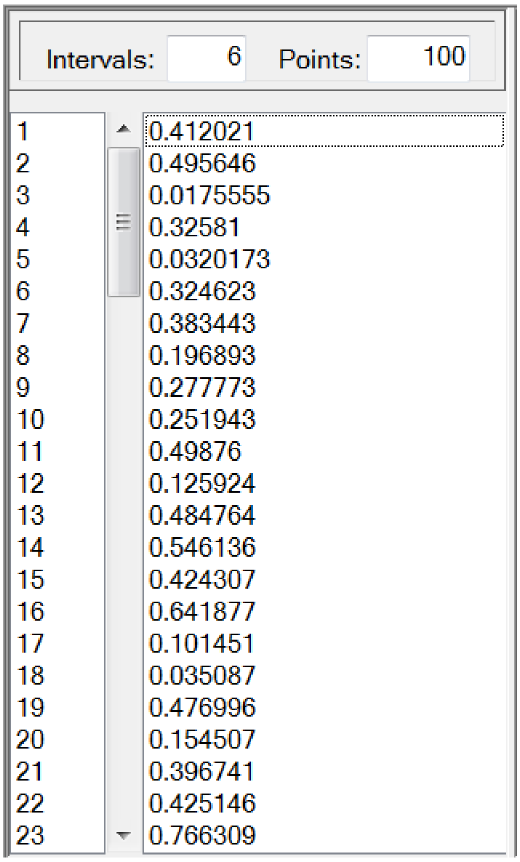
Stat::Fit®

Statistically Fit®
Software

Table of Contents

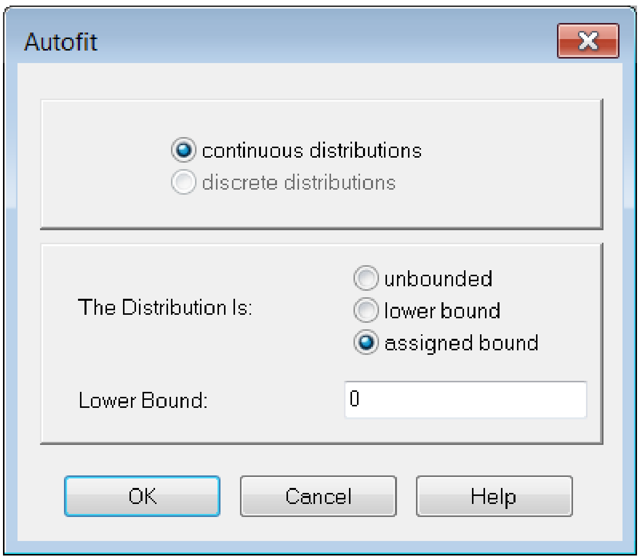
Getting Started

Stat::Fit opens with an empty Data Table. The easiest way to enter data is through the paste function usually from a spreadsheet. However, Stat::Fit can also read a text file if the numbers are delimited. The number of points is shown at the top as well the default number of intervals used for graphing.



Next the Autofit function (in the FIT menu or on the Tool Bar) should be used to fit the data to analytical distributions. Two choices must be made in the Autofit dialog:

1. Is this continuous or discrete data?
2. Is the distribution bounded and if so should a lower bound be fixed?

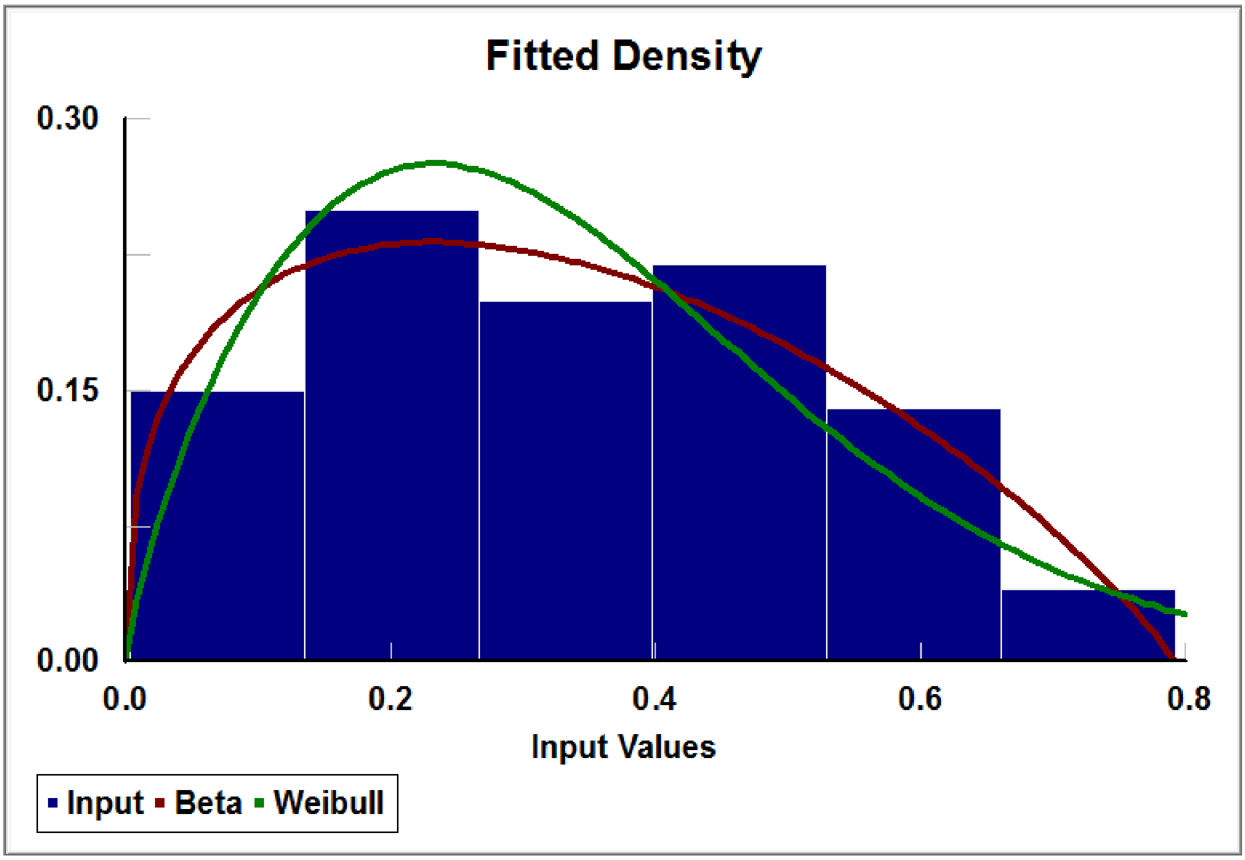


The calculations of the parameter fits and the subsequent tests for goodness of fit may take some time for 10000 data points, but are usually quick for 100 data points. The result is given in an Autofit view.

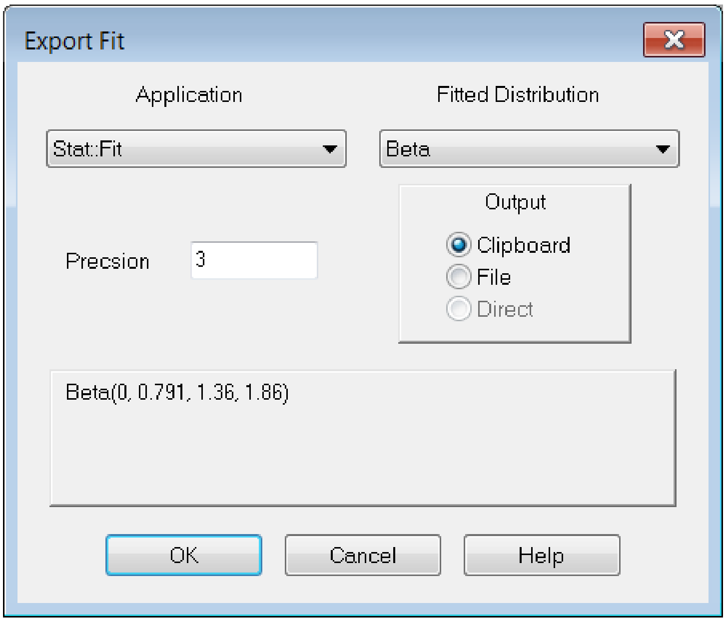
autofit of distributions

distribution	rank	acceptance
Beta(0, 0.791, 1.36, 1.86)	100	do not reject
Weibull(0, 1.75, 0.378)	43.7	do not reject
Johnson SB(0, 0.818, 0.35, 0.738)	37.1	do not reject
Rayleigh(0, 0.275)	31.8	do not reject
Triangular(0, 0.855, 0.111)	16.2	do not reject
Pearson 6(0, 1.12e+005, 2.12, 6.98e+005)	8.78	do not reject
Gamma(0, 2.12, 0.16)	8.77	do not reject
Erlang(0, 2, 0.17)	6.93	do not reject
Loglogistic(0, 2.29, 0.298)	2.88	reject
Lognormal(0, -1.33, 0.888)	0.0631	reject
Power Function(0, 0.791, 0.91)	0.0339	reject
Uniform(0, 0.791)	0	reject
Inverse Gaussian(0, 0.177, 0.34)	0	reject
Inverse Weibull(0, 0.752, 6.34)	0	reject
Pearson 5(0, 0.733, 0.0852)	0	reject
Exponential(0, 0.34)	0	reject
Chi Squared(0, 0.975)	0	reject
Pareto	no fit	reject

This view is sensitive to mouse clicks on the various distributions, and provides a quick view of the relevant density, both data and fit. Multiple distributions can be plotted over the input data.



At this point, the fitted distribution can be input into the various simulation products by using the export dialog to get the proper format. The export dialog is available in the File menu or on the Tool Bar.

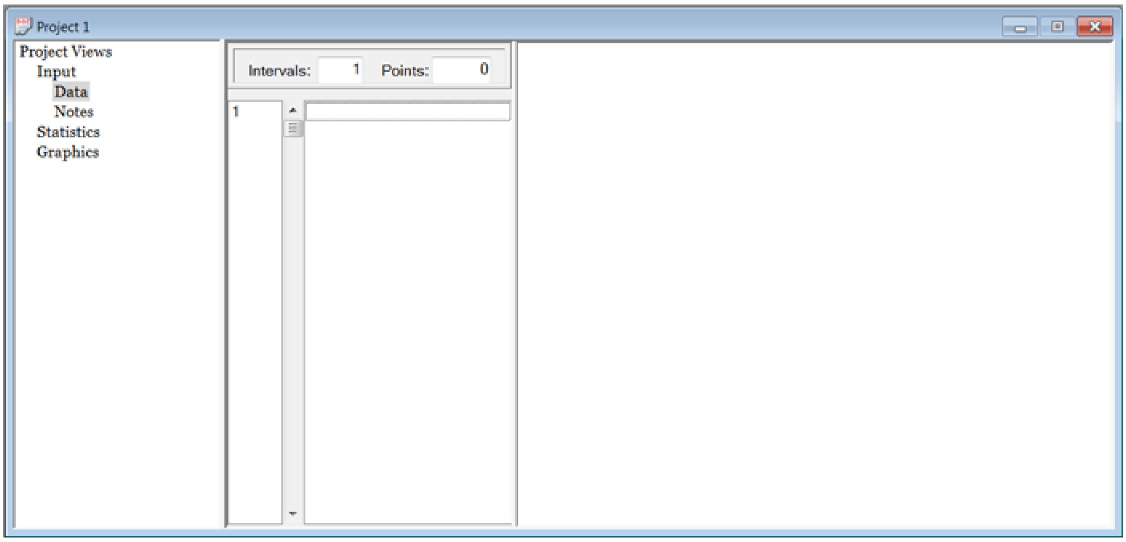


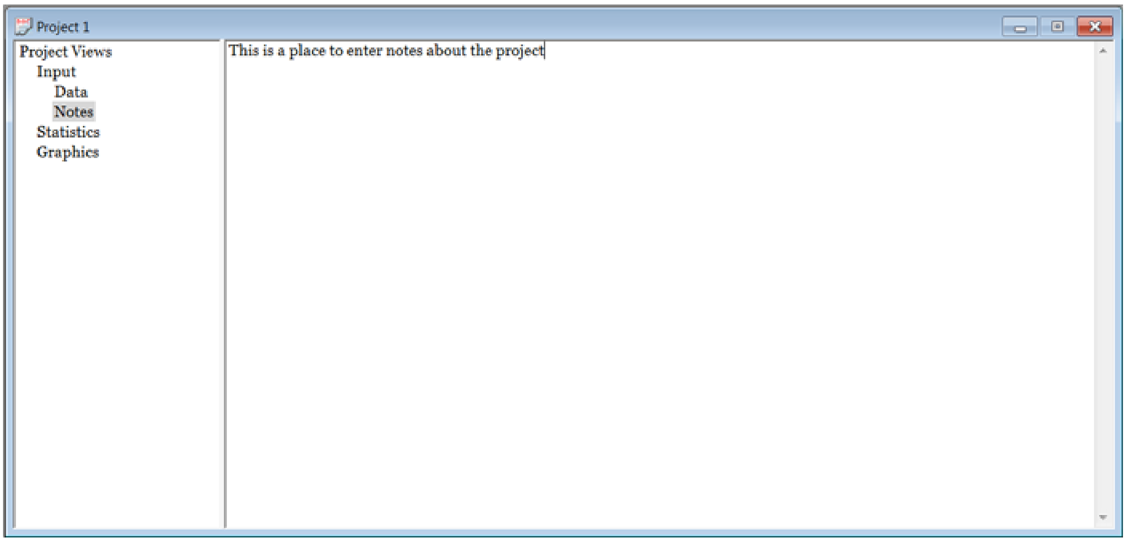
The exported distribution in the proper format can be put on the clipboard for pasting or saved in a file. Some simulation products allow direct transfer.

Please take care that the distribution actually represents the data, especially at the lower bound. Remember to ask whether the physical process is finite at the lower bound, or goes to zero or infinity.

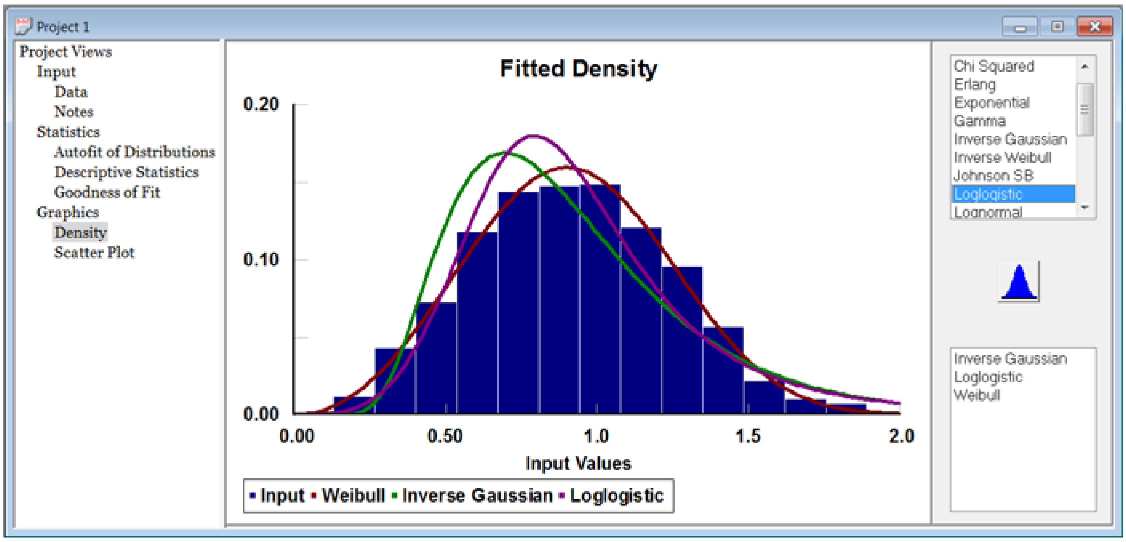
Trees And Notes

Stat::Fit uses a document structure for every project. The views on that document can be accessed from the menu, and sometimes the Tool Bar. When accessed, the views will appear in the project tree. Initially the project opens with an empty Data Table and a Notes area where project notes can be stored, both accessible by clicking on the view tree. Both can also be accessed from the Input menu.





As the project progresses, views are added to the tree, and each one can be viewed by clicking on it in the tree. As the data and setup are changed, all views are updated so that the next time a view is accessed, it will reflect the changes.



Some views may be closed if the assumptions that generated them are changed to exclude them.

Fitting distributions to user data can be done in semi-automated fashion with the Autofit command or manually by specifying each fitting parameter in turn.

Automatic fitting of continuous or discrete analytical distributions can be performed by using the Fit|Autofit command. This command follows the same procedure as manual fitting, but chooses distributions appropriate for the input data. It also ranks the distributions according to their relative goodness of fit, and gives an indication of their acceptance as good representations of the input data. For large data sets, this function can take some time to complete its calculations.

The manual fitting of analytical distributions to the input data in the Data Table takes three steps. First, distributions appropriate to the input data must be chosen in the Fit|Setup dialog along with the desired goodness of fit tests. Then, estimates of the parameters for each chosen distribution must be calculated by using either the Fit|Moment Estimates equations or the Fit|Maximum Likelihood equations. Finally, the Fit|Goodness of Fit tests are calculated for each fitted distribution in order to ascertain the relative goodness of fit. (see Breiman[1], Law&Kelton[2], Banks&Carson[3], Stuart&Ord[4])

The tests for goodness of fit are merely comparisons of the input data to the fitted distributions in a statistically significant manner. Each test makes the hypothesis that the fit is good and calculates a test statistic for comparison to a standard. The Goodness of Fit tests include:

- Chi Squared Test
- Kolmogorov Smirnov Test
- Anderson Darling Test

If the choice of test is uncertain, even after consulting the descriptions, use the Kolmogorov Smirnov test which is applicable over the widest range of data and fitted parameters.

Sometimes the user has more than 10,000 data points. In this case, the analytical distributions can be fit, but will frequently fail the goodness of fit tests. If the user data may not cover the expected range of the variate, especially in the tail of the distribution, then the best choice is the analytical distribution that visually fits the data. Splitting the large data set into subsets, and fitting each, can give a good idea of the proper distribution as well as the variation of the fitted parameters.

However, if the user data is sufficient to provide data over the expected range of the parameter, then an empirical distribution can be used. Most simulation software allows for empirical distributions, but they must be calculated from the raw data. That calculation can be done and exported through the File|Export Empirical dialog.

Sometimes the user has very little data, or is limited to estimates of range and expert opinion. In this situation, several heuristic guides can be used to get a good first approximation.

1. "Statistics: With a View Toward Applications", Leo Breiman, 1973, Houghton Mifflin

2. "Discrete-Event System Simulation", Jerry Banks, John S. Carson II, 1984, Prentice-Hall

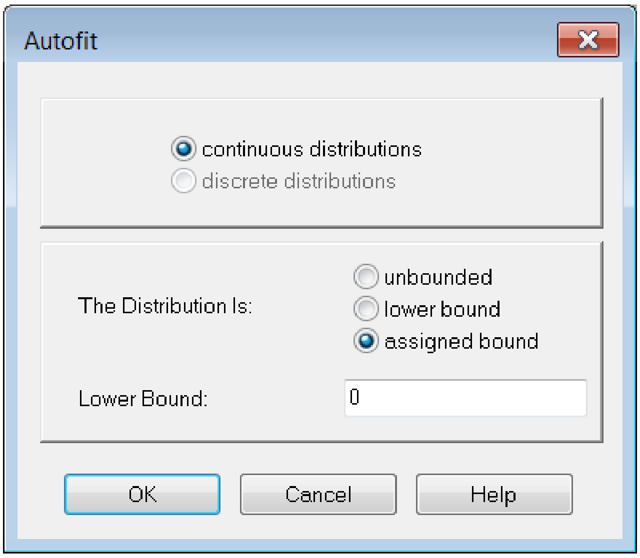
3. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill

4. "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart, J. Keith Ord, 1991, Oxford University Press

Fit|Autofit

The Autofit function will automatically choose appropriate continuous or discrete distributions to fit to the input data, calculate Maximum Likelihood Estimates for those distributions, test the results for Goodness of Fit, and display the distributions in order of their relative rank. The relative rank is determined by an empirical method which uses effective goodness of fit calculations. While a good rank usually indicates that the fitted distribution is a good representation of the input data, an absolute indication of the goodness of fit is also given by testing for outliers.

The Autofit function can be initiated from the menu (Fit|Autofit) or from the Tool Bar. The Autofit Dialog asks the user to choose certain parameters for the fit.



The first choice is between continuous and discrete distributions for the fit. The choice of discrete distributions is only possible if the data is integer.

The second choice is between various bounds. (Discrete distributions must have a lower bound.) Continuous distributions can be fit with an assigned lower bound in which case the default is listed as the lowest integer below the data. The user can change this value, but it must remain below the lowest data point. If lower bound is chosen, the lower bound is one of the parameters determined by the fit. In both cases, the list of possible distributions is limited to bounded distributions.

Continuous distributions can also be fit with unbounded distributions in addition to the set of bounded distributions. For instance, the Normal distribution would not be included unless unbounded is checked.

Finally, the list of possible distributions may be limited by the skewness of the data; some distributions cannot be fit to data with negative skewness.

The result is the Autofit view which list all the distributions fit, their parameters, and rank. This acceptance of fit usually reflects the results of the goodness of fit tests at a nominal level of significance. However the acceptance may also be modified if the fitted distribution would generate significantly more data points in the tails of the distribution than are indicated by the input data. In other words, an empirical test is done to make sure the distribution would not over represent the data in the tails.

autofit of distributions

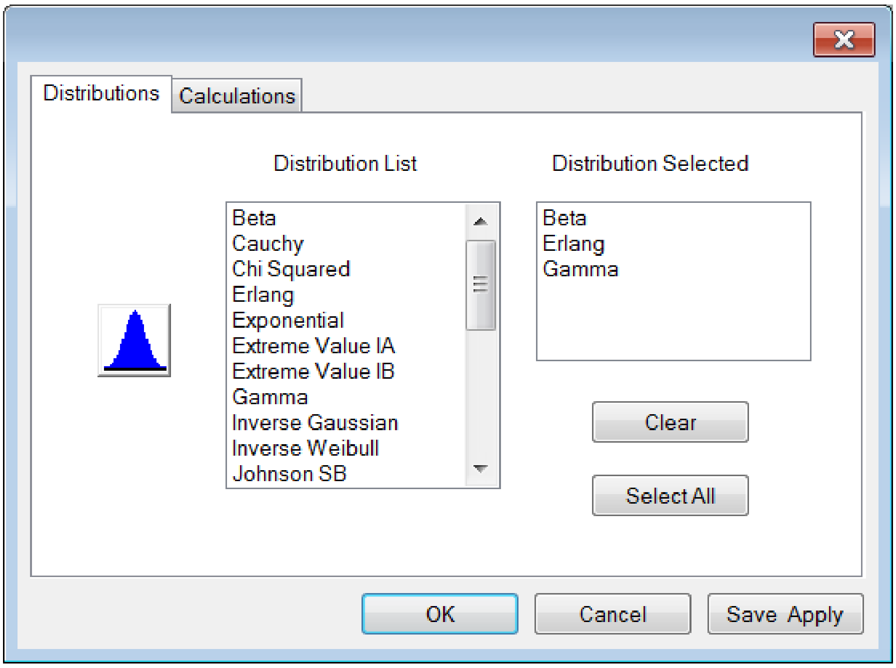
distribution	rank	acceptance
Beta(0, 0.791, 1.36, 1.86)	100	do not reject
Weibull(0, 1.75, 0.378)	43.7	do not reject
Johnson SB(0, 0.818, 0.35, 0.738)	37.1	do not reject
Rayleigh(0, 0.275)	31.8	do not reject
Triangular(0, 0.855, 0.111)	16.2	do not reject
Pearson 6(0, 1.12e+005, 2.12, 6.98e+005)	8.78	do not reject
Gamma(0, 2.12, 0.16)	8.77	do not reject
Erlang(0, 2, 0.17)	6.93	do not reject
Loglogistic(0, 2.29, 0.298)	2.88	reject
Lognormal(0, -1.33, 0.888)	0.0631	reject
Power Function(0, 0.791, 0.91)	0.0339	reject
Uniform(0, 0.791)	0	reject
Inverse Gaussian(0, 0.177, 0.34)	0	reject
Inverse Weibull(0, 0.752, 6.34)	0	reject
Pearson 5(0, 0.733, 0.0852)	0	reject
Exponential(0, 0.34)	0	reject
Chi Squared(0, 0.975)	0	reject
Pareto	no fit	reject

The Autofit function forces the setup of the document so that only continuous or only discrete distributions will be used, and both the Kolmogorov Smirnov test and the Anderson Darling test (Chi-Squared test for discrete distribution) will be calculated, and the Maximum Likelihood Estimates will be used. Because the Autofit function requires a specific setup, the Autofit view will be closed if the setup is changed.

Fit|Setup

The Fit Setup dialog provides a list of all distributions supported by Stat::Fit and the relevant choices for goodness of fit tests. At least one distribution must be chosen before the estimate, test, and graphing commands become available. Please note that the distributions shown may be limited by the data set; discrete distributions require integer data.

The Distribution page of the Fit Setup dialog provides a distribution list for the choice of distributions for subsequent fitting. All distributions chosen here will be used sequentially for estimates and goodness of fit tests. Clicking on a distribution name in the distribution list on the left chooses that distribution and moves that distribution name to the distributions selected box on the right unless it is already there. Clicking on the distribution name in the distributions selected box on the right removes the distribution. All distributions may be moved to the distributions selected box by clicking the Select All button. The distribution selected box may be cleared by clicking the Clear button.



The Calculations page of the Fit Setup dialog allows the appropriate goodness of fit tests to be chosen, along with supporting parameters.

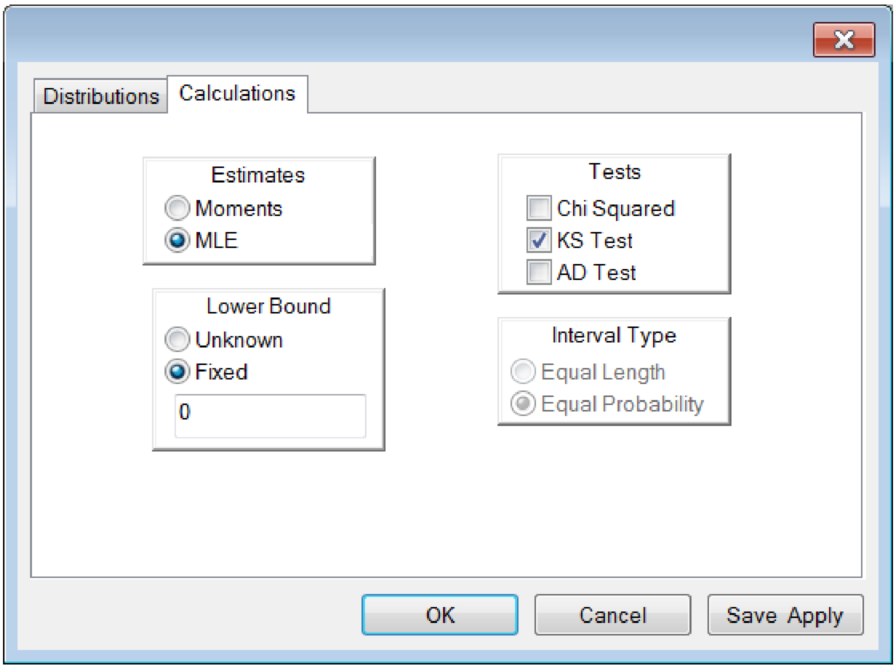
All goodness of fit tests use either the moments estimates or the maximum likelihood estimates for each selected distribution as set here. This does not prevent either set of estimates from being calculated, but designates which estimates will be used for the goodness of fit tests and result graphs. The default is the maximum likelihood estimates because they give statistically better results.

For continuous distributions with a lower bound or minimum such as the Exponential, the lower bound can be forced to assume a value at or below the minimum data value. This lower bound will be used for both the moments and maximum likelihood estimates. By default, it is pinned at the lowest integer below the minimum data point. If new data is added below a preset lower bound, the bound will be modified to assume the closest integer value below all input data. An unknown lower bound will allow the fitting routines to search for a statistically proper value.

The Level of Significance used for all goodness of fit tests is set, by default, to 0.05.

The goodness of fit tests include the Chi Squared test, the Kolmogorov Smirnov test, and the Anderson Darling test. The test selection is set, by default, to the Kolmogorov Smirnov test. The

Chi Squared test may be done with intervals of equal length or equal probability. The interval type is set to equal probability, by default, because this test gives more consistent results with peaked distributions.



Stat::Fit does not start the fitting process when this dialog is exited, but turns on the appropriate commands and waits for further instruction.

At any time the default Fit Setup dialog can be changed by checking the Save To Default check box at the bottom of the dialog and clicking on OK. This default configuration is loaded in the Fit Setup dialog when the program starts.

Fit|Moment Estimates

The Moment Estimates command can be accessed from the Fit menu. It is activated only after specific distributions have been chosen and sufficient data has been entered.

When the Moment Estimates command is chosen, the estimates of the parameters for all analytical distributions chosen in the Fit Setup dialog are calculated using the moment equations for each choice along with the sample moments from calculations on the input data in the Data Table. The parameters thus estimated are displayed in a new view. An example view shows the moment estimates for three distributions to a set of data points where the lower bound was not fixed.

```
        method of moments estimates

    Beta
        minimum     =   -0.0241336
        maximum     =   0.879844
        p           =   1.77742
        q           =   2.63967

    Gamma
        minimum     =   -0.979541
        alpha       =   47.9713
        beta        =   0.027499

    Weibull
        minimum     =   -0.131756
        alpha       =   2.66544
        beta        =   0.530295
```

Some distributions do not have moment estimates for given ranges of sample moments. This is especially evident for many of the bounded continuous distributions when the sample skewness is negative. When such situations occur, an error message rather than the parameters will be displayed with the name of the analytical distribution.

Some distributions, such as Cauchy, do not have moments at all.

The moment estimates will be used for Result Graphs and Goodness of Fit tests if they have been selected on the Calculation page of the Fit Setup dialog. At any time the choice of estimates is changed, all visible views of the Result Graphs and the Goodness of Fit tests are redisplayed with the new choice of estimates.

The moment estimates have been included as an aid to the fitting process; except for the simplest distributions, they do NOT give good estimates of the parameters of a fitted distribution.

Fit|Maximum Likelihood

When the Maximum Likelihood command is chosen, the maximum likelihood estimates of the parameters for all analytical distributions chosen in the Fit Setup dialog are calculated using the log likelihood equation and its derivatives for each choice; these equations are various summations on the input data in the Data Table. The parameters thus estimated are displayed in a new view.

Some distributions do not have maximum likelihood estimates for given ranges of sample moments because initial estimates of the distribution's parameters are unreliable. This is especially evident for many of the bounded continuous distributions when the sample skewness is negative. When such situations occur, an error message rather than the parameters will be displayed with the name of the analytical distribution. An example view shows the maximum likelihood estimates for three distributions to a set of data points where the lower bound was not fixed.

maximum likelihood estimates

Beta

minimum	=	0.00325685
maximum	=	0.79062
p	=	1.4511
q	=	1.94848

Gamma

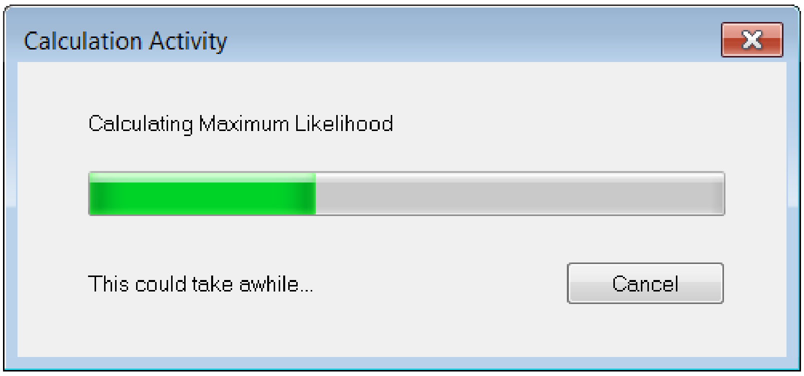
minimum	=	-0.838343
alpha	=	39.0199
beta	=	0.0301943

Weibull

minimum	=	-0.0405016
alpha	=	2.10871
beta	=	0.429027

The Maximum Likelihood estimates will be used for Result Graphs and Goodness of Fit tests if they have been selected on the Calculation page of the Fit Setup dialog. [the default condition] At any time the choice of estimates is changed, all visible views of the Result Graphs and the Goodness of Fit tests are redisplayed with the new choice of estimates.

Many of the Maximum Likelihood Estimates require significant calculation, and, therefore, significant time. Because of this, a Cancel dialog, shown below, will appear with each calculation. If The Cancel button is clicked, the calculations will cease at the next distribution and an error message will be displayed. Note that all chosen maximum likelihood estimates must be finished before the Result Graphs can be displayed or the Goodness of Fit tests can be done.


Calculation Activity
Calculating Maximum Likelihood
This could take awhile...
Cancel

Fit|Goodness of Fit

The Goodness of Fit command can be accessed in the Fit menu. It is activated only after specific distributions have been chosen.

The Goodness of Fit command calculates the goodness of fit tests, as chosen on the Calculation page of the Fit Setup dialog, for each of the analytical distributions chosen. These calculations, along with the distributions and their fitted parameters, are then shown in the Goodness of Fit view. If the Goodness of Fit command is chosen before the moments or maximum likelihood estimates have been calculated, then these calculations precede the goodness of fit calculations as necessary, but are not displayed in a separate view.

The goodness of fit tests include the Chi Squared Test, Kolmogorov Smirnov Test, and the Anderson Darling Test. All the goodness of fit tests report a REJECT or DO NOT REJECT decision based on the a comparison between a calculated test statistic and a test statistic specific to the goodness of fit test and the level of significance.

The Chi Squared Test is a test of the goodness of fit of the fitted density to the input data in the Data Table, that data appropriately separated into intervals (continuous data) or classes (discrete data). The data can be separated into intervals of equal length or equal probability by selecting the appropriate check box in the Calculation page of the Fit Setup dialog.

The Kolmogorov Smirnov Test is a test of the goodness of fit of the fitted cumulative distribution to the input data in the Data Table, point by point.

The Anderson Darling Test is a test for the goodness of fit of the fitted cumulative distribution to the input data in the Data Table, by point pairs, weighted to make the tails of the distribution more sensitive.

goodness of fit

data points	100
estimates	maximum likelihood estimates
accuracy of fit	0
level of significance	0.05

summary

distribution	Chi Squared	Kolmogorov Smirnov	Anderson Darling
Beta	0.32 (5)	0.0451	0.255
Gamma	4.4 (5)	0.0874	1.39
Weibull	4.04 (5)	0.0732	0.589

detail

Beta

minimum	=	0 [fixed]
maximum	=	0.79062
p	=	1.35992
q	=	1.86041

Chi Squared

total classes	6
interval type	equal probable
net bins	6
chi**2	0.32
degrees of freedom	5
alpha	0.05
chi**2(5,0.05)	11.1
p-value	0.997
result	DO NOT REJECT

Kolmogorov-Smirnov

data points	100
ks stat	0.0451
alpha	0.05
ks stat(100,0.05)	0.134
p-value	0.981
result	DO NOT REJECT

Anderson-Darling

data points	99
ad stat	0.255
alpha	0.05
ad stat(0.05)	2.49
p-value	0.968
result	DO NOT REJECT

Gamma

minimum	=	0 [fixed]
alpha	=	2.12419
beta	=	0.159884

Chi Squared

total classes	6
interval type	equal probable
net bins	6
chi**2	4.4
degrees of freedom	5
alpha	0.05
chi**2(5,0.05)	11.1
p-value	0.493
result	DO NOT REJECT

Kolmogorov-Smirnov

data points	100
ks stat	0.0874
alpha	0.05

Chi Squared Test

The Chi Squared test is a test of the goodness of fit of the fitted density to the input data in the Data Table, that data appropriately separated into intervals (continuous data) or classes (discrete data). This test can be chosen in the Calculation page of the Fit Setup dialog.

The test starts with the observed data in classes (intervals). While the number of classes for discrete data is set by the range of the integers, the choice of the appropriate number of intervals for continuous data is not well determined. Stat::Fit has an automatic calculation which chooses the least number of intervals which does not over smooth the data, appropriately adjusted for skewness. An empirical rule of some popularity, Sturges' rule, can also be used, or two other algorithms. If neither appears satisfactory, the number of intervals may be set manually. The intervals are set by checking the appropriate box in the Input Options dialog of the Input menu.

The test then calculates the expected value for each interval from the fitted distribution, where the expected values of the end intervals include the sum or integral to infinity(+-) or the nearest bound. In order to make the test valid, intervals(classes) with less than 5 data points are joined to neighbors until remaining intervals have at least 5 data points. Then the Chi Squared statistic for this data is calculated according to the equation:

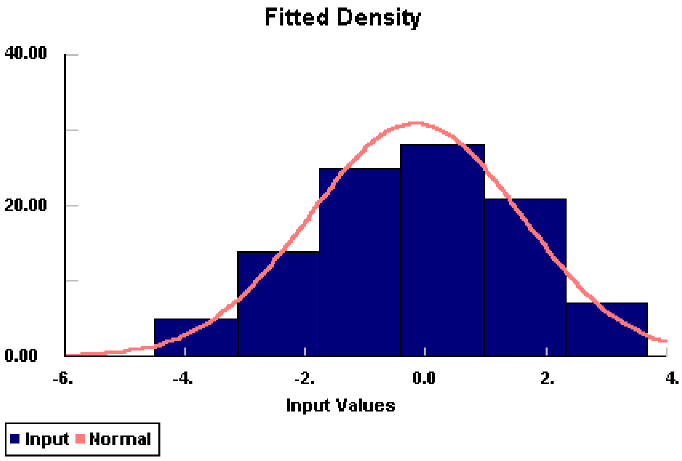
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

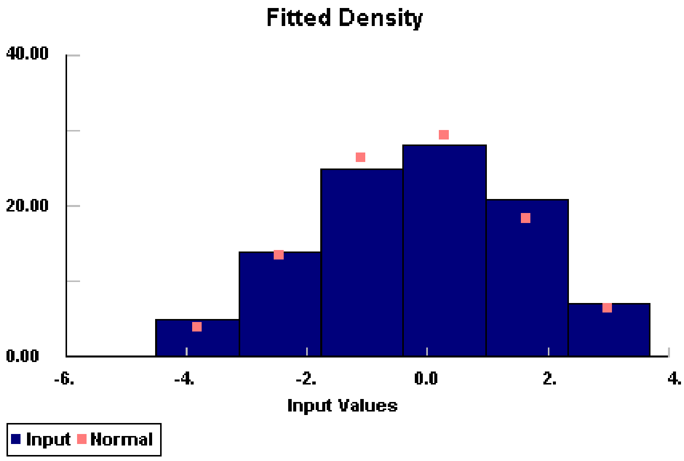
where χ^2 is the chi squared statistic, n is the total number of data points, n_i is the number of data points in the ith continuous interval or ith discrete class, k is the number of intervals or classes used, and p_i is the expected probability of occurrence in the interval or class for the fitted distribution.

The resulting test statistic is then compared to a standard value of Chi Squared with the appropriate number of degrees of freedom and level of significance, usually labeled alpha. In Stat::Fit, the number of degrees of freedom is always taken to be the net number of data bins [intervals, classes] used in the calculation minus 1, because this is the most conservative test, that is, least likely to reject the fit in error. The actual number of degrees of freedom is somewhere between this number and a similar number reduced by the number of parameters fitted by the estimating procedure. While Chi Squared test is an asymptotic test which is valid only as the number of data points gets large, it may still be used in the comparative sense. (see Law&Kelton1 ,Brunk2, Stuart&Ord3)

The goodness of fit view also reports a REJECT or DO NOT REJECT decision for each Chi Squared test based on the the comparison between the calculated test statistic and the standard statistic for the given level of significance.

To visualize this process for continuous data, consider the two graphs below. The first is the normal comparison graph of the histogram of the input data versus a continuous plot of the fitted density. Note that the frequency not the relative frequency is used; this is the actual number of data points per interval. However, for the Chi Squared test, the comparison is made between the histogram and the value of the area under the continuous curve between each interval end point. This is represented in the second graph by comparing the observed data, the top of each histogram interval, with the expected data shown as square points. Notice that the interval near 6 has fewer than 5 as an expected value and would be combined with the adjacent interval for the calculation. The result is the sum of the normalized square of the error for each interval.





In this case, the data were separated into intervals of equal length. This magnifies any error in the center interval which has more data points and a larger difference from the expected value. An alternative, and more accurate way, to separate the data is to chose intervals with equal probability so that the expected number of data points in each interval is the same. Now the resulting intervals are NOT equal length, in general, but the errors are of the same relative size for each interval. This equal probable technique gives a better test, especially with highly peaked data. The Chi Squared test can be calculated with intervals of equal length or equal probability by selecting the appropriate check box in the Calculation page of the Fit Setup dialog. The equal probable choice is the default.

While the test statistic can be useful, the p-value is more useful in determining the goodness of fit. The p-value is defined as the probability that another sample will be as unusual as the current sample given that the fit is appropriate. A small p-value indicates that the sample is highly unlikely, and, therefore, the fit should be rejected. Conversely, a high p-value indicates that the sample is likely and would be repeated and, therefore, the fit should not be rejected. Thus, the HIGHER the p-value, the more likely the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-value is likely to be the better fit.

1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 382

2. "An Introduction to Mathematical Statistics", H. D. Brunk, 1960, Ginn&Co., p261

3. "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart & J. Keith Ord, 1991, Oxford University Press, p1159

Kolmogorov Smirnov Test

The Kolmogorov Smirnov test is a statistical test of the goodness of fit of the fitted cumulative distribution to the input data in the Data Table, point by point. This test can be chosen in the Calculation page of the Fit Setup dialog

The Kolmogorov Smirnov test (KS) calculates the largest absolute difference between the cumulative distributions for the input data and for the fitted distribution according to the equations:

$$D = \max(D^+, D^-)$$

$$D^+ = \max\left(\frac{i}{n} - F(x)\right), \quad i = 1, \ldots, n$$

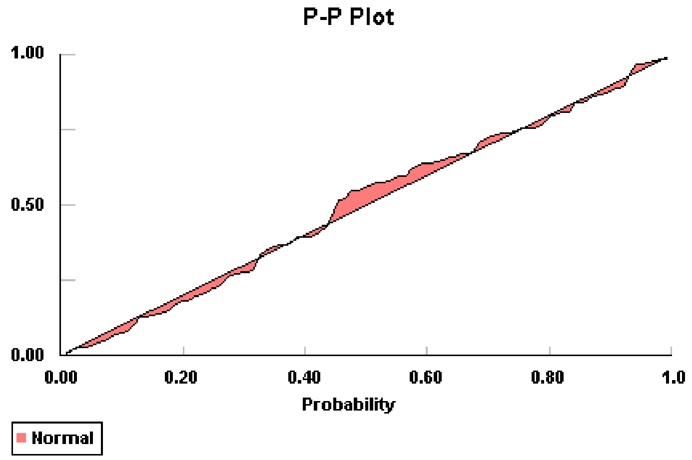
$$D^- = \max\left(F(x) - \frac{(i-1)}{n}\right), \quad i = 1, \ldots, n$$

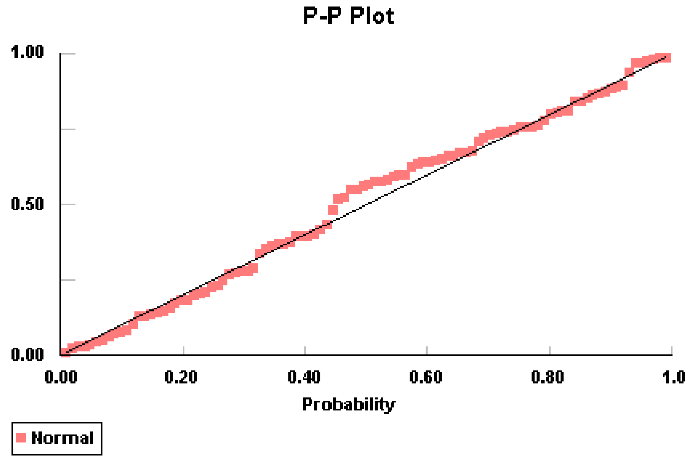
where D is the KS statistic, x is the value of the ith point out of n total data points, and F(x) is the fitted cumulative distribution. Note that the difference is determined separately for positive and negative discrepancies on a point by point basis.

The resulting test statistic is then compared to a standard value of KS statistic with the appropriate number of data points and level of significance, usually labeled alpha. The KS test is not a limiting distribution; it is appropriate for any sample size. While the KS test is only valid if none of the parameters in the test have been estimated from the data, it can be used for fitted distributions with the understanding that it is then a conservative test, that is, less likely to reject the fit in error. The validity of the KS test can be improved for some specific distributions. These more stringent tests take the form of an multiplicative adjustment to the general KS statistic, but then comparison with the more general results are erroneous. (see Law&Kelton1 ,Brunk2, Stuart&Ord3)

The goodness of fit view also reports a REJECT or DO NOT REJECT decision for each KS test based on the the comparison between the calculated test statistic and the standard statistic for the given level of significance.

To visualize this process for continuous data, consider the two graphs below. The first is the normal P-P plot, the cumulative probability of the input data versus a continuous plot of the fitted cumulative distribution. However, for the KS test, the comparison is made between the probability of the input data having a value at or below at given point and the probability of the cumulative distribution at that point. This is represented in the second graph by comparing the cumulative probability for the observed data [the straight line] with the expected probability from the fitted cumulative distribution [square points]. The KS test measures the largest difference between these, being careful to account for the discrete nature of the measurement.





Note that the KS test can be applied to discrete data in a slightly different manner, and the resulting test is even more conservative than the KS test for continuous data. Also, the test may be further strengthened for discrete data.(see Gleser4).

While the test statistic can be useful, the p-value is more useful in determining the goodness of fit. The p-value is defined as the probability that another sample will be as unusual as the current sample given that the fit is appropriate. A small p-value indicates that the sample is highly unlikely, and, therefore, the fit should be rejected. Conversely, a high p-value indicates that the sample is likely and would be repeated and, therefore, the fit should not be rejected. Thus, the HIGHER the p-value, the more likely the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-value is likely to be the better fit.

1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 382

2. "An Introduction to Mathematical Statistics", H. D. Brunk, 1960, Ginn&Co., p261

3. "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart & J. Keith Ord, 1991, Oxford University Press, p1159

4."Exact Power of Goodness-of-Fit of Kolmogorov Type for Discontinuous Distributions", Leon Jay Gleser, J.Am.Stat.Assoc., 80, (1985), p954

Anderson Darling Test

The Anderson Darling test is a test of the goodness of fit of the fitted cumulative distribution to the input data in the Data Table, weighted heavily in the tails of the distributions. This test can be chosen in the Calculation page of the Fit Setup dialog

The Anderson Darling [AD] test calculates the the integral of the squared difference between the input data and the fitted distribution, with increased weighting for the tails of the distribution, by the equation:

$$W_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1 - F(x)]} dF(x)$$

where W_n^2 is the AD statistic, n is the number of data points, F(x) is the fitted cumulative distribution, and $F_n(x)$ is the cumulative distribution of the input data. This can be reduced to the more useful computational equation:

$$W_n^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)[\log \mu_i + \log(1 - \mu_{n-i+1})]$$

where u_i is the value of the fitted cumulative distribution,$F(x_i)$, for the ith data point. (see Law&Kelton1, Anderson&Darling2,3).

The resulting test statistic is then compared to a standard value of AD statistic with the appropriate number of data points and level of significance, usually labeled alpha. The limitations of the AD test are similar to the Kolmogorov Smirnov test with the exception of the boundary conditions discussed below. The AD test is not a limiting distribution; it is appropriate for any sample size. While the AD test is only valid if none of the parameters in the test have been estimated from the data, it can be used for fitted distributions because with the understanding that it is then a conservative test, that is, less likely to reject the fit in error. The validity of the AD test can be improved for some specific distributions. These more stringent tests take the form of an multiplicative adjustment to the general AD statistic, but then comparison with the more general results are erroneous.

The goodness of fit view also reports a REJECT or DO NOT REJECT decision for each AD test based on the the comparison between the calculated test statistic and the standard statistic for the given level of significance.

The AD test is very sensitive to the tails of the distribution. For this reason, the test must be used with discretion for many of the continuous distributions with lower bounds and finite values at that lower bound. The test is inaccurate for discrete distributions as the standard statistic is not easily calculated.

While the test statistic can be useful, the p-value is more useful in determining the goodness of fit. The p-value is defined as the probability that another sample will be as unusual as the current sample given that the fit is appropriate. A small p-value indicates that the sample is highly unlikely, and, therefore, the fit should be rejected. Conversely, a high p-value indicates that the sample is likely and would be repeated and, therefore, the fit should not be rejected. Thus, the HIGHER the p-value, the more likely the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-value is likely to be the better fit.

1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 392

2. "A Test of Goodness of Fit", T.W.Anderson, D.A. Darling, J.Am.Stat.Assoc., 1954, p765

3."Asymptotic Theory of Certain "Goodness Of Fit" Criteria Based On Stochastic Processes", T.W. Anderson, D.A. Darling, Ann.Math.Stat., 1952, p193

File|Export|Export Empirical

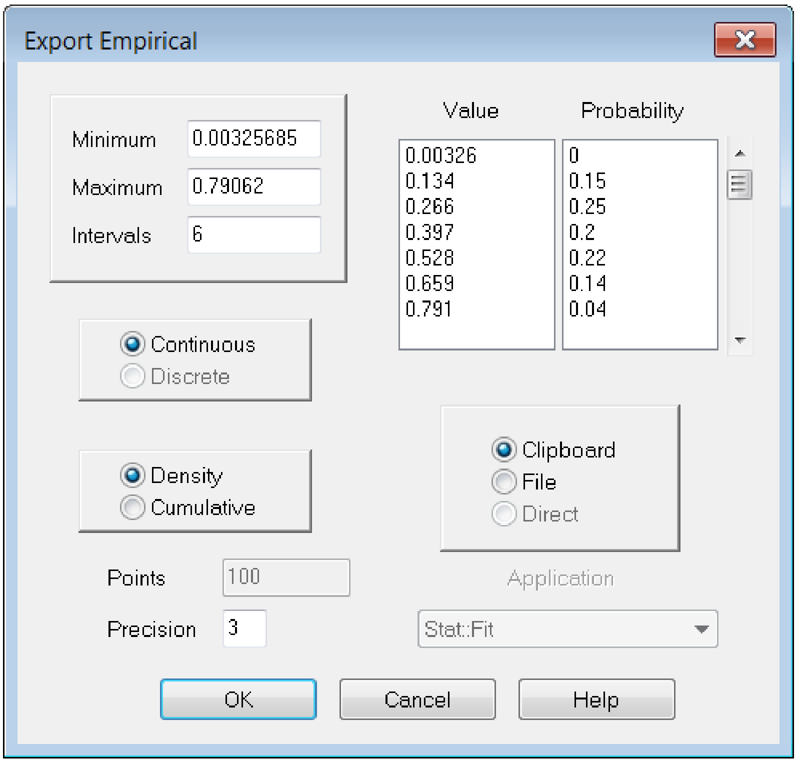
The Export Empirical command can be accessed in the Export sub menu of the File menu.

The Export Empirical command allows the export of an empirical distribution for the input data to either the Windows clipboard or a text file, or directly into Applications which have implemented Dynamic Data Exchange(DDE) with Stat::Fit. Applications with DDE activated must be running concurrently. The output is formatted with a value and a probability, delimited by a space, on each line. An empirical distribution is advisable if the input data cannot be fit to the analytical distributions or for very large data sets greater than 10,000 points.

If the any of the data are non-integers, then the exported distribution may be only a continuous distribution with the minimum and maximum values determining the range of data included in the distribution. The number of data points included is shown below. The number of intervals may be set as well, but should be small enough to avoid bins with zero data and large enough to avoid over smoothing. The default interval value is the same as the auto value set by Stat::Fit and will avoid over smoothing. The distribution may be either a cumulative or density distribution.

If the data is integer, then the exported distribution may be either a continuous or discrete distribution with the minimum and maximum values determining the range of data included in the distribution. The discrete integer range is limited to 1000. The number of data points included is shown below. The distribution may be either a cumulative or density distribution.

The precision of the numbers in the output is set by the Precision box whose default value is 3.



Guide to No Data Representations

This section is a guide to assigning an analytical distribution to a particular situation where little or no sampling data is available, but where some basic information about the continuous random variable in question can be assumed. While this guide is no substitute for actual data, it can provide a working model consistent with available information. Use this guide only when sufficient data cannot be gathered from an existing processes due to inaccessibility, or when the physical process does not yet exist but must be modeled. Once a distribution is chosen, the Distribution Viewer can be used to export the parameters to a given Application.

Frequently, this is the type of model that must be used with expert opinion as the only source of information. In general, avoid using the Triangular distribution when possible because if usually gives much greater variance than other choices. Further, some distributions can be estimated from sample simulations.

Is the random variable:

Unbounded
Bounded Above a Minimum
Bounded between a Minimum and a Maximum

No Data|Unbounded

For an unbounded continuous distribution, at least a mean and standard deviation must be estimated in order to assign an analytical distribution. The simplest, most useful, unbounded distribution is given by a Normal distribution as in:

Normal(MEAN, STANDARD DEVIATION)

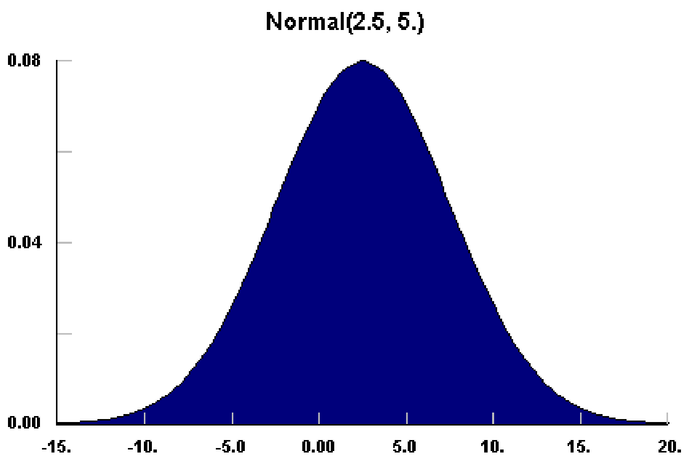
which is symmetrical about the mean.

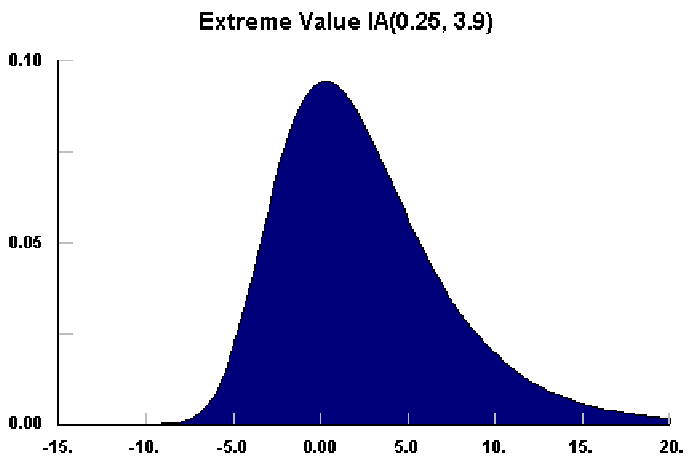
If the unbounded distribution is expected to be skewed in the one direction, that is, with a long tail in one direction, the Extreme Value distribution can provide a good representation For a positive tail, use the Extreme Value IA distribution as described. For a negative tail, use the complimentary Extreme Value IB distribution. The mean and standard deviation of the Extreme Value IA distribution can be used to calculate the parameters of the appropriate distribution:

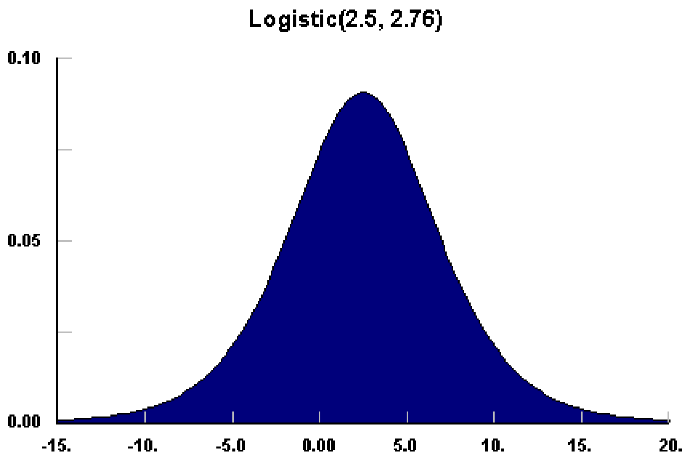
$$\beta = \text{STANDARD DEVIATION} \times 0.77970$$
$$\tau = \text{MEAN} - 0.57722\beta$$

If the unbounded distribution is expected to be symmetrical, but can have exceptionally large positive or negative values on occasion, the Logistic distribution is of more use. the mean and standard deviation can be used to calculate the parameters of the appropriate distribution:

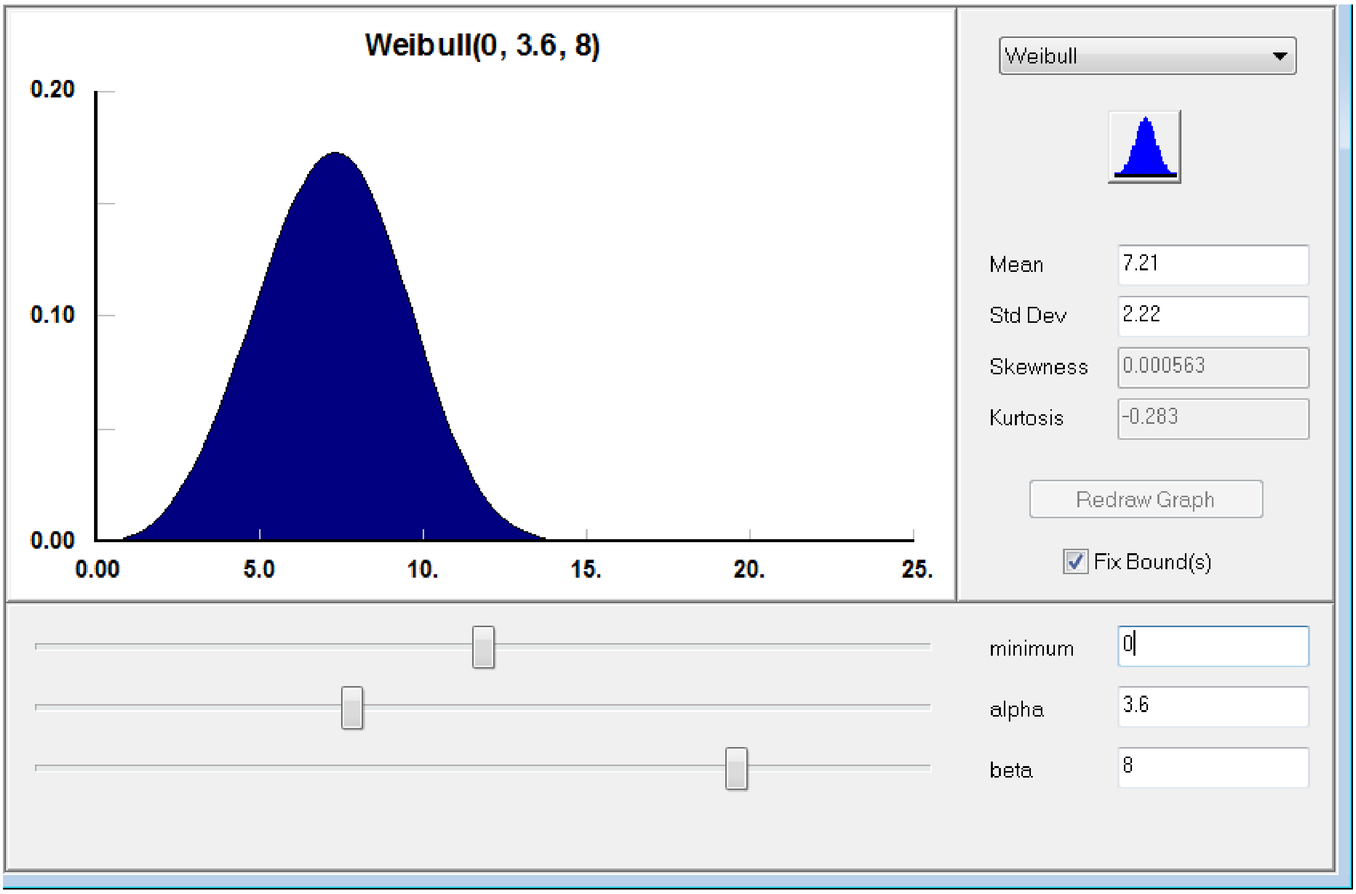
A Normal distribution, Extreme Value distribution, and Logistic distribution with a mean of 2.5 and a standard deviation of 5 are shown below. More examples can be viewed by using the Distribution Viewer capability.







If negative values are to be avoided, but the distribution is thought to resemble the Normal distribution, that is, a bell curve, then a Weibull distribution can be used with alpha near 3.6... as shown. Beta can be used to scale in order to get the desired standard deviation.



No Data|Bounded Above a Minimum

Given that the distribution is bounded above a minimum, many possible distributions exist, as can be seen from browsing through the analytical distributions available in Stat::Fit. Only a few convenient examples will be given here, tied to specific models.

- time to a random event
- time to complex event
- time to task completion

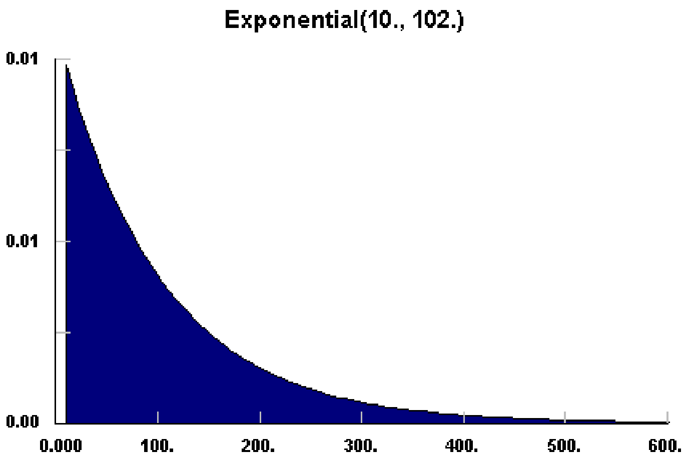
time to a random event: arrival times, catastrophic failure rates

The model for the time to a random event is Exponential with beta being the mean time to occurrence or its inverse, the mean rate of occurrence, as in

Exponential(MEAN TIME TO OCCUR)

The Exponential distribution has the advantage of being history independent, that is, the probability of occurrence is independent of the time to the last occurrence. This distribution can be used to represent the time to the arrival the next phone call, customer, etc.

A Exponential distribution with a lower bound of 10 and a mean of 102 is shown below. Note that the standard deviation is equal to the mean. More examples can be viewed by using the Distribution Viewer capability.



time to complex event: electronic component failure, inter arrival time from multiple sources

A reasonable model for the time to a complex event is a Weibull distribution with alpha less than 1. Note that alpha=1 reduces the Weibull distribution to an exponential distribution so that this distribution can also be used for random occurrences. However, the Gamma distribution has similar properties, and is somewhat easier to use because parameter estimates from estimates of the mean and standard deviation are straightforward.

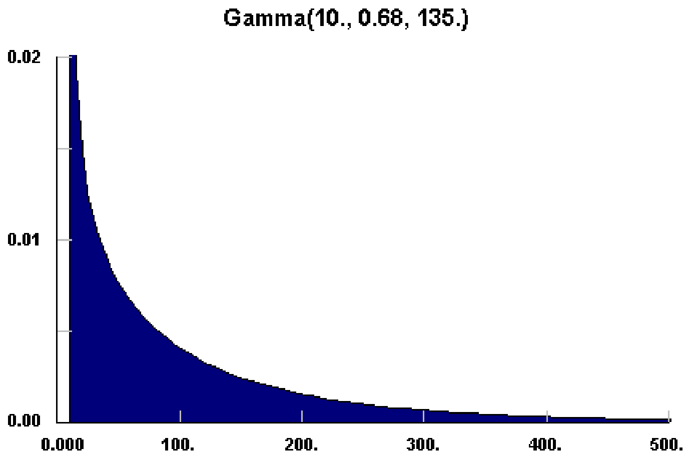
For the Gamma distribution, the parameters can be calculated by:

$$\beta = \frac{(\text{STANDARD DEVIATION})^2}{(\text{MEAN} - \text{MIN})}$$

$$\alpha = \frac{(\text{MEAN} - \text{MIN})}{\beta}$$

Gamma(MIN, ALPHA, BETA)

These distributions can be used to represent the time to a complex event such as failure times for a tool, time between arrival of a job a work station, etc. A typical value of the Gamma parameter, alpha, is 0.68, which gives a slightly larger standard deviation than the exponential distribution. A Gamma distribution with a bound of 10, a mean of 102, and a standard deviation of 135 is shown below. More examples can be viewed by using the Distribution Viewer capability.



time to task completion: repair time, service time

A reasonable model for the time to a complex event is a Weibull distribution with alpha greater than 1. However, the Gamma distribution has similar properties, and is somewhat easier to use because parameter estimates from estimates of the mean and standard deviation are straightforward. Note that these distributions can have offsets from zero to allow for minimum execution times.

For the Gamma distribution, the parameters can be calculated by:

$$\beta = \frac{(\text{STANDARD DEVIATION})^2}{(\text{MEAN} - \text{MIN})}$$

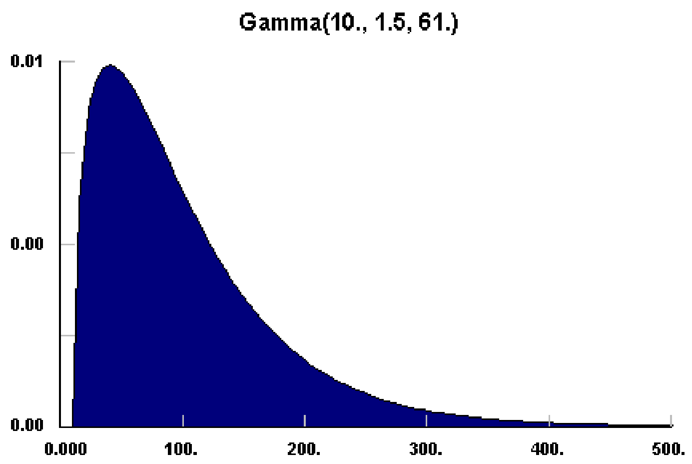
$$\alpha = \frac{(\text{MEAN} - \text{MIN})}{\beta}$$

Gamma(MIN, ALPHA, BETA)

If the most likely value [MODE] is easier to estimate than calculate alpha from;

$$\alpha = \frac{\text{MODE} - \text{MIN}}{\beta} + 1$$

These distributions can represent the time to complete a task, such as service times, repair times, etc. For these types of applications, a typical value of the Gamma parameter, alpha, is 1.5. A Gamma distribution with a bound of 10, a mean of 102, and a standard deviation of 61 is shown below. More examples can be viewed by using the Distribution Viewer capability.


Gamma(10., 1.5, 61.)
0.01
0.00
0.00
0.000
100.
200.
300.
400.
500.

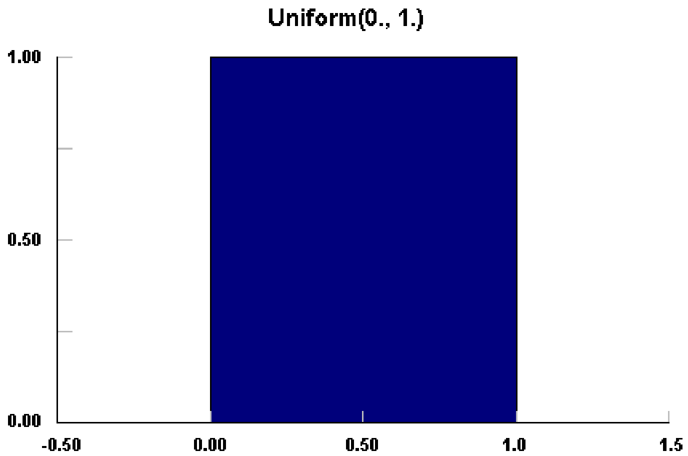
No Data|Bounded Between a Minimum and a Maximum

Please be aware that using doublely bounded distributions poses the same danger as boot strapping: the limits may be surpassed by the actual process and will not be represented. However, if very little information is available, representing the random process with some form of variability is better than using an average value. Frequently, it is best to try several choices to test the robustness of the model.

only the bounds are known
zero at or near both bounds
non-zero at one bound
non-zero at both bounds

If no other information but the limits is available, the Uniform distribution is the only choice. Then, any value between the limits is equally probable.

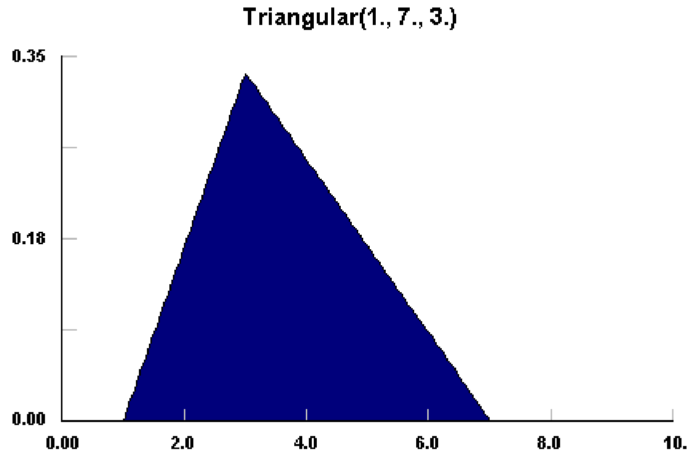
Uniform(MINIMUM, MAXIMUM)



If the process goes to zero at or near both bounds, then other information must be present. This may be either the most likely value (MODE) or the average value (MEAN) or both. If only one of these is available then the Triangular distribution can be used (unless visualization of a Beta distribution is possible in the Distribution Viewer).

MODE = 3*MEAN - MINIMUM - MAXIMUM

Triangular(MINIMUM, MAXIMUM, MODE)



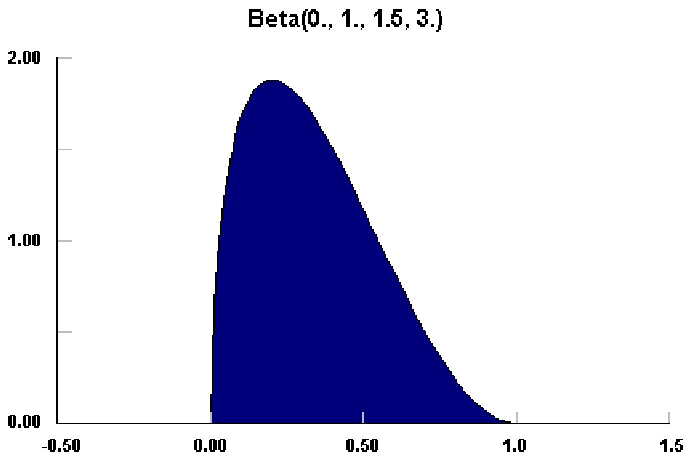
If both the MODE and MEAN are known, than the Beta distribution should be used. Then the third and fourth parameters of the distribution are determined by:

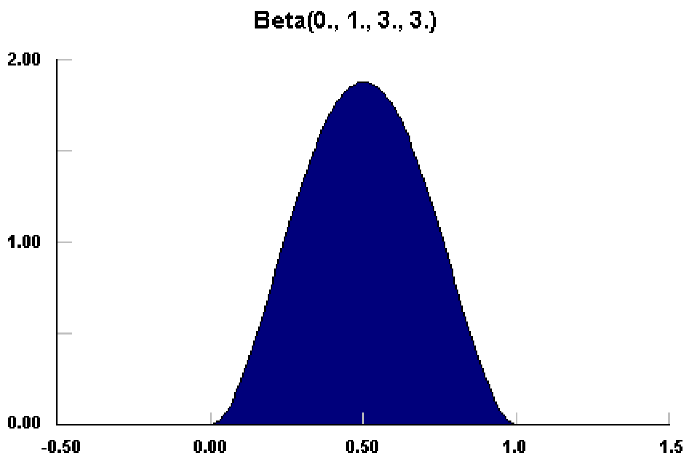
$$p = \frac{(\text{MEAN} - \text{MIN})(2 \times \text{MODE} - \text{MIN} - \text{MAX})}{(\text{MODE} - \text{MEAN})(\text{MAX} - \text{MIN})}$$

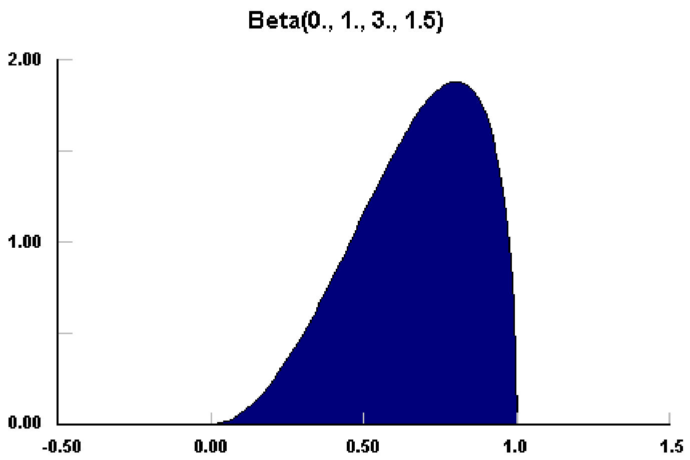
$$q = p\frac{(\text{MAX} - \text{MEAN})}{(\text{MEAN} - \text{MIN})}$$

Beta(MINIMUM, MAXIMUM, p, q)

Note that the choice of mode and mean determines the symmetry of the distribution, with the distribution taking the forms shown below, [MODE<MEAN], [MODE=MEAN], [MODE>MEAN], respectively



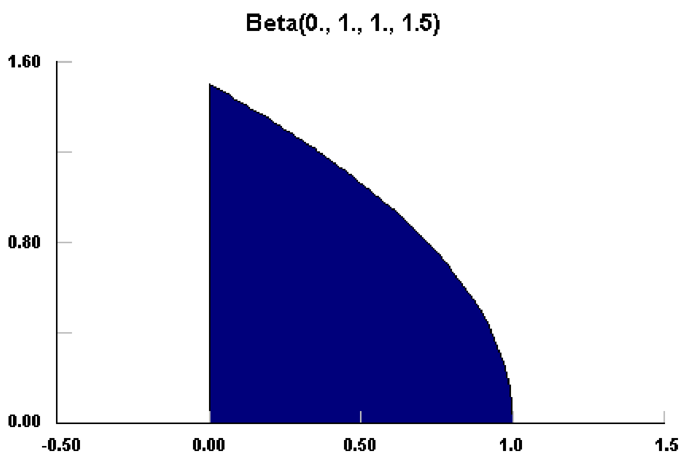


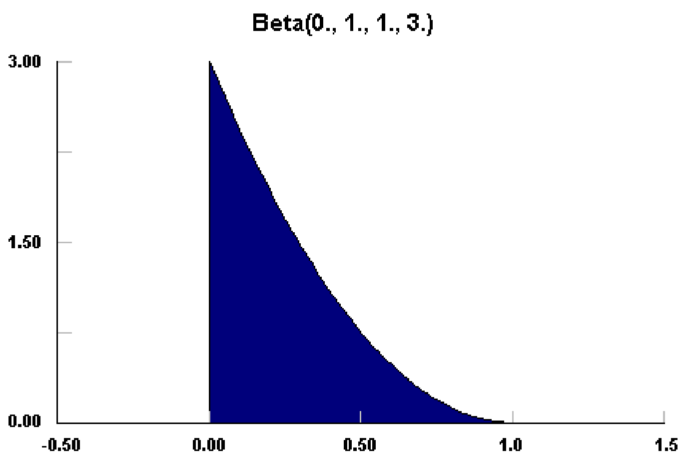


If the process is non-zero at one bound, then the Beta distribution is the best representation. Given that the distribution is bounded between a minimum and a maximum, is finite at the lower bound, and zero at the upper bound of the distribution, it is best represented by a Beta distribution with p=1 and q>1.
If the mean can be estimated, then q can be determined from:

$$q = \frac{(MAX - MEAN)}{(MEAN - MIN)}$$

Otherwise q must be estimated visually. Note again the restriction that q>1. If the calculation involving the mean gives a q<1, then consider using a distribution that is infinite at the lower bound.

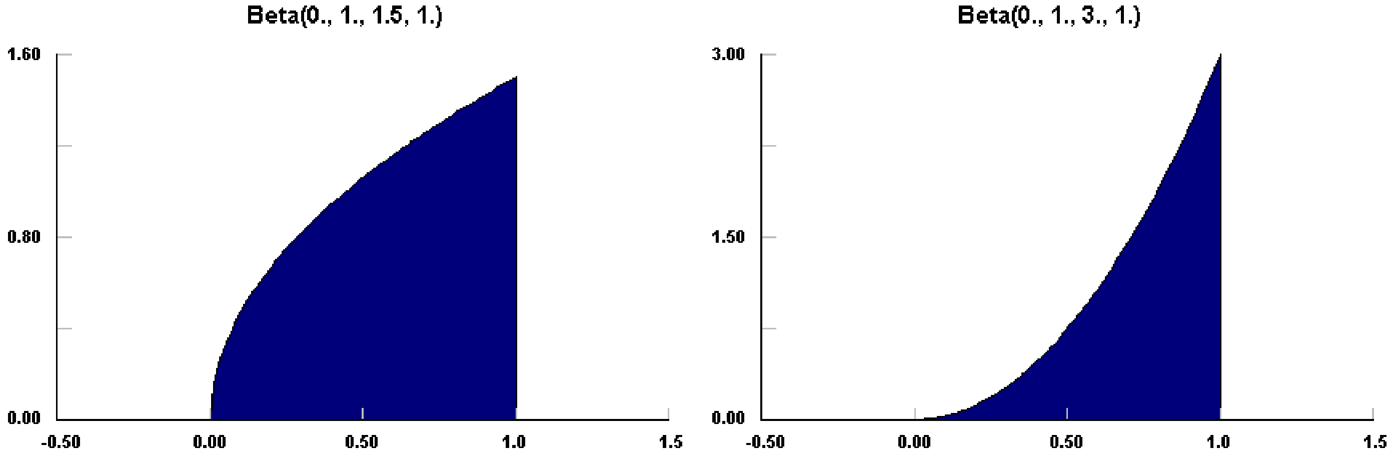




Given that the distribution is bounded between a minimum and a maximum, is finite at the upper bound, and zero at the lower bound of the distribution, it is best represented by a Beta distribution with p>1 and q=1. If the mean can be estimated, then p can be determined from:

$$p = \frac{(\text{MEAN} - \text{MIN})}{(\text{MAX} - \text{MEAN})}$$

Otherwise p must be estimated visually. Note again the restriction that p>1. If the calculation involving the mean gives a p<1, then consider using a distribution that is infinite at the upper bound.

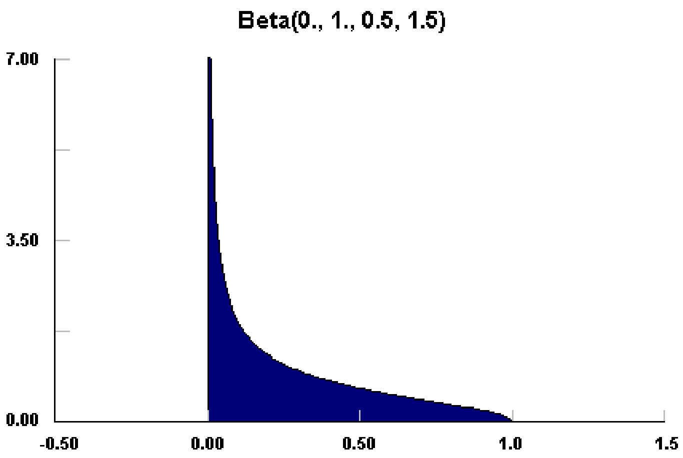


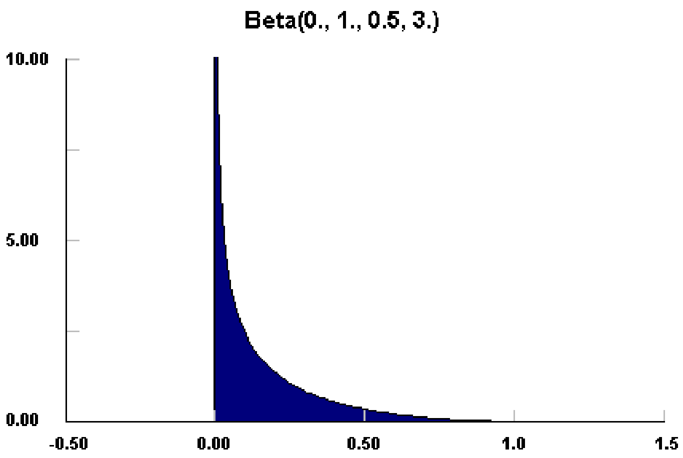
Given that the distribution is bounded between a minimum and a maximum, is infinite at the lower bound, and zero at the upper bound of the distribution, it is best represented by a Beta distribution with p<1 and q>1. If the mean and standard deviation can be estimated, then p,q can be determined from:

$$p = \frac{(\text{MEAN} - \text{MIN})^2 (\text{MAX} - \text{MEAN})}{(\text{STANDARD DEVIATION})^2 (\text{MAX} - \text{MIN})} - \frac{(\text{MEAN} - \text{MIN})}{(\text{MAX} - \text{MIN})}$$

$$q = p\frac{(\text{MAX} - \text{MEAN})}{(\text{MEAN} - \text{MIN})}$$

Note that the mean must be less than the midpoint. If the mean and standard deviation cannot be estimated, p,q must be estimated visually. Note again the restrictions that p<1 and q>1.



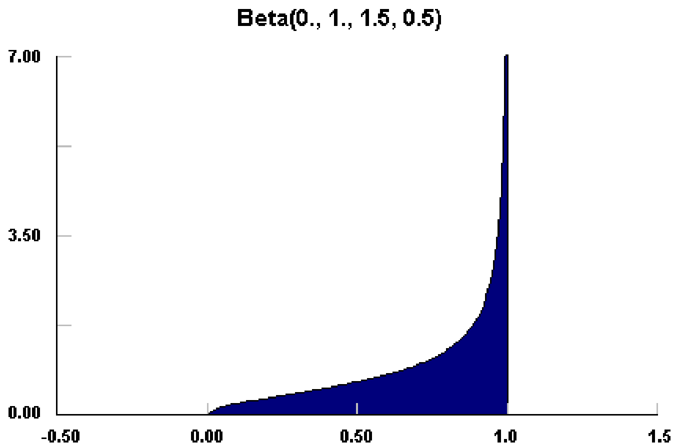


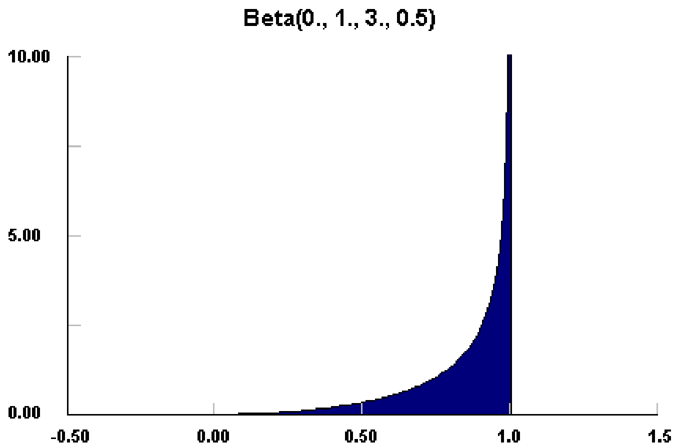
Given that the distribution is bounded between a minimum and a maximum, is infinite at the upper bound, and zero at the lower bound of the distribution, it is best represented by a Beta distribution with p>1 and q<1. If the mean and standard deviation can be estimated, then p,q can be determined from:

$$p = \frac{(\text{MEAN} - \text{MIN})^2 (\text{MAX} - \text{MEAN})}{(\text{STANDARD DEVIATION})^2 (\text{MAX} - \text{MIN})} - \frac{(\text{MEAN} - \text{MIN})}{(\text{MAX} - \text{MIN})}$$

$$q = p \frac{(\text{MAX} - \text{MEAN})}{(\text{MEAN} - \text{MIN})}$$

Note that the mean must be greater than the midpoint. If the mean and standard deviation cannot be estimated, p,q must be estimated visually. Note again the restrictions that p>1 and q<1.



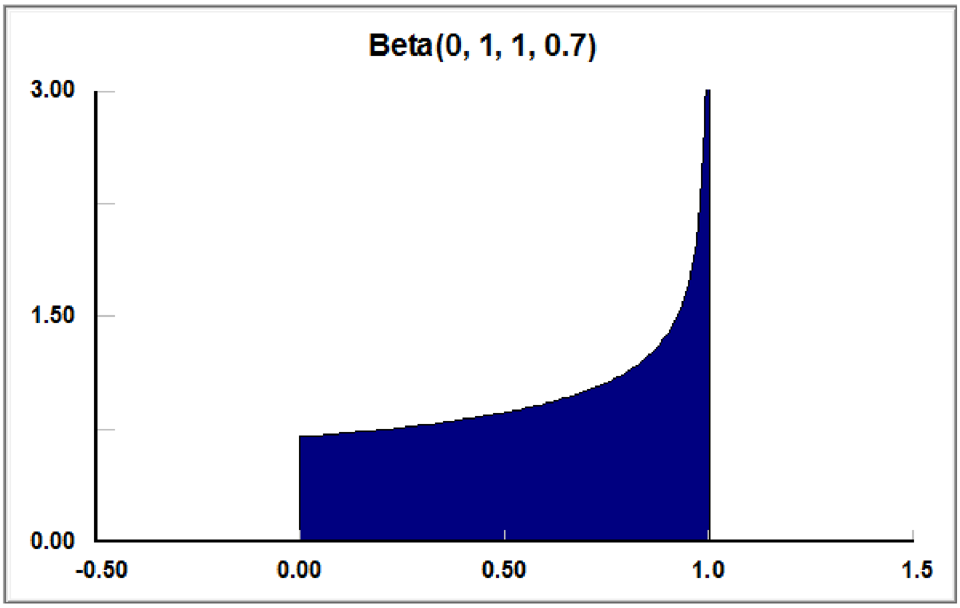


If the process is non-zero at both bounds, but finite, then the representation depends on the simulation product in use. The best representation is a truncated distribution; typically a truncated Normal distribution serves as well as any as long as a non-symmetrical truncation is allowed. Without that choice, the sum of two Beta distributions as given for single finite bounds works well.

If the process is non-zero at both bounds, and infinite at at least one, then a Beta distribution can be used. In particular, either p or q will be equal to 1 at the finite bound and less than 1 at the infinite bound. Again, if the mean is known, then parameter at the infinite bound can be estimated by:

q = (max - mean)/(mean - min)

Otherwise, the shape will have to be visualized using the Distribution Viewer.



Beta(min, max, p, q)

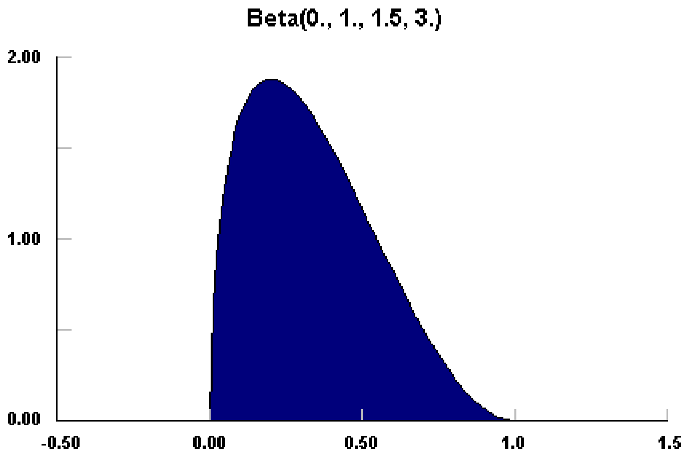
$$f(x) = \frac{1}{B(p,q)} \frac{(x - min)^{p-1}(max - x)^{q-1}}{(max - min)^{p+q-1}}$$

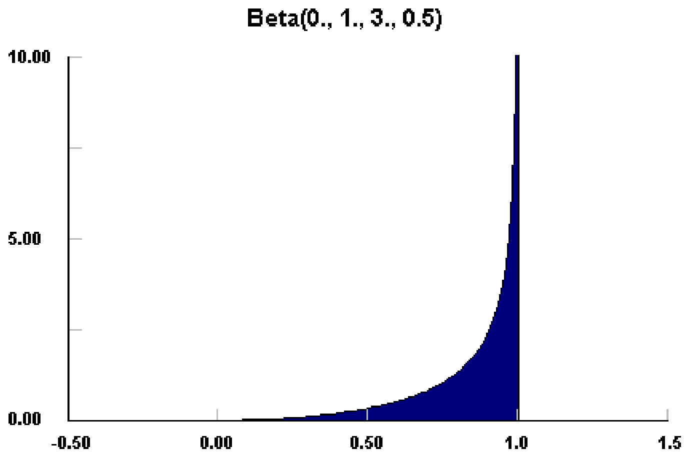
min = minimum value of x
max = maximum value of x
p = lower shape parameter > 0
q = upper shape parameter >0
B(p, q) = Beta mathematical function

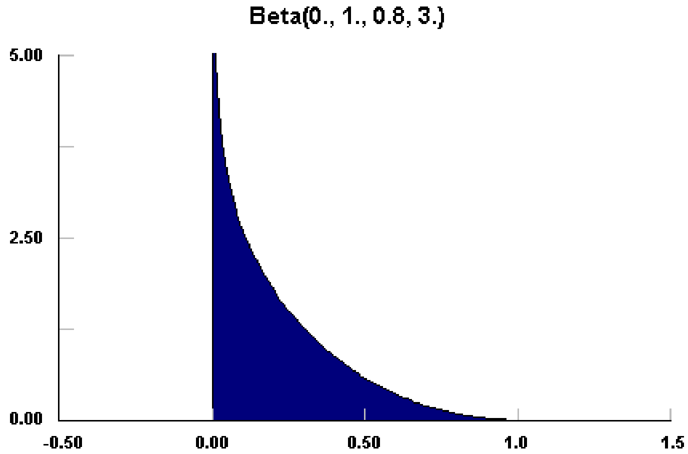
The Beta distribution is a continuous distribution that has both upper and lower finite bounds. Because many real situations can be bounded in this way, the Beta distribution can be used empirically to estimate the actual distribution before much data is available. Even when data is available, the Beta distribution should fit most data in a reasonable fashion, although it may not be the best fit. The Uniform distribution is a special case of the Beta distribution with p,q = 1. The Power Function distribution is a special case of the Beta distribution with q = 1.

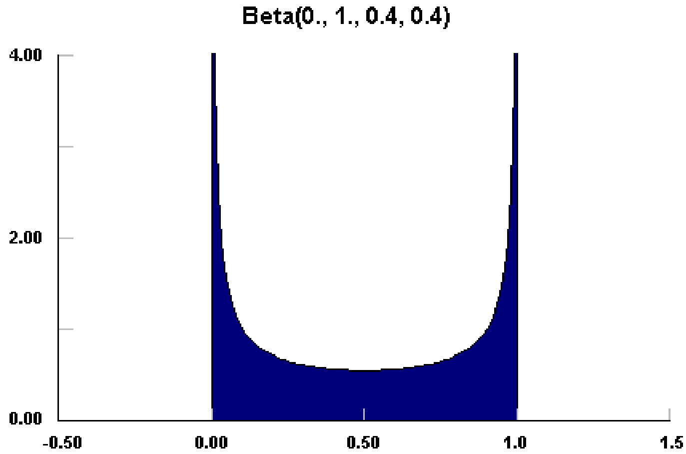
As can be seen in the following examples, The Beta distribution can approach zero or infinity at either of its bounds, with p controlling the lower bound and q controlling the upper bound. Values of p,q less than 1 cause the Beta distribution to approach infinity at that bound. Values of p,q greater than 1 cause the Beta distribution to be zero at that bound. Values of p,q exactly 1 cause the Beta distribution to be finite at that bound. More examples can be viewed by using the Distribution Viewer capability.

Beta distributions have many, many uses. As summarized in Johnson et. al.1, Beta distributions have been used to model distributions of hydrologic variables, logarithm of aerosol sizes, activity time in PERT analysis, isolation data in photovoltaic system analysis, porosity/void ratio of soil, phase derivatives in communication theory, size of progeny in Escherchia Coli, dissipation rate in breakage models, proportions in gas mixtures, sea-state reflectivity, clutter and power of radar signals, construction duration, particle size, tool wear, and others. Many of these uses occur because of the doubly bounded nature of the Beta distribution.









1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons., p236-237

Cauchy(theta, lambda)

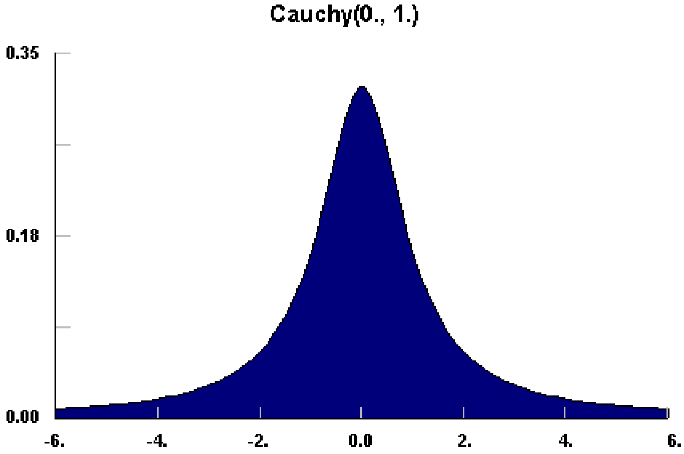
$$f(x) = \left(\frac{1}{\pi\lambda}\right)\left[1+\left(\frac{x-\theta}{\lambda}\right)^2\right]^{-1}$$

theta = mode or central peak position
lambda = scaling parameter

The Cauchy distribution is an unbounded continuous distribution that has a sharp central peak but significantly broad tails. The tails are much heavier than the tails of the Normal distribution. The Cauchy distribution does not have finite moments.

The Cauchy distribution can be used to represent the ratio of two equally distributed parameters in certain cases, e.g. the ratio of two normal parameters. This distribution has no finite moments because of its extensive tails. Thus it can also be used to generate wildly divergent data as long as the data has a central tendency. (see Johnson et.al.1)

The Cauchy distribution, as shown below, has a distinct peaked shape. It is unchanged in shape with changes in theta or lambda. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p298

Chi Squared(min, nu)

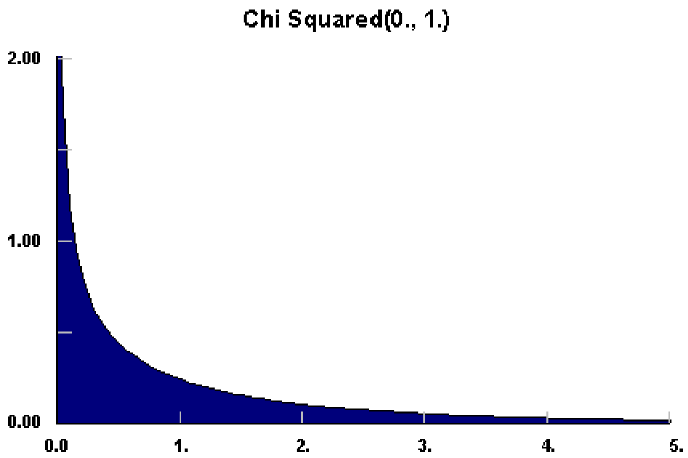
$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\exp\left(-\frac{(x-min)}{2}\right)(x-min)^{(\nu/2)-1}$$

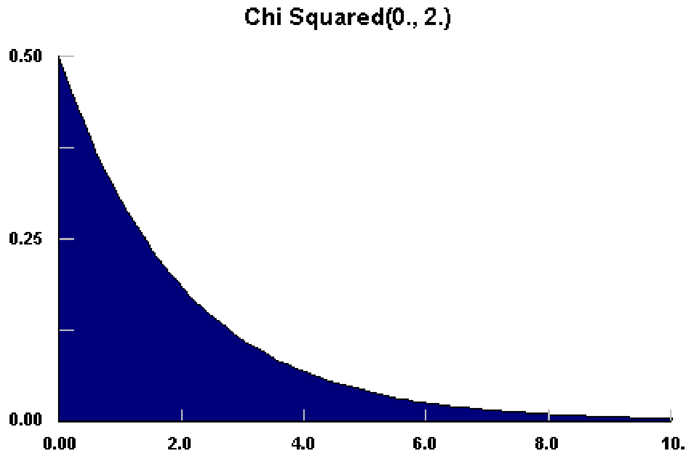
min = minimum x value
nu = shape parameter

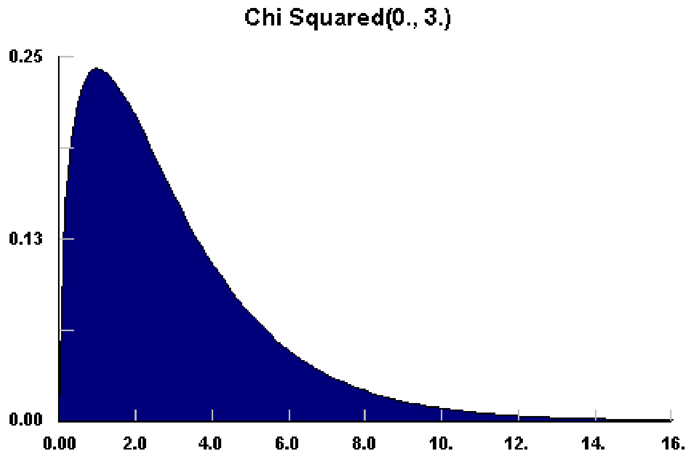
The Chi Squared distribution is a continuous distribution bounded on the left side. Note that the Chi Squared distribution is a subset of the Gamma distribution with beta = 2 and alpha = nu/2. Like the Gamma distribution, it has three distinct regions. For nu = 2, the Chi Squared distribution reduces to the Exponential distribution, starting at a finite value at minimum x and decreasing monotonically thereafter. For nu < 2, the Chi Squared distribution tends to infinity at minimum x and decreases monotonically for increasing x. For nu > 2, the Chi Squared distribution is 0 at minimum x, peaks at a value that depends on nu, decreasing monotonically thereafter.

Because the Chi Squared distribution does not having a scaling parameter, its utilization is somewhat limited. Frequently, this distribution will try to represent data with a clustered distribution with nu less than 2. However, it can be viewed as the distribution of the sum of squares of independent unit normal variables with nu degrees of freedom and is used in many statistical tests. (see Johnson et.al.1)

Examples of each of the regions of the Chi Squared distribution are shown below. Note that the peak of the distribution moves away from the minimum value for increasing nu, but with a much broader distribution. More examples can be viewed by using the Distribution Viewer capability.







1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 415

Erlang(min, m, beta)

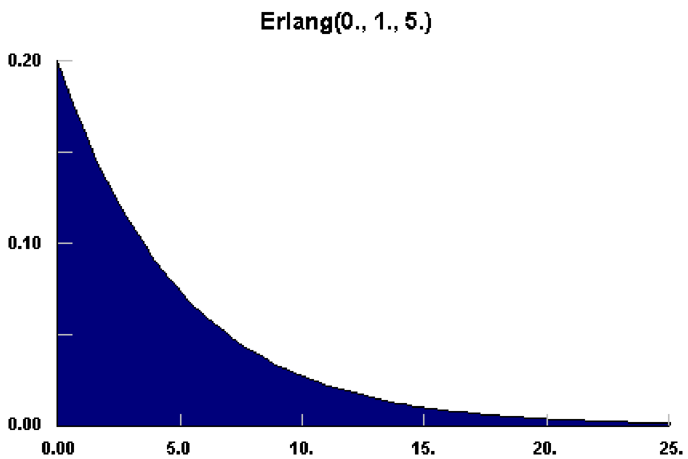
$$f(x) = \frac{(x - \min)^{m-1}}{\beta^m \Gamma(m)} \exp\left(-\frac{[x - \min]}{\beta}\right)$$

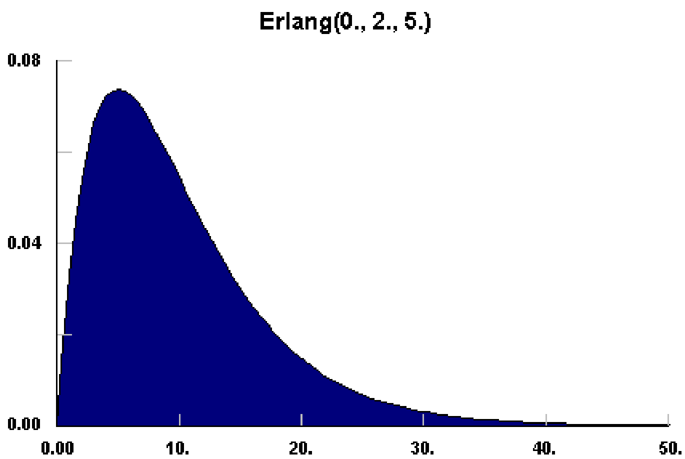
min = minimum x value
m = shape factor = positive integer
beta = scale factor > 0

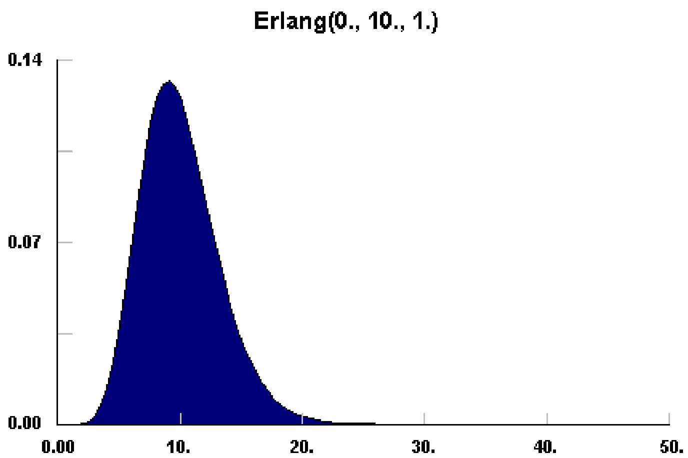
The Erlang distribution is a continuous distribution bounded on the lower side. It is a special case of the Gamma distribution where the parameter, m, is restricted to a positive integer. As such, the Erlang distribution has no region where f(x) tends to infinity at the minimum value of x [m<1], but does have a special case at m = 1, where it reduces to the Exponential distribution.

The Erlang distribution has been used extensively in reliability and in queuing theory, thus in discrete event simulation, because it can be viewed as the sum of m exponentially distributed random variables, each with mean beta. It can be further generalized. (see Johnson et. al.1, Banks & Carson2)

As can be seen in the following examples, the Erlang distribution follows the Exponential distribution at m = 1, has a positive skewness with a peak near 0 for m between 2 and 9, and tends to a symmetrical distribution offset from the minimum at larger m. More examples can be viewed by using the Distribution Viewer capability.







1. "Continuous Univariate Distributions", Volume 1, Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons.

2. "Discrete-Event System Simulation", Jerry Banks, John S. Carson II, 1984, Prentice-Hall

Exponential(min, beta)

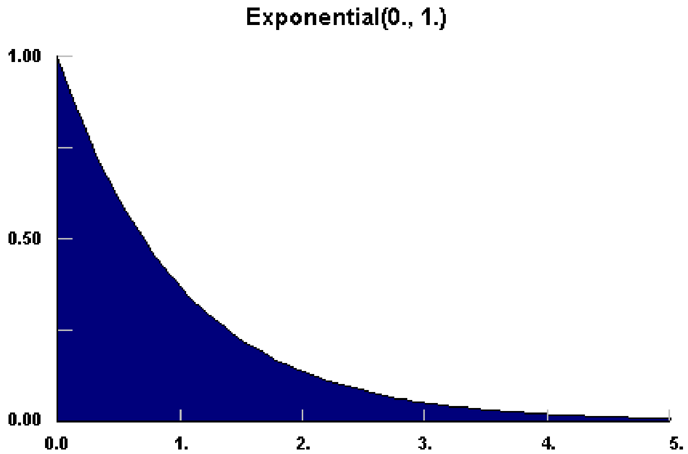
$$f(x) = \frac{1}{\beta} \exp\left(-\frac{[x - \min]}{\beta}\right)$$

min = minimum x value
beta = scale parameter = mean

The Exponential distribution is a continuous distribution bounded on the lower side. It's shape is always the same, starting at a finite value at the minimum and continuously decreasing at larger x.

The Exponential distribution is frequently used to represent the time between random occurrences, such as the time between arrivals at a specific location in a queuing model or the time between failures in reliability models. It has also been used to represent the services times of a specific operation. Further, it serves as an explicit manner in which the time dependence on noise may be treated. As such, these models are making explicit use of the lack of history dependence of the exponential distribution; it has the same set of probabilities when shifted in time. Even when Exponential models are known to be inadequate to describe the situation, their mathematical tractability provides a good starting point. Later, a more complex distribution such as Erlang or Weibull may be investigated. (see Law & Kelton[1], Johnson et. al.[2])

As shown in the example, the Exponential distribution decreases rapidly for increasing x. More examples can be viewed by using the Distribution Viewer capability.



1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 330

2. "Continuous Univariate Distributions", Volume 1, Norman L. Johnson, Samuel Kotz, N. Balakkrishnan, 1994, John Wiley & Sons, p 499

Extreme Value IA(tau, beta)

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{[x-\tau]}{\beta}\right) \exp\left(-\exp\left(-\frac{[x-\tau]}{\beta}\right)\right)$$

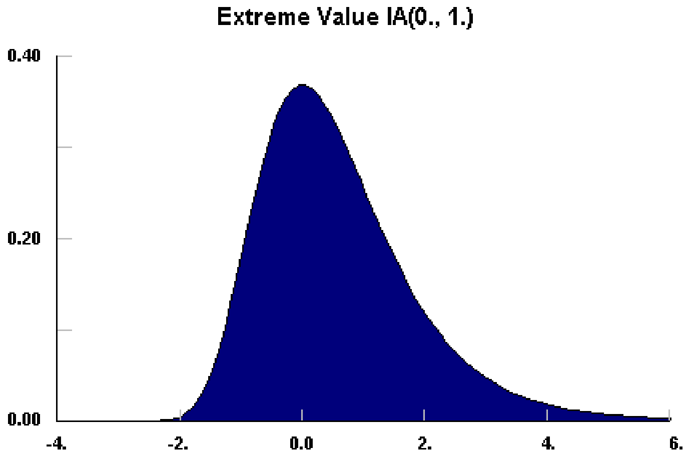
tau = threshold/shift parameter
beta = scale parameter

The Extreme Value IA distribution is an unbounded continuous distribution. It's shape is always the same but may be shifted or scaled to need. It is also called the Gumbel distribution.

The Extreme Value IA distribution describes the limiting distribution of the greatest values of many types of samples. Actually, the Extreme Value distribution given above is usually referred to as Type 1, with Type 2 and Type 3 describing other limiting cases. If X is replaced by -X, then the resulting distribution describes the limiting distribution for the least values of many types of samples. These reflected pair of distributions are sometimes referred to as Type 1A and Type 1B. Note that the complimentary distribution can be used to represent samples with negative skewness.

The Extreme Value distribution has been used to represent parameters in growth models, astronomy, human lifetimes, radioactive emissions, strength of materials, flood analysis, seismic analysis, and rainfall analysis. It is also directly related to many learning models. (see Johnson et. al.1)

The Extreme Value IA distribution starts below tau, is skewed in the positive direction peaking at tau, then decreasing monotonically thereafter. Beta determines the breadth of the distribution. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions", Volume 2, Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons

Extreme Value IB(tau, beta)

$$f(x) = \frac{1}{\beta}\exp\left(\frac{x-\tau}{\beta}\right)\exp\left(-\exp\left(\frac{x-\tau}{\beta}\right)\right)$$

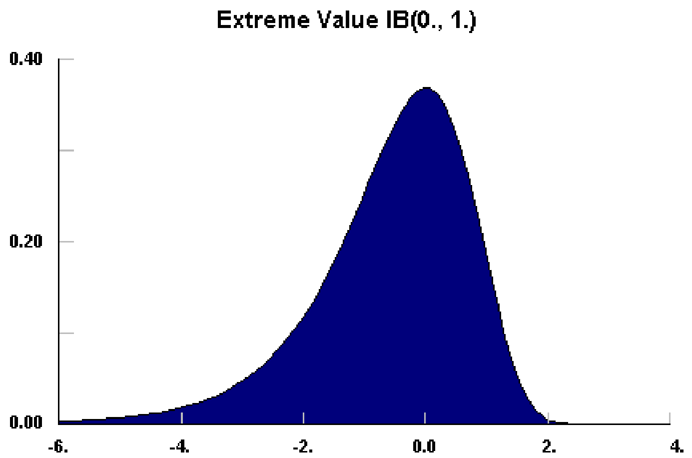
tau = threshold/shift parameter
beta = scale parameter

The Extreme Value IB distribution is an unbounded continuous distribution. It's shape is always the same but may be shifted or scaled to need.

The Extreme Value IB distribution describes the limiting distribution of the least values of many types of samples. Actually, the Extreme Value distribution given above is usually referred to as Type 1, with Type 2 and Type 3 describing other limiting cases. If X is replaced by -X, then the resulting distribution describes the limiting distribution for the greatest values of many types of samples. These reflected pair of distributions are sometimes referred to as Type 1A and Type 1B. Note that the complimentary distribution can be used to represent samples with positive skewness.

The Extreme Value distribution has been used to represent parameters in growth models, astronomy, human lifetimes, radioactive emissions, strength of materials, flood analysis, seismic analysis, and rainfall analysis. It is also directly related to many learning models. (see Johnson et. al.1)

The Extreme Value IB distribution starts below tau, is skewed in the negative direction peaking at tau, then decreasing monotonically thereafter. Beta determines the breadth of the distribution. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions", Volume 2, Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons.

Gamma(min, alpha, beta)

$$f(x) = \frac{(x - min)^{\alpha - 1}}{\beta^{\alpha}\Gamma(\alpha)} exp\left(-\frac{[x - min]}{\beta}\right)$$

min = minimum x value
alpha = shape parameter > 0
beta = scale parameter > 0

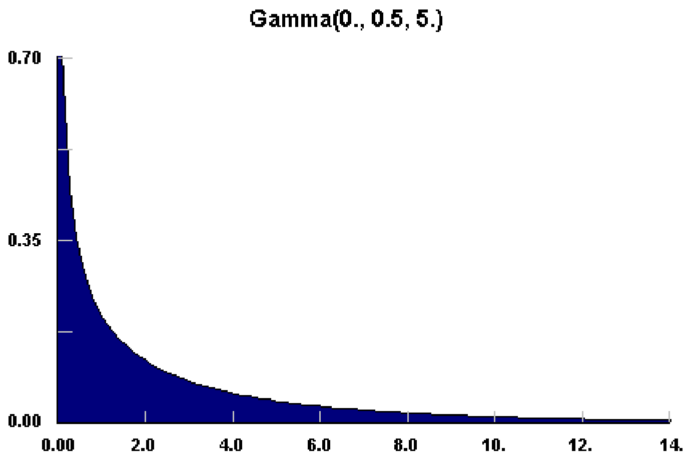
The Gamma distribution is a continuous distribution bounded at the lower side. It has three distinct regions. For alpha = 1, the Gamma distribution reduces to the Exponential distribution, starting at a finite value at minimum x and decreasing monotonically thereafter. For alpha < 1, the Gamma distribution tends to infinity at minimum x and decreases monotonically for increasing x. For alpha > 1, the Gamma distribution is 0 at minimum x, peaks at a value that depends on both alpha and beta, decreasing monotonically thereafter. If alpha is restricted to positive integers, the Gamma distribution is reduced to the Erlang distribution.

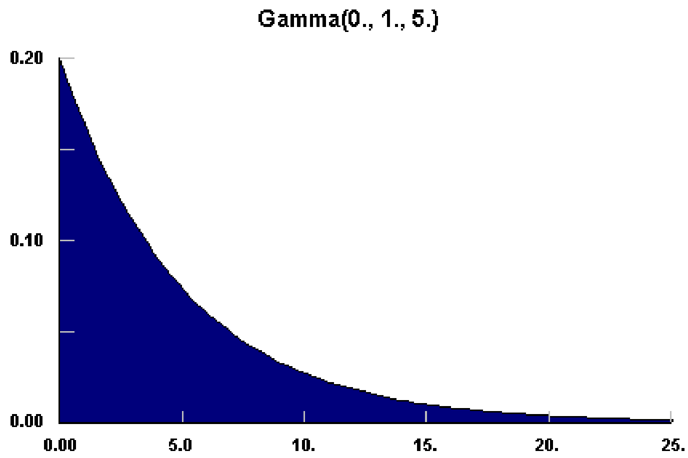
Note that the Gamma distribution also reduces to the Chi Squared distribution for min = 0, beta = 2, and alpha = nu/2. It can then be viewed as the distribution of the sum of squares of independent unit normal variables. with nu degrees of freedom and is used in many statistical tests.

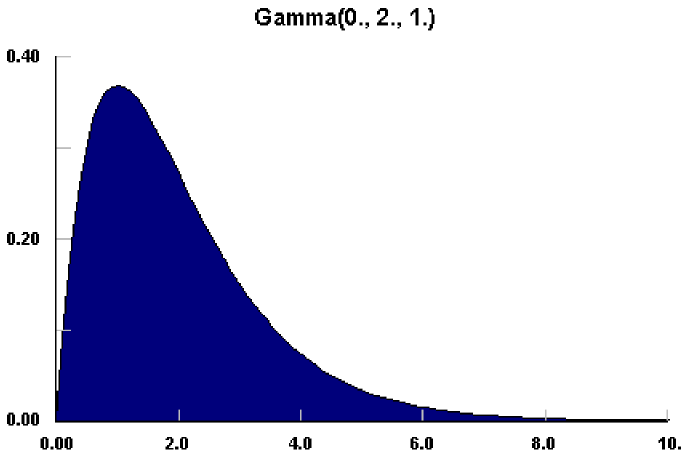
The Gamma distribution can also be used to approximate the Normal distribution, for large alpha, while maintaining its strictly positive values of x [actually (x-min)].

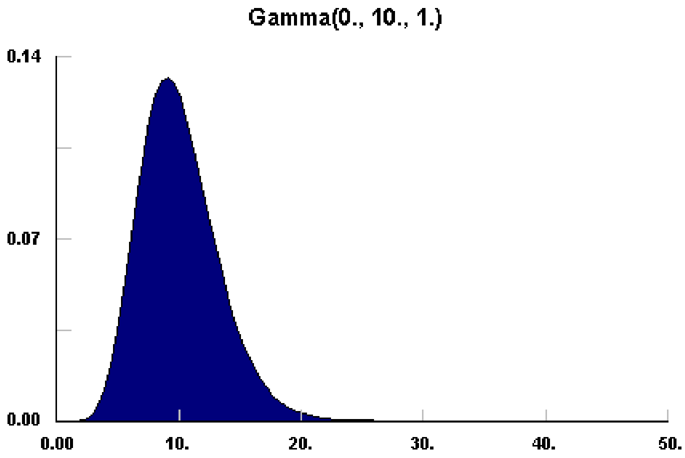
The Gamma distribution has been used to represent lifetimes, lead times, personal income data, a population about a stable equilibrium, inter arrival times, and service times. In particular, it can represent lifetime with redundancy. (see Johnson et.al1, Shooman2)

Examples of each of the regions of the Gamma distribution are shown below. Note the peak of the distribution moving away from the minimum value for increasing alpha, but with a much broader distribution. More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 343

2. "Probabilistic Reliability: An Engineering Approach", Martin L. Shooman, 1990, Robert E. Krieger

Inverse Gaussian(min, alpha, beta)

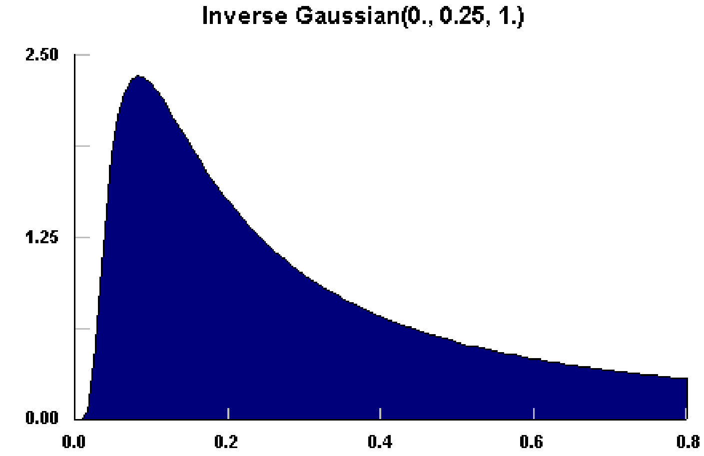
$$f(x) = \left(\frac{\alpha}{2\pi(x-\text{min})^3}\right)^{1/2} \exp\left[-\frac{\alpha(x-\text{min}-\beta)^2}{2\beta^2(x-\text{min})}\right]$$

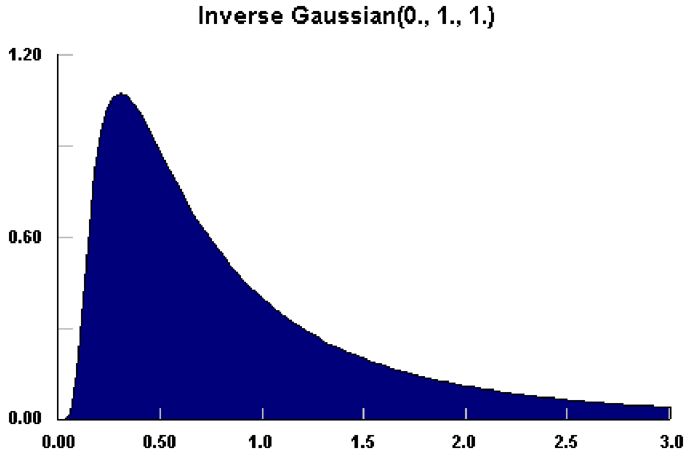
min = minimum x value
alpha = shape parameter > 0
beta = mixture of shape and scale > 0

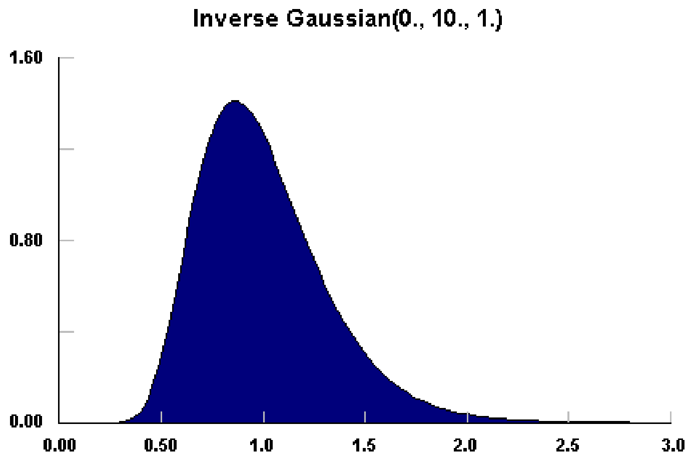
The Inverse Gaussian distribution is a continuous distribution with a bound on the lower side. It is uniquely zero at the minimum x, and always positively skewed. The Inverse Gaussian distribution is also known as the Wald distribution.

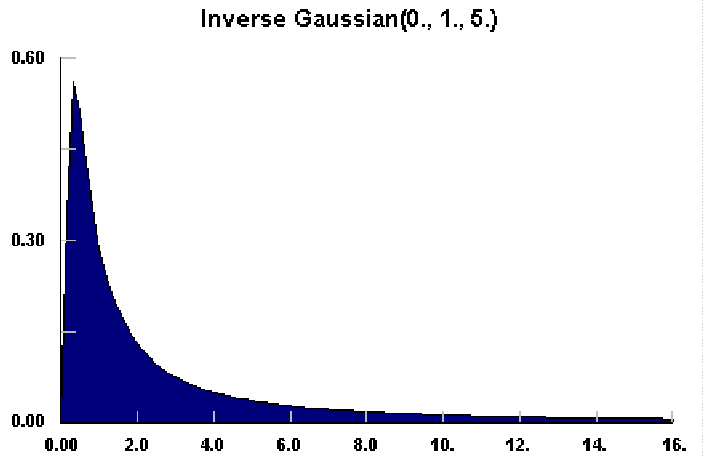
The Inverse Gaussian distribution was originally used to model Brownian motion and diffusion processes with boundary conditions. It has also been used to model the distribution of particle size in aggregates, reliability and lifetimes, and repair time. (see Johnson et. al.1)

Examples of Inverse Gaussian distribution are shown below. In particular, notice the drastically increased upper tail for increasing Beta. More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 290

Inverse Weibull(min, alpha, beta)

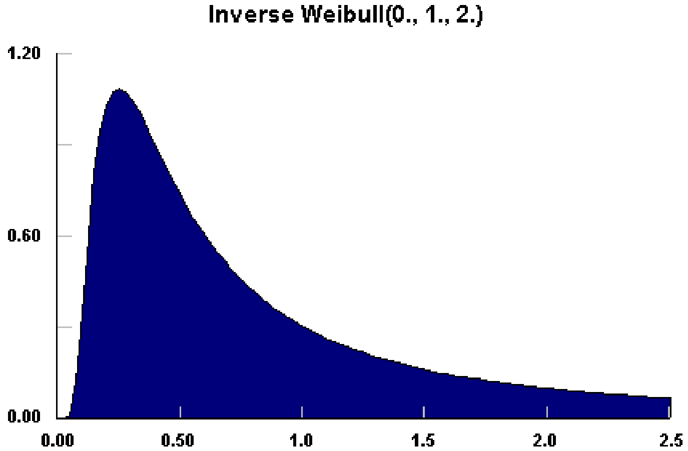
$$f(x) = \alpha\beta\left(\frac{1}{\beta(x-\min)}\right)^{\alpha+1} \exp\left(-\left(\frac{1}{\beta(x-\min)}\right)^{\alpha}\right)$$

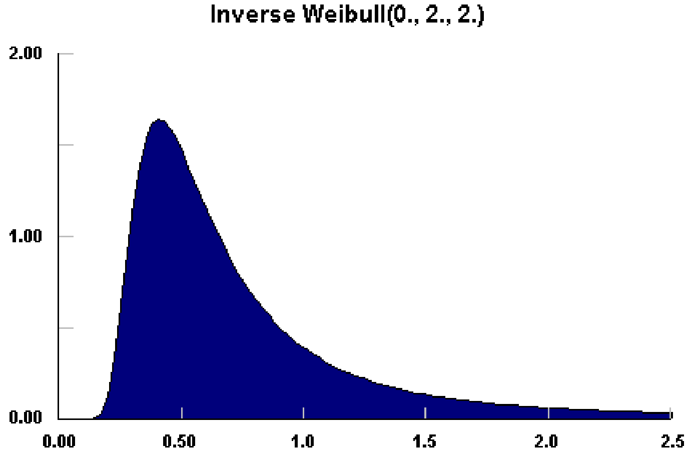
min = minimum x value
alpha = shape parameter > 0
beta = mixture of shape and scale > 0

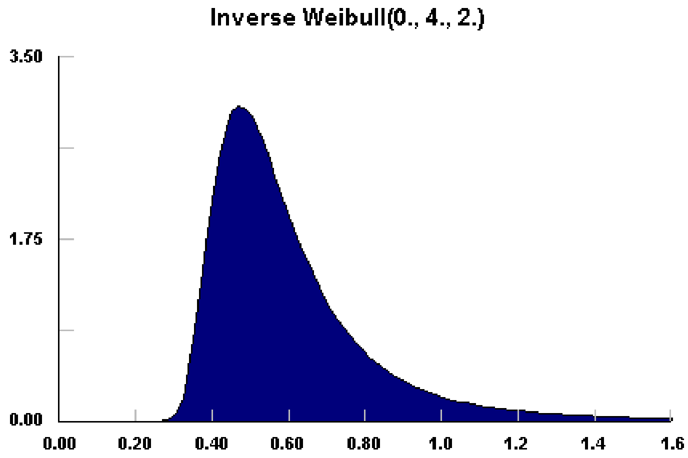
The Inverse Weibull distribution is a continuous distribution with a bound on the lower side. It is uniquely zero at the minimum x, and always positively skewed. In general, the Inverse Weibull distribution fits bounded, but very peaked, data with a long positive tail.

The Inverse Weibull distribution has been used to describe several failure processes as a distribution of lifetime. (see Calabria & Pulcini1) It can also be used to fit data with abnormal large outliers on the positive side of the peak.

Examples of Inverse Weibull distribution are shown below. In particular, notice the increased peakedness and movement from the minimum for increasing alpha. More examples can be viewed by using the Distribution Viewer capability.







1. R. Calabria, G. Pulcini, "On the maximum liklihood and least-squares estimation in the Inverse Weibull Distribution", Statistica Applicata, Vol. 2, n. 1, 1990, p53

Johnson SB(min, lambda, gamma, delta)

$$f(x) = \frac{\delta}{\sqrt{2\pi}\, y(1-y)\lambda} \exp\left(-\frac{1}{2}\left(\gamma + \delta \ln\left(\frac{y}{1-y}\right)\right)^2\right)$$

where

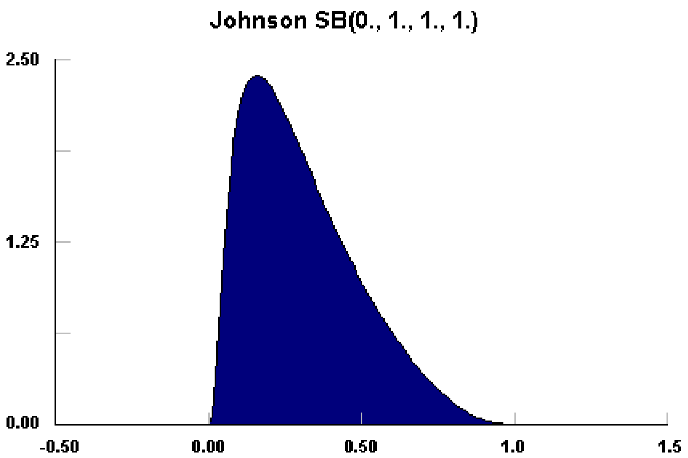
$$y = \frac{x - min}{\lambda}$$

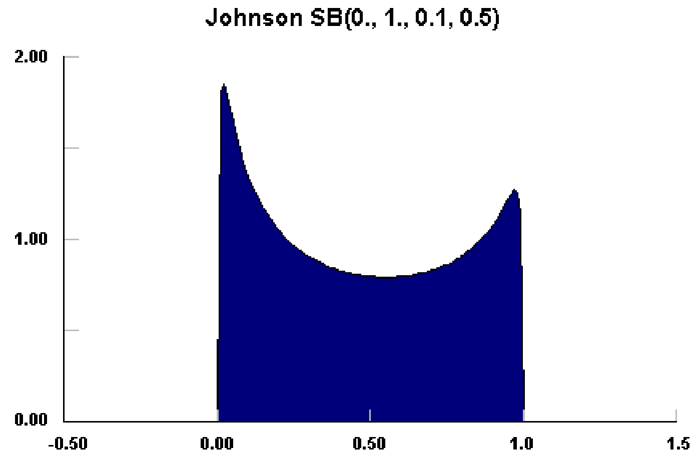
min = minimum value of x
lambda = range of x above minimum
gamma = skewness parameter
delta = shape parameter > 0

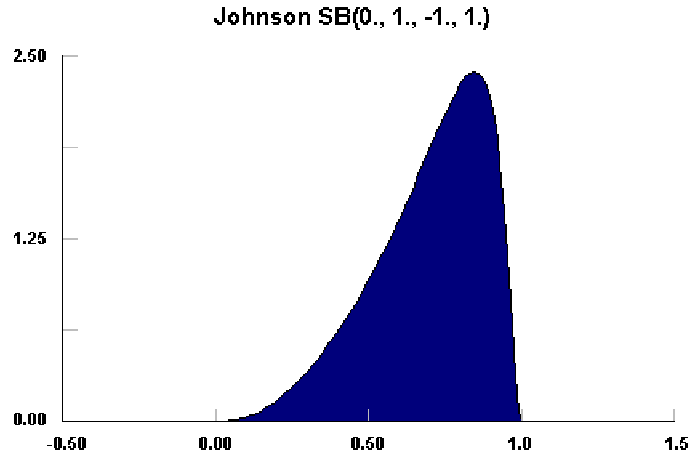
The Johnson SB distribution is a continuous distribution that has both upper and lower finite bounds, similar to the Beta distribution. The Johnson SB distribution, together with the Lognormal and the Johnson SU distributions, are transformations of the Normal distribution and can be used to describe most naturally occurring unimodal sets of data. However, the Johnson SB and SU distributions are mutually exclusive, each describing data in specific ranges of skewness and kurtosis. This leaves some cases where the natural boundedness of the population cannot be matched.

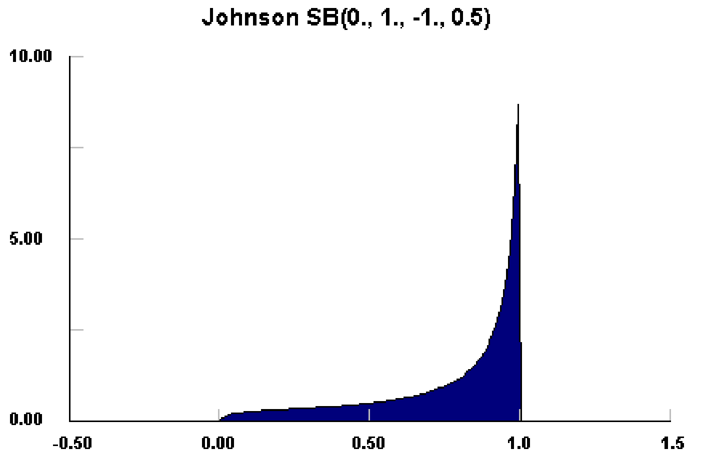
The family of Johnson distributions have been used in quality control to describe non-normal processes, which can then be transformed to the Normal distribution for use with standard tests.

As can be seen in the following examples, the Johnson SB distribution goes to zero at both of its bounds, with gamma controlling the skewness and delta controlling the shape. The distribution can be either unimodal or bimodal. (see Johnson et.al.1, and N.L. Johnson2) More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 34

2. N.L. Johnson, "Systems of frequency curves generated by methods of translation", Biometrika, Vol. 36, 1949, p149

Johnson SU(xi, lambda, gamma, delta)

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}\sqrt{y^2+1}} \exp\left(-\tfrac{1}{2}\left[\gamma+\delta\ln\left(y+\sqrt{y^2+1}\right)\right]^2\right)$$

where

$$y = \frac{x-\xi}{\lambda}$$

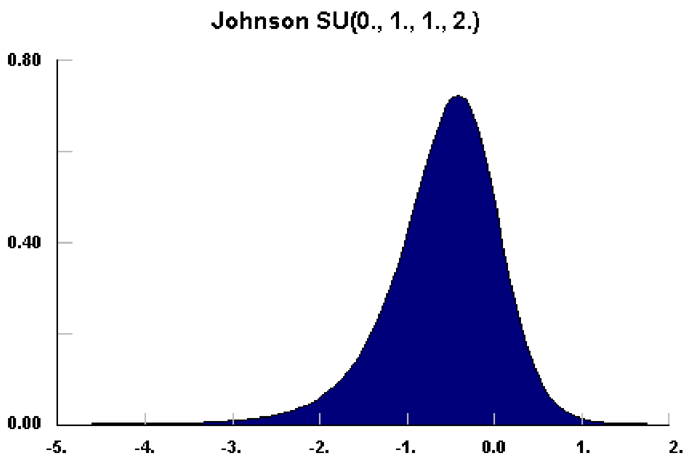
xi = location value of x
lambda = scale parameter > 0
gamma = skewness parameter
delta = kurtosis parameter > 0

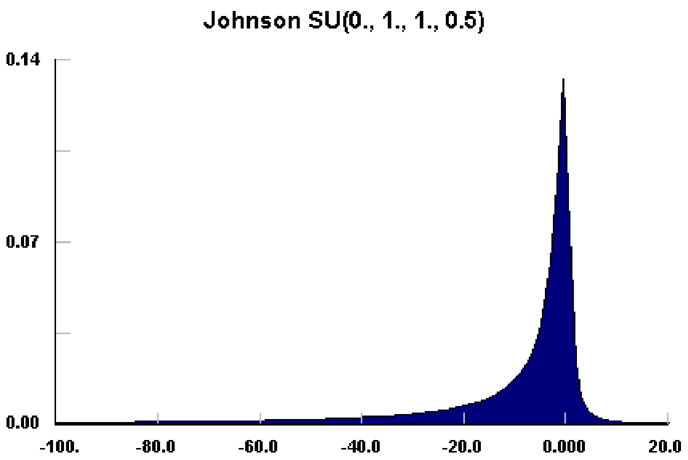
The Johnson SU distribution is an unbounded continuous distribution. The Johnson SU distribution, together with the Lognormal and the Johnson SB distributions, are transformations of the Normal distribution and can be used to describe most naturally occurring unimodal sets of data. However, the Johnson SB and SU distributions are mutually exclusive, each describing data in specific ranges of skewness and kurtosis. This leaves some cases where the natural boundedness of the population cannot be matched.

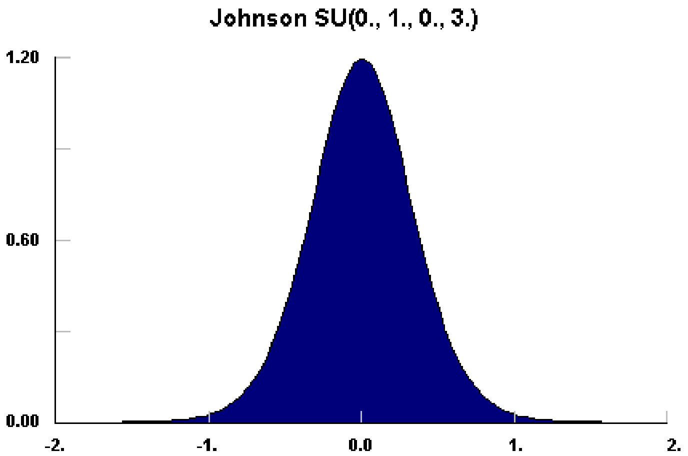
The family of Johnson distributions have been used in quality control to describe non-normal processes, which can then be transformed to the Normal distribution for use with standard tests.

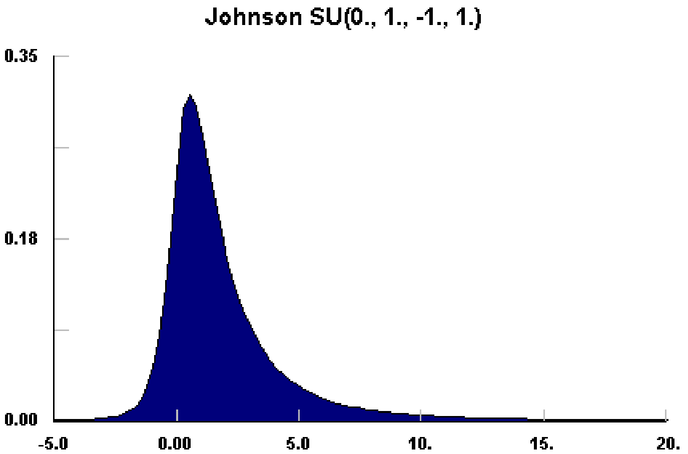
The Johnson SU distribution can be used in place with the notoriously unstable Pearson IV distribution, with reasonably good fidelity over the most probable range of values.

As can be seen in the following examples, the Johnson SU distribution is one of the few unbounded distributions that can vary its shape, with gamma controlling the skewness and delta controlling the shape. The scale is controlled by gamma, delta, and lambda. (see Johnson et.al.1, and N.L. Johnson2) More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 34

2. N.L. Johnson, "Systems of frequency curves generated by methods of translation", Biometrika, Vol. 36, 1949, p149

Laplace(theta, phi)

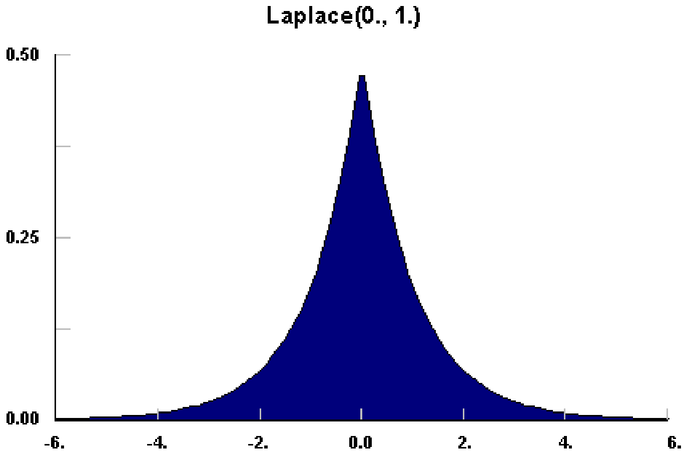
$$f(x) = \frac{1}{2\pi}\exp\left(-\left|x-\theta\right|/\varphi\right)$$

theta = mode or central peak position
phi = scale parameter

The Laplace distribution, sometimes called the double exponential distribution, is an unbounded continuous distribution that has a very sharp central peak, located at theta. The distribution scales with phi.

The Laplace distribution can be used to describe the difference of two independent, and equally distributed, exponentials..It is also used in error analysis.. (see Johnson et.al.1)

The Laplace distribution, as shown below, has a distinct spike at its mode. It is unchanged in shape with changes in theta or phi. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p164

Logistic(alpha, beta)

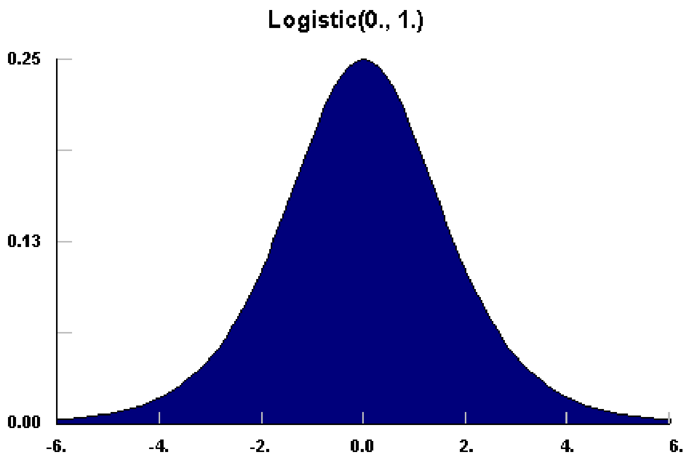
$$f(x) = \frac{\exp\left(-\frac{[x-\alpha]}{\beta}\right)}{\beta\left[1+\exp\left(-\frac{[x-\alpha]}{\beta}\right)\right]^2}$$

alpha = shift parameter
beta = scale parameter > 0

The Logistic distribution is an unbounded continuous distribution which is symmetrical about its mean [and shift parameter], alpha. As shown in the example, the shape of the Logistic distribution is very much like the Normal distribution, except that the Logistic distribution has broader tails.

The Logistic function is most often used a growth model: for populations, for weight gain, for business failure, etc.. The Logistic distribution can be can be used to test for the suitability of such a model, with transformation to get back to the minimum and maximum values for the Logistic function. Occasionally, the Logistic function is used in place of the Normal function where exceptional cases play a larger role.(see Johnson et. al.1)

More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p113

Loglogistic(min, p, beta)

$$f(x) = \frac{p\left(\frac{x-\min}{\beta}\right)^{p-1}}{\beta\left[1+\left(\frac{x-\min}{\beta}\right)^{p}\right]^{2}}$$

min = minimum x value
p = shape parameter > 0
beta = scale parameter > 0

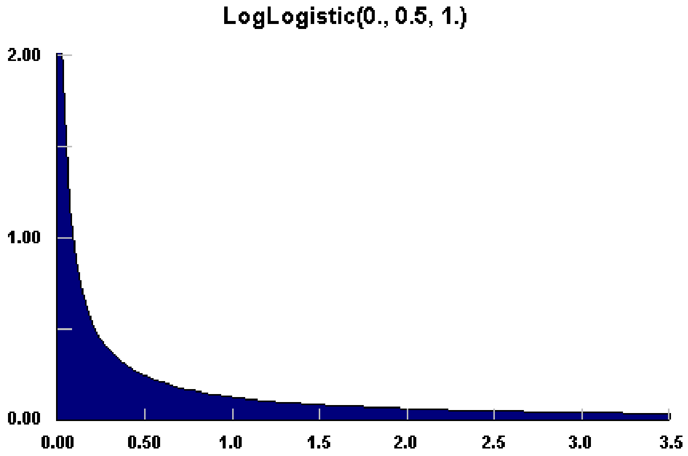
The Loglogisic distribution is a continuous distribution bounded on the lower side. Like the Gamma distribution, it has three distinct regions. For p = 1, the Loglogistic distribution resembles the Exponential distribution, starting at a finite value at minimum x and decreasing monotonically thereafter. For p < 1, the Loglogistic distribution tends to infinity at minimum x and decreases monotonically for increasing x. For p > 1, the Loglogistic distribution is 0 at minimum x, peaks at a value that depends on both p and beta, decreasing monotonically thereafter.

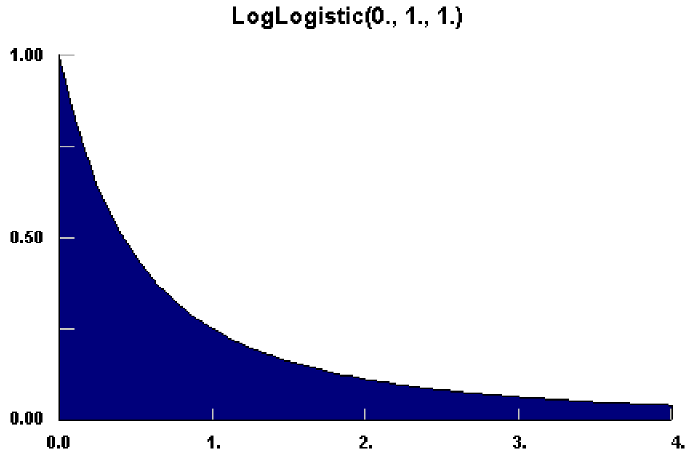
By definition, the natural logarithm of a Loglogistic random variable is a Logistic random variable, and can be related to the included Logistic distribution in much the same way that the Lognormal distribution can be related to the included Normal distribution. The parameters for the included Logistic distribution, Lalpha and Lbeta, are given in terms of the Loglogistic parameters, LLp and LLbeta, by

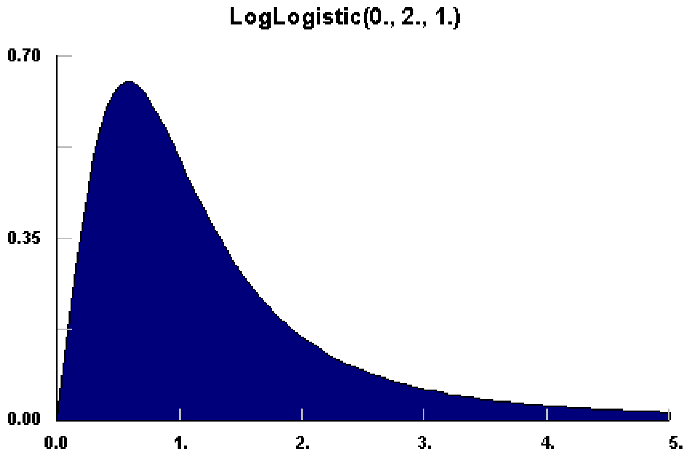
Lalpha = ln(LLbeta)
LBeta = 1/LLp

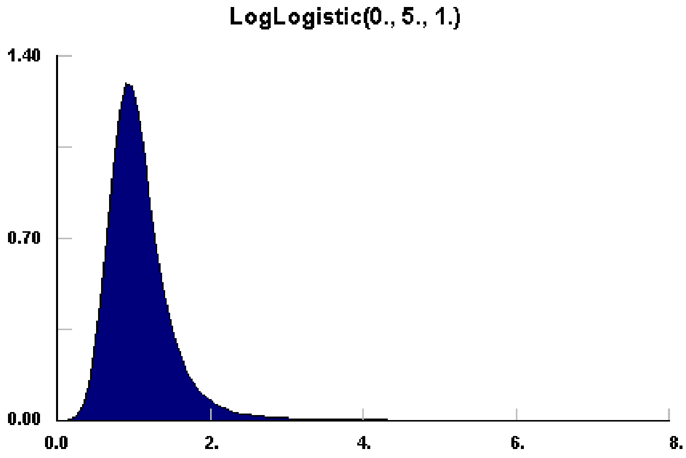
The Loglogistic distribution is used to model the output of complex processes such as business failure, product cycle time, etc. (see Johnson et. al.1).

Note for p = 1, The Loglogistic distribution decreases more rapidly than the Exponential distribution but has a broader tail. For large p, the distribution becomes more symmetrical and moves away from the minimum. More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p151

Lognormal(min, mu, sigma)

$$f(x) = \frac{1}{(x-\min)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[\ln(x-\min)-\mu]^2}{2\sigma^2}\right)$$

min = minimum x value
mu = mean of the included Normal
sigma = standard deviation of the included Normal

The Lognormal distribution is a continuous distribution bounded on the lower side. It is always 0 at minimum x, rising to a peak that depends on both mu and sigma, then decreasing monotonically for increasing x.

By definition, the natural logarithm of a Lognormal random variable is a Normal random variable. Its parameters are usually given in terms of this included Normal.

The Lognormal distribution can also be used to approximate the Normal distribution, for small sigma, while maintaining its strictly positive values of x [actually (x-min)].

The Lognormal distribution is used in many different areas including the distribution of particle size in naturally occurring aggregates, dust concentration in industrial atmospheres, the distribution of minerals present in low, concentrations, duration of sickness absence, physicians' consultant time, lifetime distributions in reliability, distribution of income, employee retention, and many applications modeling weight, height, etc.(see Johnson et. al.1)

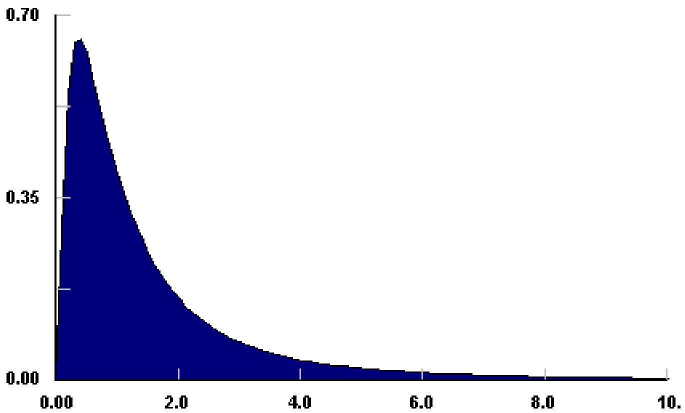
The Lognormal distribution can provide very peaked distributions for increasing sigma, indeed, far more peaked than can be easily represented in graphical form. More examples can be viewed by using the Distribution Viewer capability, which can also be used to convert the mean and variance to mu and sigma.

To convert the parameters of the included Normal to the mean and variance of the Lognormal data:

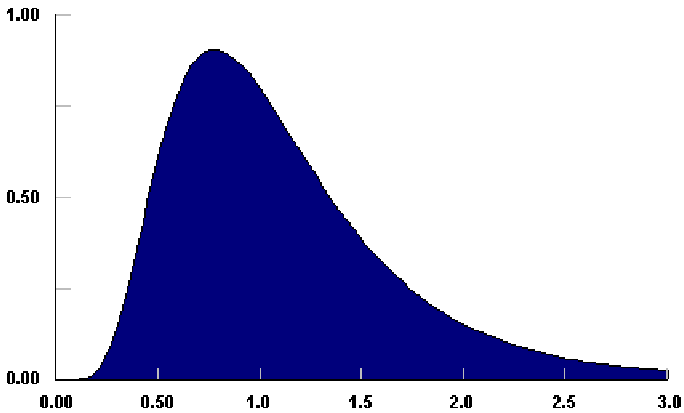
mean = min + exp(mu + sigma^2/2)
variance = exp(2*mu + sigma^2)*(exp(sigma^2) - 1)

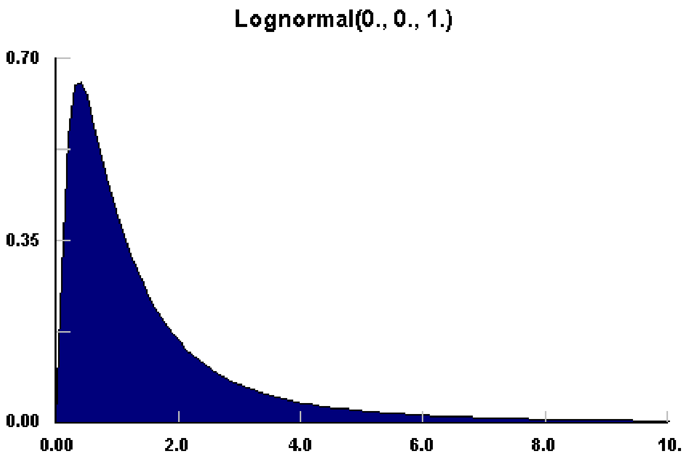
....

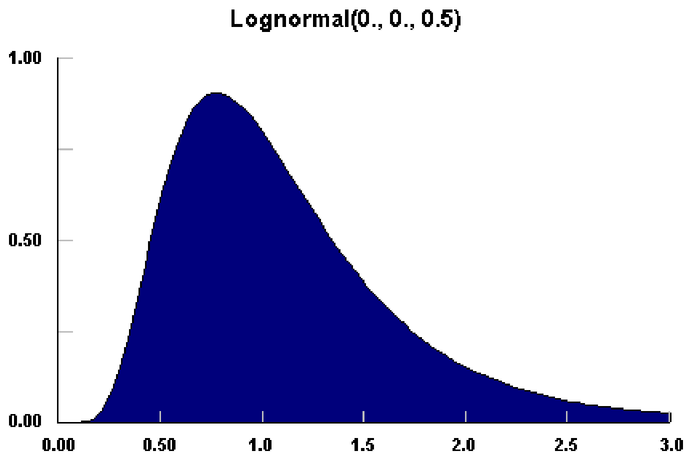
Lognormal(0., 0., 1.)



Lognormal(0., 0., 0.5)







1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p207

Normal(mu, sigma)

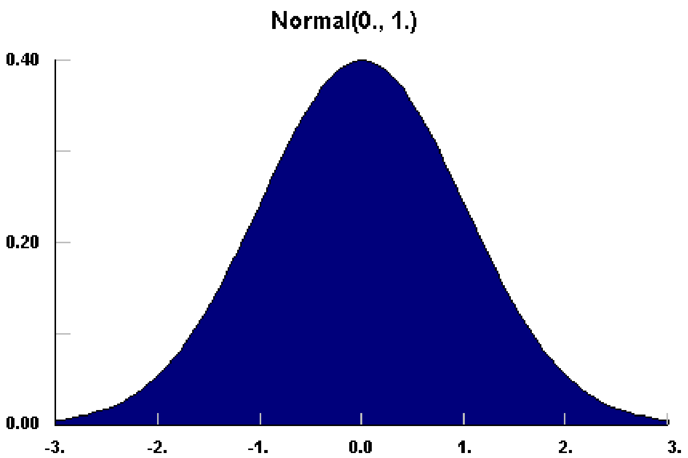
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[x-\mu]^2}{2\sigma^2}\right)$$

mu = shift parameter = mean
sigma = scale parameter = standard deviation

The Normal distribution is a unbounded continuous distribution. It sometimes called a Gaussian distribution or the bell curve. Because of its property of representing an increasing sum of small, independent errors, the Normal distribution finds many, many uses in statistics. It is wrongly used in many situations. Possibly, the most important test in the fitting of analytical distributions is the elimination of the Normal distribution as a possible candidate. (see Johnson et. al.1)

The Normal distribution is used as an approximation for the Binomial distribution when the values of n,p are in the appropriate range. The Normal distribution is frequently used to represent symmetrical data, but suffers from being unbounded in both directions. If the data is known to have a lower bound, it may be better represented by suitable parametrization of the Lognormal, Weibull, or Gamma distributions. If the data is known to have both upper and lower bounds, the Beta distribution can be used, although much work has been done on truncated Normal distributions (not supported in Stat::Fit).

The Normal distribution, as shown below, has the familiar bell shape. It is unchanged in shape with changes in mu or sigma. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p80

Pareto(min, alpha)

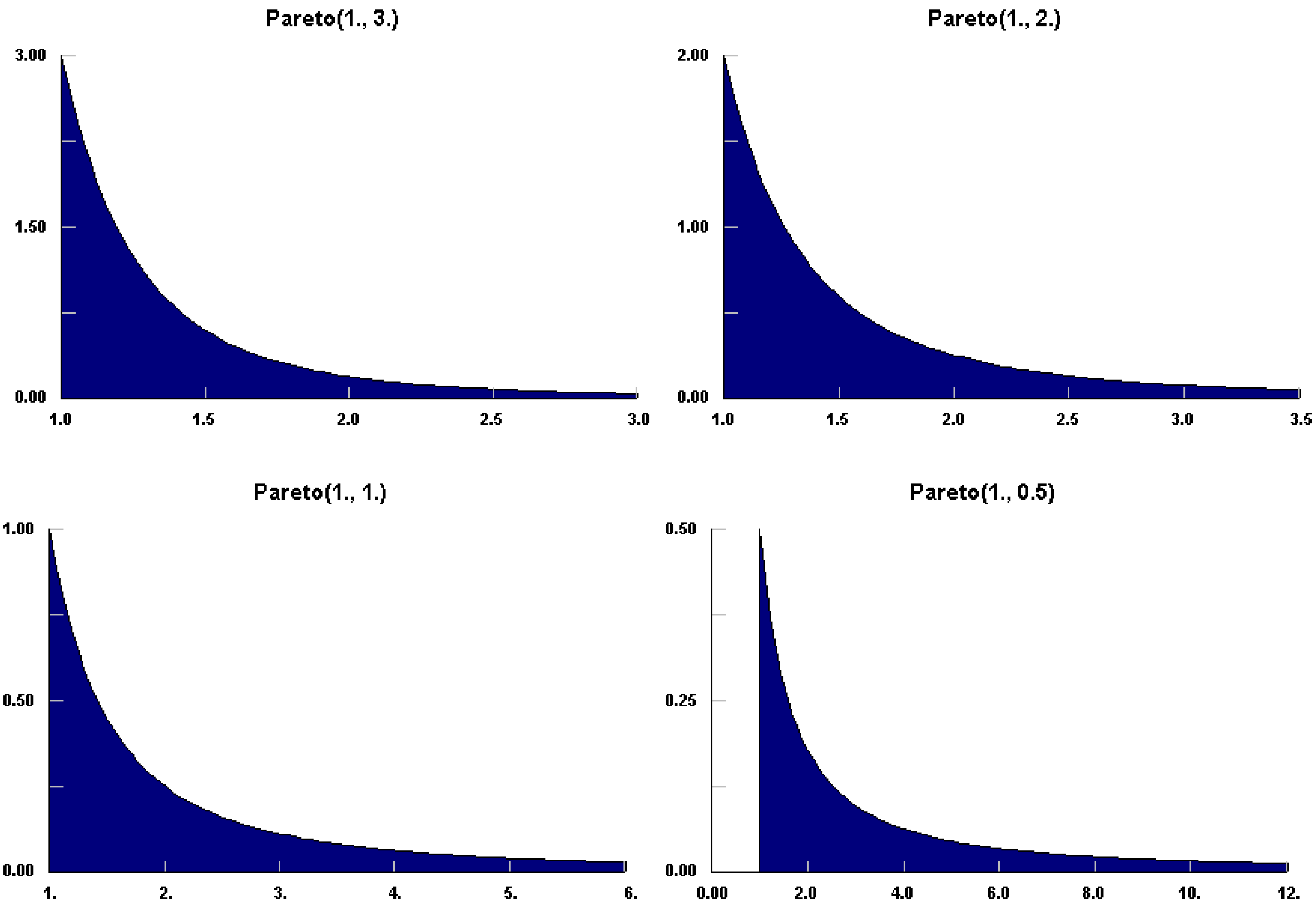
$$f(x) = \frac{\alpha \ \min^{\alpha}}{x^{\alpha+1}}$$

min = minimum x value > 0
alpha = scale parameter > 0

The Pareto distribution is a continuous distribution bounded on the lower side. It has a finite value at the minimum x and decreases monotonically for increasing x. A Pareto random variable is the exponential of an Exponential random variable, and possesses many of the same characteristics.

The Pareto distribution has, historically, been used to represent the income distribution of a society. It is also used to model many empirical phenomena with very long right tails, such as city population sizes, occurrence of natural resources, stock price fluctuations, size of firms, brightness of comets, and error clustering in communication circuits. (see Johnson et. al.1).

The shape of the Pareto curve changes slowly with alpha, but the tail of the distribution increases dramatically with decreasing alpha. More examples can be viewed by using the Distribution Viewer capability.



1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p 607

Pearson 5(min, alpha, beta)

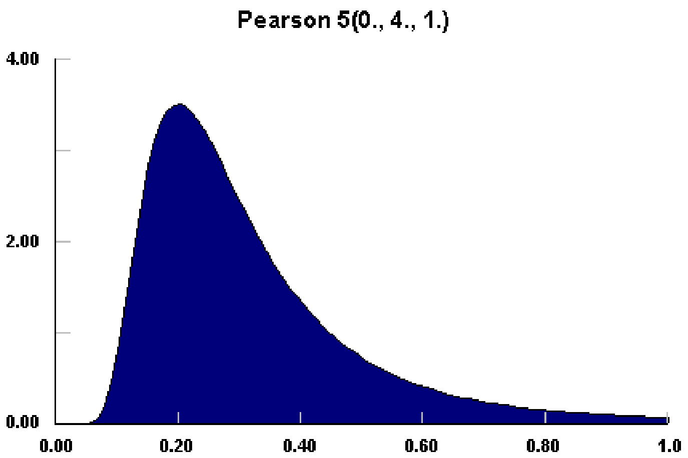
$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)(x - \min)^{\alpha+1}} \exp\left(-\frac{\beta}{[x - \min]}\right)$$

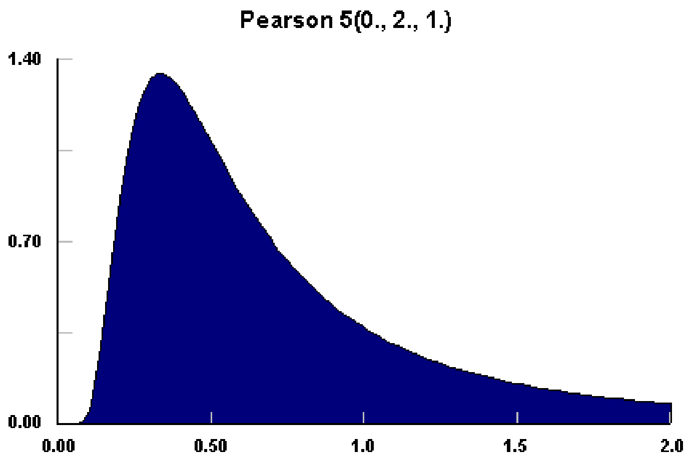
min = minimum x value
alpha = shape parameter > 0
beta = scale parameter > 0

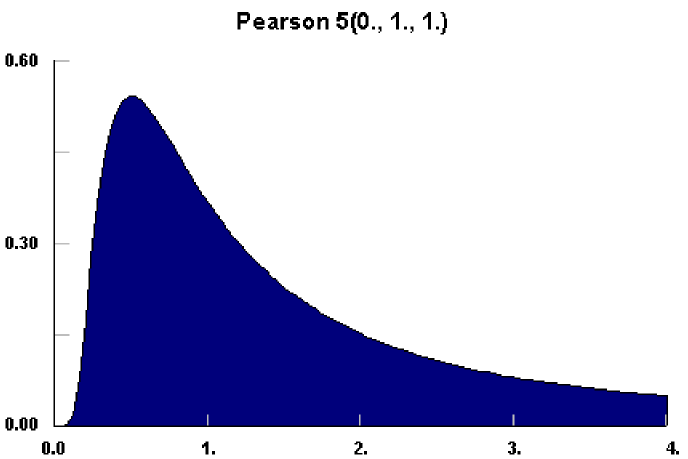
The Pearson 5 distribution is a continuous distribution with a bound on the lower side. The Pearson 5 distribution is sometimes called the Inverse Gamma Distribution due to the reciprocal relationship between a Pearson 5 random variable and a Gamma random variable.

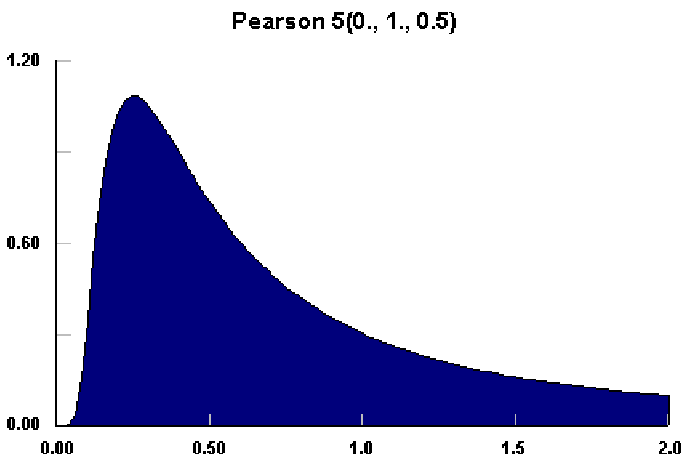
The Pearson 5 distribution is useful for modeling time delays where some minimum delay value is almost assured and the maximum time is unbounded and variably long, such as time to complete a difficult task, time to respond to an emergency, time to repair a tool, etc.. Similar space situations also exist such as manufacturing space for a given process. (see Law&Kelton1).

The Pearson 5 distribution starts slowly near its minimum and has a peak slightly removed from it, as shown below. With decreasing alpha, the peak gets flatter [see vertical scale] and the tail gets much broader. More examples can be viewed by using the Distribution Viewer capability.









1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 339

Pearson 6(min, beta, p, q)

$$f(x) = \frac{\left(\frac{x - min}{\beta}\right)^{p-1}}{\beta\left[1+\left(\frac{x-min}{\beta}\right)\right]^{p+q} B(p,q)}$$

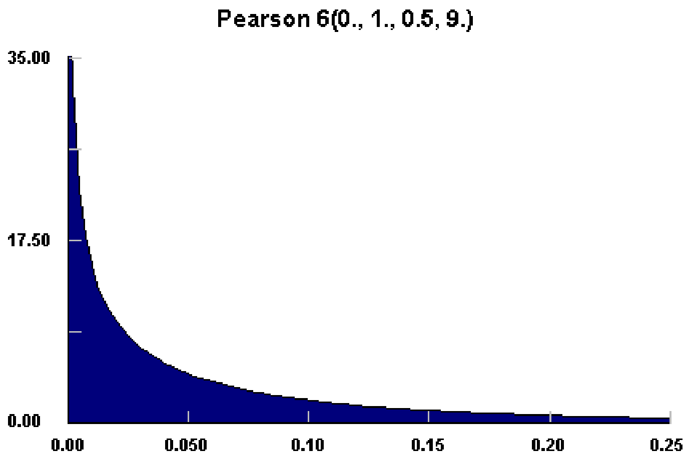
min = minimum x value
beta = scale parameter > 0
p = shape parameter > 0
q = shape parameter > 0

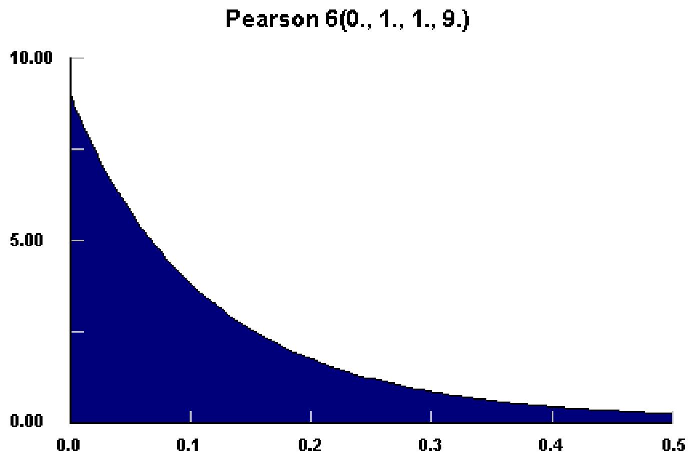
The Pearson 6 distribution is a continuous distribution bounded on the low side. The Pearson 6 distribution is sometimes called the Beta distribution of the second kind due to the relationship of a Pearson 6 random variable to a Beta random variable. When min = 0., beta = 1, p = nu1/2 and q = nu2/2, the the Pearson 6 distribution reduces to the F distribution of nu1, nu2 which is used for many statistical tests of goodness of fit. (see Johnson et. al.1)

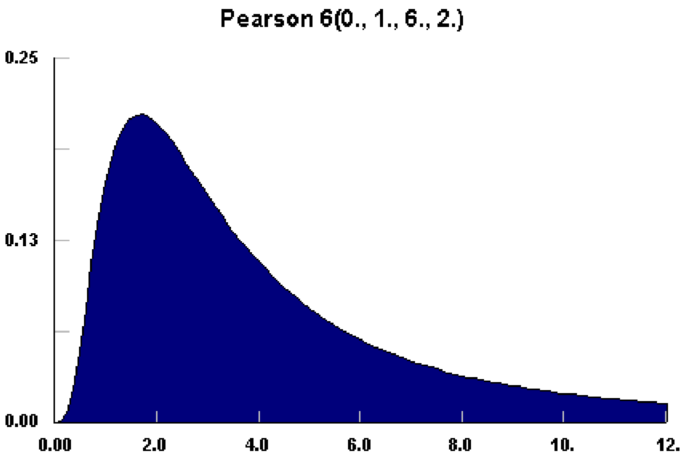
Like the Gamma distribution, it has three distinct regions. For p = 1, the Pearson 6 distribution resembles the Exponential distribution, starting at a finite value at minimum x and decreasing monotonically thereafter. For p < 1, the Pearson 6 distribution tends to infinity at minimum x and decreases monotonically for increasing x. For p > 1, the Pearson 6 distribution is 0 at minimum x, peaks at a value that depends on both p and q, decreasing monotonically thereafter.

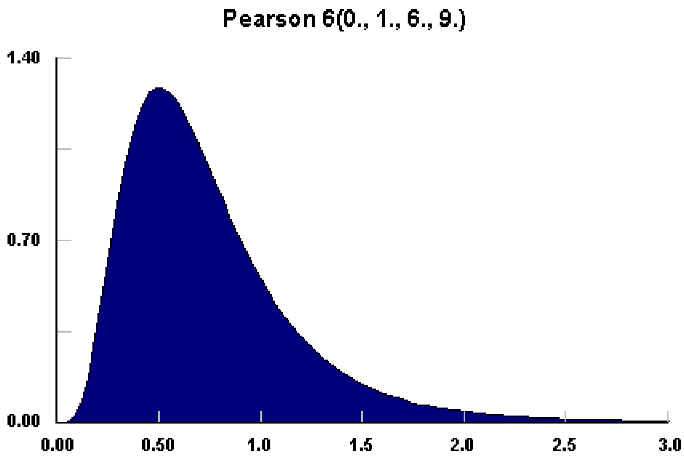
The Pearson 6 distribution appears to have found little direct use, except in its reduced form as the F distribution where it serves as the distribution of the ratio of independent estimators of variance and provides the final test for the analysis of variance.

The three regions of the Pearson 6 distribution are shown below. Also note that the distribution becomes sharply peaked just off the minimum for increasing q. More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p322

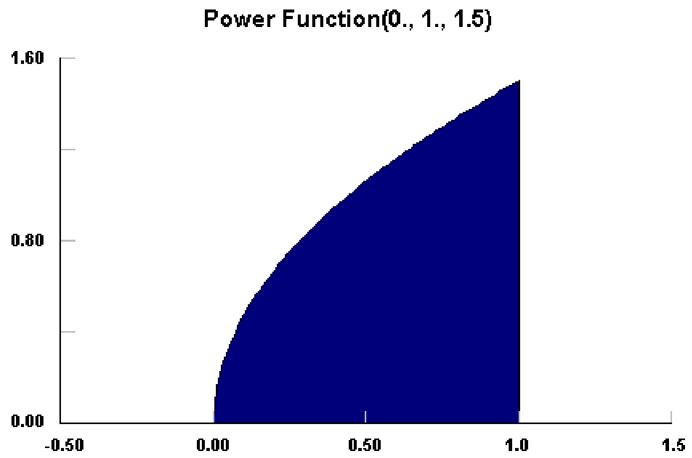
Power Function(min, max, alpha)

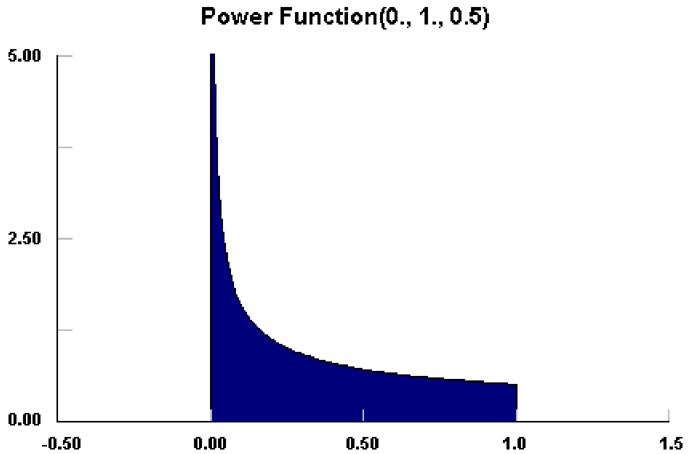
$$f(x) = \frac{\alpha (x - \min)^{\alpha - 1}}{(\max - \min)^{\alpha}}$$

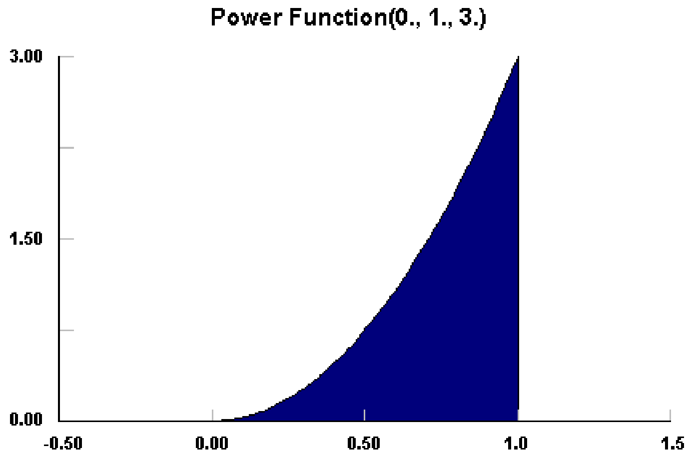
min = minimum x value
max = maximum x value
alpha = shape parameter > 0

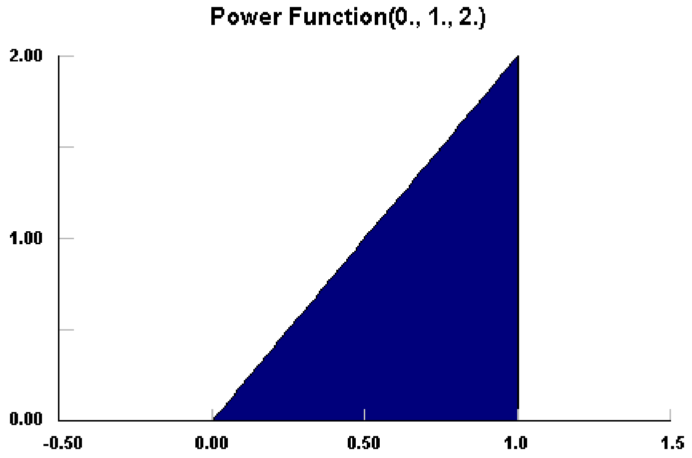
The Power Function distribution is a continuous distribution that has both upper and lower finite bounds, and is a special case of the Beta distribution with q = 1. (see Johnson et.al.1) The Uniform distribution is a special case of the Power Function distribution with p = 1.

As can be seen in the following examples, The Power Function distribution can approach zero or infinity at its lower bound, but always has a finite value at its upper bound. Alpha controls the value at the lower bound as well as the shape. More examples can be viewed by using the Distribution Viewer capability.









1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons., p210

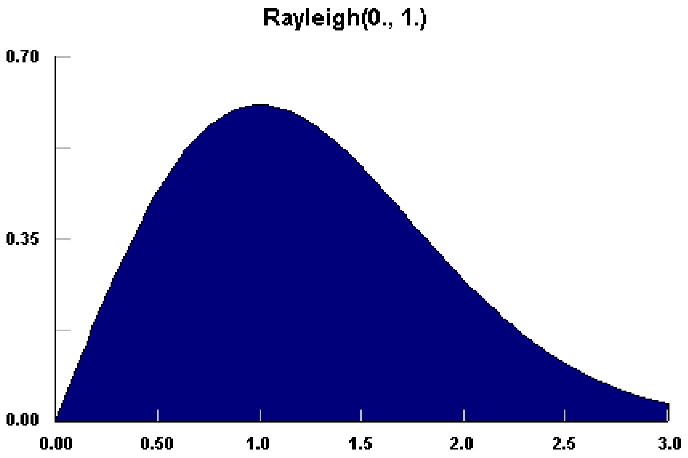
Rayleigh(min, sigma)

$$f(x) = \frac{(x - min)}{\sigma^2} \exp\left(-\frac{(x - min)^2}{2\sigma^2} \right)$$

min = minimum x value
sigma = scale parameter > 0

The Rayleigh distribution is a continuous distribution bounded on the lower side. It is a special case of the Weibull distribution with alpha = 2 and beta/sqrt(2) = sigma. Because of the fixed shape parameter, the Rayleigh distribution does not change shape although it can be scaled. More examples can be viewed by using the Distribution Viewer capability.

The Rayleigh distribution is frequently used to represent lifetimes because its hazard rate increases linearly with time, e.g. the lifetime of vacuum tubes. This distribution also finds application in noise problems in communications. (see Johnson et.al.1 and Shooman2)



1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p456

2. "Probabilistic Reliability: An Engineering Approach", Martin L. Shooman, 1990, Robert E. Krieger, p48

Triangular(min, max, mode)

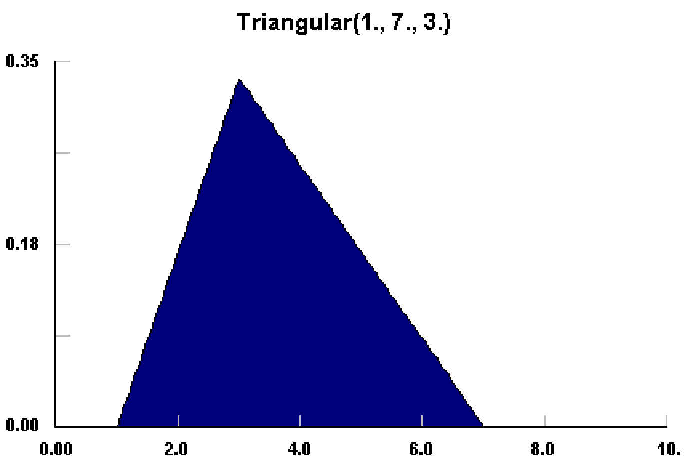
$$f(x) = \begin{cases} \dfrac{2(x - \min)}{(\max - \min)(\text{mode} - \min)} \\ \\ \dfrac{2(\max - x)}{(\max - \min)(\max - \text{mode})} \end{cases}$$

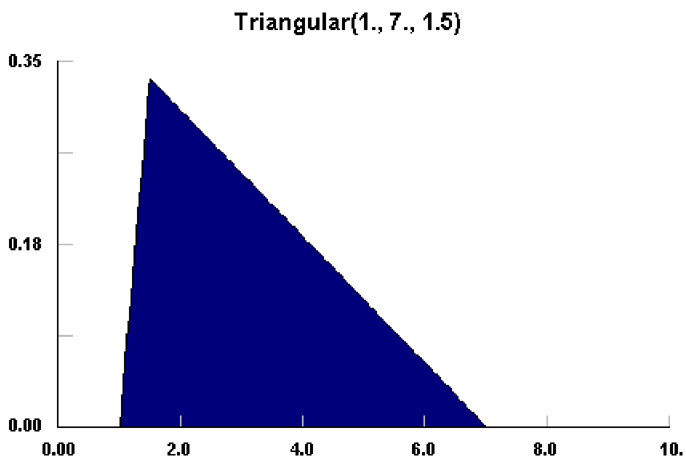
min = minimum x value
max = maximum x value
mode = most likely x

The Triangular distribution is a continuous distribution bounded on both sides.

The Triangular distribution is often used when no or little data is available; it is rarely an accurate representation of a data set. (see Law&Kelton1) As a substitute for situations with little data, it will often have excess variance. However, it is employed as the functional form of regions for fuzzy logic due to its ease of use.

The Triangular distribution can take on very skewed forms, as shown below, including negative skewness. For the exceptional cases where the mode is either the min or max, the Triangular distribution becomes a right triangle. More examples can be viewed by using the Distribution Viewer capability.





1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 341

Uniform(min, max)

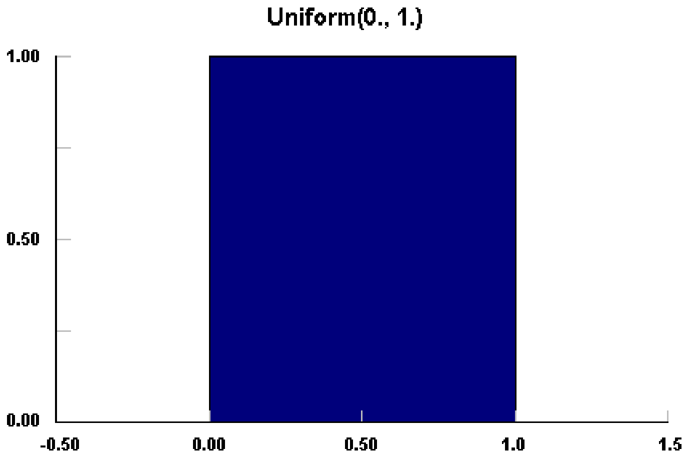
$$f(x) = \frac{1}{max - min}$$

min = minimum x value
max = maximum x value

The Uniform distribution is a continuous distribution bounded on both sides. Its density does not depend on the value of x. It is a special case of the Beta distribution. It is frequently called the rectangular distribution. (see Johnson et. al.1). Most random number generators provide samples from the Uniform distribution on (0,1) and then convert these samples to random variates from other distributions.

The Uniform distribution is used to represent a random variable with constant likelihood of being in any small interval between min and max. Note that the probability of either the min or max value is 0; the end points do NOT occur. If the end points are necessary, try the sum of two opposing right Triangular distributions.

More examples can be viewed by using the Distribution Viewer capability, but the view doesn't change..



1. "Continuous Univariate Distributions, Volume 2", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p276

Weibull(min, alpha, beta)

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\min}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{[x-\min]}{\beta}\right)^{\alpha}\right)$$

min = minimum x value
alpha = shape parameter > 0
beta = scale parameter > 0

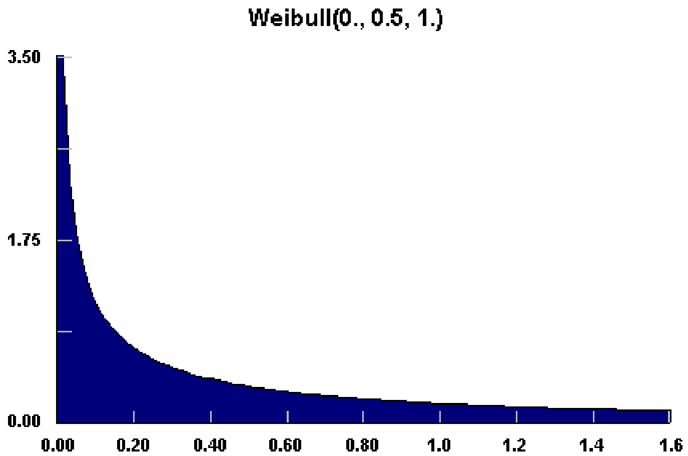
The Weibull distribution is a continuous distribution bounded on the lower side. Because it provides one of the limiting distributions for extreme values, it is also referred to as the Frechet Distribution and the Weibull-Gnedenko distribution. Unfortunately, the Weibull distribution has been given various functional forms in the many engineering references; the form above is the standard form given in Johnson et. al.1.

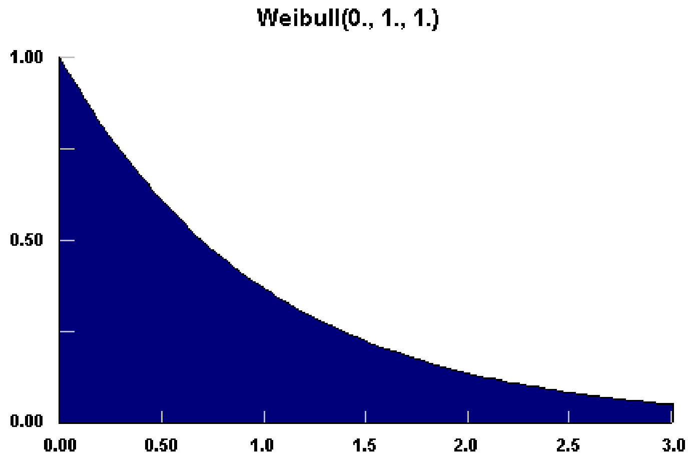
Like the Gamma distribution, it has three distinct regions. For alpha = 1, the Weibull distribution is reduced to the Exponential distribution, starting at a finite value at minimum x and decreasing monotonically thereafter. For alpha < 1, the Weibull distribution tends to infinity at minimum x and decreases monotonically for increasing x. For alpha > 1, the Weibull distribution is 0 at minimum x, peaks at a value that depends on both alpha and beta, decreasing monotonically thereafter. Uniquely, the Weibull distribution has negative skewness for alpha > 3.6.

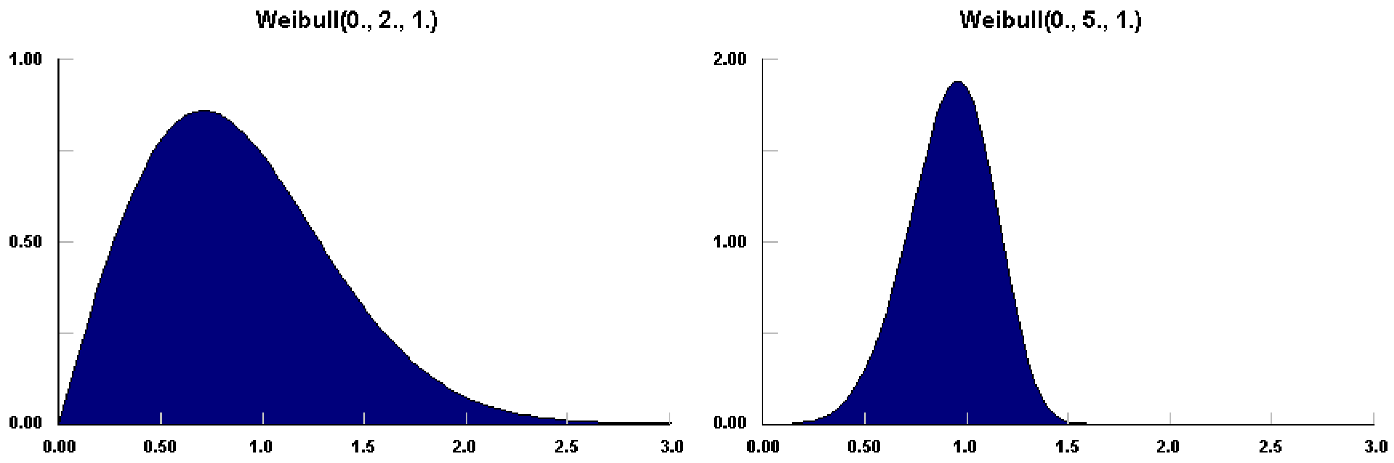
The Weibull distribution can also be used to approximate the Normal distribution for alpha = 3.6..., while maintaining its strictly positive values of x [actually (x-min)], although the kurtosis is slightly smaller than 3, the Normal value.

The Weibull distribution derived its popularity from its use to model the strength of materials, and has since been used to model just about everything. In particular, the Weibull distribution is used to represent wear out lifetimes in reliability, wind speed, rainfall intensity, health related issues, germination, duration of industrial stoppages, migratory systems, and thunderstorm data. (see Johnson et. al.1 and Shooman2)

Examples of each of the regions of the Weibull distribution are shown below, along with an example of the region of negative skewness. More examples can be viewed by using the Distribution Viewer .







1. "Continuous Univariate Distributions, Volume 1", Norman L. Johnson. Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p628

2. "Probabilistic Reliability: An Engineering Approach", Martin L. Shooman, 1990, Robert E. Krieger, p190

Binomial(n, p)

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

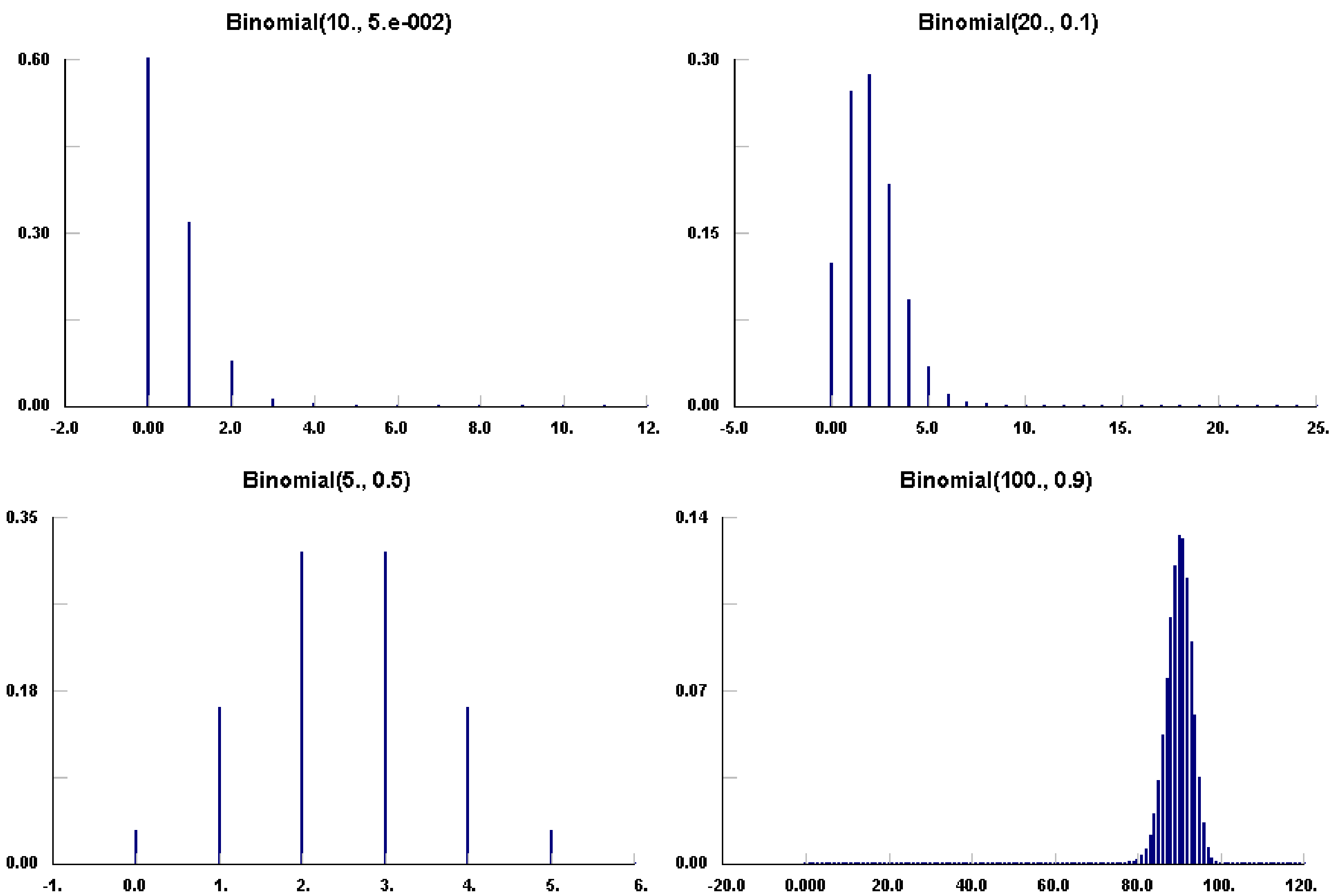
n = number of trials
p = probability of event occurring

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

The Binomial distribution is a discrete distribution bounded by [0,n]. Typically, it is used where a single trial is repeated over and over, such as the tossing of a coin. The parameter, p, is the probability of the event, either heads or tails, either occurring or not occurring. Each single trial is assumed to be independent of all others. For large n, the Binomial distribution may be approximated by the Normal distribution, for example when np>9 and p <0.5 or when np(1-p)>9.

As shown in the examples, low values of p give high probabilities for low values of x and visa versa, so that the peak in the distribution may approach either bound. More examples can be viewed by using the Distribution Viewer capability.

The binomial distribution has had extensive use in games, but is also useful in genetics, sampling of defective parts in a stable process, and other event sampling tests where the probability of the event is known to be constant or nearly so. (see Johnson, et. al.2)



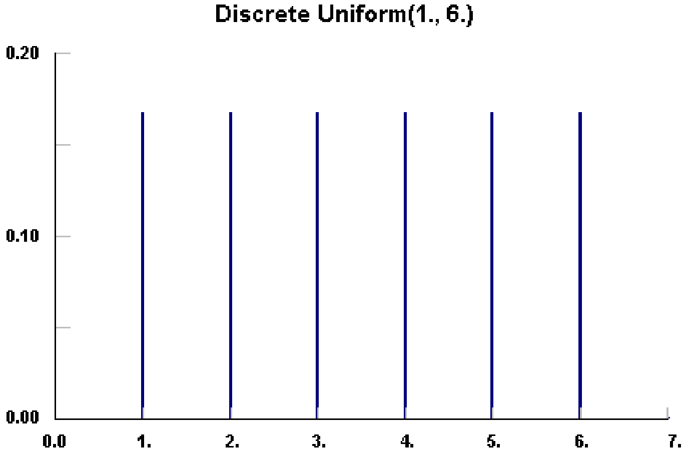
2. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.134

Discrete Uniform(min, max)

$$p(x) = \frac{1}{max - min + 1}$$

min = minimum x value
max = maximum x value

The Discrete Uniform distribution is a discrete distribution bounded on [min, max] with constant probability at every value on or between the bounds. Sometimes called the discrete rectangular distribution, it arises when an event can have a finite and equally probable number of outcomes. (see Johnson et. al.3). More examples can be viewed by using the Distribution Viewer capability.



3. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.272

Geometric(p)

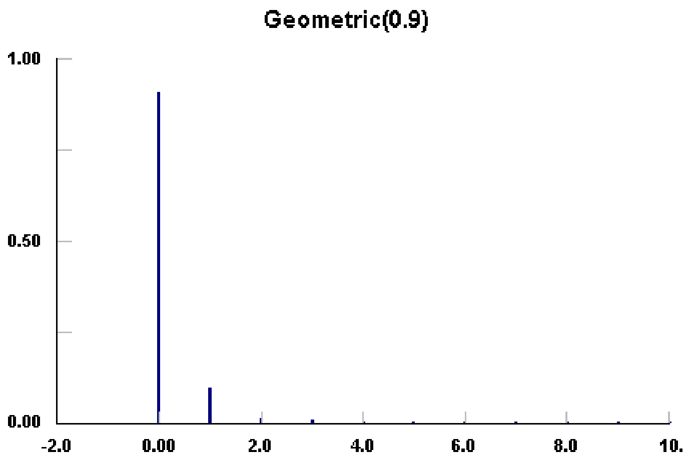
$$p(x) = p(1-p)^x$$

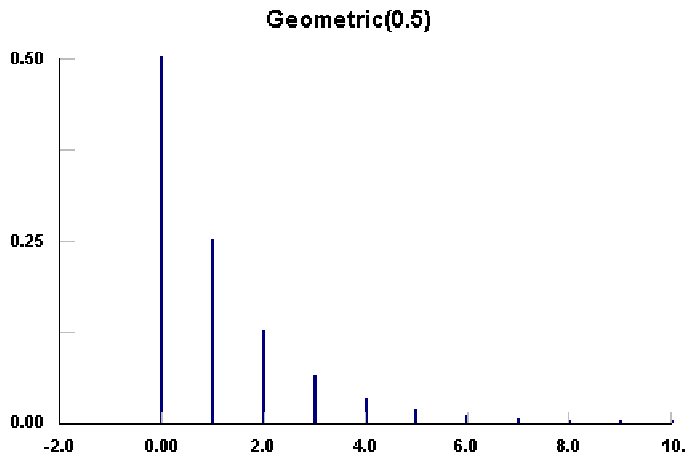
p = probability of occurrence

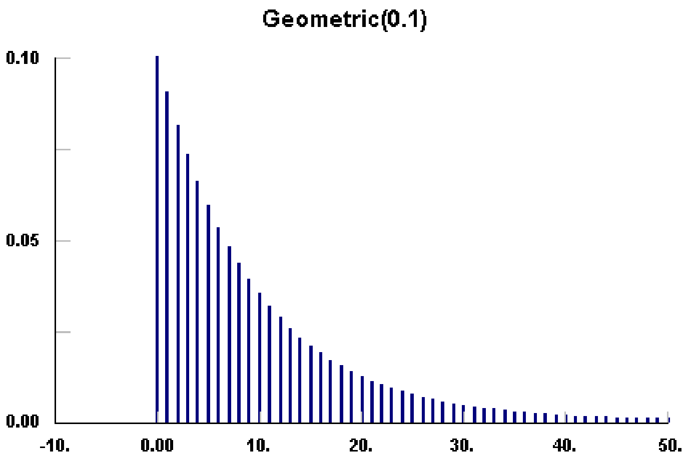
The Geometric distribution is a discrete distribution bounded at 0 and unbounded on the high side. It is a special case of the Negative Binomial distribution. In particular, it is the direct discrete analog for the continuous Exponential distribution. The Geometric distribution has no history dependence, its probability at any value being independent of a shift along the axis.

The Geometric distribution has been used for inventory demand, marketing survey returns, a ticket control problem, and meteorological models. (see Johnson et. al.1, Law&Kelton2)

Several examples with decreasing probability are shown below. More examples can be viewed by using the Distribution Viewer capability.







1. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.201

2. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 366

Hypergeometric(s, m, N)

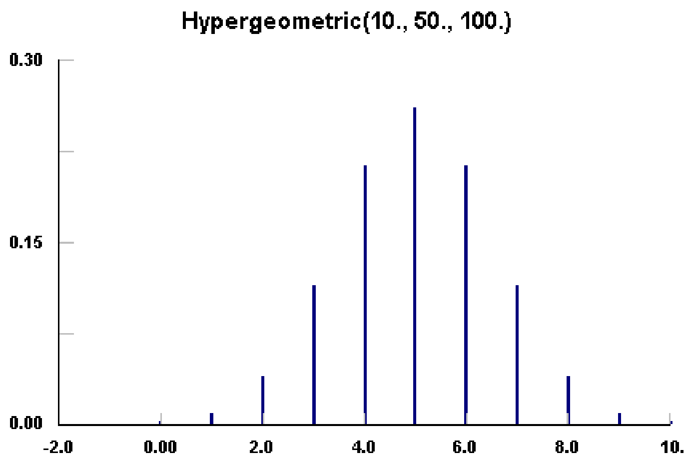
$$p(x) = \frac{m!(N-m)!s!(N-s)!}{x!(m-x)!N!(s-x)!(N-m-s+x)!}$$

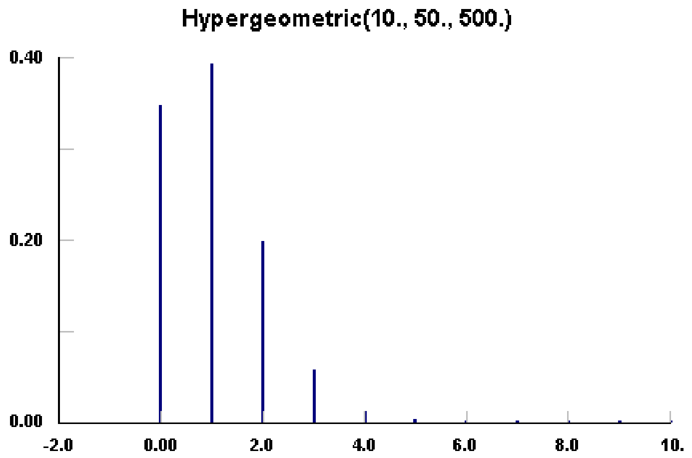
s = sample size
m = number of defects in a population
N = size of the population

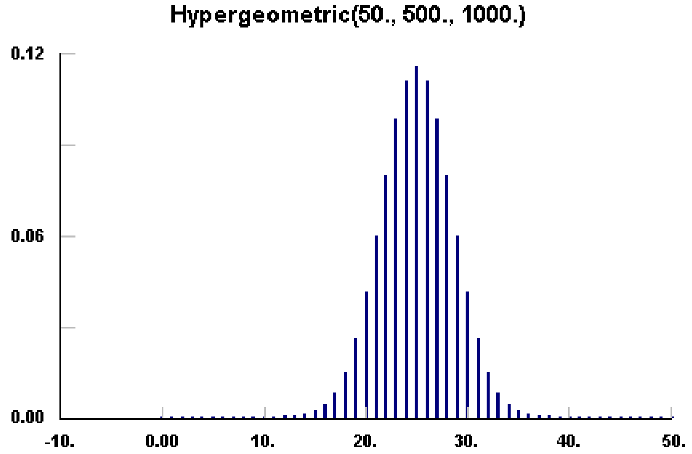
The Hypergeometric distribution is a discrete distribution bounded by [0,s]. It describes the number of defects, x, in a sample of size s from a population of size N which has m total defects. The ratio of m/N = p is sometimes used rather than m to describe the probability of a defect. Note that defects may be interpreted as successes, in which case x is the number of failures until (s-x) successes. The sample is taken without replacement.

The Hypergeometric distribution is used to describe sampling from a population where an estimate of the total number of defects is desired. It has also been used to estimate the total population of species from a tagged subset. However, estimates of all three parameters from a data set are notoriously fickle and error prone, so use of these parameters to estimate a physical quantity without specifying at least one of the parameters is not recommended. (see Johnson et. al.1)

As shown in the examples, low values of p give high probabilities for low values of x and visa versa, so that the peak in the distribution may approach either bound. More examples can be viewed by using the Distribution Viewer capability.







1. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.237

Logarithmic(theta)

$$p(x) = \frac{a\,\Theta^x}{x}$$

where

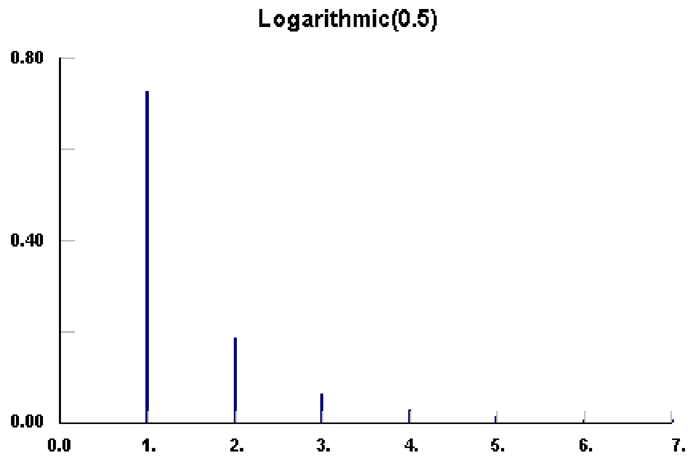
$$a = -1/\ln(1-\Theta)$$

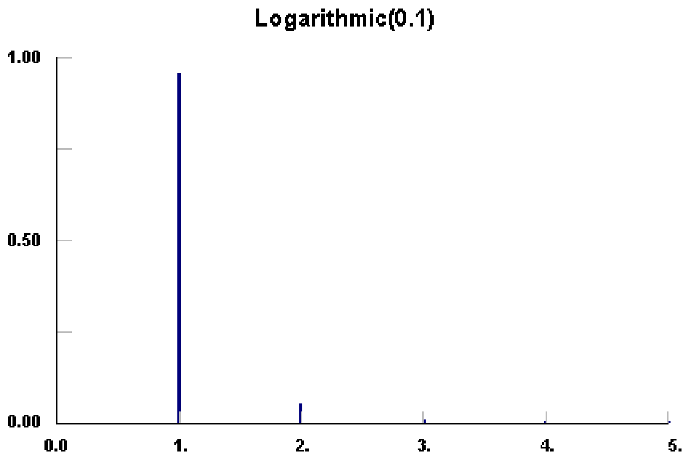
theta = shape/scale parameter 0 < theta < 1

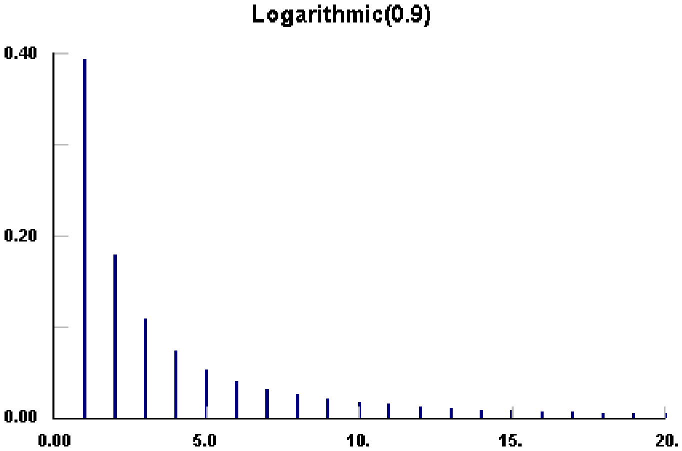
The Logarithmic distribution is a discrete distribution bounded by [1,...]. Typically, if the data is bounded by [0,...], then translating the data before fitting is required. Theta is related to the sample size and the mean.

The Logarithmic distribution is used to describe the diversity of a sample, that is, how many of a given type of thing are contained in a sample of things. For instance, this distribution has been used to describe the number of individuals of a given species in a sampling of mosquitoes, or the number of parts of a given type in a sampling of inventory. (see Johnson et.al.1)

As shown in the examples, low values of theta give high probabilities for low values of x with the distribution expanding as theta nears 1. More examples can be viewed by using the Distribution Viewer capability.







1. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.285

Negative Binomial(x, p, k)

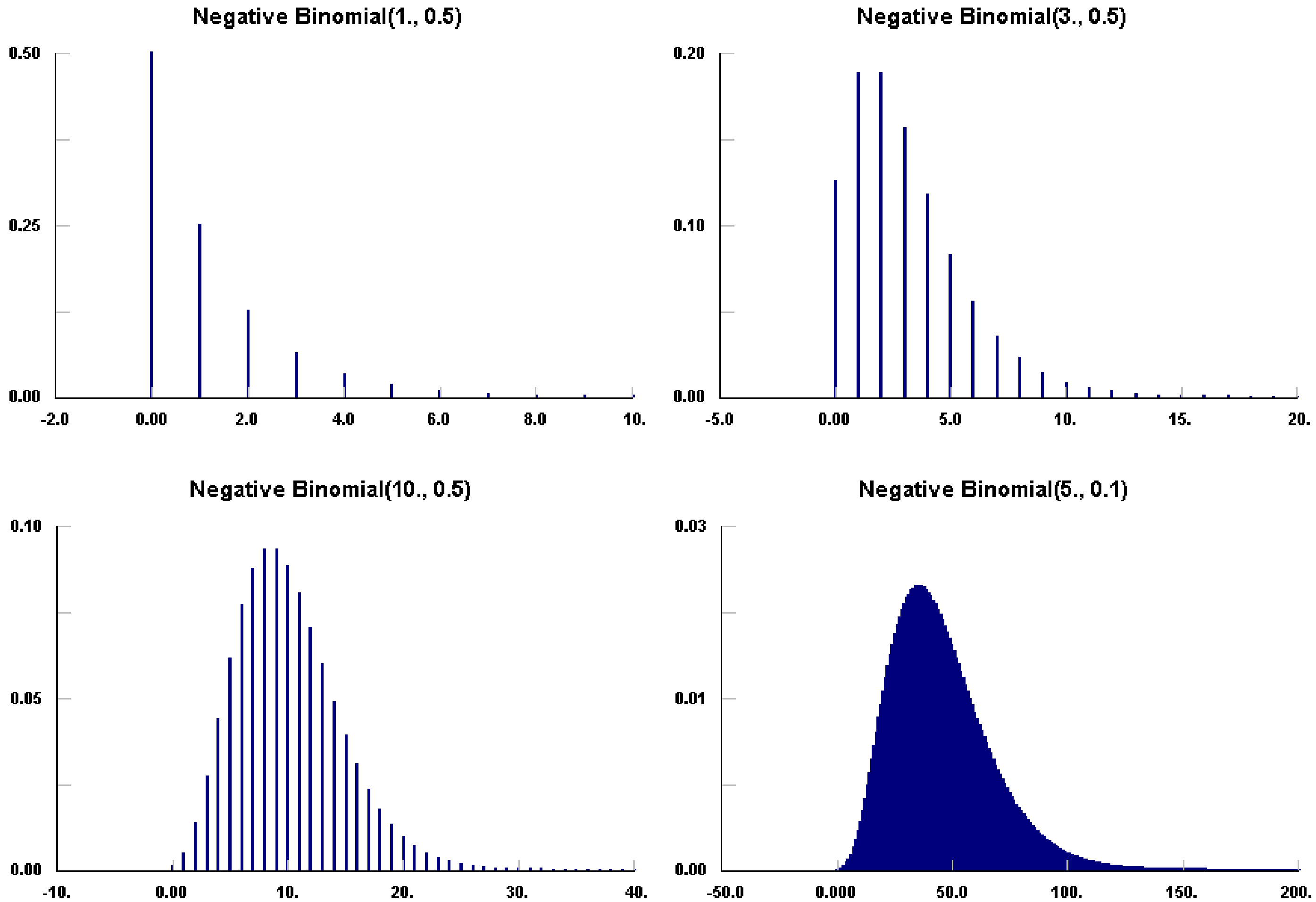
$$p(x) = \binom{k + x - 1}{x} p^k (1 - p)^x$$

x = number of trials to get k events
p = probability of event
k = number of desired events

The Negative Binomial distribution is an discrete distribution bounded on the low side at 0 and unbounded on the high side. The Negative Binomial distribution reduces to the Geometric distribution for k = 1. The Negative Binomial distribution gives the total number of trials, x, to get k events (failures...), each with the constant probability, p, of occurring.

The Negative Binomial distribution has many uses; some occur because it provides a good approximation for the sum or mixing of other discrete distributions. By itself, it is used to model accident statistics, birth-and-death processes, market research and consumer expenditure, lending library data, biometrical data, and many others (see Johnson et. al.1).

Several examples with increasing k are shown below. With smaller probability, p, the number of classes is so large that the distribution is best plotted as a filled polygon. Note that the probabilities are actually weights at each integer, but are represented by broader bars for visibility. More examples can be viewed by using the Distribution Viewer capability.



1. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.223

Poisson(lambda)

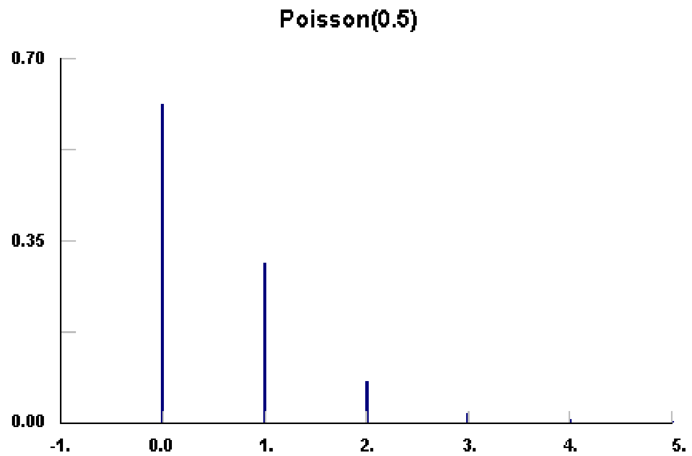
$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

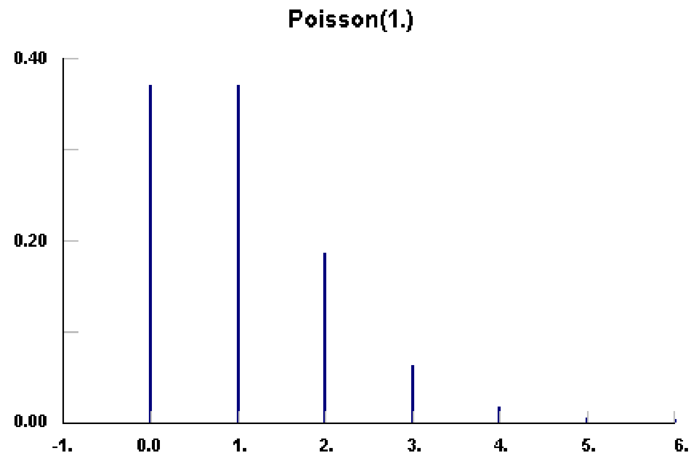
lambda = rate of occurrence

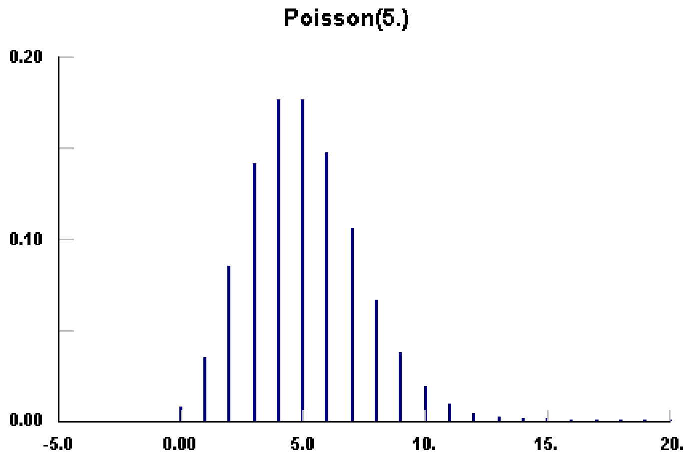
The Poisson distribution is a discrete distribution bounded at 0 on the low side and unbounded on the high side. The Poisson distribution is a limiting form of the Hypergeometric distribution.

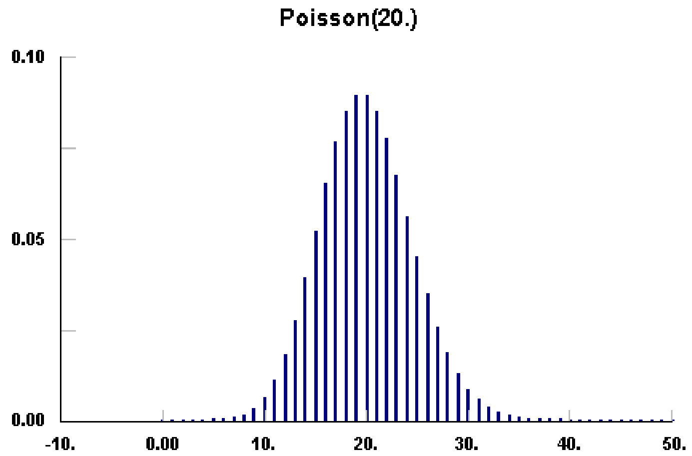
The Poisson distribution finds frequent use because it represents the infrequent occurrence of events whose rate is constant. This includes many types of events in time or space such as arrivals of telephone calls, defects in semiconductors manufacturing, defects in all aspects of quality control, molecular distributions, stellar distributions, geographical distributions of plants, shot noise, etc..It is an important starting point in queuing theory and reliability theory. Note that the time between arrivals [defects] is Exponentially distributed, which makes this distribution a particularly convenient starting point even when the process is more complex.

The Poisson distribution peaks near lambda and falls off rapidly on either side. More examples can be viewed by using the Distribution Viewer capability.









1. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p.151

File|Export|Export Fit

The Export Fit command provides the fitted distribution in the form required by the Application in order to generate random variates from that distribution. The Export Fit dialog allows a choice of Applications in order to determine the format of the output and the choice of the analytical distribution. After both choices have been made, the expected output is shown near the bottom of the dialog.

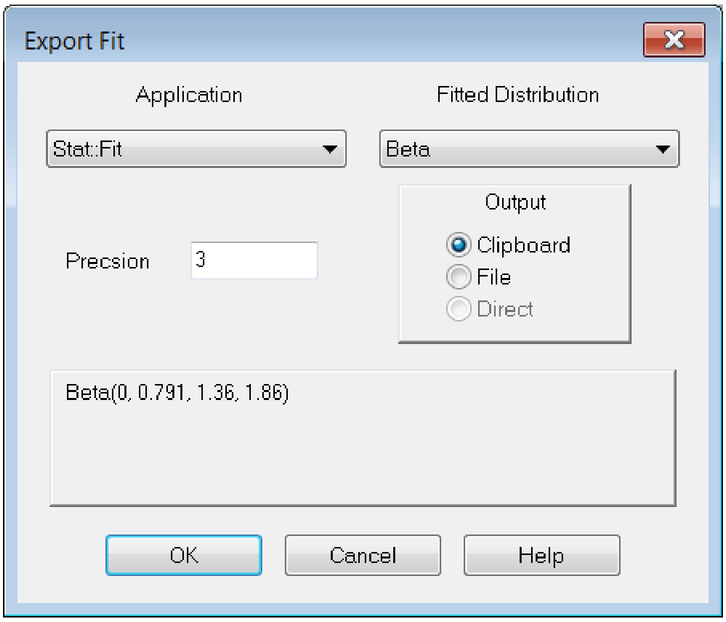
The description of a Distribution Viewer can also be exported by choosing the Export Fit command while the Distribution Viewer is the active window. In this way, no-data or minimal data descriptions can be translated from the form of the distribution in Stat::Fit to the form of the distribution in a particular application.

For some Applications, the requested distribution may not be supported. If the generator for the unsupported distribution can be formed from a Uniform distribution, the analytical form of this generator is given instead. If the generator is not straightforward, no output will be provided.

The precision of the parameters in the output is set by the Precision box whose default value is 3.

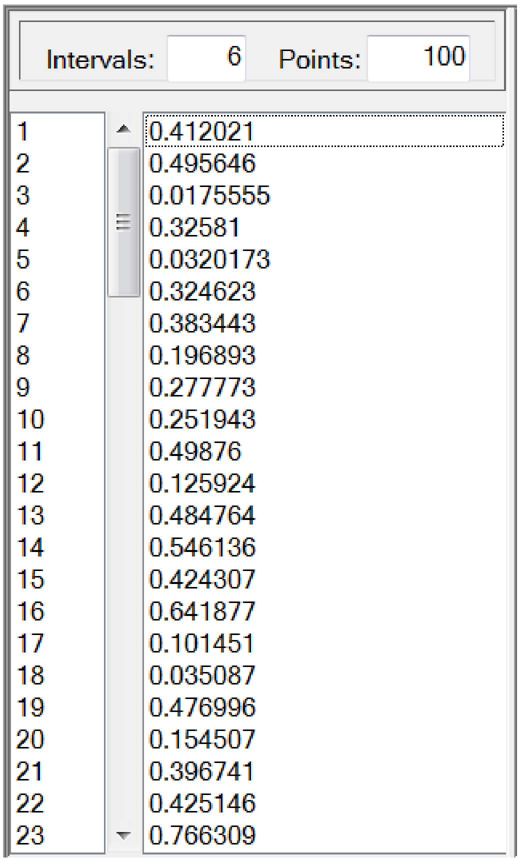
The output can be directed to the Windows clipboard, a text file, or directly into Applications which have implemented Dynamic Data Exchange(DDE) with Stat::Fit. Applications with DDE activated must be running concurrently.

The Export Fit command requires that the distribution be fully specified. For a fitted distribution, this requires parameters estimates be calculated, either manually with one of the estimating commands or automatically with the Autofit command.



Input

The Data Table holds the input data as shown.   All data are entered as single measurements, not cumulative data.   The numbers on the left are aides for location and scroll with the data.   The total number of data points and intervals for continuous data are shown at the top.



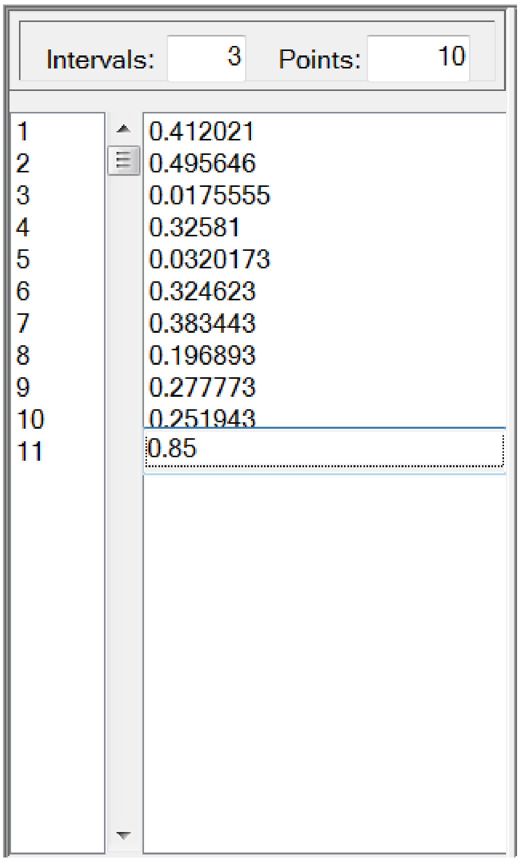
The Data Table can be edited with the standard functions: Cut, Copy, Paste, Insert, Delete, Clear, and Select All which can be accessed from the menu or a right click on the table.

All data can be viewed by using the central scroll bar or the keyboard.   The scroll bar handle can be dragged to get to a data area quickly, or the scroll bar can be clicked above or below the handle to step up or down a page of data.   The arrows can be clicked to step up or down one data point.

The Page Up and Page Down keys can be used to step up or down a data page.   The up and down arrow keys can be used to step up or down a data point.   The Home key forces the Data Table to the top of the data, the End key, to the bottom.

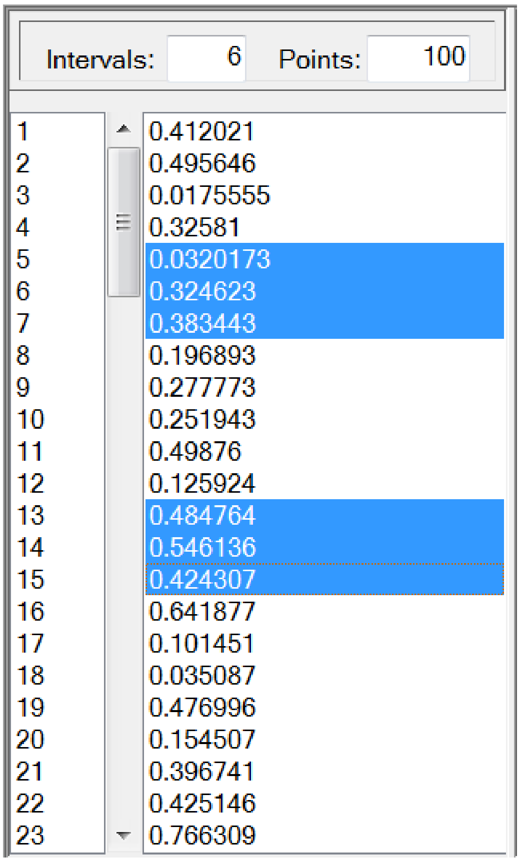
Manual data entry requires that the Data Table be the currently active window which requires clicking on the view until the input focus dotted lines are visible.   Manual data entry begins when a number is typed.   The current data in the Data Table is grayed and an input box is opened, as shown.   The input box will remain open until the Enter key is typed.   The Esc key can be used to abort the entry.

All numbers are floating point, and can be entered in straight decimal fashion, such as 0.972, or scientific notation, 9.72e-1 where exx stands for the power of ten to be multiplied by the preceding number.   Integers are stored as floating point numbers.



If Insert is off, the default condition, the data point is entered at the current highlighted location [the dashed line box, not the colored box]. A number may be highlighted with a click of the mouse at that location. Note that the number is also selected [the colored box] although this does not affect manual data entry. If Insert is on, the data point is entered before the data point in the highlighted box except at the end of the data set. If data point is entered in the highlighted box at the end of the data set, the data point is appended to the data set and the highlighted box is moved to the next empty location. In this way data may be entered continuously without relocating the data entry point. The empty position at the end of the data set can be easily reached by using the End key unless the Data Table is full, 50000 numbers.

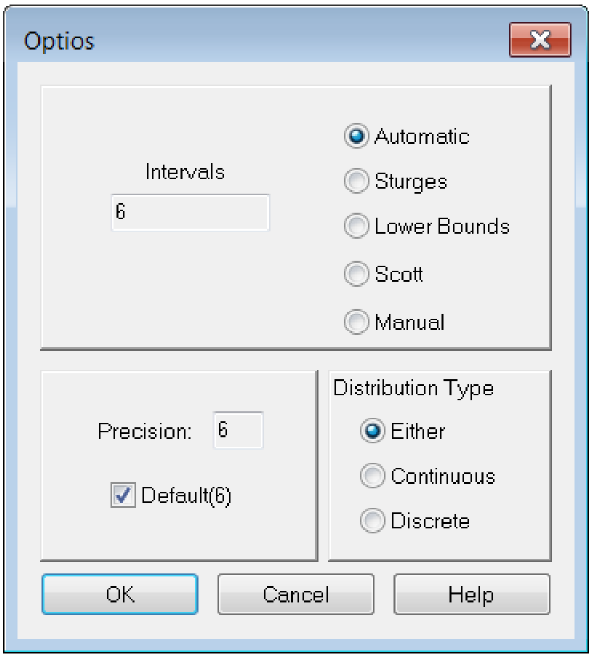
A single number or group of numbers may be selected in the Data Table by clicking the mouse. The selected numbers are highlighted in a color, usually blue, as shown. If the shift key is used with a mouse click, a range of numbers is selected from the last selected number to the current position. If the ctrl key is used with a mouse click, the current position is added to the current selections unless it was already selected, in which case, it is deselected.



The Delete key deletes the currently selected area [the colored area] which can be a single number or group of numbers. There is no undelete. The Delete command in the Edit menu may also be used. The Cut command in the Edit menu deletes the selected numbers and places them in the Clipboard. The Copy command in the Edit Menu copies the currently selected numbers into the Clipboard. The Paste command pastes the numbers in the Clipboard before the number in the current highlighted [dashed box] location, not the selected location.

Input|Options

The Input Options command can be accessed from the Input menu. The Input Options dialog allows several data handling options to be set.



First, the number of intervals can be set. The number of intervals specifies the number of bins into which the input data will be sorted for graphing and calculation. These bins are used only for continuous distributions; discrete distributions are collected at integer values.   If the input data is forced to be treated as discrete, this choice will be grayed.   Note that the name "intervals" is used in Stat::Fit to represent the classes for continuous data in order to separate its use from the integer classes used for discrete data.

The number of intervals are used to display continuous data in a histogram and to compare the input data with the fitted data through a Chi Squared Test.   Please note that the intervals will be equal length for display, but may be of either equal length or of equal probability for the Chi Squared Test. Also, the number of intervals for a continuous representation of discrete data may well bunch the data across integers and look odd. Further, the plotting of discrete distibutions against continuous data will be unnormalized.

Auto:

Automatic mode uses the minimum number of intervals possible without losing information. Then the intervals are increased if the skewness of the sample is large

Sturges:

Sturges mode is an empirical rule for assessing the desirable number of intervals. If N is the number of data points and k is the number of intervals, then

$$k = 1 + 3.3 \log_{10} N$$

Lower Bounds:

Lower Bounds mode uses the minimum number of intervals possible without losing information. If N is the number of data points and k is the number of intervals, then

$$k = (2N)^{1/3}$$

Scott:

Scott mode is based on using the Normal density as a reference density for constructing histograms. If N is the number if data points, sigma is the standard deviation of the sample, and k is the number of intervals, then

$$k = (N)^{1/3}(\max - \min)/(3.5\sigma)$$

Manual:

Manual mode allows arbitrary setting of the number of intervals.

Second, the displayed precision of the data can be set. The precision of the data is the number of decimal places shown for the input data and all subsequent calculations. The default precision is 6 decimal places and is initially set on. The precision can be set between 0 and 15. Note that all discrete data is stored as a floating point number.

IMPORTANT: While all calculations are performed at maximum precision, the input data and calculations will be written to file with the precision chosen here. If the data has greater precision then the precision chosen here, it will be rounded when stored in a file.

Third, the type of distribution can be set. The type of an analytical distribution can be either continuous or discrete. In general, all distributions will be used by default, with the appropriate check for integer values. However, the analysis may be forced to either continuous distributions or discrete distributions by checking the appropriate box in the Input Options dialog.

In particular, discrete distributions are forced to be distributions with integer values only. If the input data is discrete, but the data points are multiples of continuous values, divide the data by the smallest common denominator before attempting to analyze it. Input truncation to eliminate small round-off errors is also useful.

The maximum number of classes for a discrete distribution is limited to 5000. If the number of classes to support the input data is greater than this, the analysis will be limited to continuous distributions.

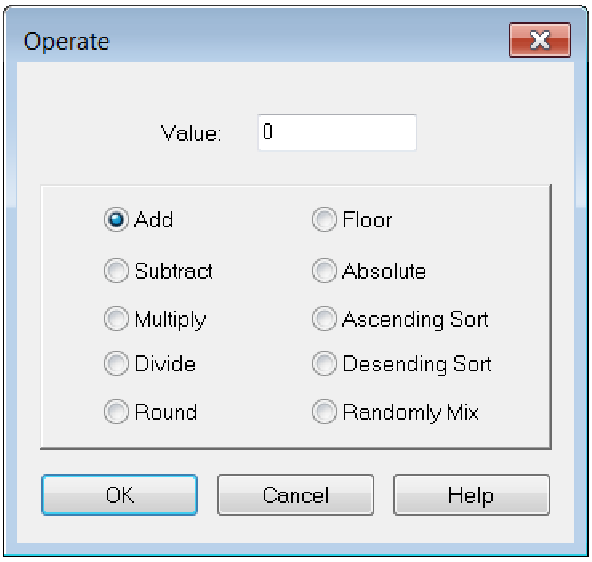
Most of the discrete distributions start at 0. If the data has negative values, an offset should be added to it before analysis.

Input|Operate

Mathematical operations and sorting can be accessed through the Operate command in the Input menu.

The Operate dialog allows the choice of a single standard mathematical operation on the input data. The operation will affect all input data regardless of whether a subset of input data is selected. Mathematical overflow, under flow or other error will cause an error message and the all the input data will be restored. The operation is irrevocable with no undo. The operations are:

add
adds the value to all input data
subtract
subtracts the value to all the input data
multiply
multiplies all the input data by value
divide
divides all the input data by value, zero not allowed
round
rounds all the input data to the nearest integer
floor
truncates or removes the decimal part from all input data
absolute
takes the absolute value of all input data by making all negative values positive
ascend sort
sorts all the input data from the smallest number to the largest
descend sort
sorts all the input data from the largest to the smallest
randomly mix
randomly mixes the data to remove correlation

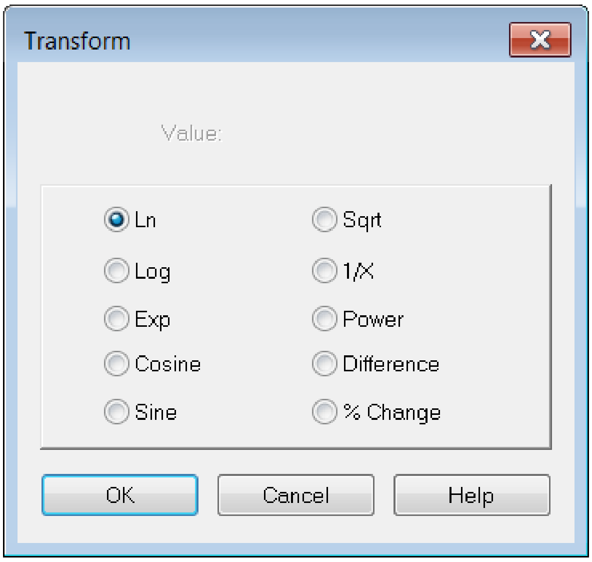


Input|Transform

Mathematical functions can be accessed through the Transform command in the Input menu.

The Transform dialog allows the choice of a single standard mathematic function to be used on the input data. The operation will affect all input data regardless of whether a subset of input data is selected. Mathematical overflow, under flow or other error will cause an error message and the all the input data will be restored. The operation is irrevocable with no undo. The operations are:

ln
the natural logarithm is taken of each data point, all data points must be greater than 0
log
the logarithm to the base 10 is taken of each data point, all data points must be greater than 0
exp
the exponential is taken of each data point
cosine
the cosine is taken for each data point
sine
the sine is taken for each data point
sqrt
the square root is taken for each data point, all data points must be 0 or greater
1/x
the reciprocal is taken for each data point, no data point can be 0
power
each data point will be raised to the value
difference
takes the difference between adjacent data points with the lower data point first
% change
calculates the percentage change between adjacent data points with the lower data point first and divided by the upper data point. No data point can be zero.



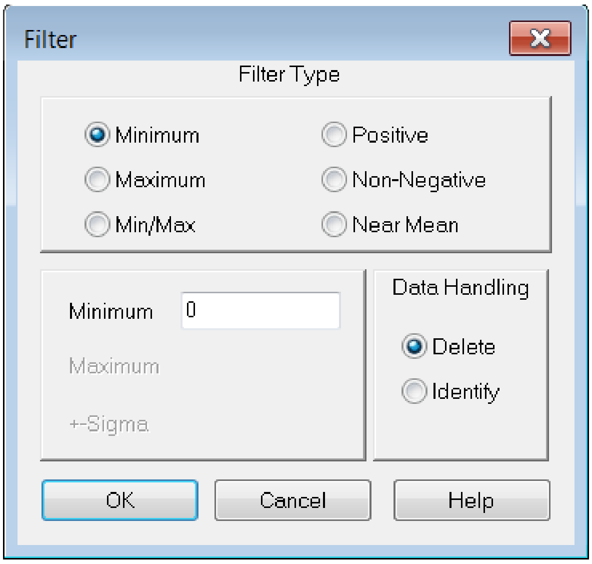
Input|Filter

Filters can be accessed through the Filter command in the Input Menu.

The Filter dialog allows the choice of a single filter to be applied to the input data, discarding data outside the constraints of the filter. All filters DISCARD unwanted data and change the statistics. The appropriate input boxes are opened with each choice of filter. With the exception of the positive filter which excludes zero, all filters are inclusive, that is, they always include numbers at the filter boundary. The filters are:

minimum
filters the data points, excluding all data below the minimum
maximum
filters the data points, excluding all data above the maximum
min/max
filters the data points, excluding all data below the minimum and above the maximum
positive
filters the data points, excluding all negative and zero data points
non-negative
filters the data points, excluding all negative data points
near mean
filters the data points, excluding all data points below the mean minus the standard deviation times the multiplier and excluding all data points above the mean plus the standard deviation times the multiplier

Alternatively, filtered data may be identified rather than discarded by Identify in the Data Handling section of the dialog.



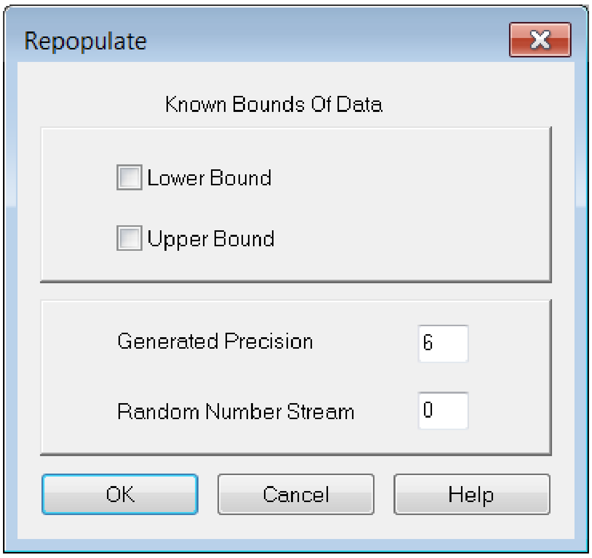
Input|Repopulate

Repopulation of the input data, or unrounding, can be accessed through the Repopulation command in the Input menu

The Repopulation command allows the user to expand rounded data about each integer. Each point is randomly positioned about the integer with its relative value weighted by the existing shape of the input data distribution. If lower or upper bounds are known, the points are restricted to regions above and below these bounds, respectively. The Repopulation command is restricted to integer data only, and limited to -1000 to +1000 in range.

The new data points will have a number of decimal places specified by the generated precision. The Goodness of Fit tests , the Maximum Likelihood Estimates, and the Moment Estimates require at least three digits to give reasonable results. The sequence of numbers is repeatable if the same random number stream is used [e.g. stream 0]. However, the generated numbers, and the resulting fit, can be varied by choosing a different random number stream, 0-99.

Important: This Repopulation of the decimal part of the data is not the same as the original data was or would have been, but only represents the information not destroyed by rounding. The parameter estimates are not as accurate as would be obtained with unrounded original data. In order to get an estimate of the variation of fitted parameters, try regenerating the data set with several random number streams.



Input|Generate

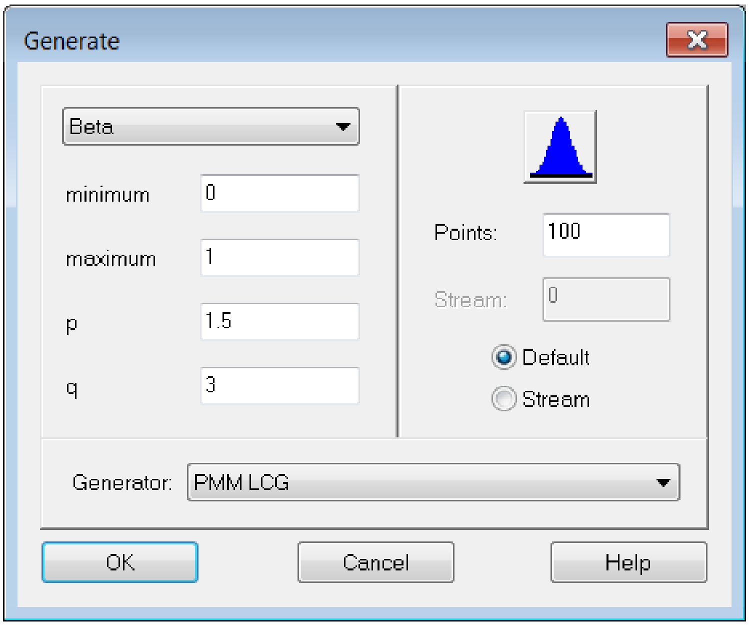
Random variate generation can be accessed through the Generate command in the Input menu or from the Generate button on the Tool Bar.

The Generate dialog provides the choice of distribution, parameters, and random number stream for the generation of random variates from each of the distributions covered by Stat::Fit. The generation is limited to 50000 points maximum, the limit of the input table used by Stat::Fit. The sequence of numbers is repeatable for each distribution because the same random number stream is used [stream 0]. However, the sequence of numbers can be varied by choosing a different random number stream, or a different generator.

The generator will not change existing data in the Data Table, but will append the generated data points up to the limit of 50000 points. In this manner the sum of two or more distributions may be tested. Sorting will not be preserved.

This generator can be used to provide a file of random numbers for another program as well as to test the variation of the distribution estimates once the input data has been fit.

By default, the generator is a Prime Modulus Multiplicative Linear Congruential Generator set up with 100 streams, each 100,000 numbers long with no overlap. Another generator, using Mixed Prime Modulus Multiplicative Linear Generators (L'Ecuyer) with shuffling added, may also be used. It has 1000 streams, each 10,000,000 numbers long, again with no overlap. A third generator, the Mersenne Twister, may be used with 1000 streams. Non overlap for the Mersenne Twister is not guaranteed, but is extremely unlikely. For data files longer than 50000 numbers, use the Utilities|Generate Variates File menu item.

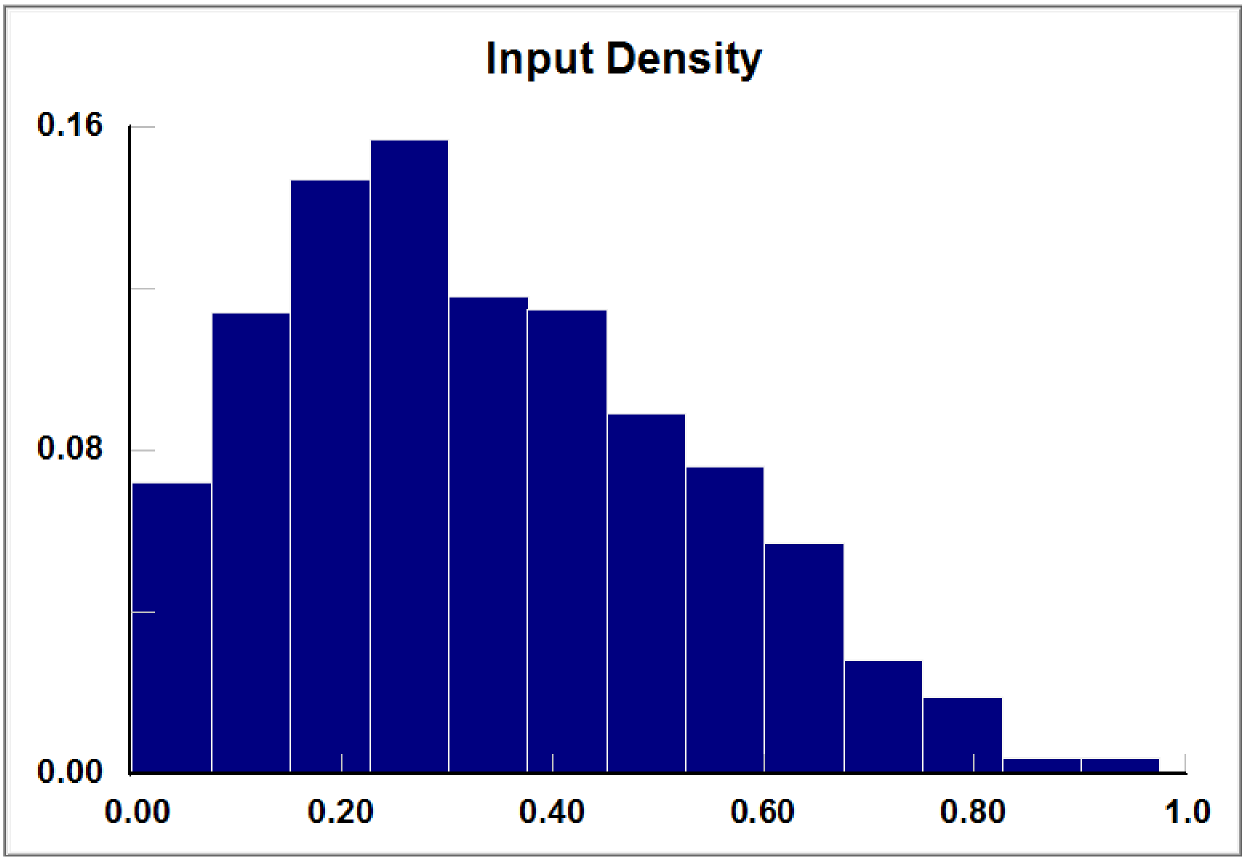


Input|Input Graph

The Input Graph can be accessed from the Input menu as well as the Input Graph button in the Tool Bar.

If the input data in the Data Table is continuous data, or is forced to be treated as continuous in the Input Options dialog, the input graph will be a histogram with the number of intervals being given by the choice of interval type in the Input Options. If the data is discrete, or forced to be treated as discrete, the input graph will be a line graph with the number of classes being determined by the minimum and maximum values. Note that discrete data must be integer values. The data used to generate the Input Graph can be viewed by using the Binned Data command in the Statistics menu.

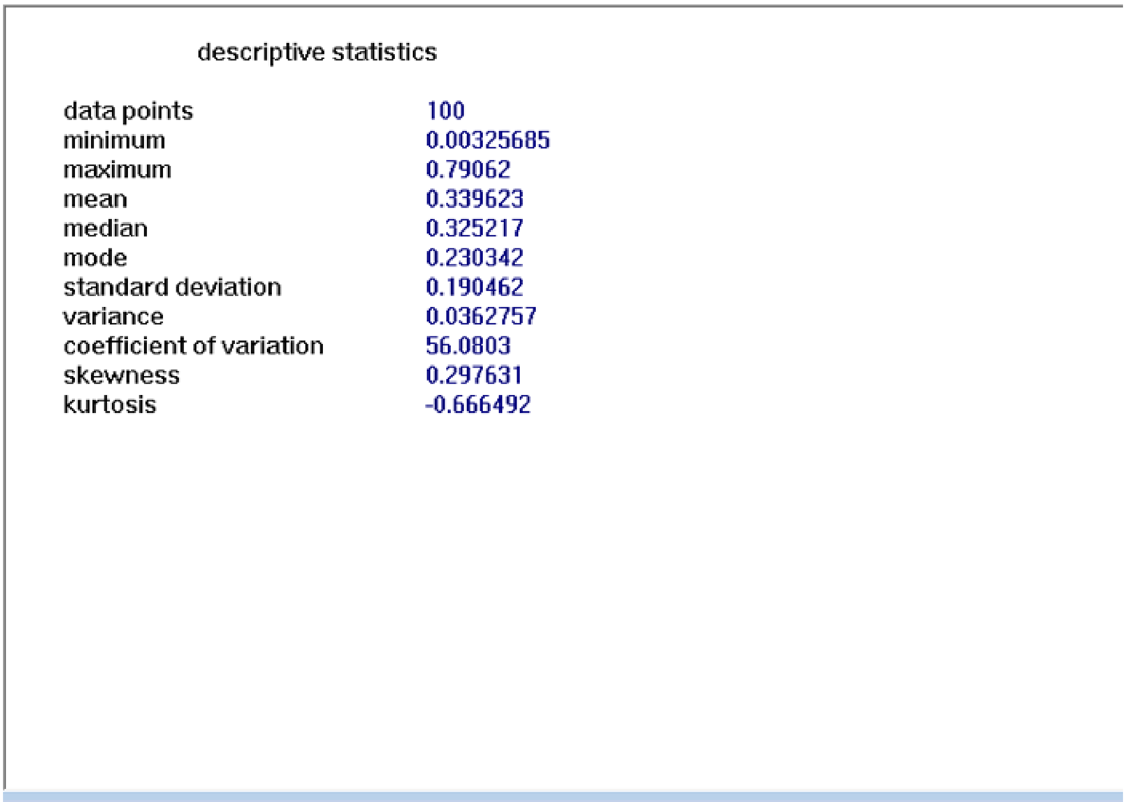
This graph, as with all input data graphs in Stat::Fit, may be modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.



Statistics|Descriptive

The Descriptive statistics command can be accessed in the Statistics menu.

The Descriptive statistics command provides the basic statistical observations and calculations on the input data, and presents these in a simple view as shown. The observations and calculations include: number of data points, minimum, maximum, mean, median, mode, variance, standard deviation, coefficient of variation, skewness, and kurtosis. As long as the view is displayed, the values are updated as the input in the Data Table is changed. Note that the kurtosis is relative to normal, that is, the absolute kurtosis minus 3.



Statistics|Binned Data

The Binned Data command can be accessed in the Statistics menu.

The Binned Data command produces a view of the calculated values for the histogram [continuous data] and line graph [discrete data]. The number of intervals used for continuous data is determined by the interval option in the Input Options dialog. By default, this number is determined automatically from the total number of data points. For convenience, both the frequency and relative frequency are given. If the data is sensed to be discrete [all integer], then the classes for the discrete representation are also given, at least up to 1000 classes. The availability of interval or class data can also be affected by forcing the distribution type to be either continuous or discrete.

The Binned Data can be exported using the Export Empirical command, modifying bounds as desired.

binned data

data points	100
precision	6

continuous relative frequency

intervals	6

end points	mid points	density	ascending cumulative	descending cumulative
0.00325685				1
	0.0688705	0.15		
0.134484			0.15	0.85
	0.200098	0.25		
0.265711			0.4	0.6
	0.331325	0.2		
0.396939			0.6	0.4
	0.462552	0.22		
0.528166			0.82	0.18
	0.593779	0.14		
0.659393			0.96	0.04
	0.725007	0.04		
0.79062			1	

continuous frequency

intervals	6

end points	mid points	density	ascending cumulative	descending cumulative
0.00325685				100
	0.0688705	15		
0.134484			15	85
	0.200098	25		
0.265711			40	60
	0.331325	20		
0.396939			60	40
	0.462552	22		
0.528166			82	18
	0.593779	14		
0.659393			96	4
	0.725007	4		
0.79062			100	

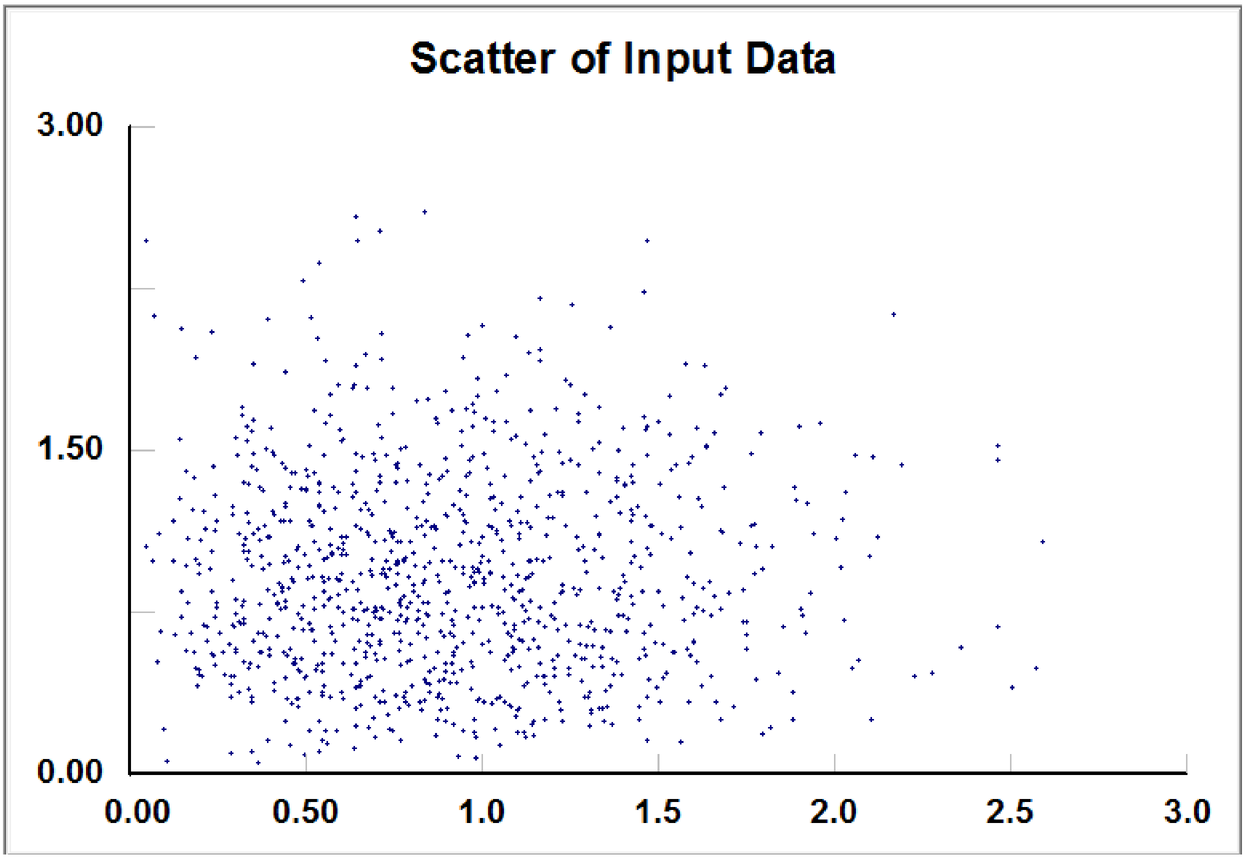
Statistics|Independence|Scatter Plot

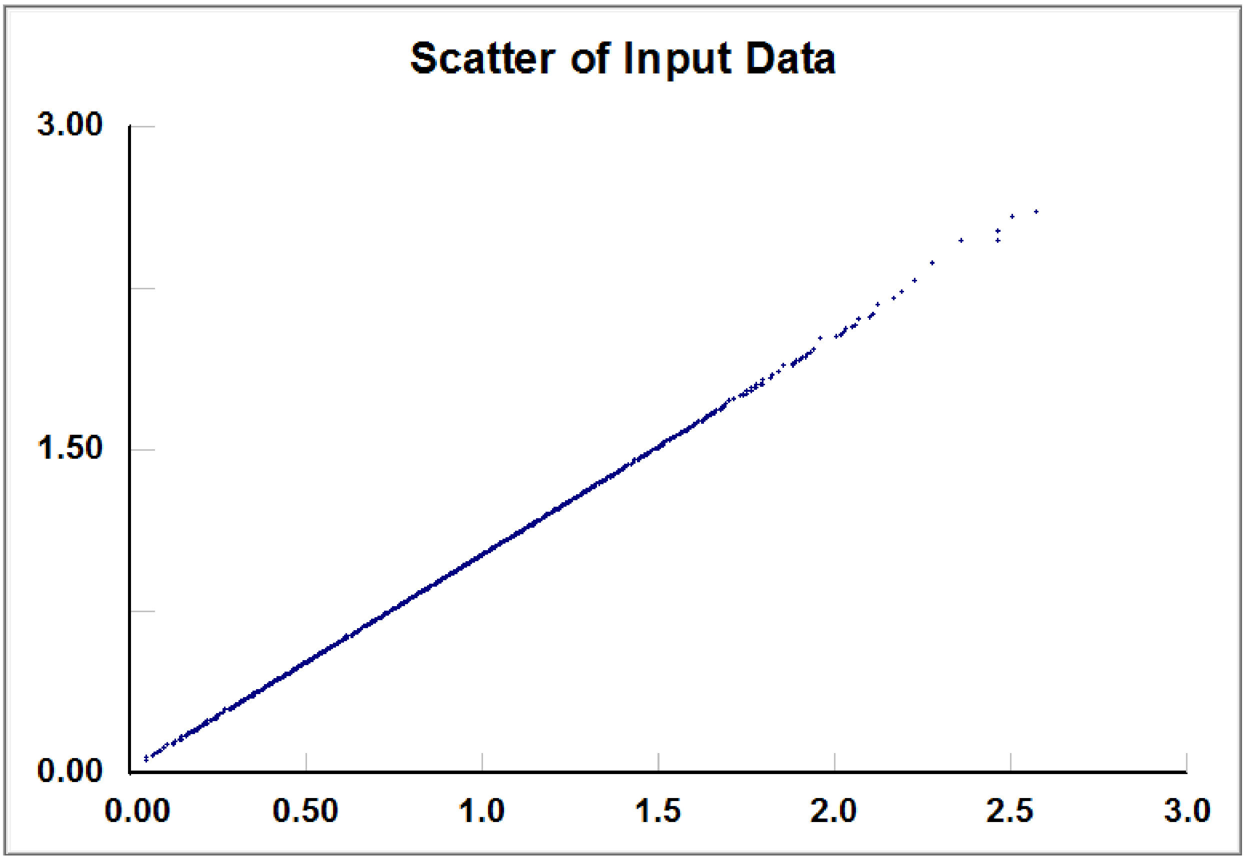
The Scatter Plot command can be accessed from the Independence sub menu of the Statistics menu.

The Scatter Plot command plots adjacent points in the sequence of input data against each other. Thus, each plotted point represents a pair of data points, $[X_{i+1}, X_i]$. This is repeated for all pairs of adjacent data points. If the input data points are somewhat dependent on each other, then this plot will exhibit that dependence. Time series, where the current data point may depend on the nearest previous value(s), will show that pattern here as a structured curve rather than a seemingly independent scatter of points. While the Autocorrelation gives a more complete answer, this plot can quickly show strongly dependent behavior.

As an example, the two plots shown below are 1000 random variates from a distribution before and after sorting. Sorting the data gives the direct appearance of linear dependence for this distribution.

As with all graphs in Stat::Fit, the Scatter Plot may be customized by using the Graphics Style dialog in the Graphics menu. The graphs may also be copied to the Clipboard or saved as graphic files [.BMP] by using the Copy or Save As commands in the Graphics menu. Note that , while the graph view currently open can still be modified, the copied or saved version is a fixed bitmap.





Scatter of Input Data
3.00
1.50
0.00
0.00
0.50
1.0
1.5
2.0
2.5
3.0

Statistics|Independence|Autocorrelation

The Autocorrelation command can be accessed in the Independence sub menu of the Statistics menu.

The autocorrelation calculation used here assumes that the data are taken from a stationary process, that is, the data would appear the same [statistically] for any reasonable subset of the data. In the case of a time series, this implies that the time origin may be shifted without affecting the statistical characteristics of the series. Thus the variance for the whole sample can be used to represent the variance of any subset. For a simulation study, this may mean discarding an early warm-up period (see Law & Kelton1). In many other applications involving ongoing series, including financial, a suitable transformation of the data might have to be made. If the process being studied is not stationary, the calculation and discussion of autocorrelation is more complex. (see Box et. al.2).

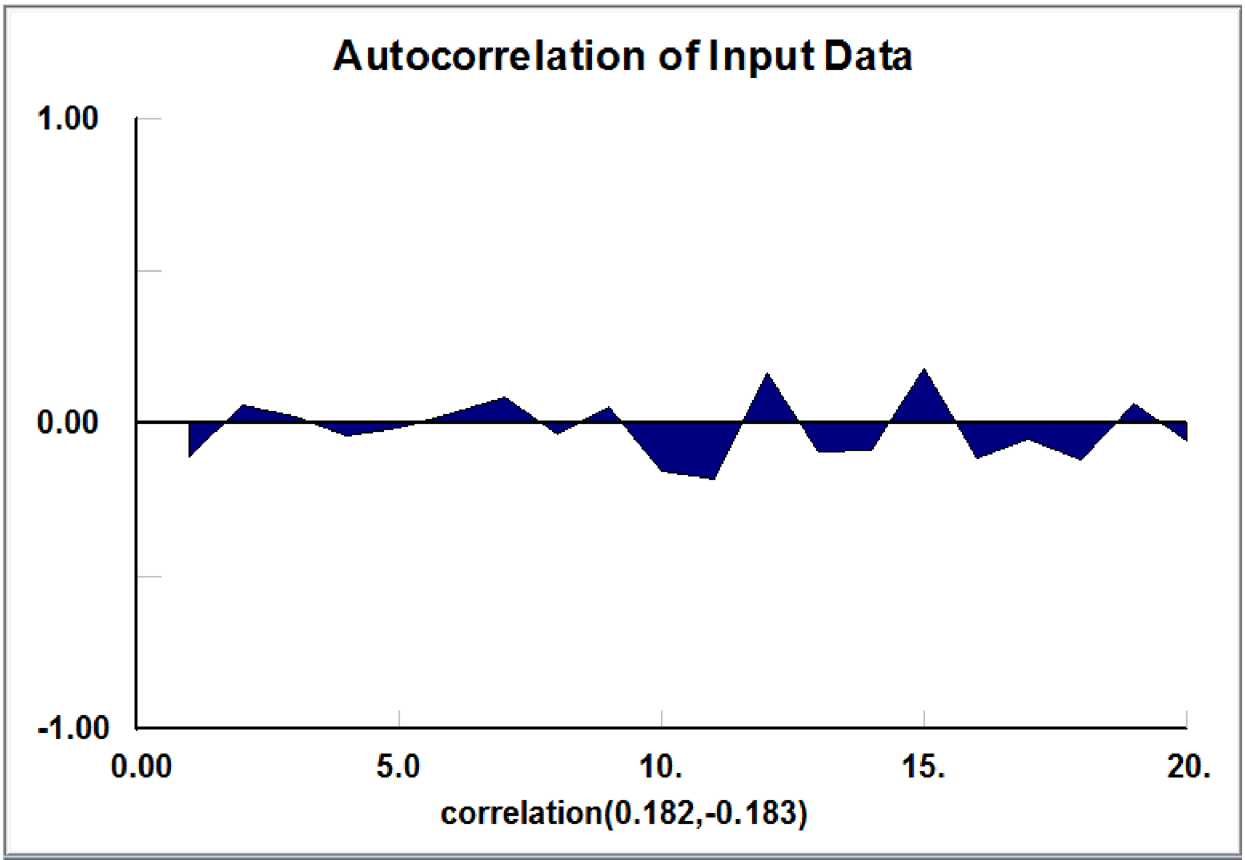
A graphical view of the autocorrelation can be displayed by plotting the scatter of related data points. The Scatter Plot available in the Statistics menu is a plot of adjacent data points, that is, of separation or lag 1. Scatter plots for data points further removed from each other in the series, that is, for lag j, could also be plotted, but the autocorrelation is more instructive. The autocorrelation, rho, is calculated from the equation:

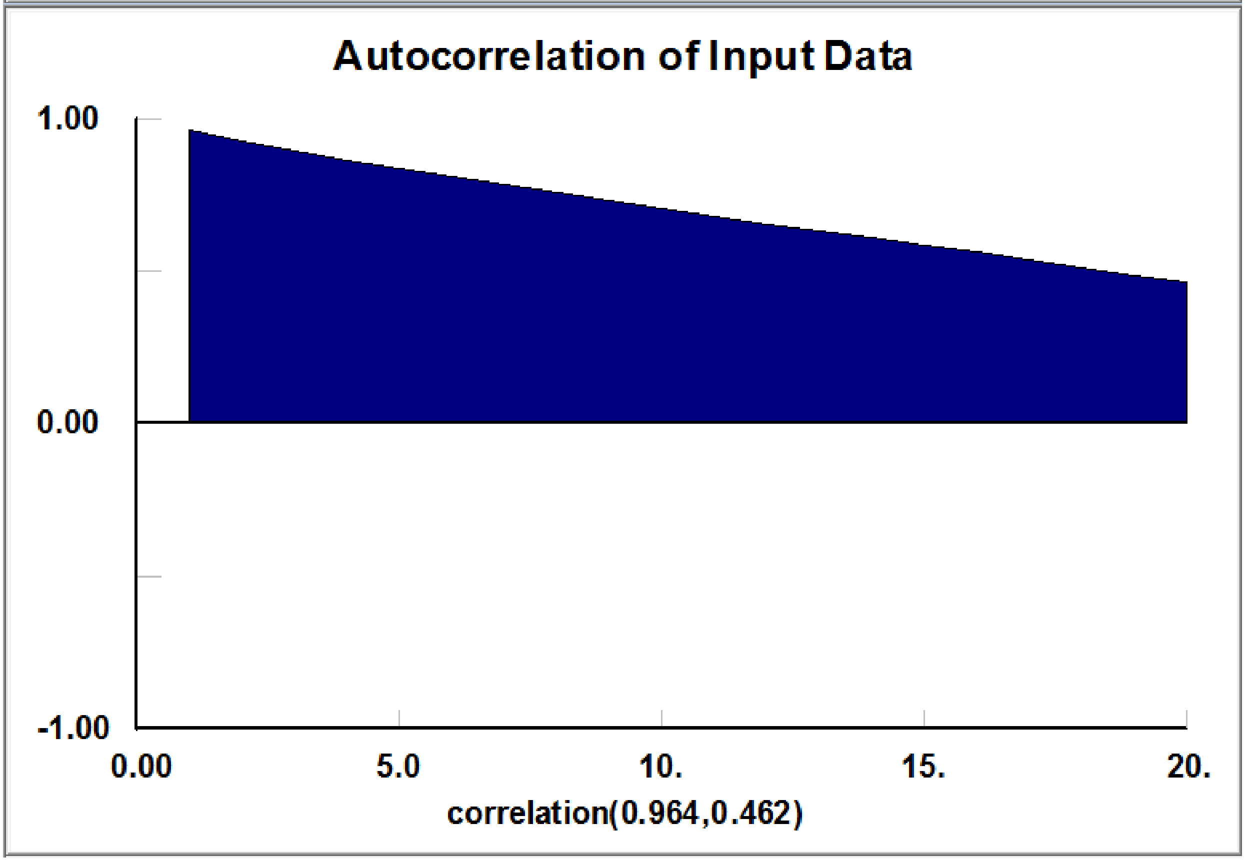
$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_{i+j} - \bar{x})}{\sigma^2 (n - j)}$$

where j is the lag between data points, sigma is the standard deviation of the population, approximated by the standard deviation of the sample, and xbar is the sample mean. The calculation is carried out by Fast Fourier Transform.

The autocorrelation varies between 1 and -1, between positive and negative correlation. If the autocorrelation is near either extreme, the data are autocorrelated. Note, however, that the autocorrelation can assume finite values due to the randomness of the data even though no significant autocorrelation exists.

As an example, the two plots shown below are 100 random variates from a distribution before and after sorting. Note that sorting the data gives the direct appearance of broad autocorrelation. The numbers after correlation along the x axis are the maximum autocorrelation in both the positive and negative directions.





As with all input data graphs in Stat::Fit, the Autocorrelation Plot may be customized by using the Graphics Style dialog in the Graphics menu. The graphs may also be copied to the Clipboard or saved as graphic files [.BMP] by using the Copy or Save As commands in the Graphics menu. Note that , while the graph view currently open can still be modified, the copied or saved version is a fixed bitmap.

1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p293

2. "Time Series Analysis", George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, 1994,

Prentice-Hall

Statistics|Independence|Runs Tests

The Runs Test command can be accessed in the Independence sub menu of the Statistics menu.

The Runs Test command calculates two different runs tests for randomness of the data and displays a view of the results. The result of each test is either DO NOT REJECT the hypothesis that the series is random or REJECT that hypothesis with the level of significance given. The level of significance is the probability that a rejected hypothesis is actually true, that is, that the test rejects the randomness of the series when the series is actually random.

The first runs test is a median test which measures the number of runs, that is, sequences of numbers, above and below the median (see Brunk1). The run can be a single number above or below the median if the numbers adjacent to it are in the opposite direction. If there are too many or too few runs, the randomness of the series is rejected. This median runs test uses a normal approximation for acceptance/rejection which requires that the number of data points above/below the median be greater than 10. An error message will be printed if this condition is not met.

The second runs test is a turning point test which measures the number of times the series changes direction (see Johnson et. al.2). Again, if there are too many turning points or too few, the randomness of the series is rejected. This turning point runs test uses a normal approximation for acceptance/rejection which requires that the total number of data points be greater than 12. An error message will be printed if this condition is not met.

While there are many other runs tests for randomness, some of the most sensitive require larger data sets, in excess of 4000 numbers. (see Knuth3).

runs test on input

runs test (above/below median)

data points	100
points above median	50
points below median	50
total runs	50
mean runs	51
standard deviation runs	4.97468
runs statistic	0.201018
level of significance	0.05
runs statistic(0.025)	1.95996
p-value	0.840685
result	DO NOT REJECT

runs test (turning points)

data points	100
turning points	71
mean turnings	66.3333
standard deviation turnings	4.17798
turnings statistic	1.11697
level of significance	0.05
turnings statistic(0.025)	1.95996
p-value	0.264009
result	DO NOT REJECT

1. "An Introduction to Mathematical Statistics", H.D. Brunk, 1960, Ginn

2. "Univariate Discrete Distributions", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p 425

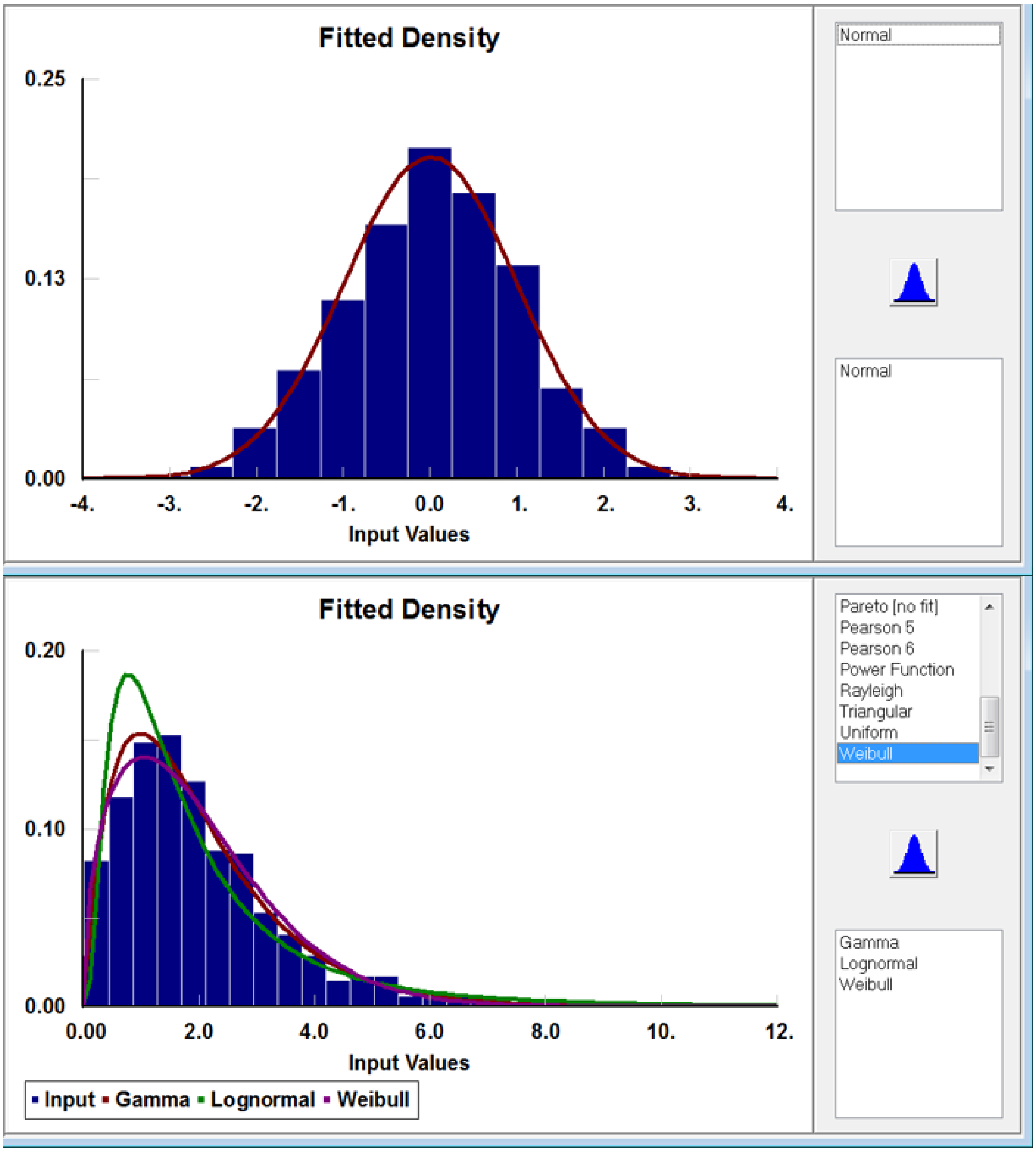
3. "Seminumerical Algorithms", Donald E. Knuth, 1981, Addison-Wesley

Fit|Result Graphs|Density

The Density graph can be accessed from the Result Graphs sub menu of the Fit menu, or the Density graph button in the Tool Bar.   It is activated only after specific distributions have been chosen.

The Density graph, as shown below, is a plot of the input data in the Data Table versus a line drawing of the fitted densities.   If more than one distribution were fit to the input data, the list at the top right will allow a choice of available plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the densities plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all result graphs in Stat::Fit, may be partially modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.

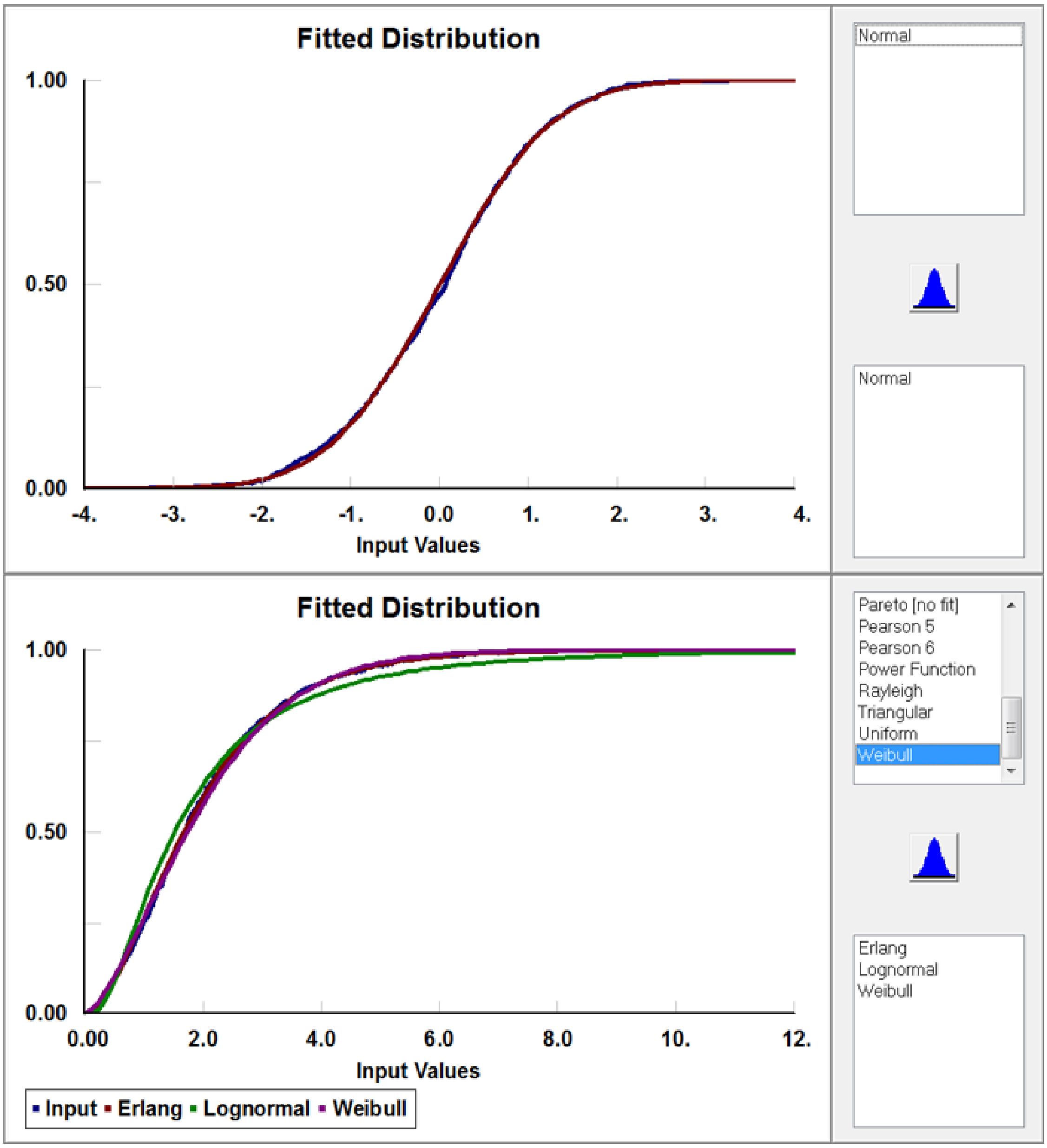


Fit|Result Graphs|Distribution

The Distribution graph can be accessed from the Result Graphs sub menu of the Fit menu.   It is activated only after specific distributions have been chosen.

The Distribution graph, as shown below, is a plot of the input data in the Data Table versus a line drawing of the fitted cumulative distributions.   If more than one distribution were fit to the input data, the list at the top right will allow a choice of available plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the cumulative distributions plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all result graphs in Stat::Fit, may be partially modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.

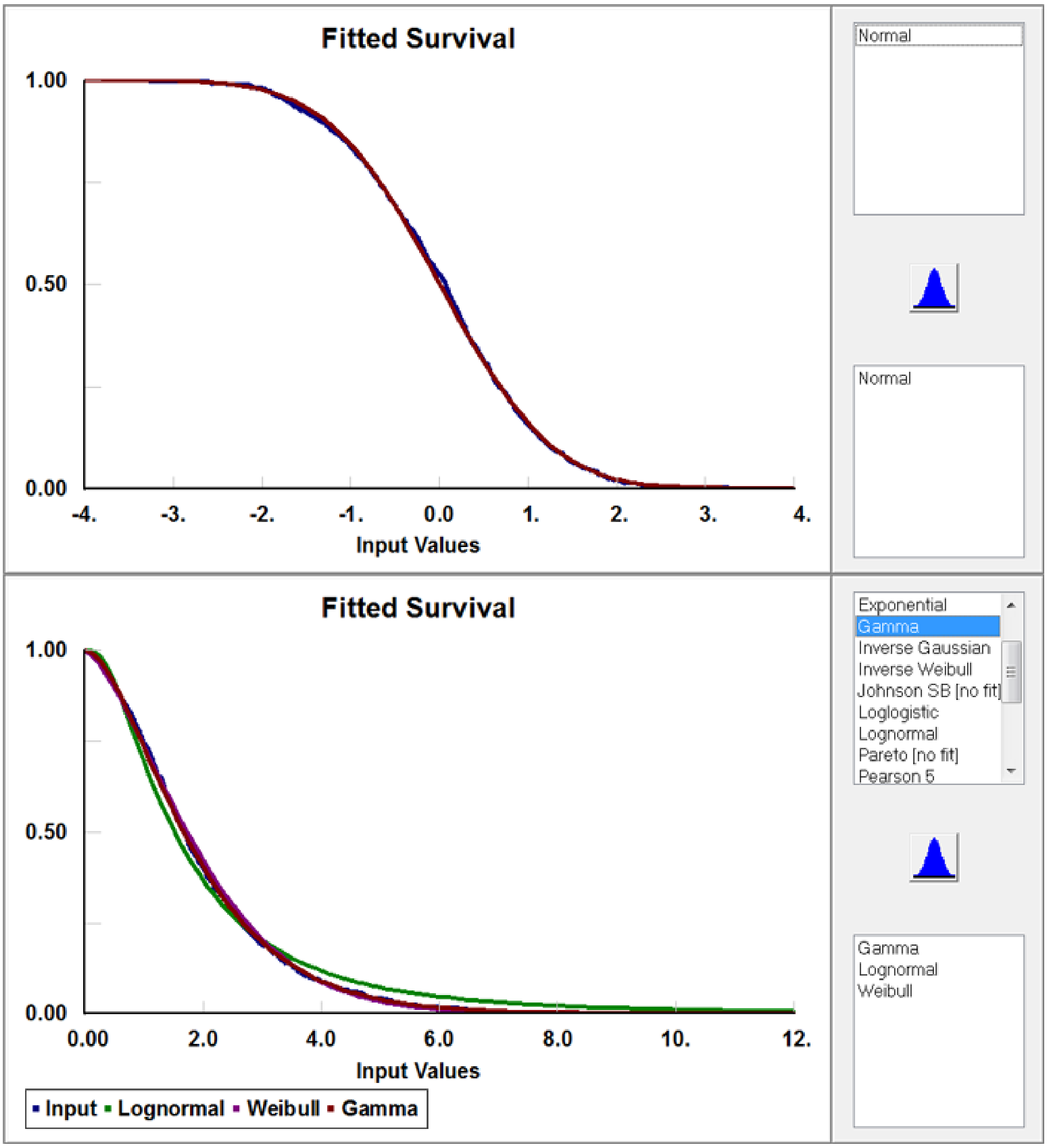


Fit|Result Graphs|Survival

The Survival graph can be accessed from the Result Graphs sub menu of the Fit menu. It is activated only after specific distributions have been chosen.

The Survival graph, as shown below, is a plot of the input data in the Data Table versus a line drawing of the fitted cumulative distributions. If more than one distribution were fit to the input data, the list at the top right will allow a choice of available plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the cumulative distributions plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all result graphs in Stat::Fit, may be partially modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.

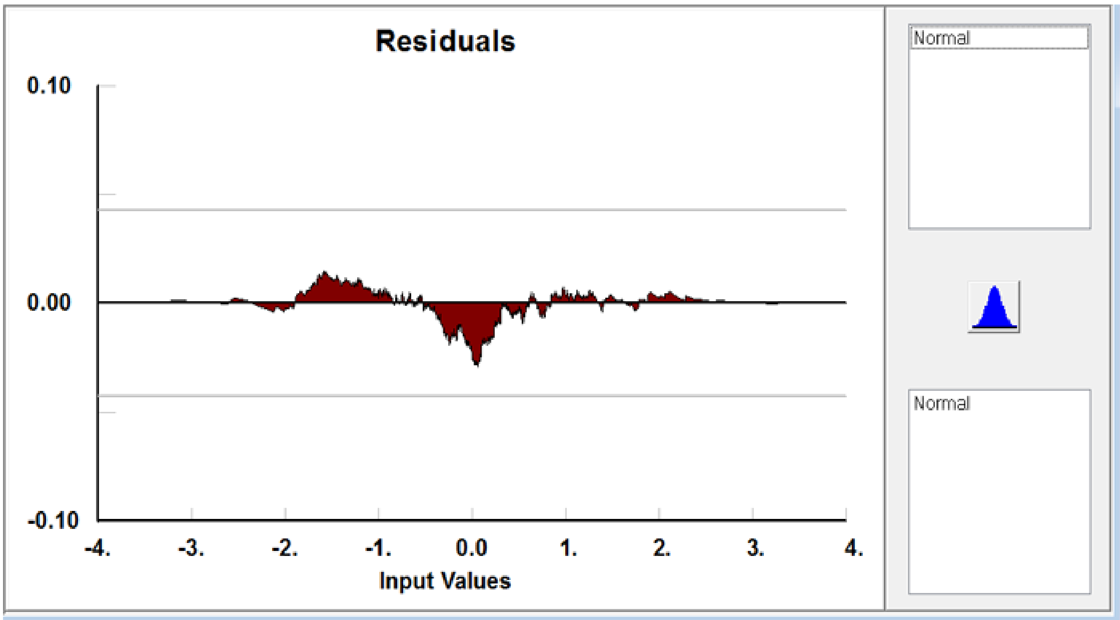


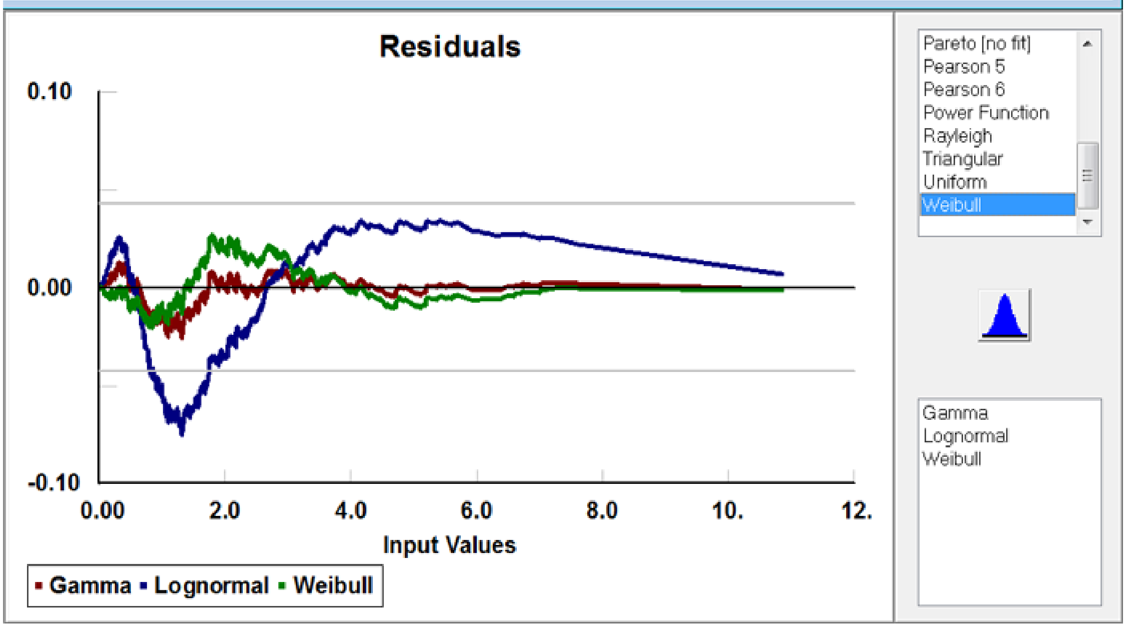
Fit|Result Graphs|Difference

The Difference graph can be accessed from the Result Graphs sub menu of the Fit menu combination. It is activated only after specific distributions have been chosen.

The Difference graph, as shown below, is a plot of the difference between the cumulative input data in the Data Table minus the fitted cumulative distribution, point by point. Note that conservative error bars shown in the graph are not a function of the number of intervals for continuous data. The fit may be rejected for the Kolmogorov Smirnov Test even though the histogram plot of the data would not exceed the error bars. If more than one distribution were fit to the input data, the list at the top right will allow a choice of available plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the densities plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all graphs in Stat::Fit, may be modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.





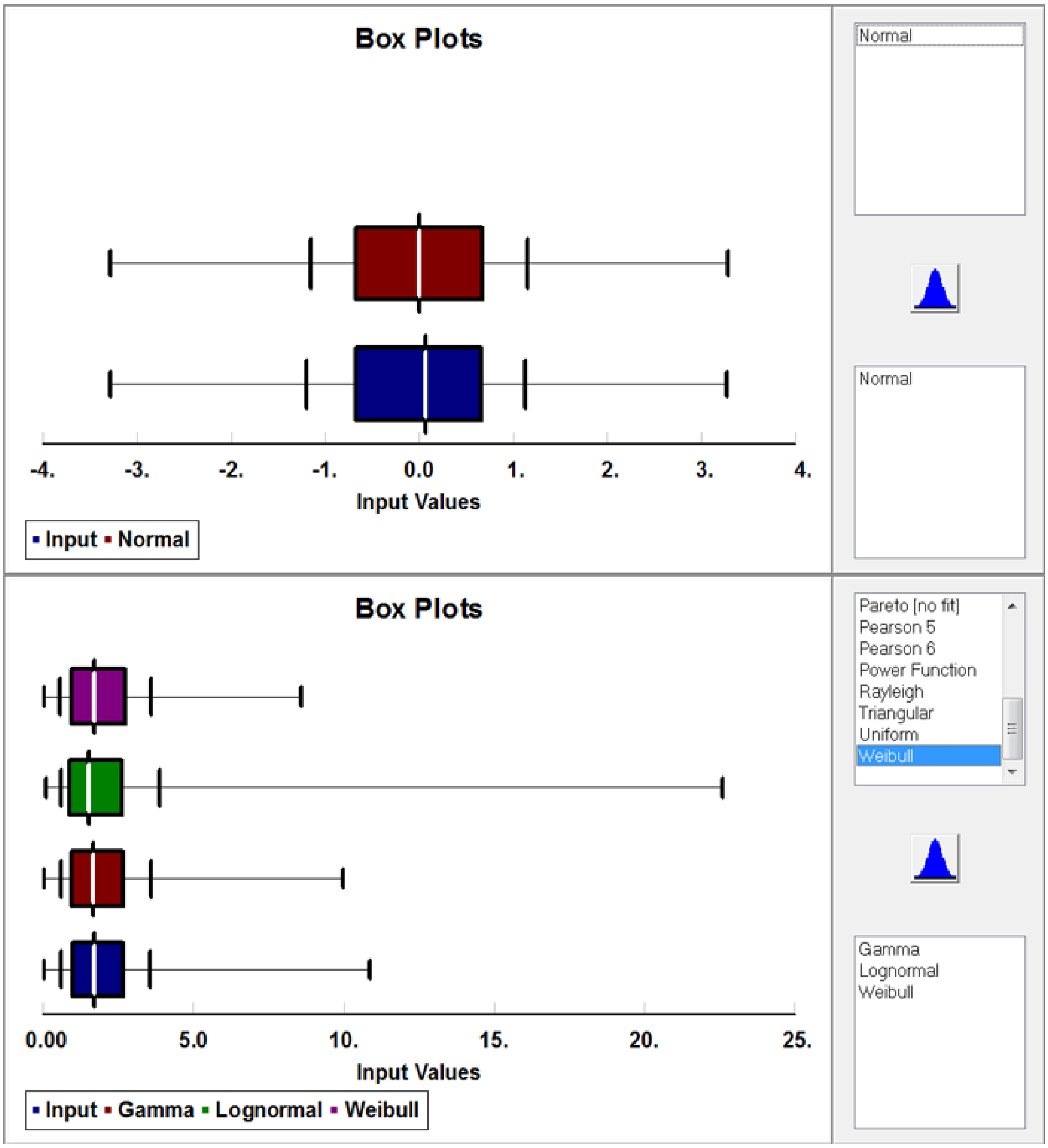
Fit|Result Graphs|Box Plot

The Box Plot can be accessed from the Result Graphs sub menu of the Fit menu.   It is activated only after specific distributions have been chosen.

The Box Plot, as shown below, is a plot of the input data in the Data Table versus a summary data of the fitted cumulative distributions.   The central colored section of the plot delineates the central quartiles of the data or the fitted distribution. The central line is the median. The lines just outside the central section delineate the octile points. The outermost lines indicate the extremes of the data related to the corresponding extremes of the fitted distributions for equal probability of occurrence. Occasionally, the extreme of a distribution is so large that the rest of the plot is obscured.

If more than one distribution were fit to the input data, the list at the top right will allow a choice of available box plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the distributions plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input.

This graph, as with all result graphs in Stat::Fit, may be partially modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.

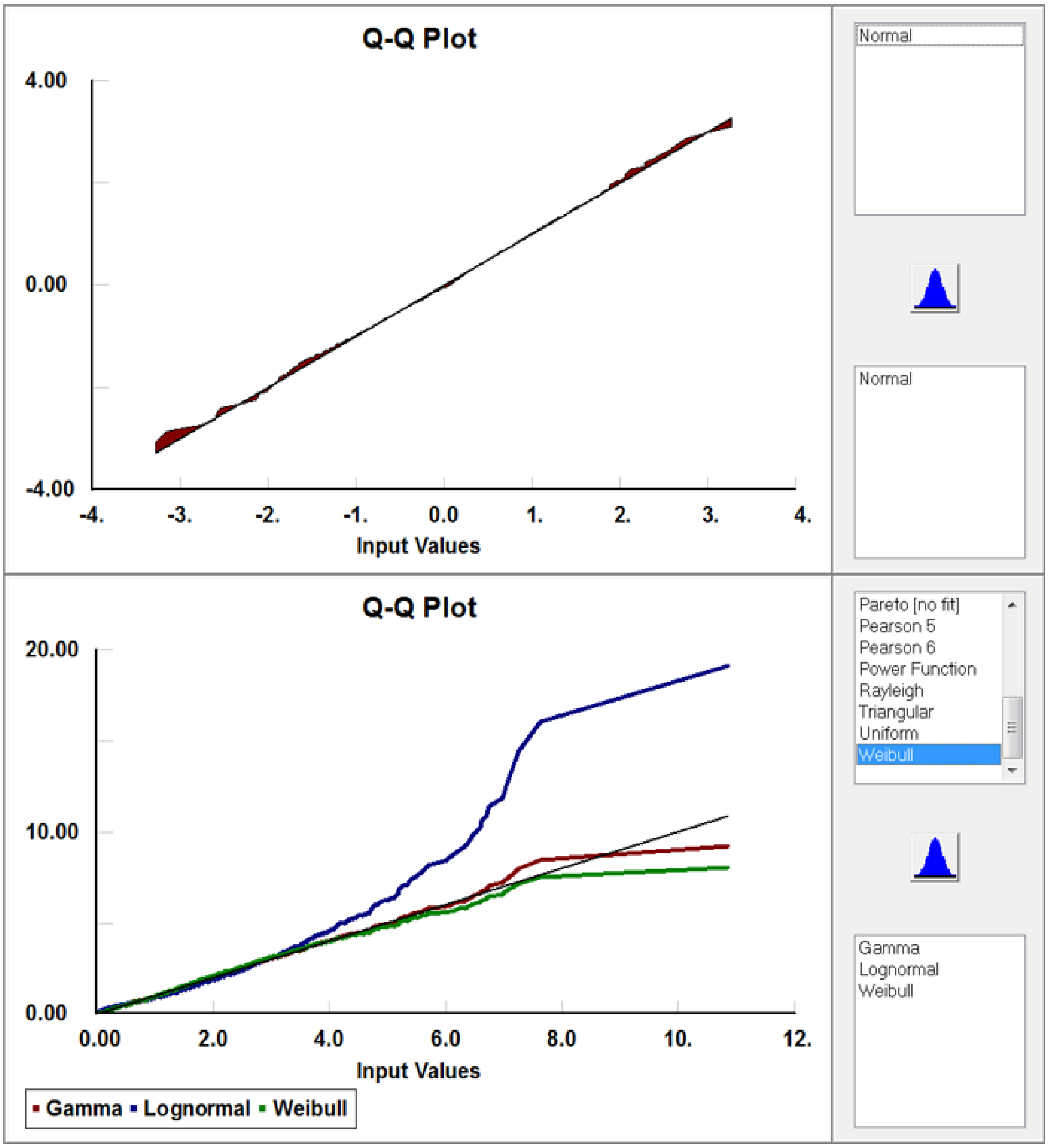


Fit|Result Graphs|Q-Q Plot

The Q-Q Plot can be accessed from the Result Graphs sub menu of the Fit menu. It is activated only after specific distributions have been chosen.

The Q-Q plot, as shown below, is a plot of the input data [straight line] in the Data Table versus the value of x that the fitted distribution must have in order to give the same probability of occurrence. This plot tends to be sensitive to variations of the input data in the tails of the distribution. (see Law&Kelton[1]) If more than one distribution were fit to the input data, the list at the top right will allow a choice of available Q-Q plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the densities plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all graphs in Stat::Fit, may be modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.



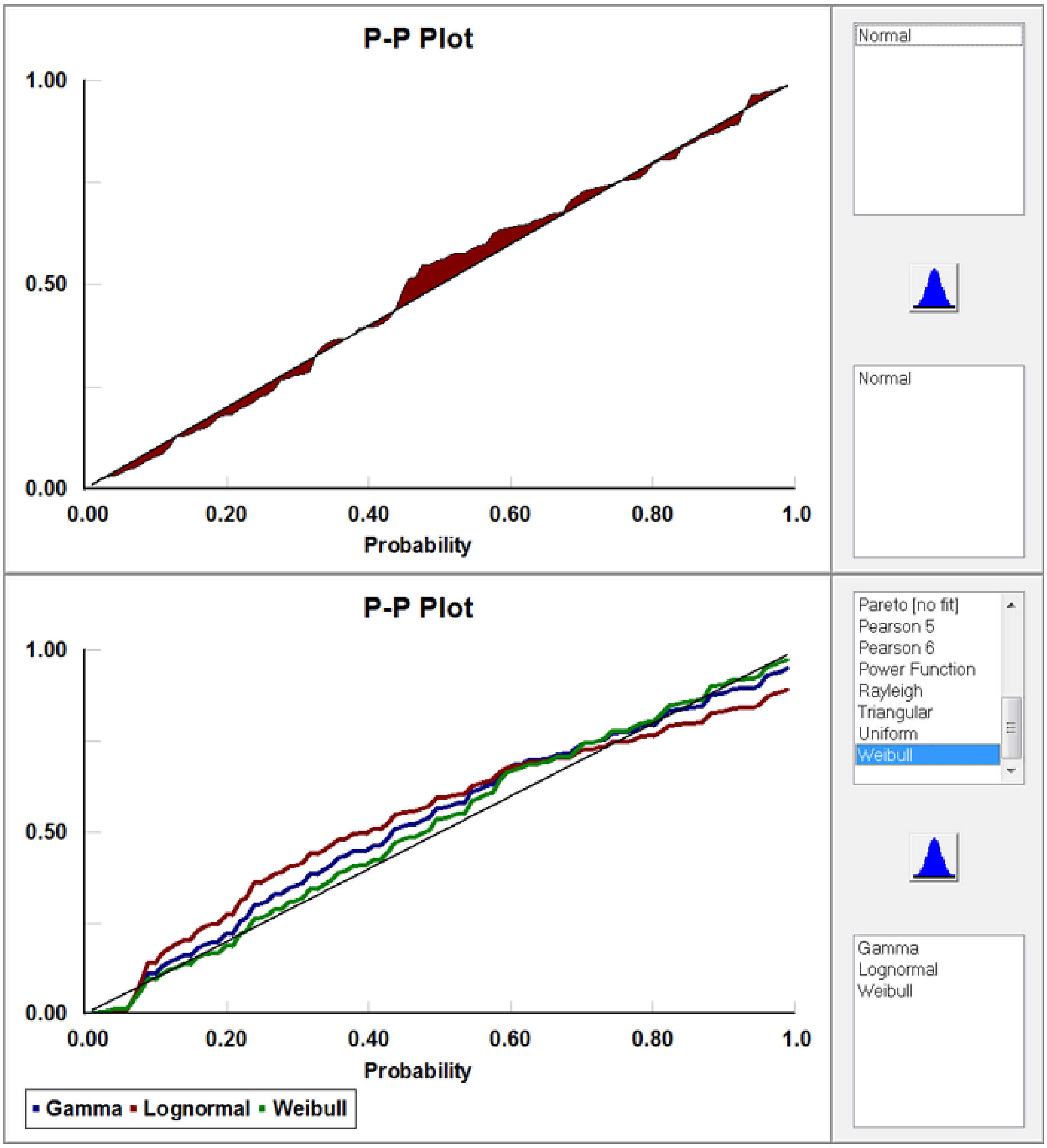
1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 374

Fit|Result Graphs|PP Plot

The P-P Plot can be accessed from the Result Graphs sub menu of the Fit menu.   It is activated only after specific distributions have been chosen.

The P-P Plot, as shown below, is a plot of the probability of the ith data point in the input data [straight line] from the Data Table versus the probability of that point from the fitted cumulative distribution.   This plot tends to be sensitive to variations in the center of the fitted data. (see Law&Kelton1)   If more than one distribution were fit to the input data, the list at the top right will allow a choice of available P-P plots of fitted distributions. Clicking a distribution will add that distribution to the plot. The list at the bottom right contains the densities plotted for comparison. Clicking a distribution in this list will remove that distribution from the plot. The legend will contain color-coded identification of the plotted distributions and input as shown in the second plot.

This graph, as with all graphs in Stat::Fit, may be modified, saved, copied, or printed with options generally given in the Graphics Style , Save As, and Copy commands in the Graphics menu.



1. "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p 339

Utilities|Distribution Viewer

The Distribution Viewer command can be accessed in the Utilities menu or from the Tool Bar.

The Distribution Viewer command creates a graph of any analytical distribution supported by Stat::Fit. The distribution can be changed using the distribution list in the upper right corner. This graph is not connected to any input data or document.

The Distribution Viewer initially sets default parameters for the chosen distribution. This graph of the distribution may then be modified using the sliders for each parameter, specifying the parameter value, or specifying one of the moments of the distribution. The number of moments which may be modified is limited to the number of free parameters for that distribution.

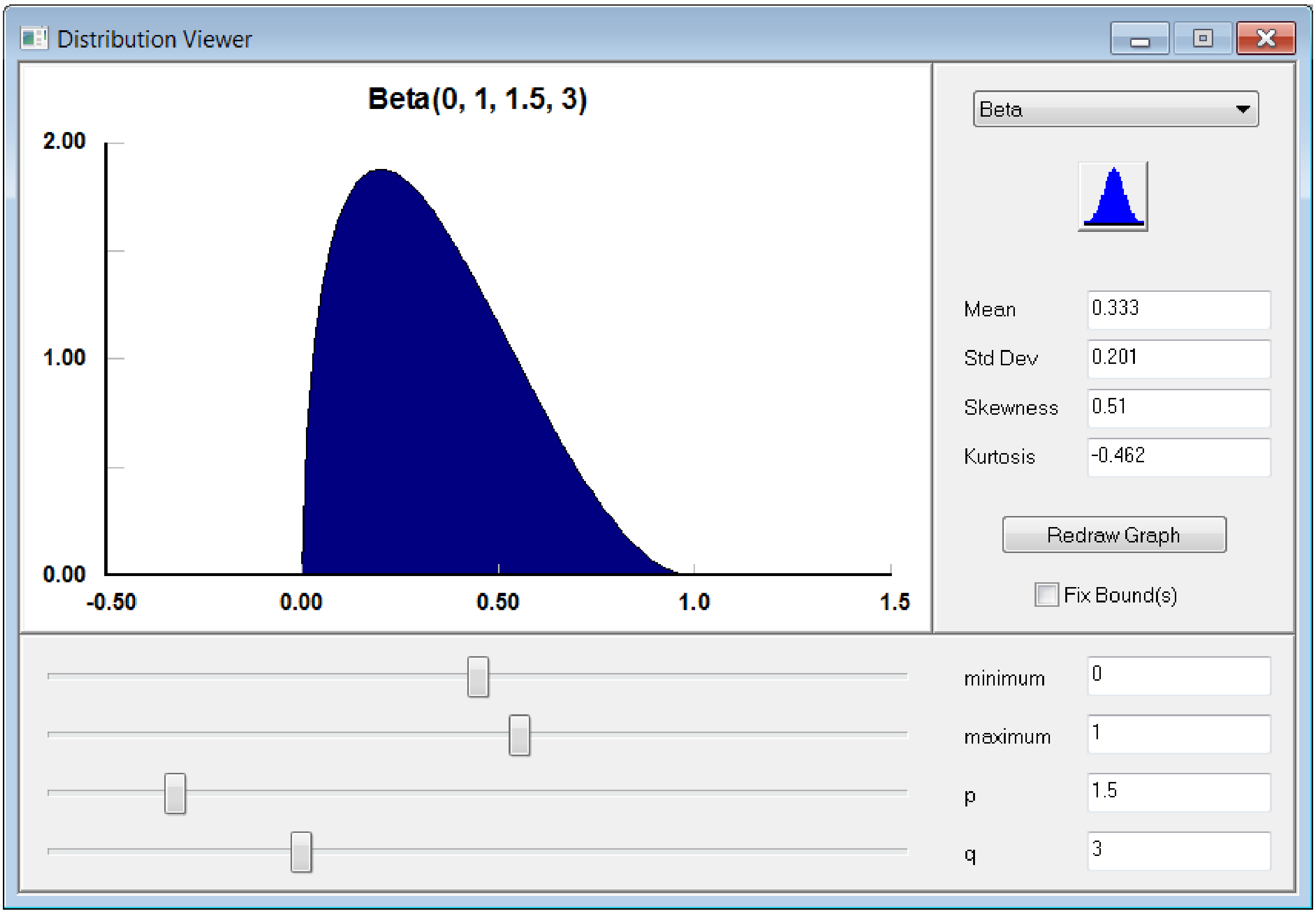
As the value of each parameter or moment is changed, the graph is frozen at its previous representation. The graph is updated when the slider is released, or the edited value is entered with a Return, Tab, or mouse click in another area. The graph may also be updated with the Redraw Graph button when active.

The bounds of the distribution, if any, can be fixed; however this reduces the number of moments that can be modified. A grayed moment box can be viewed, but not modified.

Occasionally, the specified moments cannot be calculated with the given parameters, such as when the mean is beyond one of the bounds. In this situation, an error message is given and the moments are recalculated from the parameters. Also, some distributions do not have finite moments for all values of the parameters and the appropriate moment boxes are shown empty.

As with all graphs in Stat::Fit, each Distribution Viewer may be customized by using the Graphics Style dialog in the Graphics menu. The graphs may also be copied to the Clipboard or saved as graphic files [.BMP] by using the Copy or Save As commands in the Graphics menu. Note that ,while the graph view currently open can still be modified, the copied or saved version is a fixed bitmap. The bitmap contains only the graph, and excludes parameter boxes and sliders.

The distribution in the Distribution Viewer may also be exported to another application by choosing the Export Fit command while the Distribution Viewer is the active window. In this way, no-data or minimal data descriptions can be translated from the form of the distribution in Stat::Fit to the form of the distribution in a particular application.



Distribution Viewer
Beta(0, 1, 1.5, 3)
2.00
1.00
0.00
-0.50
0.00
0.50
1.0
1.5
Beta
Mean
0.333
Std Dev
0.201
Skewness
0.51
Kurtosis
-0.462
Redraw Graph
Fix Bound(s)
minimum
0
maximum
1
p
1.5
q
3

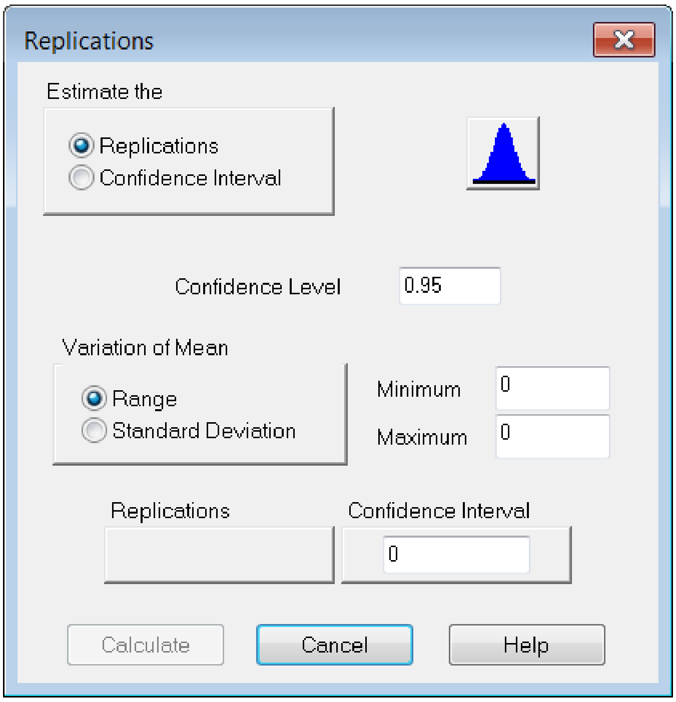
Utilities|Replications

The Replications command can be accessed in the Utilities menu.

The Replications command allows the user to calculate the number of independent data points, or replications, of an experiment necessary to provide a given range, or confidence interval, for the estimate of a parameter. The confidence interval is given for the confidence level specified, with a default of 0.95. The resulting number of replications is calculated using the t distribution.(see Banks et. al.1)

The expected variation of the parameter must be specified by either its expected maximum range or its expected standard deviation. Quite frequently, this variation is calculated by pilot runs of the experiment or simulation, but can be chosen by experience if necessary. Be aware that this is just an initial value for the required replications, and should be refined as further data are available.

Alternatively, the confidence interval for a given estimate of a parameter can be calculated from the known number of replications and the expected or estimated variation of the parameter.

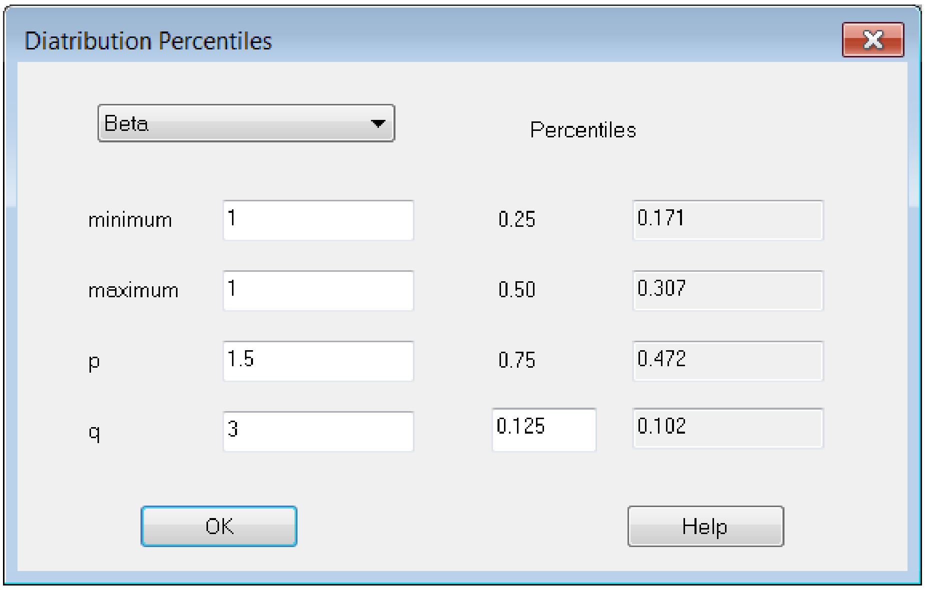


1. "Discrete-Event System Simulation, Second Edition", Jerry Banks, John S. Carson II,, Barry L. Nelson 1996, Prentice-Hall, p447

Utilities|Distribution Percentiles

The Distribution Percentiles dialog can be accessed in the Utilities menu.

The Distribution Percentiles dialog allows the user to get the various percentiles of any distribution allowed in Stat::Fit. First a distribution type must be chosen, the the parameters for that distribution entered. The quartile values of the distribution are given, with an extra line for any other value to be chosen.

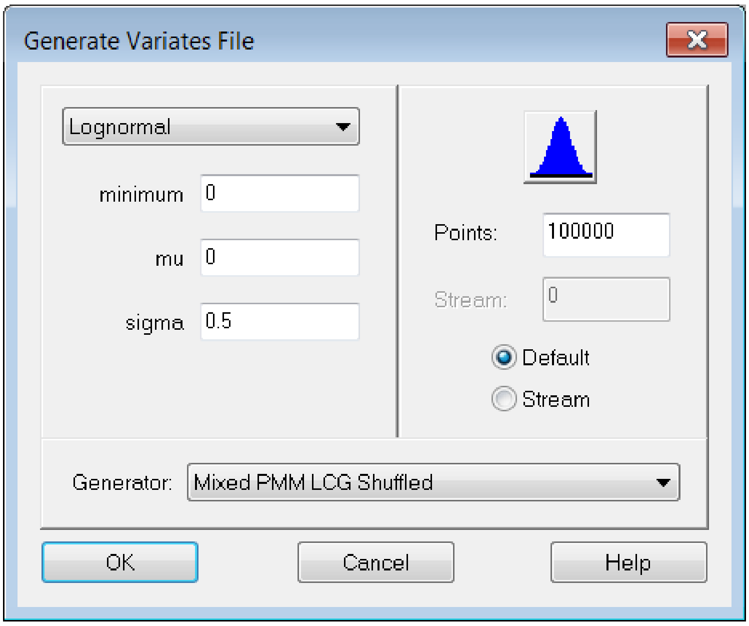


Utilities|Generate Variates File

Random variates can be generated to a file by accessing the Generate Variates File in the Utilities menu which allows up to 10,000,000 variates to be output.

The Generate Variates dialog provides the choice of distribution, parameters, and random number stream for the generation of random variates from each of the distributions covered by Stat::Fit, similar to the Generate dialog. The generated variates will not be entered into the data table, but instead will access a File Save As dialog so the a text file of the numbers can be saved.

By default, the generator is a Prime Modulus Multiplicative Linear Congruential Generator set up with 100 streams, each 100,000 numbers long with no overlap. Another generator, using Mixed Prime Modulus Multiplicative Linear Generators (L'Ecuyer) with shuffling added, may also be used. It has 1000 streams, each 10,000,000 numbers long, again with no overlap. A third generator, the Mersenne Twister, may be used with 1000 streams. Non overlap for the Mersenne Twister is not guaranteed, but is extremely unlikely.



Note that large numbers of variates may take awhile to generate, so be patient.

Graphics|Graphics Style

The Graphics Style command can be accessed from the Graphics menu only when a graph is the current view.

The Graphics Style dialog allows the current graph to be customized by changing the graph character, the graph scales, the title texts, the graph fonts, and the graph colors. The graphics style is limited for some graphs.

This graph remains modified as long as the document is open, and the graph is listed in the tree. It will also be saved with the project as modified. Note that any changes are singular to that particular graph; they do not apply to any other graph in that document or any other document.

If a special style is always desired, the default values may be changed by changing any graph to suit, checking the Save and Apply check box at the bottom of the dialog, and clicking OK. The resulting style becomes the default for all new graphs in Stat::Fit, with the exception of default titles, and some default appearances.

graphics text
graphics scale
graphics graph
graphics fonts
graphics colors

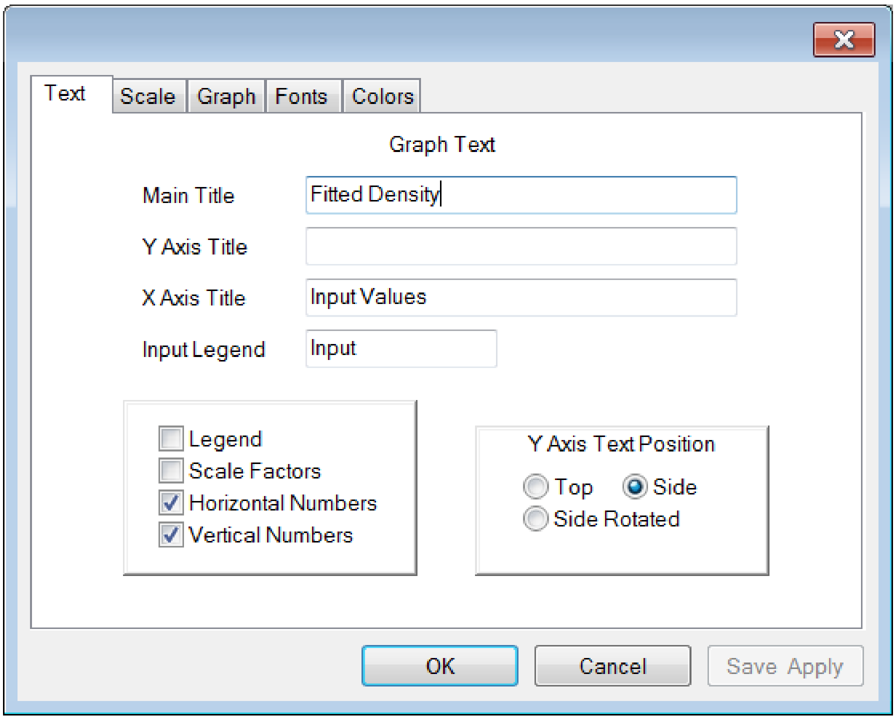
graphics text

The Text page of the Graphics Graph Style dialog allows title text to be added or changed on both axes, the legends, and the graph heading itself.

The Graph Titles set the main title, the x axis title, and the y axis title. The main title will always have a default value specific to the type of graph, unless this has been erased and saved by checking the Save and Apply check box. The x axis title will usually exist and contain information specific to the graph being displayed. The y axis title is usually blank. All the titles can be changed, but the changes are only for the graph being displayed. If the y axis title exists, it can be displayed at the top, side or side rotated.

The Graph Scales can be turned on or off for the x and y axes independently with the check boxes in the lower left of the page. The scale factor, the power of 10 necessary to reduce the scale to a manageable number, can be turned on or off with the check box in the lower left of the page.

The Graph Legends can be specified in the edit boxes as desired. The legends can be turned on with the check box in the lower left of the page, as shown below. The legends are specific to each graph as displayed and cannot be saved as the default.



graphics scale

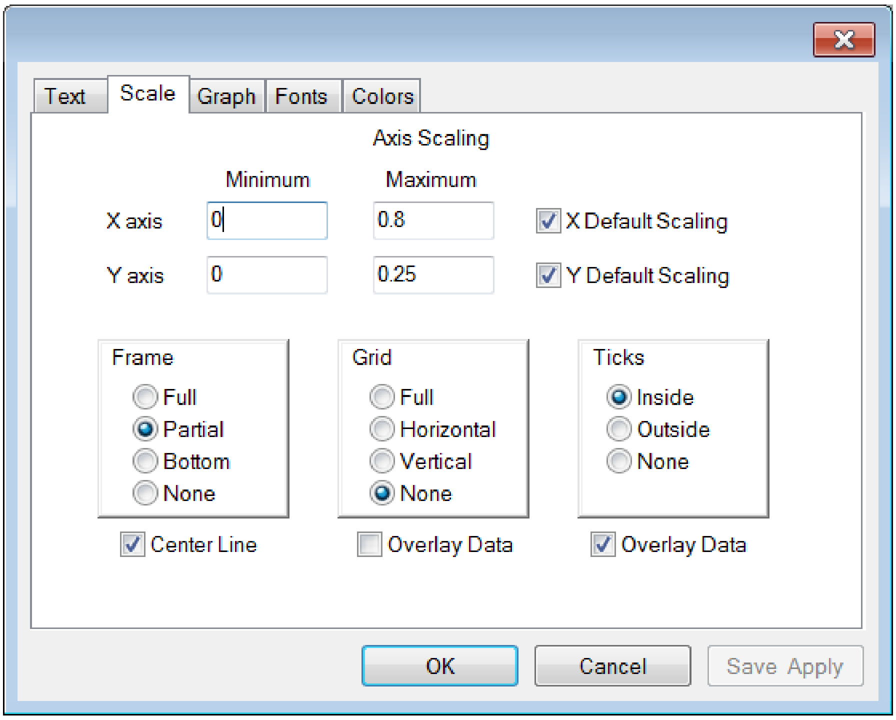
The Scale page of the Graphics Graph Style dialog allows the x and y axes to be scaled in various ways, as well as modifying the use of a graph frame, a grid, or tick marks.

The Graph Scale sets the minimum and maximum along each axis. If the default is selected for either the X or Y axis, Stat::Fit automatically chooses the minimum and maximum values to be plotted for that axis. These relative values can be different for different types of plots. Alternatively, the minimum and maximum values may be manually set for either axis by unchecking the default check box for that axis and providing the appropriate values in the input boxes. For this option, the closest minimum and maximum values will be chosen. If the default box is rechecked, the automatic values are returned and the manual values are lost. The default check boxes are normally on.

The Graphics Frame can be turned on in parts using the check boxes, full, partial, bottom or none, as shown below. The default is partial frame.

The Graphics Grid can be turned on by parts using the check boxes, full, horizontal, vertical ,or none, as shown below. The default is none. The grid may also be under or over the data by using the check box, Overlay Data. The default is under the data.

The Graphics Ticks can be turned on by parts using the check boxes, inside, outside,or none, as shown below. The default is inside. The ticks may also be under or over the data by using the check box, Overlay Data. The default is over the data.

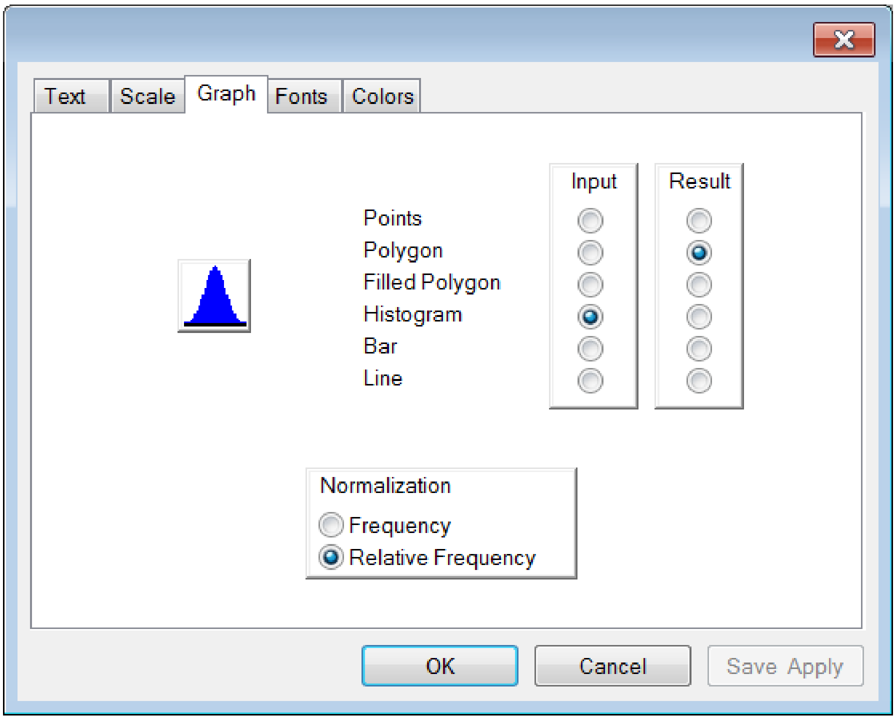


graphics graph

The Graph Appearance changes the way in which the data is plotted in the graph. For most graphs, the choices usually include: Points, Polygon, Filled Polygon, Histogram, Bar, Line. For Scatter Plots, the choices are modified and limited to: Points, Cross, Dots.

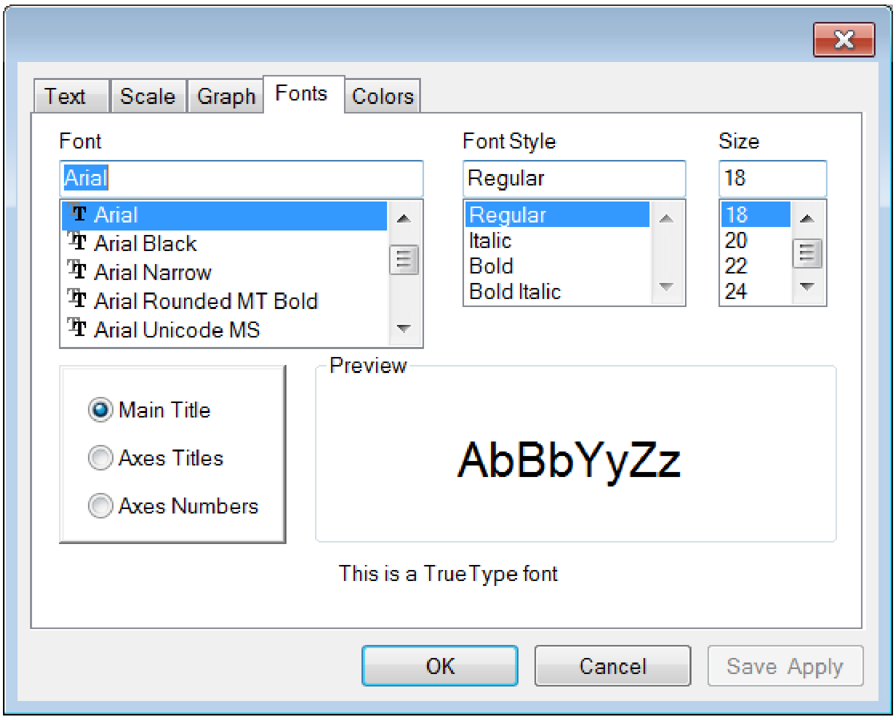
On graphs with comparisons, both the input and the result can be changed independently. Some graphs use straight lines for the comparison. Note that not every possibility gives a pleasing or even correct looking graph.

Normalization indicates whether the graph represents actual counts or a relative fraction of the total counts. Frequency represents actual counts for each interval [continuous random variable] or class [discrete random variable]. Relative Frequency represents the relative fraction of the total counts for each interval [continuous random variable] or class [discrete random variable]. Normalization is only available for the distribution graph types, such as Density, Distribution and Difference graphs.



graphics fonts

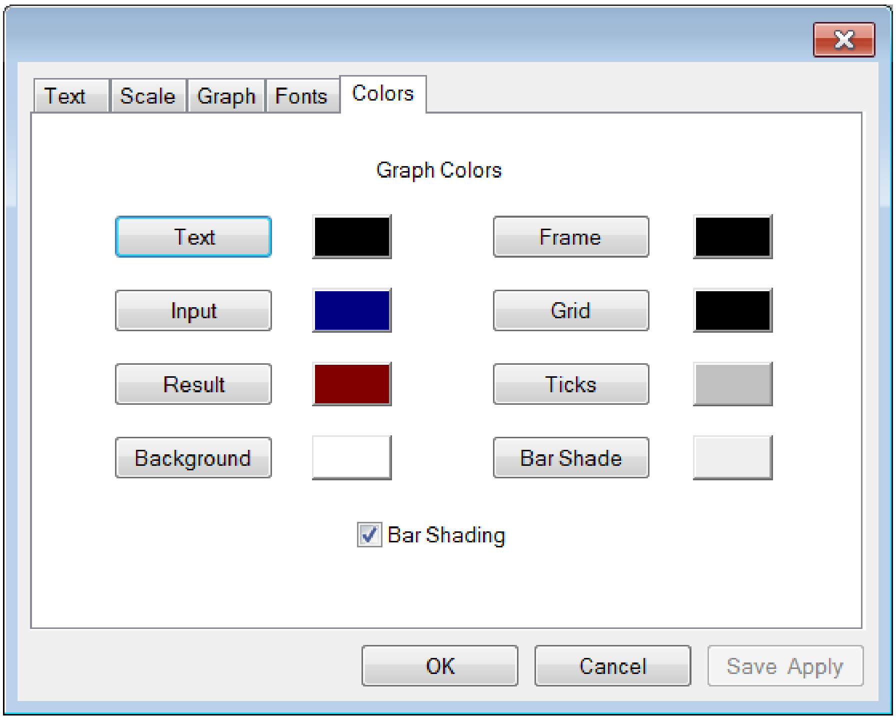
The Fonts page of the Graphics Graph Style dialog provides font selection for the text titles and scales in the the currently active graph. The Font type is restricted to True Type fonts that can be scaled on the display. The Font size is limited to a range that can be contained in the same window as the graph. Text colors can be changed in the Color page; no underlining or strikeouts are available.



graphics colors

The Colors page of the Graphics Graph Style dialog provides color options for all the fields of the currently active graph. For each object in the graph, a button to call the color dialog is located to the left and a color patch is located on the right. Text refers to all text including scales. Input refers to the first displayed graph, the input data in comparison graphs. Result refers to fitted data, with variation added as more results are added. Bar Shade refers to the left and bottom of histogram boxes and requires the check box be set on as well. Background refers to the background color; full white does not print.

Note that the colors are chosen to display well on a monitor. If a laser printer with gray scales is used, the colors should be changed to brighter colors or grays in order to generate appropriate gray levels. Some of the colors will default to the nearest of the 16 basic Windows colors in order to display properly.



Graphics|Copy

The Graphics Copy command can be accessed from the Graphics menu only when a graph is the current view. The Graphics Copy command places a copy of the current graph in the Clipboard as a bitmap. From there, it can be pasted into another application if that application supports pasting of bitmaps. Note that the copy can no longer be modified with the Graphics Style dialog.

Graphics|Save As

The Graphics Save As command can be accessed only from the Graphics menu only when a graph is the current view.   The File Save As command is only for use on the Stat::Fit projects and the data in the Data Table.

The Graphics Save As command saves a bitmap [.BMP] file of the current graph.   From there, it can be loaded into another application if that application supports the display of bitmaps. Note that the copy can no longer be modified with the Graphics Style dialog.

File|Save, File|Open

The Save command in the File menu saves the Stat::Fit document to its project file. The existing file is overwritten. If a project file does not exist [the project window will have a Project xx name], the Save As command will be called.

The Save command does NOT save the input data in a text file, but saves the full document, that is, input data, calculations, and view information, to a binary project file, "your project name".SFPX. This binary file can be reopened in Stat::Fit, but cannot be imported into other applications. If a text file of the input data is desired, the Save Input command should be used instead.

The Open command in the File menu prompts for a filename using the Open dialog. If the filename has a .SFPX extension indicating a Stat::Fit project file, the project file is opened in a new document and associated with that document. If the filename has a .SFP extension, indicating a saved project from a previous version of Stat::Fit, the project is opened and converted to the current version.

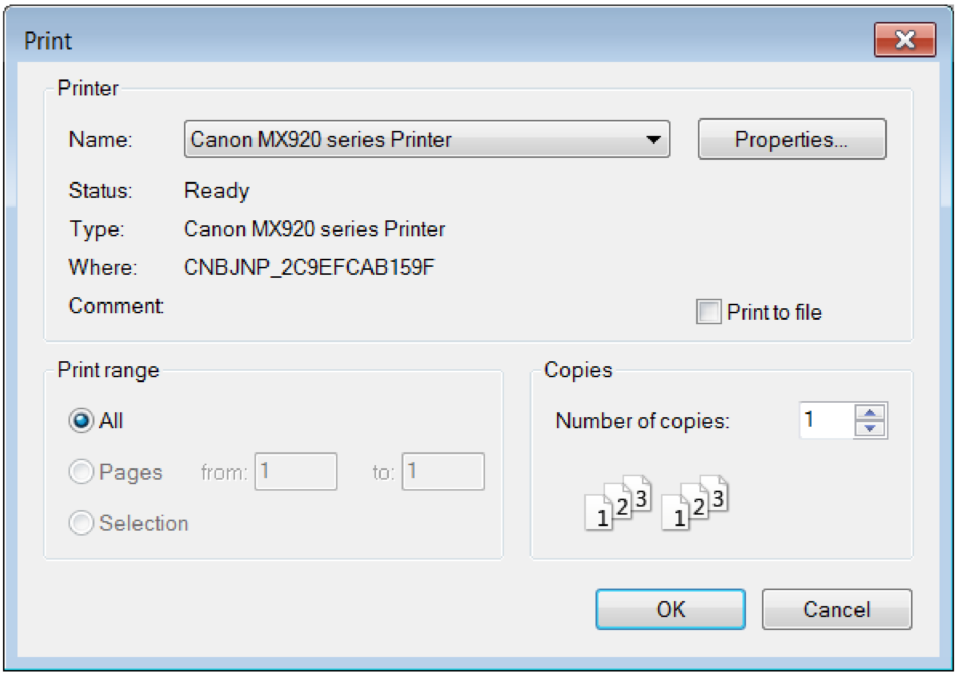
Otherwise, the file is assumed to be a text file with numbers. A new project is opened by reading the file for input data. The document created from a text file has an association with a project file named after the text file but has not been saved as a project yet.

File|Print

The Print command can be accessed in the File menu.

The Print command initiates printing of the output specified by the Print Style command, checks the printer and asks final permission to print with a standard dialog.   This dialog also gives access to the Printer Setup dialog to specify the printer type, paper size and orientation, as well as the printer Options dialog specific to the chosen printer.

If you are uncertain of the expected output, exit this dialog with the Cancel button and use the Print Preview command to view a screen copy.

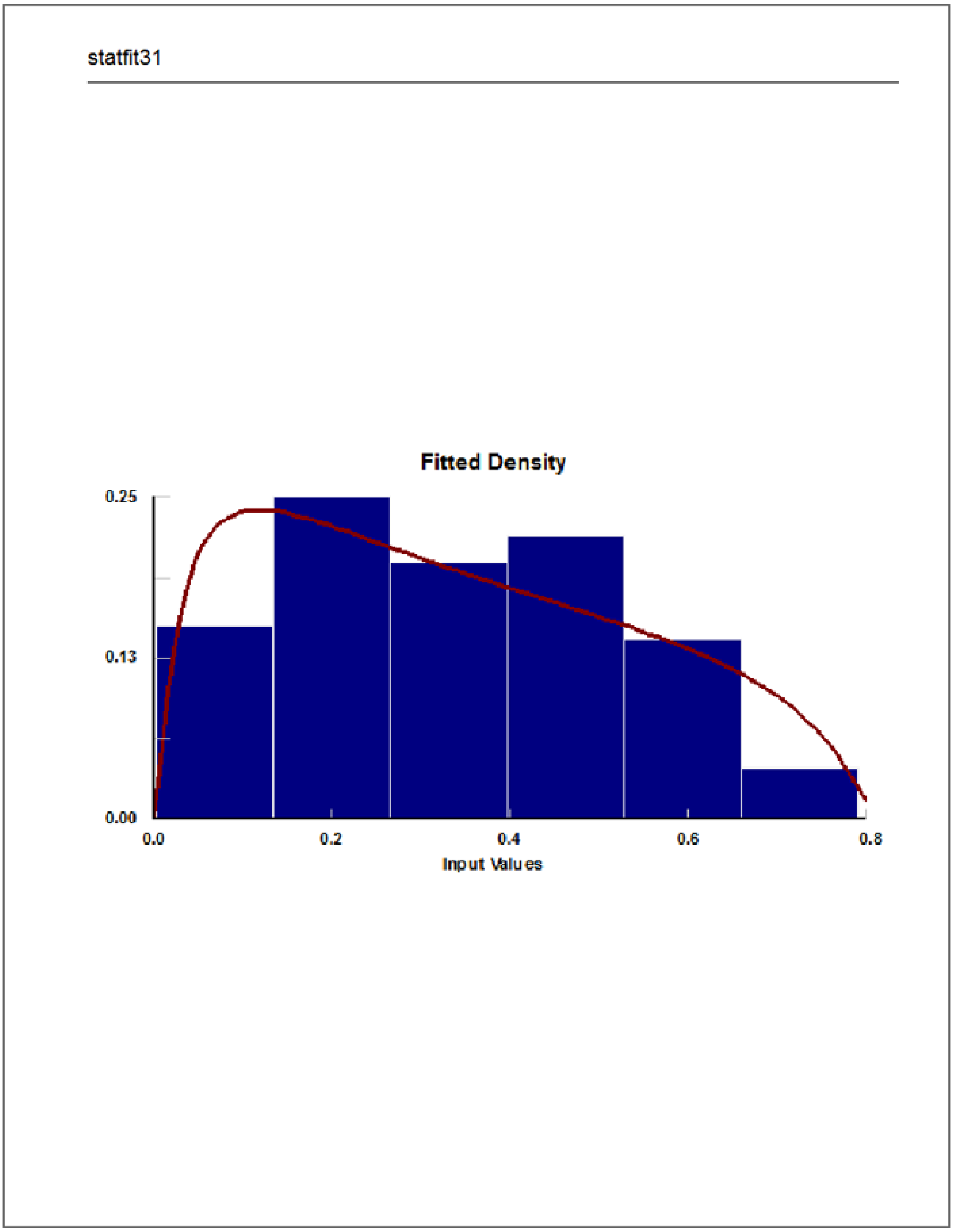


File|Print Preview

The Print Preview command can be accessed from the File menu.

The Print Preview command opens a separate window to display each expected page of the print output, using the options specified in the Print Style command.   These windows can be closed by clicking on the Close button.

The Print Preview windows give a scaled version of the output to be generated by the Print command as shown.



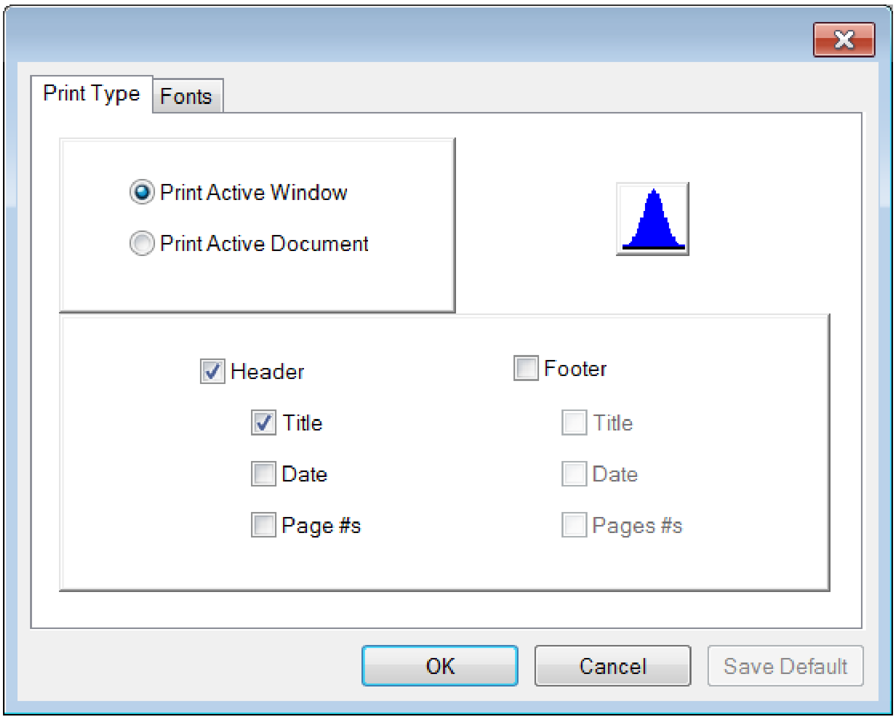
File|Print Style

The File Print Style dialog provides printer output options for the project currently displayed (the currently active view is part of this document). Both the File Print command and the File Print Preview command will follow these options when building print output.

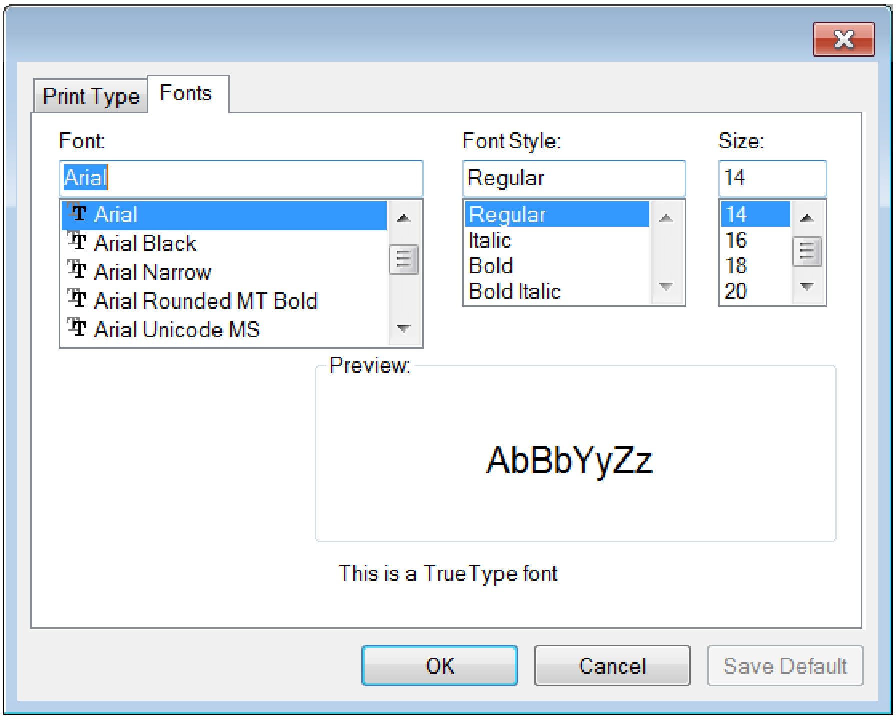
The Select page of the File Print Style dialog provides choices for setup of the print output. The print type chosen here will determine what kind of print output will be generated by the Print and Print Preview commands.

The Print Type determines which data views of the Stat::Fit document get printed. Print Active Window prints just the currently active view, and can be a single view of a Stat::Fit document or a Distribution Viewer. Print Active Document prints all open views of the currently active Stat::Fit project, one by one.

Print Labels can be header, footer, or absent as selected. They can include a title [document name], date, and page number as selected. The left margin is left intentionally larger than the right in order to allow binding.



The Fonts page of the File Print Style dialog provides font choice for the text pages to be printed. This choice does not change the font(s) chosen for the graphics pages.


Print Type
Fonts
Font:
Arial
Arial
Arial Black
Arial Narrow
Arial Rounded MT Bold
Arial Unicode MS
Font Style:
Regular
Regular
Italic
Bold
Bold Italic
Size:
14
14
16
18
20
Preview:
AbBbYyZz
This is a TrueType font
OK
Cancel
Save Default

File|Printer Setup

The Printer Setup command can be accessed from the File menu.

This standard Print Setup dialog will allow specification of the printer, paper size, and orientation of printed output. It will also allow access to the Options dialog of the chosen printer. This setup will subsequently be used by the Print command.