1	1
818	7	2	821.654	-1	1	1
819	7	1	730.706	-1	1	1
820	7	2	684.490	-1	1	1
821	7	1	763.124	-1	1	1
822	7	2	600.427	-1	1	1
823	7	1	724.193	-1	1	1
824	7	2	686.023	-1	1	1
825	7	1	630.352	-1	1	1
826	7	2	628.109	-1	1	1
827	7	1	750.338	-1	1	1
828	7	2	605.214	-1	1	1
829	7	1	752.417	-1	1	1
830	7	2	640.260	-1	1	1
831	7	1	707.899	-1	1	1
832	7	2	700.767	-1	1	1
833	7	1	715.582	-1	1	1
834	7	2	665.924	-1	1	1
835	7	1	728.746	-1	1	1
836	7	2	555.926	-1	1	1
837	7	1	591.193	-1	1	1
838	7	2	543.299	-1	1	1
839	7	1	592.252	-1	1	1
840	7	2	511.030	-1	1	1
901	8	1	740.833	-1	-1	1
902	8	2	583.994	-1	-1	1
903	8	1	786.367	-1	-1	1
904	8	2	611.048	-1	-1	1
905	8	1	712.386	-1	-1	1
906	8	2	623.338	-1	-1	1

907	8	1	738.333	-1	-1	1
908	8	2	679.585	-1	-1	1
909	8	1	741.480	-1	-1	1
910	8	2	665.004	-1	-1	1
911	8	1	729.167	-1	-1	1
912	8	2	655.860	-1	-1	1
913	8	1	795.833	-1	-1	1
914	8	2	715.711	-1	-1	1
915	8	1	723.502	-1	-1	1
916	8	2	611.999	-1	-1	1
917	8	1	718.333	-1	-1	1
918	8	2	577.722	-1	-1	1
919	8	1	768.080	-1	-1	1
920	8	2	615.129	-1	-1	1
921	8	1	747.500	-1	-1	1
922	8	2	540.316	-1	-1	1
923	8	1	775.000	-1	-1	1
924	8	2	711.667	-1	-1	1
925	8	1	760.599	-1	-1	1
926	8	2	639.167	-1	-1	1
927	8	1	758.333	-1	-1	1
928	8	2	549.491	-1	-1	1
929	8	1	682.500	-1	-1	1
930	8	2	684.167	-1	-1	1
931	8	1	658.116	1	-1	-1
932	8	2	672.153	1	-1	-1
933	8	1	738.213	1	-1	-1
934	8	2	594.534	1	-1	-1
935	8	1	681.236	1	-1	-1
936	8	2	627.650	1	-1	-1
937	8	1	704.904	1	-1	-1
938	8	2	551.870	1	-1	-1
939	8	1	693.623	1	-1	-1
940	8	2	594.534	1	-1	-1
941	8	1	624.993	1	-1	-1
942	8	2	602.660	1	-1	-1
943	8	1	700.228	1	-1	-1
944	8	2	585.450	1	-1	-1
945	8	1	611.874	1	-1	-1
946	8	2	555.724	1	-1	-1
947	8	1	579.167	1	-1	-1
948	8	2	574.934	1	-1	-1
949	8	1	720.872	1	-1	-1
950	8	2	584.625	1	-1	-1
951	8	1	690.320	1	-1	-1
952	8	2	555.724	1	-1	-1

953	8	1	677.933	1	-1	-1
954	8	2	611.874	1	-1	-1
955	8	1	674.600	1	-1	-1
956	8	2	698.254	1	-1	-1
957	8	1	611.999	1	-1	-1
958	8	2	748.130	1	-1	-1
959	8	1	530.680	1	-1	-1
960	8	2	689.942	1	-1	-1

1.4.2.10.2. Analysis of the Response Variable

Numerical Summary

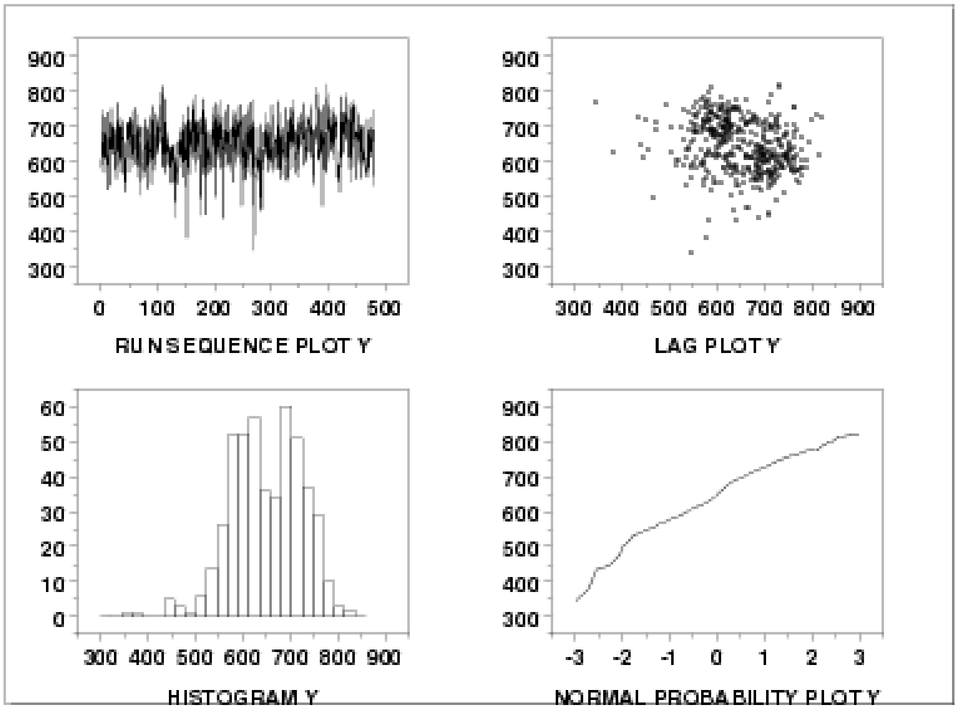
As a first step in the analysis, a table of summary statistics is computed for the response variable. The following table, generated by Dataplot, shows a typical set of statistics.

```
                         SUMMARY

                NUMBER OF OBSERVATIONS =      480


***********************************************************************
*        LOCATION MEASURES         *       DISPERSION MEASURES        *
***********************************************************************
*  MIDRANGE     =   0.5834740E+03  *  RANGE        =   0.4763600E+03  *
*  MEAN         =   0.6500773E+03  *  STAND. DEV.  =   0.7463826E+02  *
*  MIDMEAN      =   0.6426155E+03  *  AV. AB. DEV. =   0.6184948E+02  *
*  MEDIAN       =   0.6466275E+03  *  MINIMUM      =   0.3452940E+03  *
*               =                  *  LOWER QUART. =   0.5960515E+03  *
*               =                  *  LOWER HINGE  =   0.5959740E+03  *
*               =                  *  UPPER HINGE  =   0.7084220E+03  *
*               =                  *  UPPER QUART. =   0.7083415E+03  *
*               =                  *  MAXIMUM      =   0.8216540E+03  *
***********************************************************************
*       RANDOMNESS MEASURES        *     DISTRIBUTIONAL MEASURES      *
***********************************************************************
*  AUTOCO COEF  =  -0.2290508E+00  *  ST. 3RD MOM. =  -0.3682922E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.3220554E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =   0.3877698E+01  *
*               =                  *  UNIFORM PPCC =   0.9756916E+00  *
*               =                  *  NORMAL  PPCC =   0.9906310E+00  *
*               =                  *  TUK -.5 PPCC =   0.8357126E+00  *
*               =                  *  CAUCHY  PPCC =   0.5063868E+00  *
***********************************************************************
```

From the above output, the mean strength is 650.08 and the standard deviation of the strength is 74.64.

4-Plot

The next step is generate a 4-plot of the response variable.



This 4-plot shows:

1. The run sequence plot (upper left corner) shows that the location and scale are relatively constant. It also shows a few outliers on the low side. Most of the points are in the range 500 to 750. However, there are about half a dozen points in the 300 to 450 range that may require special attention.

   A run sequence plot is useful for designed experiments in that it can reveal time effects. Time is normally a nuisance factor. That is, the time order on which runs are made should not have a significant effect on the response. If a time effect does appear to exist, this means that there is a potential bias in the experiment that needs to be investigated and resolved.
2. The lag plot (the upper right corner) does not show any significant structure. This is another tool for detecting any potential time effect.
3. The histogram (the lower left corner) shows the response appears to be reasonably symmetric, but with a bimodal distribution.
4. The normal probability plot (the lower right corner) shows some curvature indicating that distributions other than the normal may provide a better fit.

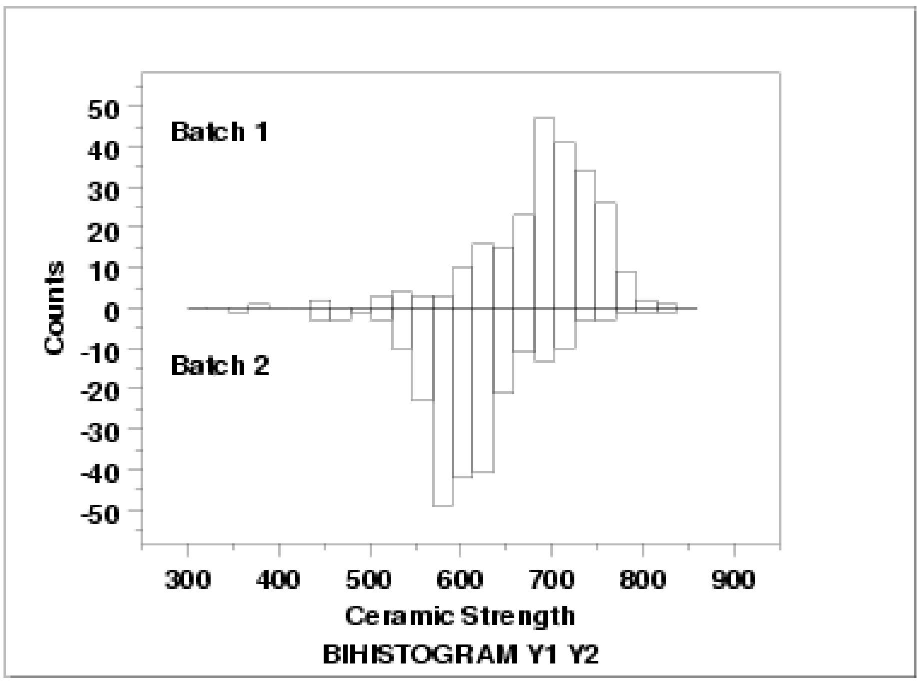
1.4.2.10.3. Analysis of the Batch Effect

Batch is a Nuisance Factor

The two nuisance factors in this experiment are the batch number and the lab. There are 2 batches and 8 labs. Ideally, these factors will have minimal effect on the response variable.

We will investigate the batch factor first.

Bihistogram

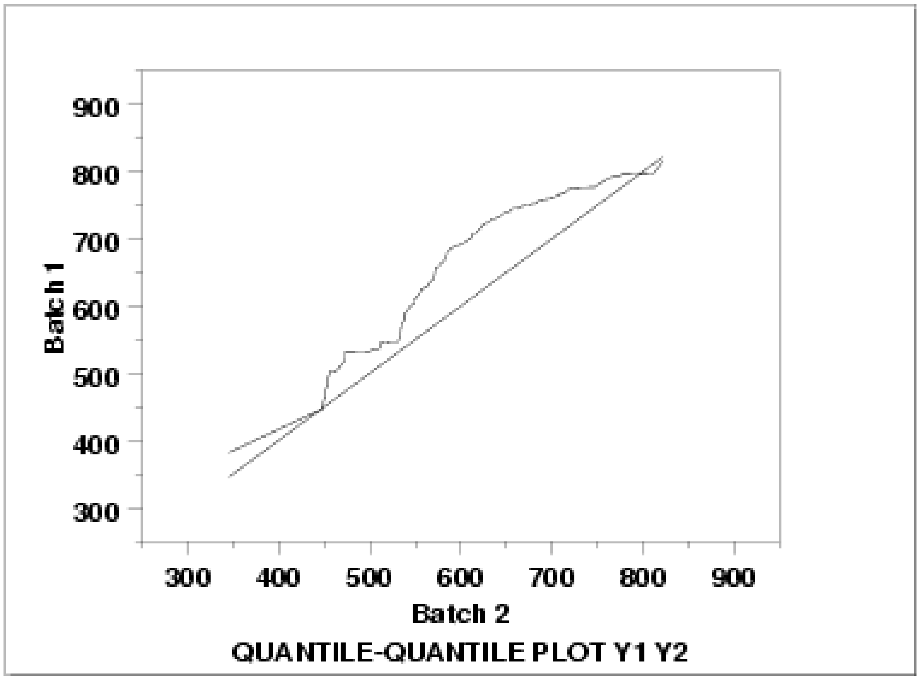


This bihistogram shows the following.

1. There does appear to be a batch effect.
2. The batch 1 responses are centered at 700 while the batch 2 responses are centered at 625. That is, the batch effect is approximately 75 units.
3. The variability is comparable for the 2 batches.
4. Batch 1 has some skewness in the lower tail. Batch 2 has some skewness in the center of the distribution, but not as much in the tails compared to batch 1.
5. Both batches have a few low-lying points.

Although we could stop with the bihistogram, we will show a few other commonly used two-sample graphical techniques for comparison.

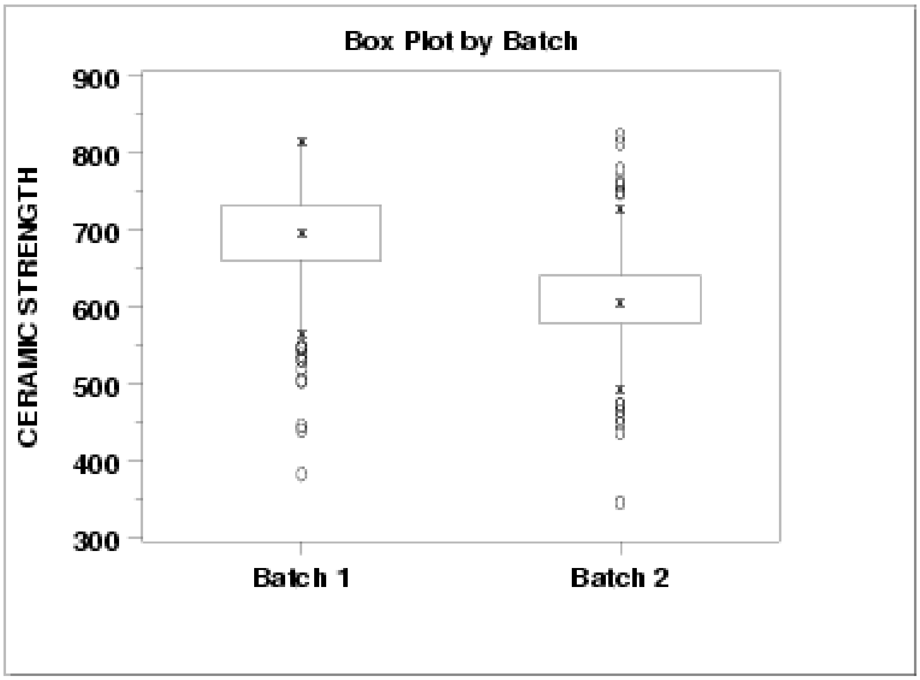
Quantile-Quantile Plot



This q-q plot shows the following.

1. Except for a few points in the right tail, the batch 1 values have higher quantiles than the batch 2 values. This implies that batch 1 has a greater location value than batch 2.
2. The q-q plot is not linear. This implies that the difference between the batches is not explained simply by a shift in location. That is, the variation and/or skewness varies as well. From the bihistogram, it appears that the skewness in batch 2 is the most likely explanation for the non-linearity in the q-q plot.

Box Plot



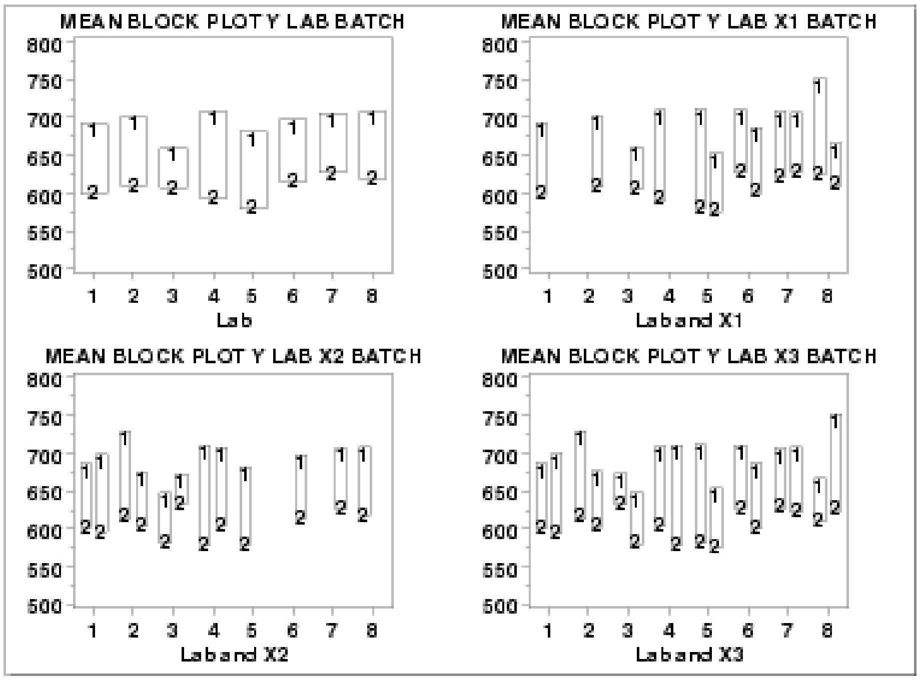
This box plot shows the following.

1. The median for batch 1 is approximately 700 while the median for batch 2 is approximately 600.

2. The spread is reasonably similar for both batches, maybe slightly larger for batch 1.
3. Both batches have a number of outliers on the low side. Batch 2 also has a few outliers on the high side. Box plots are a particularly effective method for identifying the presence of outliers.

Block Plots

A block plot is generated for each of the eight labs, with "1" and "2" denoting the batch numbers. In the first plot, we do not include any of the primary factors. The next 3 block plots include one of the primary factors. Note that each of the 3 primary factors (table speed = X1, down feed rate = X2, wheel grit size = X3) has 2 levels. With 8 labs and 2 levels for the primary factor, we would expect 16 separate blocks on these plots. The fact that some of these blocks are missing indicates that some of the combinations of lab and primary factor are empty.



These block plots show the following.

1. The mean for batch 1 is greater than the mean for batch 2 in **all** of the cases above. This is strong evidence that the batch effect is real and consistent across labs and primary factors.

Quantitative Techniques

We can confirm some of the conclusions drawn from the above graphics by using quantitative techniques. The two sample t-test can be used to test whether or not the means from the two batches are equal and the F-test can be used to test whether or not the standard deviations from the two batches are equal.

Two Sample T-Test

The following is the Dataplot output from the two sample t-test.

```
                       T-TEST
                     (2-SAMPLE)
 NULL HYPOTHESIS UNDER TEST--POPULATION MEANS MU1 = MU2

 SAMPLE 1:
    NUMBER OF OBSERVATIONS      =      240
    MEAN                        =    688.9987
    STANDARD DEVIATION          =    65.54909
    STANDARD DEVIATION OF MEAN  =    4.231175

 SAMPLE 2:
    NUMBER OF OBSERVATIONS      =      240
    MEAN                        =    611.1559
    STANDARD DEVIATION          =    61.85425
    STANDARD DEVIATION OF MEAN  =    3.992675

 IF     ASSUME SIGMA1 = SIGMA2:
    POOLED STANDARD DEVIATION   =    63.72845
    DIFFERENCE (DELTA) IN MEANS =    77.84271
    STANDARD DEVIATION OF DELTA =    5.817585
    T-TEST STATISTIC VALUE      =    13.38059
    DEGREES OF FREEDOM          =    478.0000
    T-TEST STATISTIC CDF VALUE  =    1.000000

 IF NOT ASSUME SIGMA1 = SIGMA2:
    STANDARD DEVIATION SAMPLE 1 =    65.54909
    STANDARD DEVIATION SAMPLE 2 =    61.85425
    BARTLETT CDF VALUE          =    0.629618
    DIFFERENCE (DELTA) IN MEANS =    77.84271
    STANDARD DEVIATION OF DELTA =    5.817585
    T-TEST STATISTIC VALUE      =    13.38059
    EQUIVALENT DEG. OF FREEDOM  =    476.3999
    T-TEST STATISTIC CDF VALUE  =    1.000000

                   ALTERNATIVE-         ALTERNATIVE-
 ALTERNATIVE-      HYPOTHESIS           HYPOTHESIS
 HYPOTHESIS        ACCEPTANCE INTERVAL  CONCLUSION
 MU1 <> MU2         (0,0.025) (0.975,1)   ACCEPT
 MU1 < MU2          (0,0.05)              REJECT
 MU1 > MU2          (0.95,1)              ACCEPT
```

The t-test indicates that the mean for batch 1 is larger than the mean for batch 2 (at the 5% confidence level).

F-Test

The following is the Dataplot output from the F-test.

```
                              F-TEST
 NULL HYPOTHESIS UNDER TEST--SIGMA1 = SIGMA2
 ALTERNATIVE HYPOTHESIS UNDER TEST--SIGMA1 NOT EQUAL SIGMA2

 SAMPLE 1:
    NUMBER OF OBSERVATIONS      =      240
    MEAN                        =    688.9987
    STANDARD DEVIATION          =    65.54909

 SAMPLE 2:
    NUMBER OF OBSERVATIONS      =      240
    MEAN                        =    611.1559
    STANDARD DEVIATION          =    61.85425

 TEST:
    STANDARD DEV. (NUMERATOR)   =    65.54909
    STANDARD DEV. (DENOMINATOR) =    61.85425
    F-TEST STATISTIC VALUE      =    1.123037
    DEG. OF FREEDOM (NUMER.)    =    239.0000
    DEG. OF FREEDOM (DENOM.)    =    239.0000
    F-TEST STATISTIC CDF VALUE  =    0.814808

   NULL          NULL HYPOTHESIS        NULL HYPOTHESIS
   HYPOTHESIS    ACCEPTANCE INTERVAL    CONCLUSION
 SIGMA1 = SIGMA2    (0.000,0.950)         ACCEPT
```

The F-test indicates that the standard deviations for the two batches are not significantly different at the 5% confidence level.

Conclusions

We can draw the following conclusions from the above analysis.

1. There is in fact a significant batch effect. This batch effect is consistent across labs and primary factors.
2. The magnitude of the difference is on the order of 75 to 100 (with batch 2 being smaller than batch 1). The standard deviations do not appear to be significantly different.
3. There is some skewness in the batches.

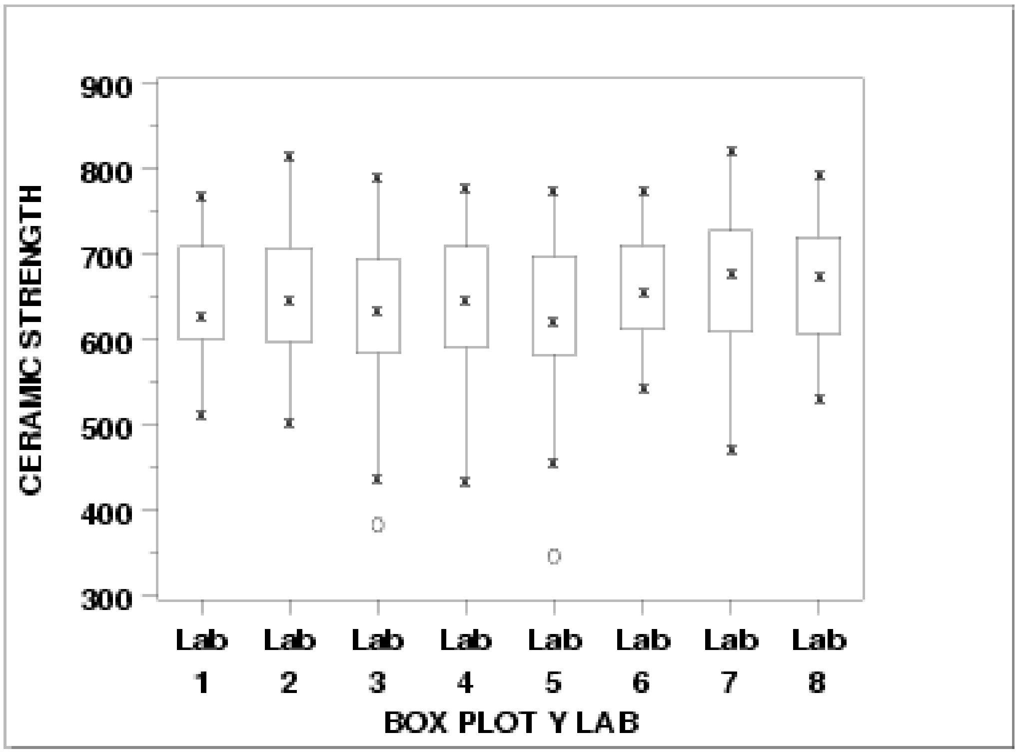
This batch effect was completely unexpected by the scientific investigators in this study.

Note that although the quantitative techniques support the conclusions of unequal means and equal standard deviations, they do not show the more subtle features of the data such as the presence of outliers and the skewness of the batch 2 data.

1.4.2.10.4. Analysis of the Lab Effect

Box Plot

The next matter is to determine if there is a lab effect. The first step is to generate a box plot for the ceramic strength based on the lab.

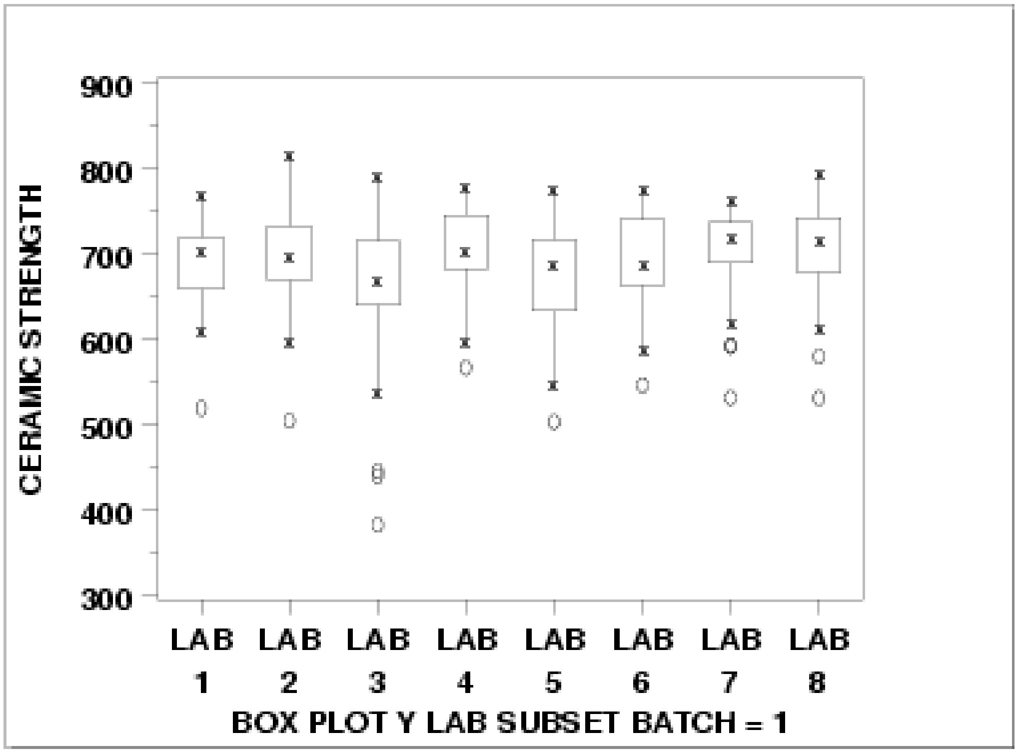


This box plot shows the following.

1. There is minor variation in the medians for the 8 labs.
2. The scales are relatively constant for the labs.
3. Two of the labs (3 and 5) have outliers on the low side.

Box Plot for Batch 1

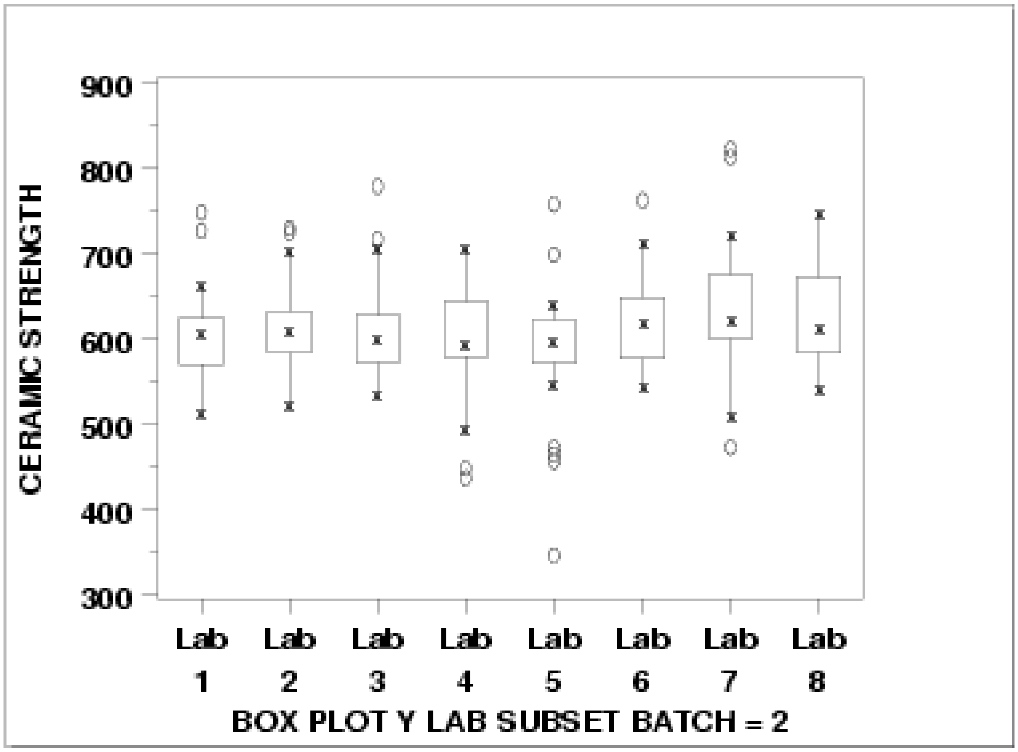
Given that the previous section showed a distinct batch effect, the next step is to generate the box plots for the two batches separately.



This [box plot] shows the following.

1. Each of the labs has a median in the 650 to 700 range.
2. The variability is relatively constant across the labs.
3. Each of the labs has at least one outlier on the low side.

Box Plot for Batch 2



This box plot shows the following.

1. The medians are in the range 550 to 600.
2. There is a bit more variability, across the labs, for batch2 compared to batch 1.
3. Six of the eight labs show outliers on the high side. Three of the labs show outliers on the low side.

Conclusions

We can draw the following conclusions about a possible lab effect from the above box plots.

1. The batch effect (of approximately 75 to 100 units) on location dominates any lab effects.
2. It is reasonable to treat the labs as homogeneous.

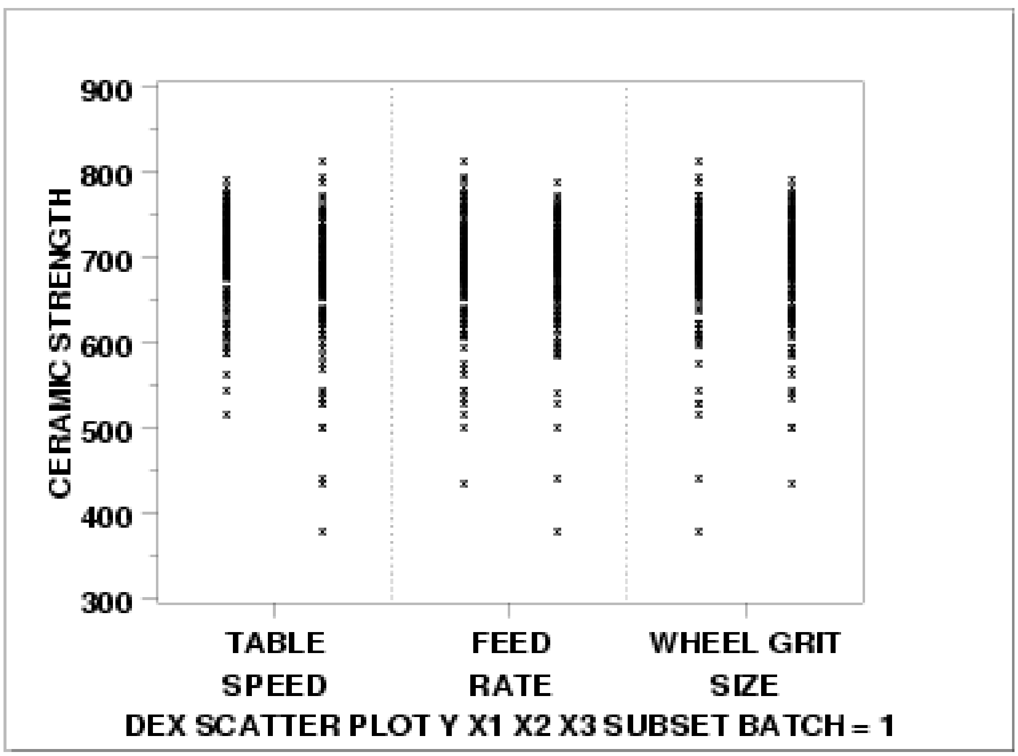
1.4.2.10.5. Analysis of Primary Factors

Main effects

The first step in analyzing the primary factors is to determine which factors are the most significant. The dex scatter plot, dex mean plot, and the dex standard deviation plots will be the primary tools, with "dex" being short for "design of experiments".

Since the previous pages showed a significant batch effect but a minimal lab effect, we will generate separate plots for batch 1 and batch 2. However, the labs will be treated as equivalent.

Dex Scatter Plot for Batch 1

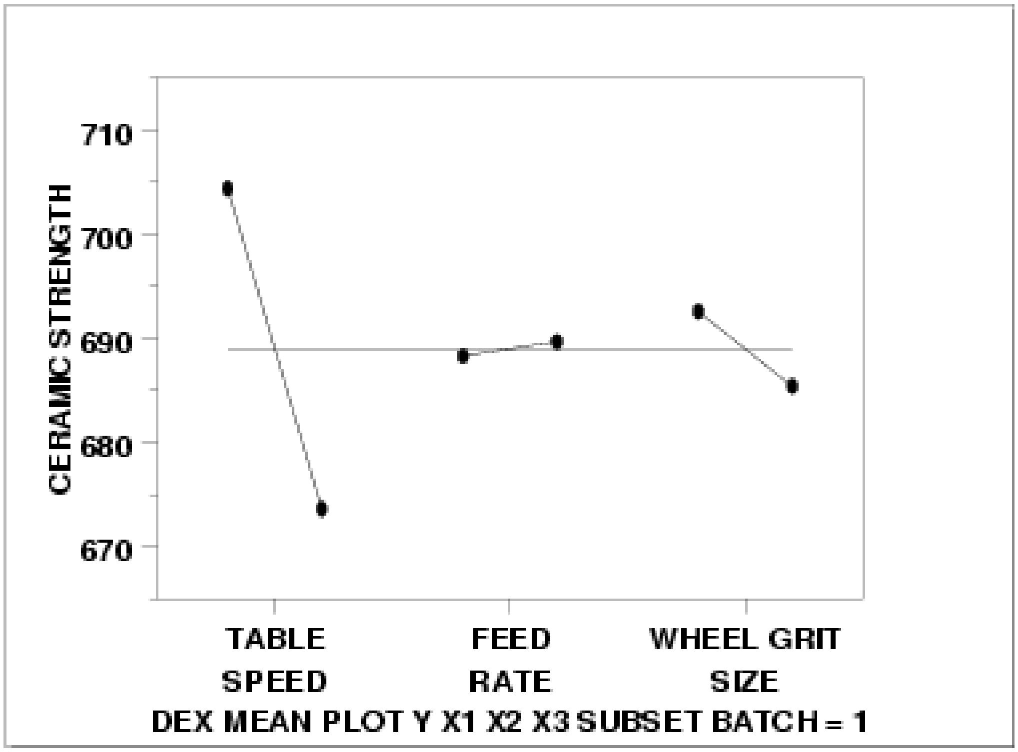


This dex scatter plot shows the following for batch 1.

1. Most of the points are between 500 and 800.
2. There are about a dozen or so points between 300 and 500.
3. Except for the outliers on the low side (i.e., the points between 300 and 500), the distribution of the points is comparable for the

3 primary factors in terms of location and spread.

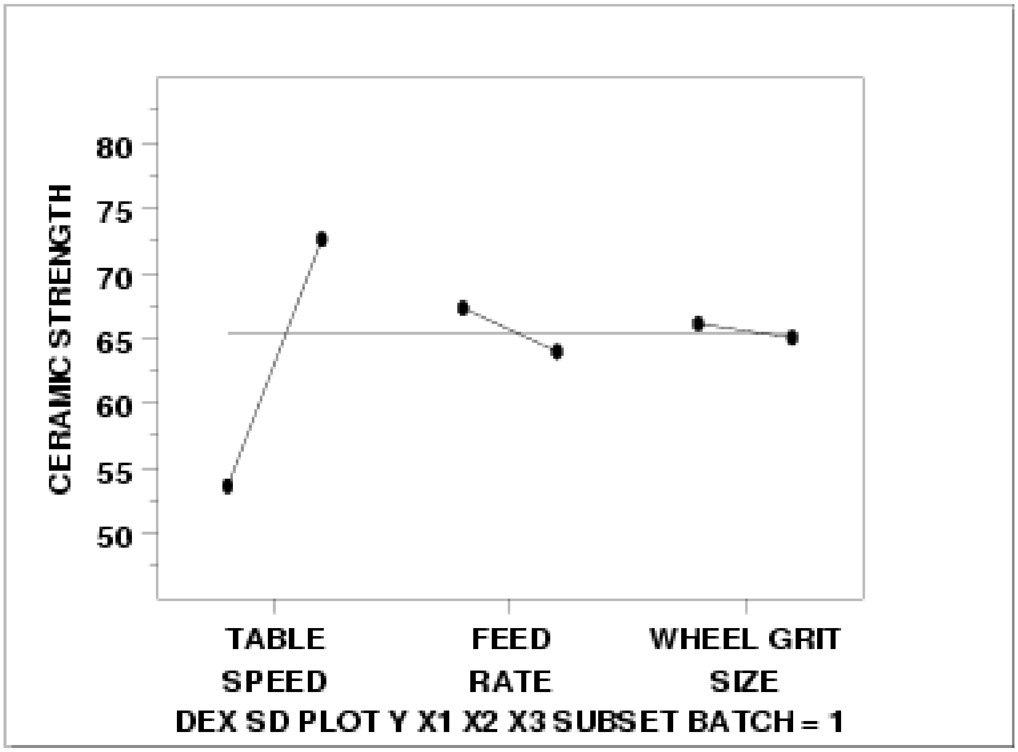
Dex Mean Plot for Batch 1



This dex mean plot shows the following for batch 1.

1. The table speed factor (X1) is the most significant factor with an effect, the difference between the two points, of approximately 35 units.
2. The wheel grit factor (X3) is the next most significant factor with an effect of approximately 10 units.
3. The feed rate factor (X2) has minimal effect.

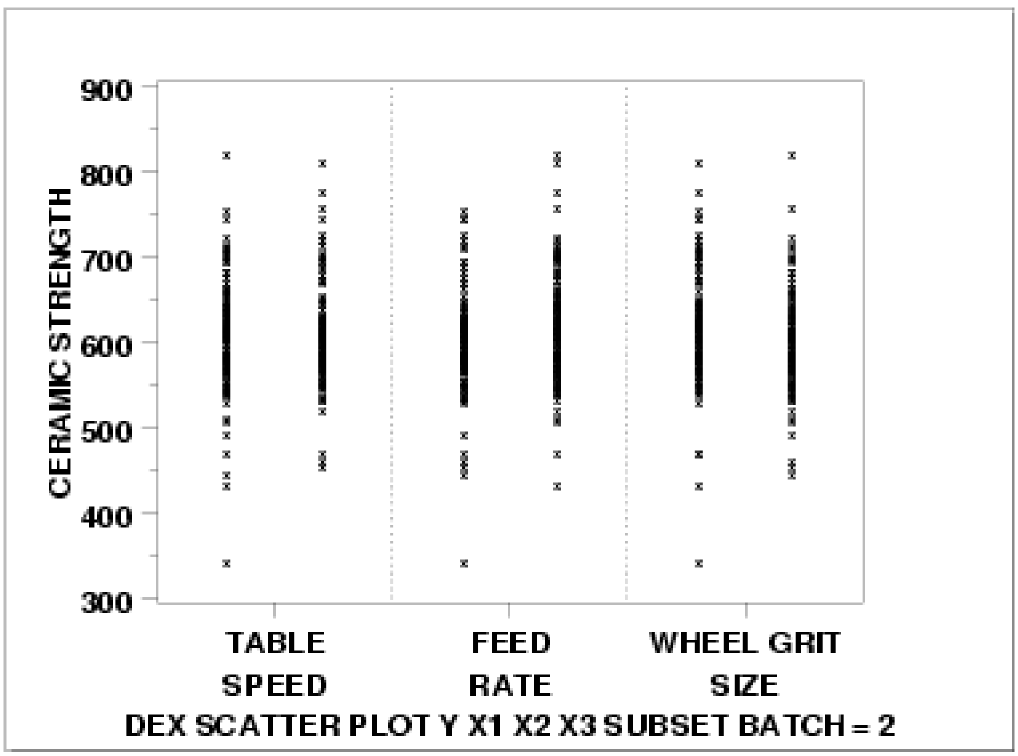
Dex SD Plot for Batch 1



This dex standard deviation plot shows the following for batch 1.

1. The table speed factor (X1) has a significant difference in variability between the levels of the factor. The difference is approximately 20 units.
2. The wheel grit factor (X3) and the feed rate factor (X2) have minimal differences in variability.

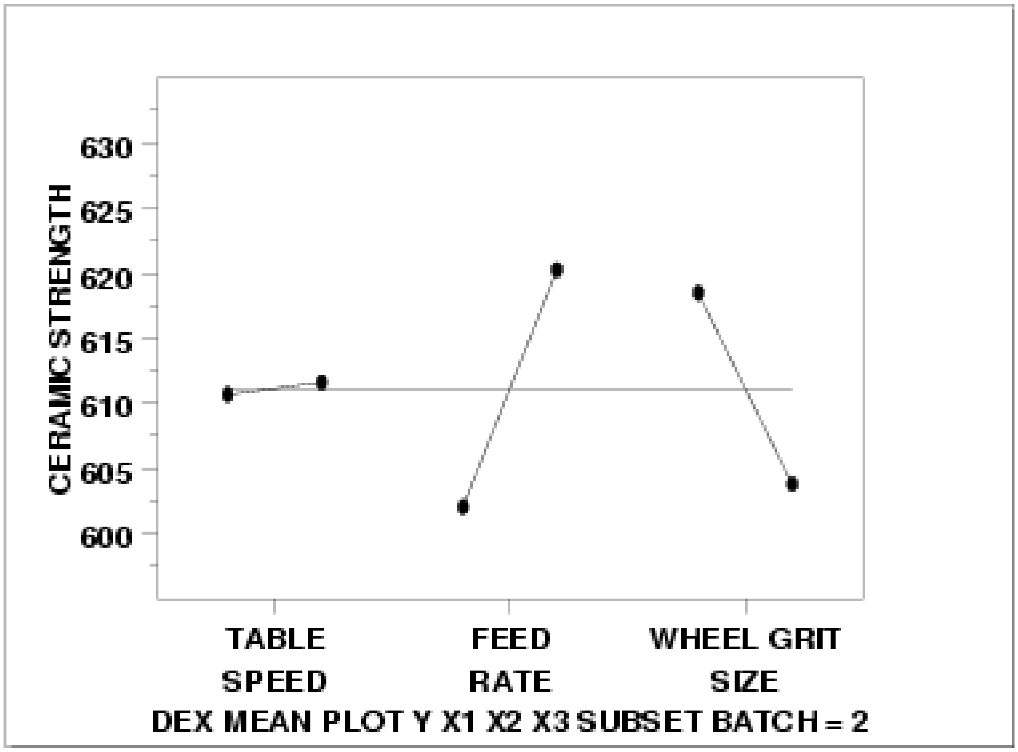
Dex Scatter Plot for Batch 2



This dex scatter plot shows the following for batch 2.

1. Most of the points are between 450 and 750.
2. There are a few outliers on both the low side and the high side.
3. Except for the outliers (i.e., the points less than 450 or greater than 750), the distribution of the points is comparable for the 3 primary factors in terms of location and spread.

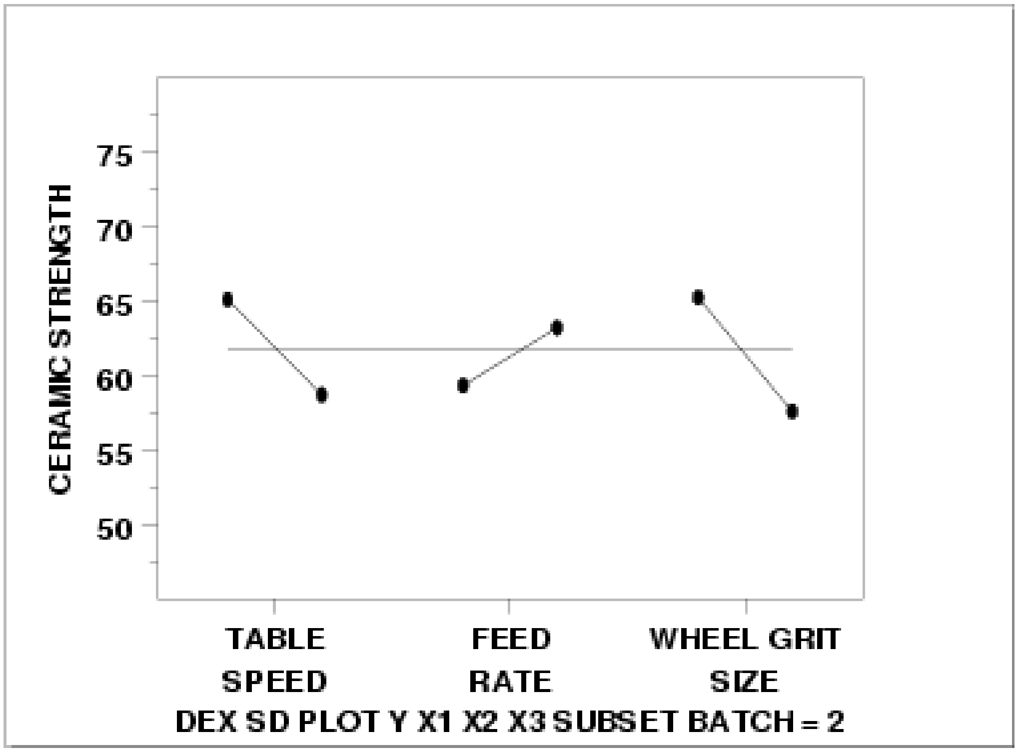
Dex Mean Plot for Batch 2



This dex mean plot shows the following for batch 2.

1. The feed rate (X2) and wheel grit (X3) factors have an approximately equal effect of about 15 or 20 units.
2. The table speed factor (X1) has a minimal effect.

Dex SD Plot for Batch 2



This dex standard deviation plot shows the following for batch 2.

1. The difference in the standard deviations is roughly comparable for the three factors (slightly less for the feed rate factor).

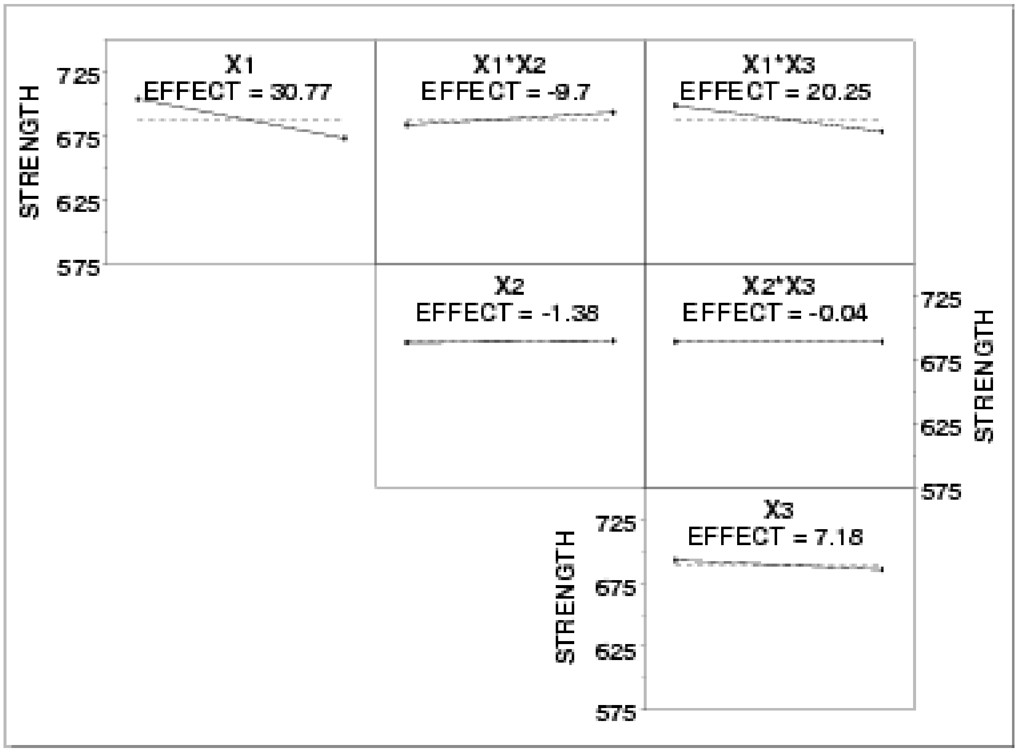
Interaction Effects

The above plots graphically show the main effects. An additonal concern is whether or not there any significant interaction effects.

Main effects and 2-term interaction effects are discussed in the chapter on Process Improvement.

In the following dex interaction plots, the labels on the plot give the variables and the estimated effect. For example, factor 1 is TABLE SPEED and it has an estimated effect of 30.77 (it is actually -30.77 if the direction is taken into account).

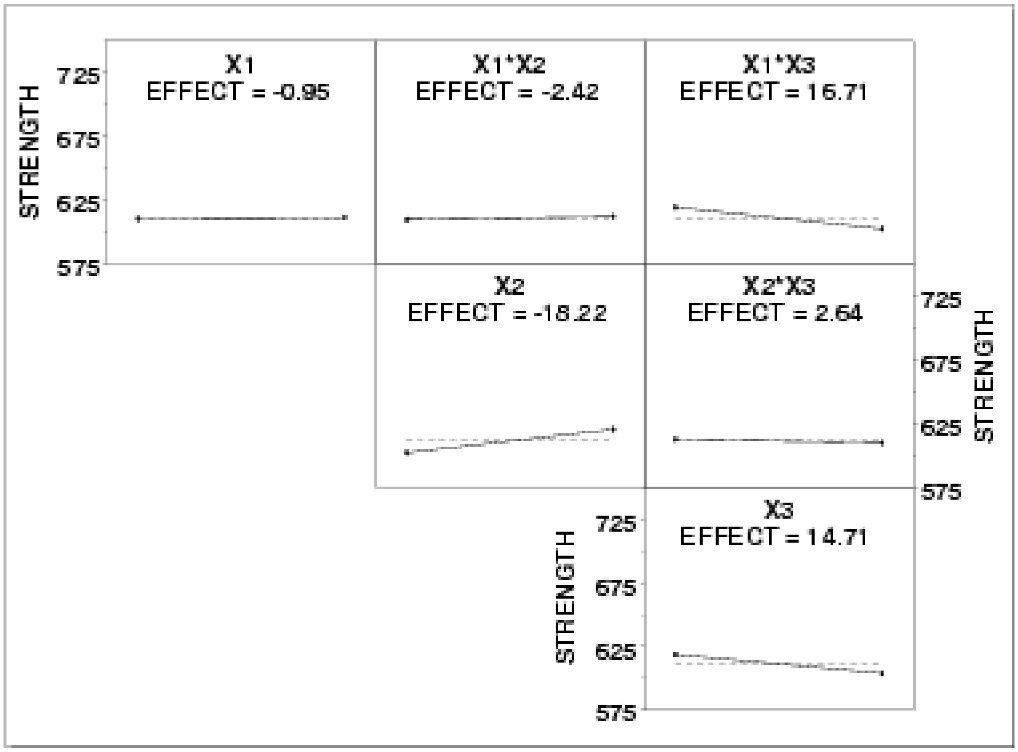
DEX Interaction Plot for Batch 1



The ranked list of factors for batch 1 is:

1. Table speed (X1) with an estimated effect of -30.77.
2. The interaction of table speed (X1) and wheel grit (X3) with an estimated effect of -20.25.
3. The interaction of table speed (X1) and feed rate (X2) with an estimated effect of 9.7.
4. Wheel grit (X3) with an estimated effect of -7.18.
5. Down feed (X2) and the down feed interaction with wheel grit (X3) are essentially zero.

DEX Interaction Plot for Batch 2



The ranked list of factors for batch 2 is:

1. Down feed (X2) with an estimated effect of 18.22.
2. The interaction of table speed (X1) and wheel grit (X3) with an estimated effect of -16.71.
3. Wheel grit (X3) with an estimated effect of -14.71
4. Remaining main effect and 2-factor interaction effects are essentially zero.

Conclusions

From the above plots, we can draw the following overall conclusions.

1. The batch effect (of approximately 75 units) is the dominant primary factor.
2. The most important factors differ from batch to batch. See the above text for the ranked list of factors with the estimated effects.

1.4.2.10.6. Work This Example Yourself

View Dataplot Macro for this Case Study

This page allows you to use Dataplot to repeat the analysis outlined in the case study description on the previous page. It is required that you have already downloaded and installed Dataplot and configured your browser. to run Dataplot. Output from each analysis step below will be displayed in one or more of the Dataplot windows. The four main windows are the Output window, the Graphics window, the Command History window, and the data sheet window. Across the top of the main windows there are menus for executing Dataplot commands. Across the bottom is a command entry window where commands can be typed in.

Data Analysis Steps	Results and Conclusions
Click on the links below to start Dataplot and run this case study yourself. Each step may use results from previous steps, so please be patient. Wait until the software verifies that the current step is complete before clicking on the next step.	*The links in this column will connect you with more detailed information about each analysis step from the case study description.*

`1. Invoke Dataplot and read data.`
 	
`1. Read in the data.`
	`1. You have read 1 column of numbers into Dataplot, variable Y.`

`2. Plot of the response variable`
 	
`1. Numerical summary of Y.`
 	
`2. 4-plot of Y.`
	`1. The summary shows the mean strength is 650.08 and the standard deviation of the strength is 74.64.`
	
	`2. The 4-plot shows no drift in the location and scale and a bimodal distribution.`

3. Determine if there is a batch effect.	
1. Generate a bihistogram based on the 2 batches.	1. The bihistogram shows a distinct batch effect of approximately 75 units.
2. Generate a q-q plot.	2. The q-q plot shows that batch 1 and batch 2 do not come from a common distribution.
3. Generate a box plot.	3. The box plot shows that there is a batch effect of approximately 75 to 100 units and there are some outliers.
4. Generate block plots.	4. The block plot shows that the batch effect is consistent across labs and levels of the primary factor.
5. Perform a 2-sample t-test for equal means.	5. The t-test confirms the batch effect with respect to the means.
6. Perform an F-test for equal standard deviations.	6. The F-test does not indicate any significant batch effect with respect to the standard deviations.
4. Determine if there is a lab effect.	
1. Generate a box plot for the labs with the 2 batches combined.	1. The box plot does not show a significant lab effect.
2. Generate a box plot for the labs for batch 1 only.	2. The box plot does not show a significant lab effect for batch 1.
3. Generate a box plot for the labs for batch 2 only.	3. The box plot does not show a significant lab effect for batch 2.

5. Analysis of primary factors.	
1. Generate a dex scatter plot for batch 1.	1. The dex scatter plot shows the range of the points and the presence of outliers.
2. Generate a dex mean plot for batch 1.	2. The dex mean plot shows that table speed is the most significant factor for batch 1.
3. Generate a dex sd plot for batch 1.	3. The dex sd plot shows that table speed has the most variability for batch 1.
4. Generate a dex scatter plot for batch 2.	4. The dex scatter plot shows the range of the points and the presence of outliers.
5. Generate a dex mean plot for batch 2.	5. The dex mean plot shows that feed rate and wheel grit are the most significant factors for batch 2.
6. Generate a dex sd plot for batch 2.	6. The dex sd plot shows that the variability is comparable for all 3 factors for batch 2.
7. Generate a dex interaction effects matrix plot for batch 1.	7. The dex interaction effects matrix plot provides a ranked list of factors with the estimated effects.
8. Generate a dex interaction effects matrix plot for batch 2.	8. The dex interaction effects matrix plot provides a ranked list of factors with the estimated effects.

1.4.3. References For Chapter 1: Exploratory Data Analysis

Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Anscombe, Francis and Tukey, J. W. (1963), *The Examination and Analysis of Residuals*, Technometrics, pp. 141-160.

Bloomfield, Peter (1976), *Fourier Analysis of Time Series*, John Wiley and Sons.

Box, G. E. P. and Cox, D. R. (1964), *An Analysis of Transformations*, Journal of the Royal Statistical Society, 211-243, discussion 244-252.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley and Sons.

Box, G. E. P., and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day.

Bradley, (1968). *Distribution-Free Statistical Tests*, Chapter 12.

Brown, M. B. and Forsythe, A. B. (1974), *Journal of the American Statistical Association*, 69, 364-367.

Chakravarti, Laha, and Roy, (1967). *Handbook of Methods of Applied Statistics, Volume I*, John Wiley and Sons, pp. 392-394.

Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), *Graphical Methods for Data Analysis*, Wadsworth.

Cleveland, William (1985), *Elements of Graphing Data*, Wadsworth.

Cleveland, William and Marylyn McGill, Editors (1988), *Dynamic Graphics for Statistics*, Wadsworth.

Cleveland, William (1993), *Visualizing Data*, Hobart Press.

Devaney, Judy (1997), *Equation Discovery Through Global Self-Referenced Geometric Intervals and Machine Learning*, Ph.d thesis, George Mason University, Fairfax, VA. Coefficient Test for Normality , Technometrics, pp. 111-117.

Draper and Smith, (1981). *Applied Regression Analysis, 2nd ed.*, John Wiley and Sons.

du Toit, Steyn, and Stumpf (1986), *Graphical Exploratory Data Analysis*, Springer-Verlag.

Evans, Hastings, and Peacock (2000), *Statistical Distributions*, 3rd. Ed., John Wiley and Sons.

Everitt, Brian (1978), *Multivariate Techniques for Multivariate Data*, North-Holland.

Efron and Gong (February 1983), *A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation*, The American Statistician.

Filliben, J. J. (February 1975), *The Probability Plot Correlation Coefficient Test for Normality* , Technometrics, pp. 111-117.

Gill, Lisa (April 1997), *Summary Analysis: High Performance Ceramics Experiment to Characterize the Effect of Grinding Parameters on Sintered Reaction Bonded Silicon Nitride, Reaction Bonded Silicon Nitride, and Sintered Silicon Nitride* , presented at the NIST - Ceramic Machining Consortium, 10th Program Review Meeting, April 10, 1997.

Granger and Hatanaka (1964). *Spectral Analysis of Economic Time Series*, Princeton University Press.

Grubbs, Frank (February 1969), *Procedures for Detecting Outlying Observations in Samples*, Technometrics, Vol. 11, No. 1, pp. 1-21.

Harris, Robert L. (1996), *Information Graphics*, Management Graphics.

Jenkins and Watts, (1968), *Spectral Analysis and Its Applications*, Holden-Day.

Johnson, Kotz, and Balakrishnan, (1994), *Continuous Univariate Distributions, Volumes I and II*, 2nd. Ed., John Wiley and Sons.

Johnson, Kotz, and Kemp, (1992), *Univariate Discrete Distributions*, 2nd. Ed., John Wiley and Sons.

Kuo, Way and Pierson, Marcia Martens, Eds. (1993), *Quality Through Engineering Design"*, specifically, the article Filliben, Cetinkunt, Yu, and Dommenz (1993), *Exploratory Data Analysis Techniques as Applied to a High-Precision Turning Machine*, Elsevier, New York, pp. 199-223.

Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, pp. 278-292.

McNeil, Donald (1977), *Interactive Data Analysis*, John Wiley and Sons.

Mosteller, Frederick and Tukey, John (1977), *Data Analysis and Regression*, Addison-Wesley.

Nelson, Wayne (1982), *Applied Life Data Analysis*, Addison-Wesley.

Neter, Wasserman, and Kunter (1990). *Applied Linear Statistical Models, 3rd ed.*, Irwin.

Nelson, Wayne and Doganaksoy, Necip (1992), *A Computer Program POWNOR for Fitting the Power-Normal and -Lognormal Models to Life or Strength Data from Specimens of Various Sizes*, NISTIR 4760, U.S. Department of Commerce, National Institute of Standards and Technology.

The RAND Corporation (1955), *A Million Random Digits with 100,000 Normal Deviates*, Free Press.

Ryan, Thomas (1997). *Modern Regression Methods*, John Wiley.

Scott, David (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization* , John Wiley and Sons.

Snedecor, George W. and Cochran, William G. (1989), *Statistical Methods, Eighth Edition*, Iowa State University Press.

Stefansky, W. (1972), *Rejecting Outliers in Factorial Designs*, Technometrics, Vol. 14, pp. 469-479.

Stephens, M. A. (1974). *EDF Statistics for Goodness of Fit and Some Comparisons*, Journal of the American Statistical Association, Vol. 69, pp. 730-737.

Stephens, M. A. (1976). *Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters*, Annals of Statistics, Vol. 4, pp. 357-369.

Stephens, M. A. (1977). *Goodness of Fit for the Extreme Value Distribution*, Biometrika, Vol. 64, pp. 583-588.

Stephens, M. A. (1977). *Goodness of Fit with Special Reference to Tests for Exponentiality* , Technical Report No. 262, Department of Statistics, Stanford University, Stanford, CA.

Stephens, M. A. (1979). *Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function*, Biometrika, Vol. 66, pp. 591-595.

Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley.

Tufte, Edward (1983), *The Visual Display of Quantitative Information*, Graphics Press.

Velleman, Paul and Hoaglin, David (1981), *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury.

Wainer, Howard (1981), *Visual Revelations*, Copernicus.