Stat::Fit®

Version 2

Statistically Fit®
Software

geer mountain software corporation

Stat::Fit®

104 Geer Mountain Road, South Kent, Ct. 06785
Telephone: 860-927-4328

Printed in the United States of America.

Geer Mountain Software Corp.
Software License and Warranty Agreement

This document is a legal agreement between you, the end user, and Geer Mountain Software Corp..
BY OPENING THE SEALED DISK PACKAGE OR CONTINUING WITH THIS SOFTWARE INSTALLATION, YOU ARE AGREEING TO BE BOUND BY THE TERMS OF THIS AGREEMENT. IF YOU DO NOT AGREE TO THE TERMS OF THIS AGREEMENT WHICH INCLUDE THE LICENSE AND LIMITED WARRANTY, PROMPTLY RETURN THE PACKAGE UNOPENED AND ALL OF THE ACCOMPANYING ITEMS (including documentation) FOR A FULL REFUND.

License

Geer Mountain Software grants to you, the end user, a non-exclusive license to use the enclosed computer program (the "SOFTWARE") on a single computer system, subject to the terms and conditions of this License and limited Warranty Agreement.

Copyright and Permitted Use

The SOFTWARE is owned by Geer Mountain Software and is protected by United States copyright law and international treaty provisions. Treat the SOFTWARE exactly as if it were a book, with one exception: You may make archival copies of the SOFTWARE to protect it from loss. The SOFTWARE may be moved from one computer to another, as long as there is no possibility of two persons using it at the same time.

You may transfer the complete SOFTWARE and the accompanying written materials together on a permanent basis provided you do not retain any copies and the recipient agrees to the terms of this Agreement.

Other Restrictions

You may not lease, rent or sub-license the SOFTWARE. You may not transfer the SOFTWARE or the accompanying written materials except as provided above. You may not reverse engineer, decompile, disassemble or create derivative works from the SOFTWARE. If you later receive an update to this SOFTWARE or if this SOFTWARE is an update to a prior version, any transfer must include both the update and all accessible prior versions of the SOFTWARE.

Warranty and Liability

Geer Mountain Software warrants that (a) the SOFTWARE will perform substantially in accordance with the accompanying written materials and (b) the SOFTWARE is properly recorded on the disk media.

Your failure to return the enclosed registration card may result in Geer Mountain Software's inability to provide you with updates to the SOFTWARE and you assume the entire risk of performance and result in such event. This Warranty extends for sixty (60) days from the date of purchase. The above Warranty is in lieu of all other warranties, whether written, express, or implied. Geer Mountain Software specifically excludes all implied warranties, including, but not limited to implied warranties of merchantability and fitness for a particular purpose.

Geer Mountain Software shall not be liable with respect to the SOFTWARE or otherwise for special, incidental, consequential, punitive, or exemplary damages even if advised of the possibility of such damages. In no event shall liability for any reason and upon any cause of action whatsoever exceed the purchase price of the software.

U. S. Government Restricted Rights

If you are acquiring the SOFTWARE on behalf of any unit or agency of the United States Government, the following provisions apply:

The Government acknowledges Geer Mountain Software's representation that the SOFTWARE and its documentation were developed at private expense and no part of them is in the public domain. The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subparagraphs ©(1)(Iii) of The Rights in Technical Data and Computer Software clause of DFARS 252.227-7013 or subparagraphs (c)(1) and (2) of the Commercial Computer Software-Restricted Rights at 48 CFR 52.227-19, as applicable. Manufacturer is Geer Mountain Software Corp., 104 Geer Mountain Road, South Kent, CT 06785.

This Agreement is governed by the laws of the State of Connecticut.

Chapter 1 – Introduction

Stat::Fit, a *Statistically Fit* application which fits analytical distributions to user data, is meant to be easy to use. Hopefully its operation is so intuitive that you never need to use this manual. However, just in case you want to look up an unfamiliar term, or a specific operation, or enjoy reading software manuals, we provide a carefully organized document with the information easily accessible.

Organization of this manual

Chapter 2 lists the system requirements and installation procedure.

Chapter 3 summarizes a Quick Start for using Stat::Fit. An overview of the basic operations using the default settings is given.

Chapter 4 provides the options for bringing data into Stat::Fit and for their manipulation.

Chapter 5 describes the distribution fitting process, the statistical calculations and the goodness of fit tests.

Chapter 6 goes into the numerous options available for the types of graphs and graph styles.

Chapter 7 provides details on how to print graphs and reports.

Chapter 8 is a tutorial with an example.

Appendix A: Distributions

Appendix B: Reference Books

Terms and conventions

This manual uses Windows-specific terminology and assumes that you know how to use Windows. For help with Windows, see your Windows documentation.

The terminology in this manual should be familiar to anyone with basic statistics knowledge.

Technical Support

If **Stat::Fit** was purchased directly from Geer Mountain Software, support is available through:

Internet:	statfit@geerms.com
Fax:	860-927-1614
Phone:	860-927-4328
	Monday-Friday, 9am to 6pm EST

Otherwise, support is available through the technical support area where **Stat::Fit** was purchased.

Chapter 2 – Installation

Installation procedure

Insert the **Stat::Fit®** CD and follow directions.

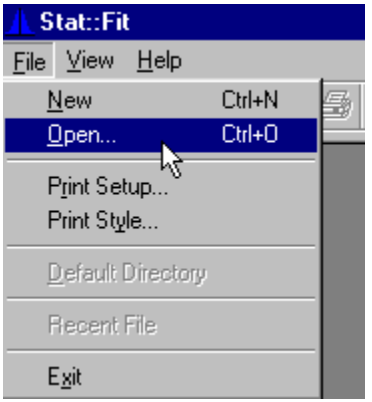
If the AutoRun function on your computer is turned off, use the Start button, click Run and go to the directory corresponding to your CD ROM and run **setup.exe**.

Chapter 3 – Overview of Stat::Fit

Basic Operation

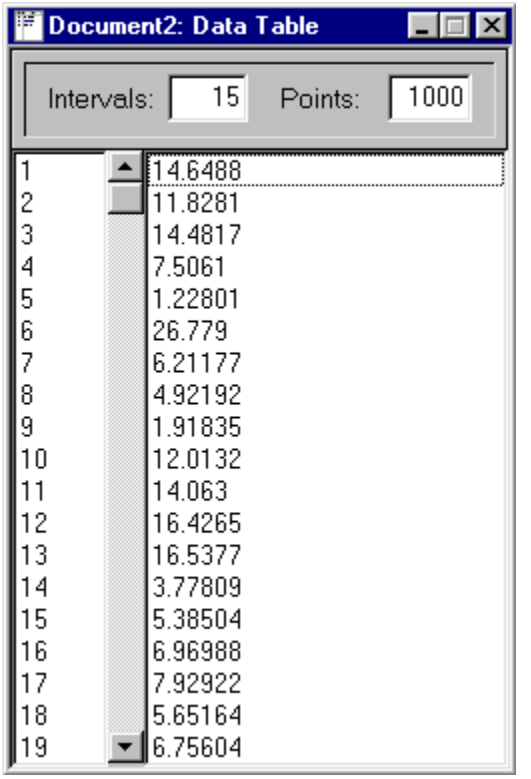
This section describes the basic operation of Stat::Fit using the program's default settings. For this example, we assume that the data is available in a text file.

The data is loaded by clicking on the Open File icon, or selecting File on the menu bar and then Open from the Submenu, as shown below. All icon commands are available in the menu.

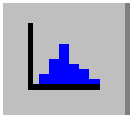


A standard Windows dialog box allows a choice of drives, directories and files.

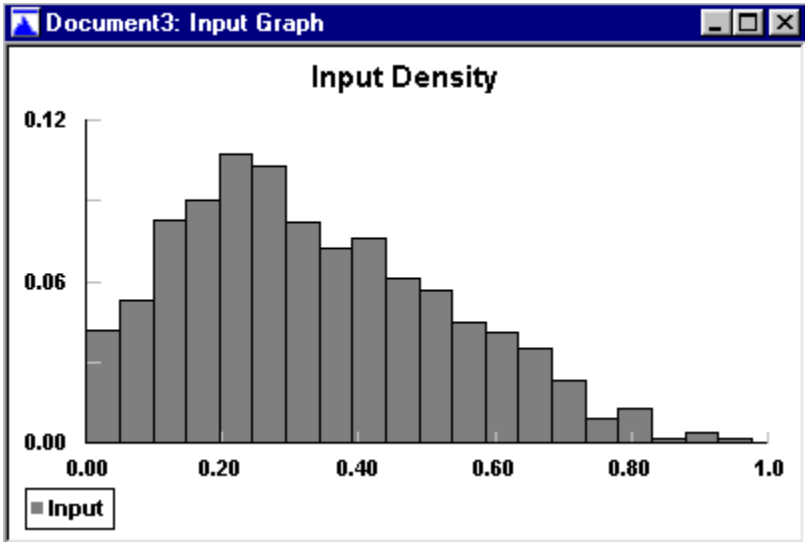
The data in an existing text file loads sequentially into a Data Table (see Chapter 4 for features of the Data Table). Data may also be entered manually. **Stat::Fit** allows up to 8000 numbers.



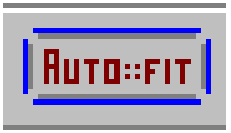
The number of data points is shown on the upper right; the number of intervals for binning the data on the upper left. By default, **Stat::Fit** automatically chooses the minimum number of intervals to avoid data smoothing. Also by default, the data precision is 6 decimal places. (See Chapter 4 for other interval and precision options.)



A histogram of the input data is displayed by clicking on the Input Graph icon. (For additional information on graph styles and options, see Chapter 6.)



To Fit a Distribution:



Continuous and discrete analytical distributions can be automatically fit to the input data by using the Auto::Fit command. This command follows nearly the same procedure described below for manual fitting, but chooses all distributions appropriate for the input data. The distributions are ranked according to their relative goodness of fit. An indication of their acceptance as good representations of the input data is also given. A table, as shown below provides the results of the Auto::Fit procedure.

Auto::Fit of Distributions

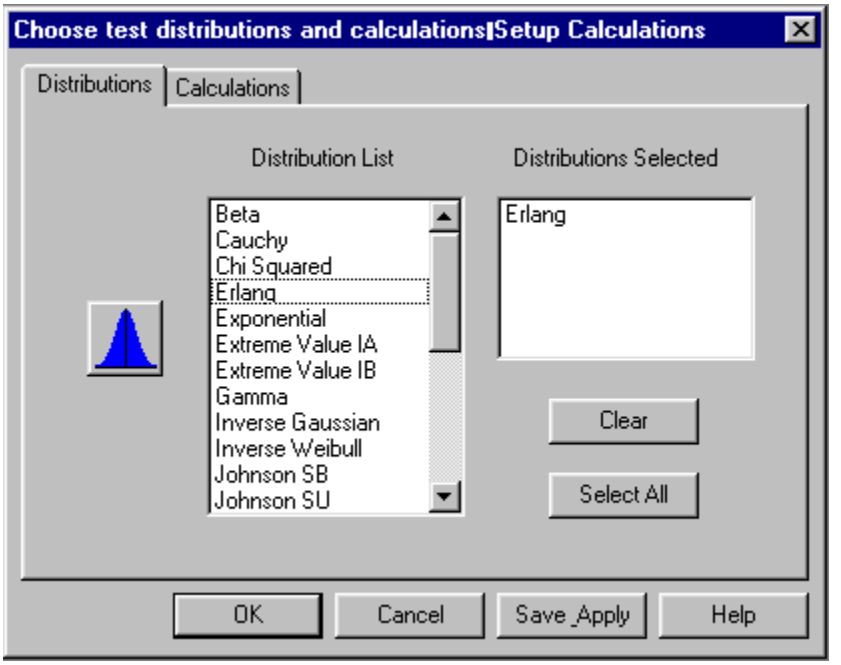
distribution	rank	acceptance
Inverse Gaussian(-2.48304, 36.8058, 12.4311)	98.9	do not reject
Lognormal(-2.00183, 2.32172, 0.572285)	96.4	do not reject
Pearson 5(-5.40605, 6.02755, 77.6132)	78.7	do not reject
Gamma(0.270107, 1.88582, 5.13074)	66.4	do not reject
Inverse Weibull(-21.125, 6.05196, 3.62105e-002)	61.7	do not reject
Erlang(0.270107, 2., 4.83921)	47.2	do not reject
LogLogistic(-1.08811, 2.69717, 9.25314)	45.7	do not reject
Pearson 6(0.44382, 98.9852, 1.80798, 19.7105)	36.3	do not reject
Weibull(0.410585, 1.41541, 10.4908)	21.	do not reject
Beta(0.44382, 59.3738, 1.52761, 7.89056)	8.75	do not reject
Chi Squared(-7.5199, 17.2907)	1.95e-003	reject
Rayleigh(-1.37887, 9.37822)	4.02e-007	reject
Exponential(0.44382, 9.50422)	0.	reject
Pareto(0.44382, 0.351087)	0.	reject
Power Function(0.443199, 38.2082, 0.588391)	0.	reject
Triangular(0.316432, 38.2564, 1.2507)	0.	reject
Uniform(0.44382, 38.1042)	0.	reject
Johnson SB	no fit	reject

Manual fitting of analytical distributions to the input data requires a sequence of steps starting with a setup of the intended calculations.

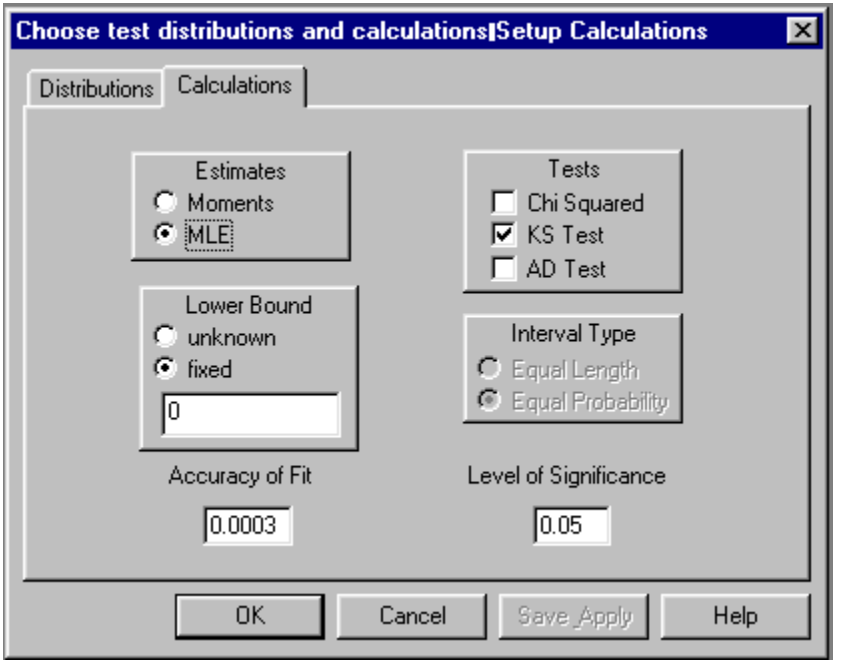
The setup dialog is entered by clicking on the Setup icon or selecting Fit from the Menu bar and Setup from the Submenu.

The first page of the setup dialog presents a list of analytical distributions. A distribution, say Erlang, is chosen by clicking on its name in the list on the left. The selected distribution then

appears in the list on the right. The setup is selected for use by clicking OK.



The goodness of fit tests are calculated by clicking on the Fit icon. By default, only the Kolmogorov Smirnov test is performed; other tests and options may be selected on the Calculations page of the setup dialog, as shown below. (For details of the Chi Squared, Kolmogorov Smirnov and Anderson Darling tests, see Chapter 5.)



A summary of the goodness of fit tests appears in a table, as shown below:

goodness of fit

data points	1000
estimates	maximum likelihood estimates
accuracy of fit	3.e-004
level of significance	5.e-002

summary

distribution	Kolmogorov Smirnov
Erlang(-0.116, 2., 5.03)	2.82e-002

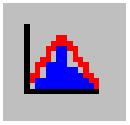
detail

Erlang

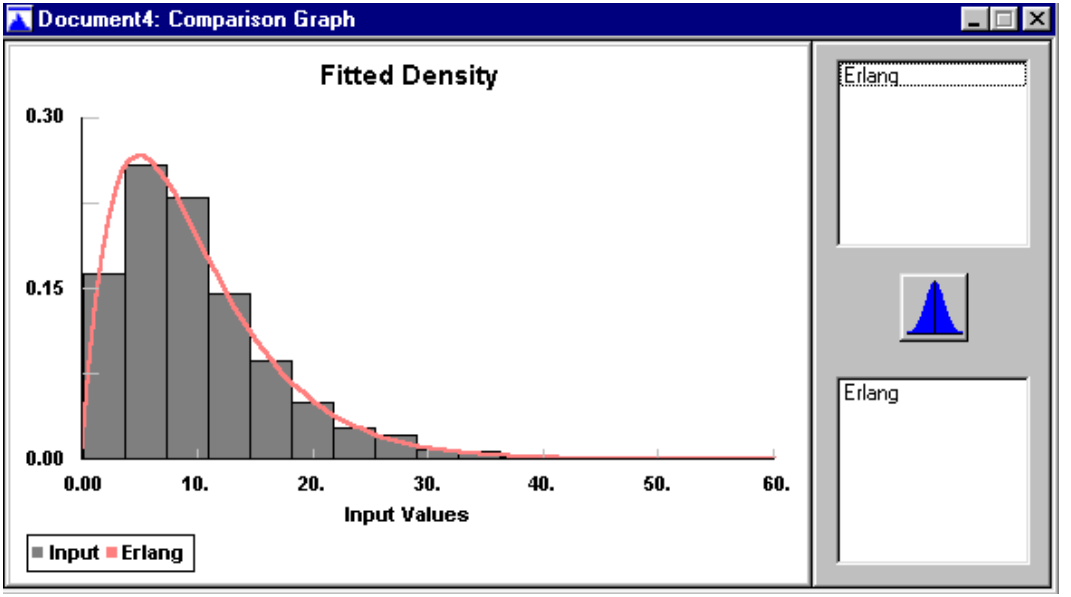
minimum	=	-0.115697
m	=	2.
beta	=	5.02752

Kolmogorov-Smirnov

data points	1000
ks stat	2.82e-002
alpha	5.e-002
ks stat(1000,5.e-002)	4.28e-002
p-value	0.393
result	DO NOT REJECT



A graph comparing the fitted distribution to the input data is viewed by clicking on the Graph Fit icon. (Other results graphs as well as modifications to each graph are described in Chapter 6.)



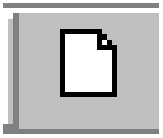
The **Stat::Fit** project is saved by clicking on the Save icon which records not only the input data but also all calculations and graphs.

Congratulations! You have mastered the **Stat::Fit** basics.

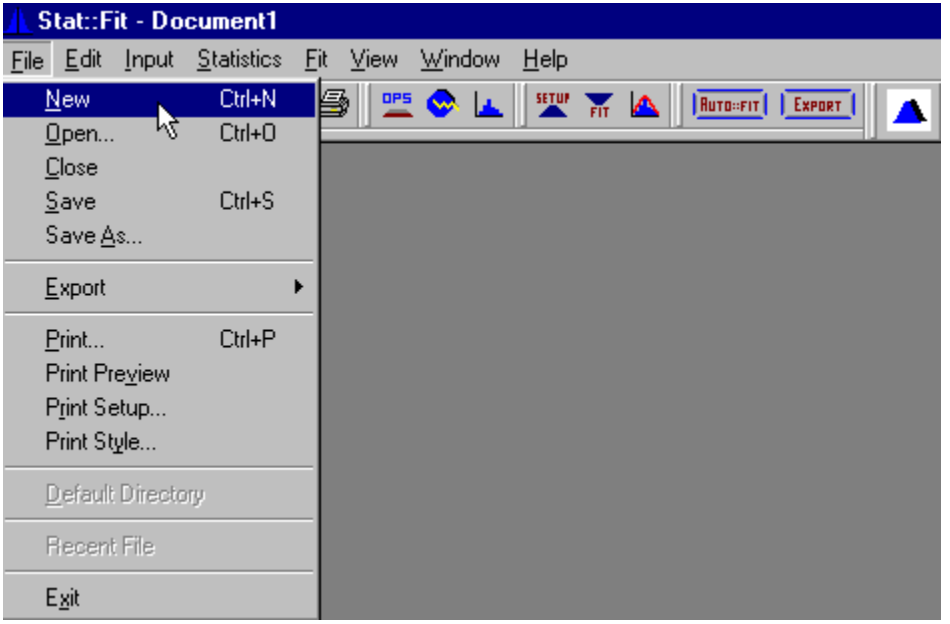
Chapter 4 – Data Entry and Manipulation

This chapter describes in more detail the options available to bring data into **Stat::Fit** and manipulate it.

Create a New Project



A New Project is created by clicking on the New Project icon on the Control Bar or by selecting File from the menu bar and then New from the Submenu.

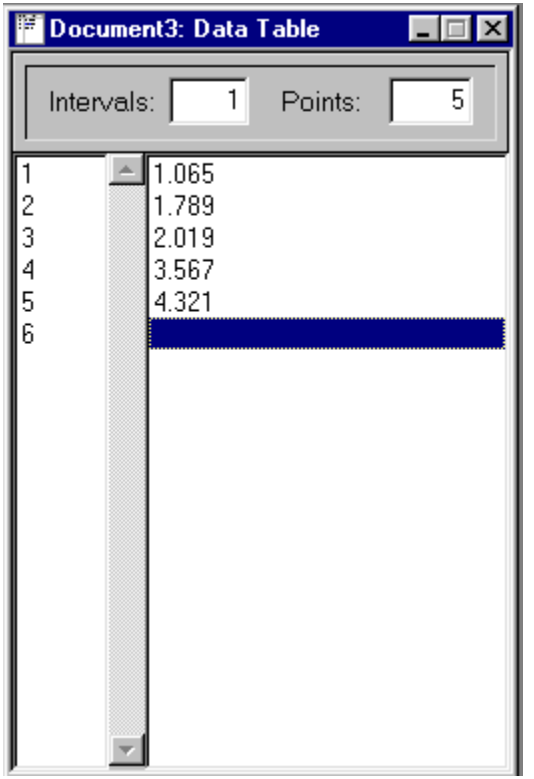


The New Project command generates a new **Stat::Fit** document, and shows an empty Data Table with the caption, *document xx*, where *xx* is a sequential number depending on the number of

previously generated documents. The document may be named by invoking the *Save As* command and naming the project file. Thereafter, the document will be associated with this stored file.

The new document does not close any other document. **Stat::Fit** allows multiple documents to be open at any time. The only limit is the confusion caused by the multitude of views that may be opened.

An input table appears, as shown below, which allows manual data entry.



Alternatively, data may be pasted from the Clipboard.

Opening an Existing Project

An existing project is opened by choosing File on the Menu bar and then Open from the Submenu, or by clicking on the Open icon on the Control Bar.

An Open Project Dialog box allows a choice of drives, directories and files.

Stat::Fit accepts 4 types of files:

.SFP – **Stat::Fit** project file

.txt – Input data

.* - User specified designation for input data

.bmp – Graphics bitmap file

Select the appropriate file type and click on OK.

If the filename has a .SFP 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 .bmp extension indicating a saved bitmap (graph…), the bitmap is displayed. Otherwise, a text file is assumed and 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 with the .SFP extension. *The project file has not been saved.*

If the number text contains non-numeric characters, they cause the number just prior to the non-numeric text to be entered. For example, 15.45% would be entered as 15.45 but 16,452,375 would be entered as three numbers: 16, 452 and 375.

Saving Files

The project file, the input data, or any graph are saved through one of the Save commands in the File submenu.

When input data is entered into **Stat::Fit** whether through manual entry in a new document, opening a data file, pasting data from the Clipboard, or reopening a **Stat::Fit** project file, a **Stat::Fit** document is created which contains the data and all subsequent calculations and graphs. If the document is initiated from an existing file, it assumes the name of that file and the document can be saved automatically as a **Stat::Fit** project [.SFP extension] with the Save command.

The **Save** command saves the **Stat::Fit** document to its project file. The existing file is overwritten. If a project file does not exist (the document window will have a document 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*".*SFP*. 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.

The **Save As** command is multipurpose. If the document is unnamed, it can be saved as either a **Stat::Fit** project of a text data file with the Save As command. If a document is named, it's name can be changed by saving either the project or the input data to a file with a new name. (In any situation, the document assumes the name of the filename used.)

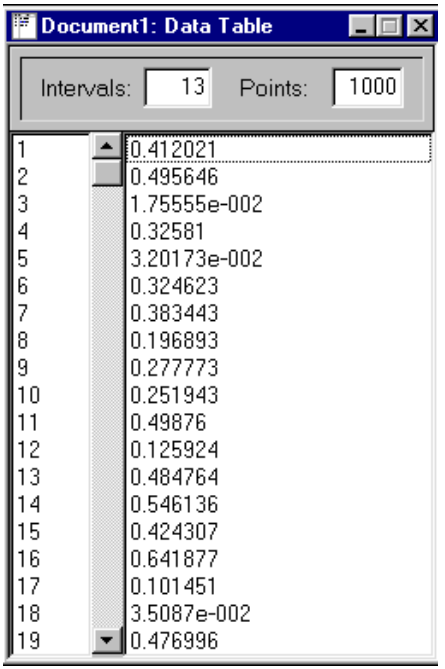
The **Save Input** command saves the input data in a separate text file, with each data point separated with a carriage return. This maintains the integrity of your data separate from the **Stat::Fit** project files and calculations. If an existing association with a text

file exists, a prompt will ask for overwrite permission. Otherwise, a *Save As* dialog will prompt for a file name, save to that file, and associate that text file with the document. If no extension is specified, the file will be saved with the extension *.txt*.

The Save icon on the Control Bar saves the current document to its project file.

The Data Table

All data entry in **Stat::Fit** occurs through the Data Table. After a project is opened, data may be entered manually, by pasting from the Clipboard, or by generating data points from the random variate generator. An existing **Stat::Fit** project may be opened and data may be added manually. An example of the Data Table is shown below:



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.

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 window if it does not already have the colored title bar. Manual data entry begins when a number is typed. The current data in the Data Table is grayed and an input box is opened. The input box will remain open until the Enter key is hit unless the Esc key is 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 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 a 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, 8000 numbers.

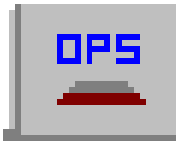
A single number or group of numbers may be selected in the Data Table. The selected number(s) are highlighted in a color, usually blue. To select multiple numbers, use the *shift key* with a mouse click to get a range of numbers from the last selected number to the current position. If the *ctrl key* 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 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.

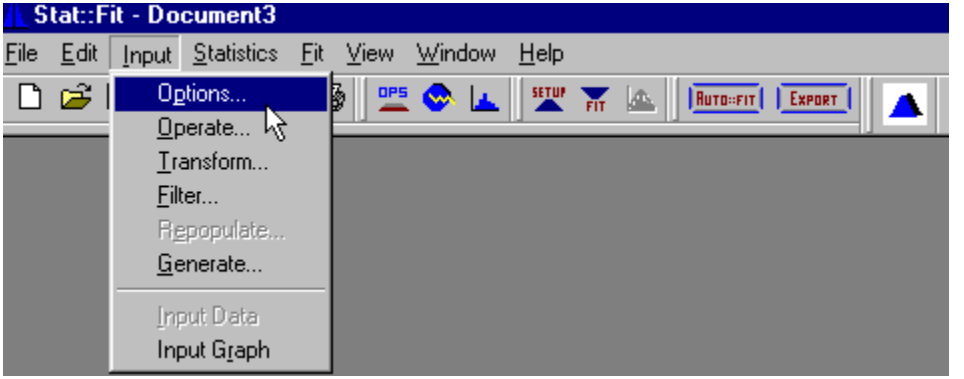
The Clear command clears all input data and calculations in the current document, after a confirming dialog. All views which depend on these data and calculations are closed. An empty Data Table is left open and the document is left open. The underlying **Stat::Fit** project file, if any, is left intact, but a Save command will clear it as well. *Use this command carefully.* This command is NOT the same as the New command because it maintains the document's connection to the disk file associated with it, if any.

Input Options

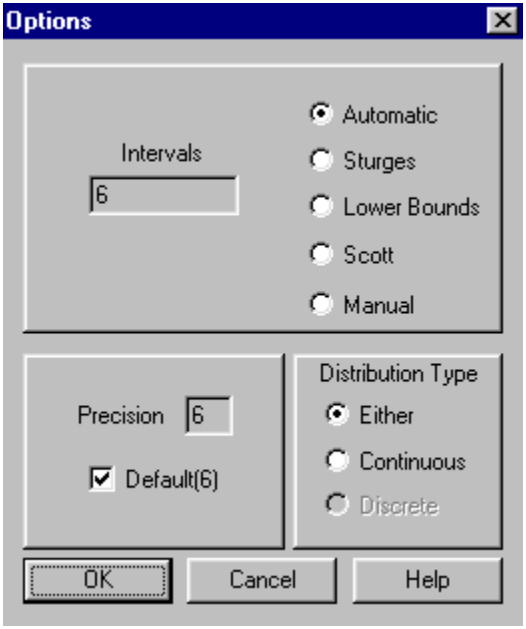
Input Options allows several data handling options to be set: the number of intervals for the histogram and the chi-squared goodness of fit test, the precision with which the data will be shown and stored, and the distribution types which will be allowed.



The Input Options dialog is entered by clicking on the Input Options icon or by selecting Input from the menu bar and then Options from the Submenu.



An Input Options Dialog box is shown below:



The number of **intervals** specifies the number of bins into which the input data will be sorted. 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 will always default to the maximum number of discrete classes for the same data.

The five choices for deciding on the number of intervals are:

Auto – Automatic mode uses the minimum number of intervals possible without losing information[1]. Then the intervals are increased if the skewness of the sample is large.

Sturges – An empirical rule for assessing the desirable number of intervals into which the distribution of observed data should be classified. If N is the number of data points and k 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 of 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 – Allows arbitrary setting of the number of intervals, up to a limit of 1000.

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.

[1] "Oversmoothed Nonparametric Density Estimates", George R. Terrell & David W. Scott, J. American Statistical Association, Vol. 80, No. 389, March 1985, p. 209-214

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 than the precision here, it will be rounded when stored.

Distribution Type: The type of analytical distribution can be either continuous or discrete. In general, all distributions will be treated as either type by default. However, the analysis may be forced to either continuous distributions or discrete distributions by checking the appropriate box in the Input Options dialog.

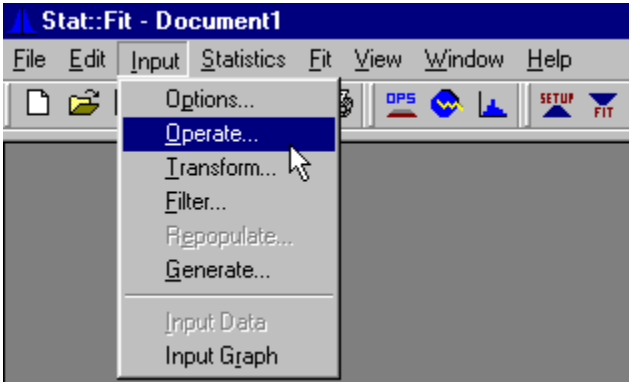
In particular, discrete distributions are forced to be distributions with integer values only. If the input data is discrete, but the data points are multiples of continuous values, divide the data by the smallest common denominator before attempting to analyze it. Input truncation to eliminate small 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 or 1. If the data has negative values, an offset should be added to it before analysis.

Operate

Mathematical operations on the input data are chosen from the Operate dialog by selecting Input from the Menu bar and then Operate from the Submenu.



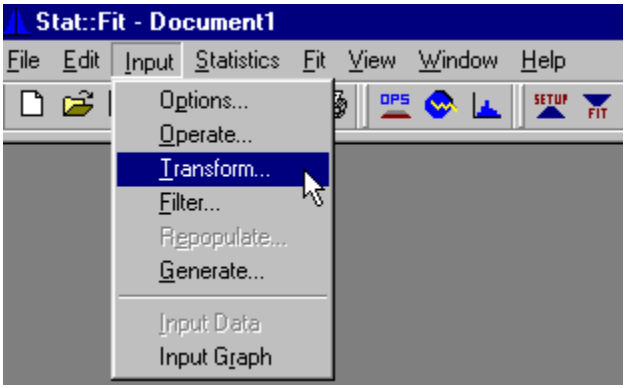
The Operate dialog allows the choice of a single standard mathematical operation on the input data. The operation will affect all input data regardless of whether a subset of input data is selected. Mathematical overflow, underflow or other error will cause an error message and all the input data will be restored.

The operations of addition, subtraction, multiplication, division, rounding, floor and absolute value can be performed. The operation of rounding will round the input data points to their

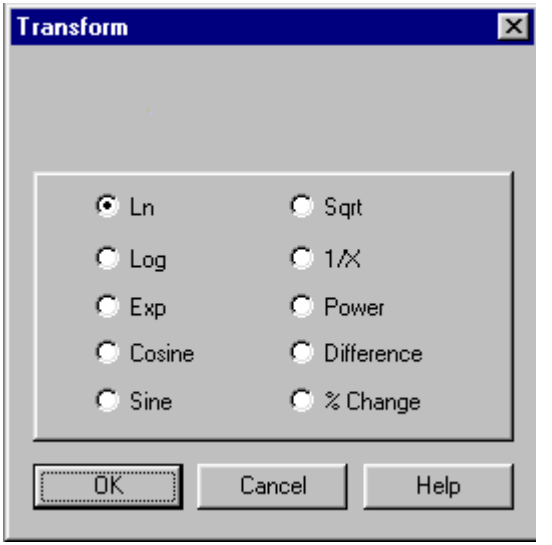
nearest integer. The data can also be sorted into ascending or descending order, or unsorted with randomly mix.

Transform

Data transformations of the input data are chosen from the Transform dialog by selecting Input from the Menu bar and then Transform from the Submenu.



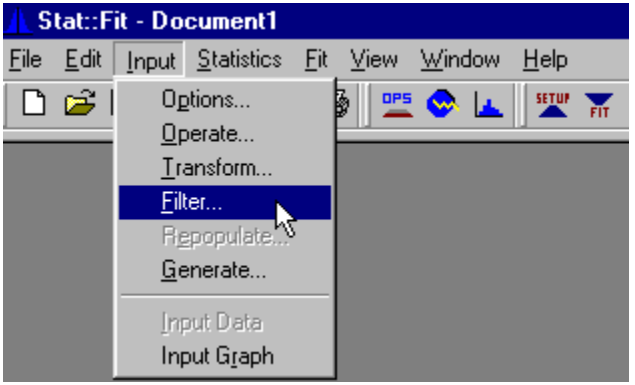
The Transform dialog allows the choice of a single standard mathematics function to be used on the input data. The operation will affect all input data regardless of whether a subset of input data is selected. Mathematical overflow, underflow or other error will cause an error message and all the input data will be restored.



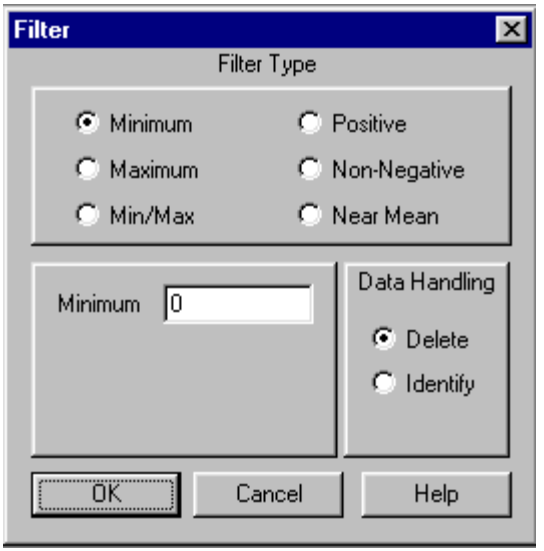
The transform functions available are: natural logarithm, log to base 10, exponential, cosine, sine, square root, reciprocal, raise to any power, difference and % change. *Difference* takes the difference between adjacent data points with the lower data point first. The total number of resulting data points is reduced by one. *% change* calculates the percent change of adjacent data points by dividing the difference, lower point first, by the upper data point and then multiplying by 100. The total number of data points is reduced by one.

Filter

Filtering of the input data can be chosen from the Filter dialog by selecting Input from the Menu bar and then filter from the Submenu.



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.* Alternatively, data to be filtered can be selected in the data set by choosing identify in the Data Handling box. 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.

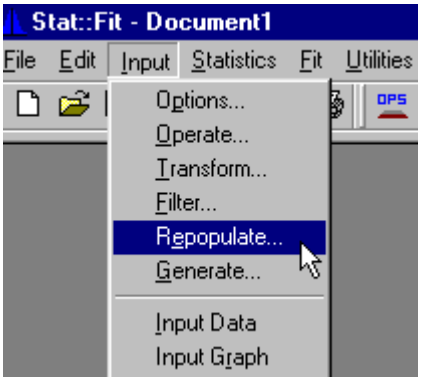


The filters include a minimum cutoff, a maximum cutoff, both minimum and maximum cutoffs, keeping only positive numbers (a negative and zero cutoff), a non-negative cutoff, and a near mean cutoff. The near mean filters all data points, excluding all data points less than the mean minus the standard deviation times the indicated multiplier or greater than the mean plus the standard deviation times the indicated multiplier

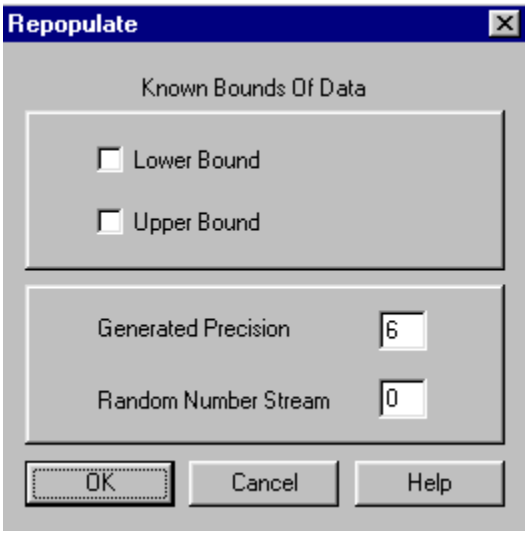
Repopulate

The Reopulation 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 in range from –1000 to +1000.

To use the repopulation function, select Input from the Menu bar and Repopulate from the Submenu.



The following dialog will be displayed.

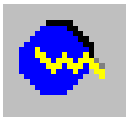


The new data points will have a number of decimal places specified by the generated precision. The goodness of fit tests, the Maximum Likelihood Estimates and the Moment Estimates require at least three digits to give reasonable results. The sequence of numbers is repeatable if the same random number 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.

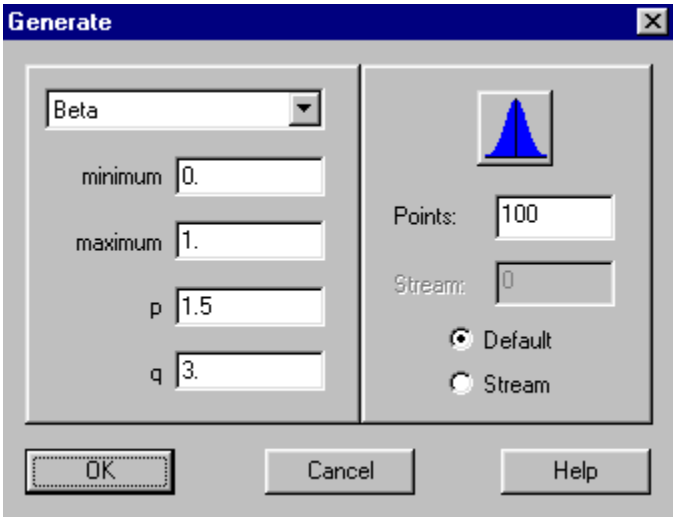
> *Important:* This repopulation of the decimal part of the data is not the same as the original data was or would have been, but only represents the information not destroyed by rounding. The parameter estimates are not as accurate as would be obtained with unrounded original data. In order to get an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

Generate

Random variates can be generated from the Generate dialog by selecting Input from the Menu bar and then Generate from the Submenu,



or by clicking on the Generate icon.



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 8000 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, 0-99.

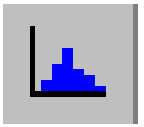
The generator will not change existing data in the Data Table, but will append the generated data points up to the limit of 8000

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.

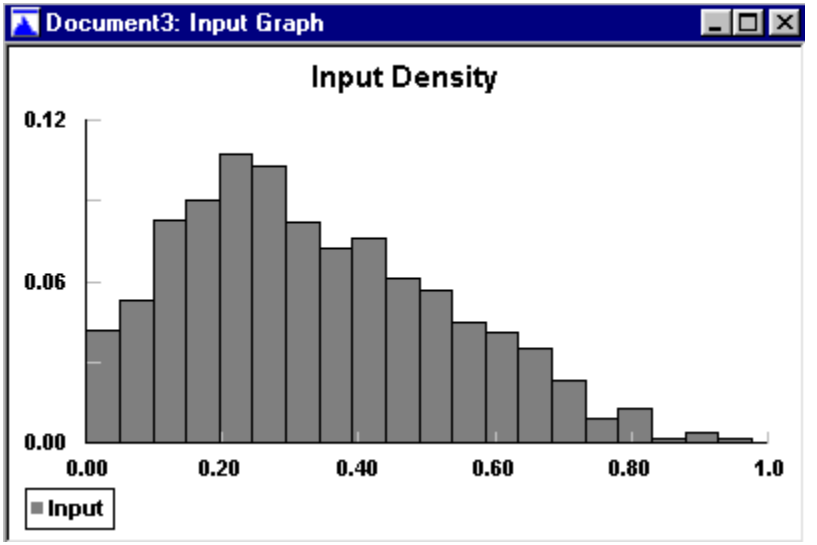
Input Graph

A graph of the input data can be viewed by selecting Input from the Menu bar and then Input Graph from the Submenu,



or by clicking on the Input Graph icon.

A histogram of your data will be displayed. An example is shown below for Gamma [0,2,1] data from the generator.



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 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 (see Chapter 5).

This graph, as with all graphs in **Stat::Fit**, may be modified, saved, copied, or printed with options generally given in the Graph Style, Save As, and Copy commands in the Graphics menu. See Chapter 6 for information on Graph Styles.

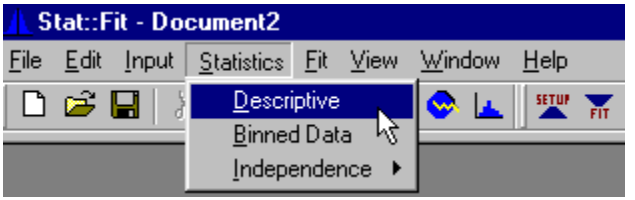
Input Data

If the Data Table has been closed, then it can be redisplayed by selecting Input from the Menu bar and Input Data from the Submenu.

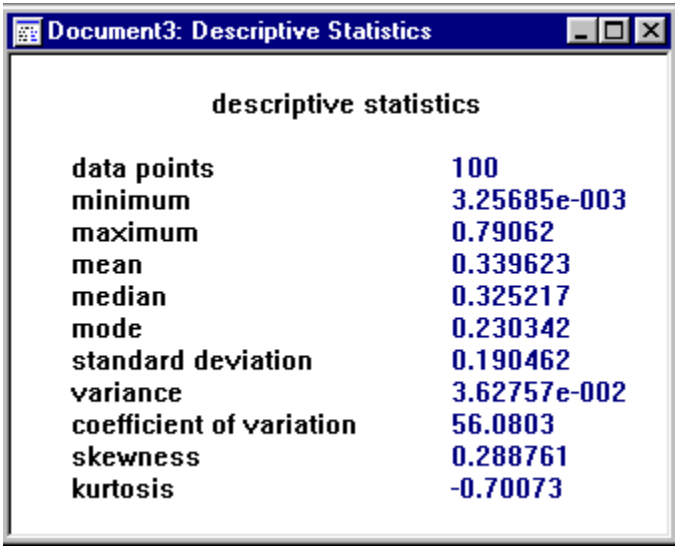
Chapter 5 – Statistical Analysis

This chapter describes the descriptive statistics, the statistical calculations on the input data, the distribution fitting process, and the goodness of fit tests. This manual is not meant as a textbook on statistical analysis. For more information on the distributions, see Appendix A. For further understanding, see the books referenced in Appendix B.

Descriptive Statistics



The descriptive statistics for the input data can be viewed by selecting Statistics on the Menu bar and then Descriptive from the Submenu. The following window will appear:



The Descriptive Statistics command provides the basic statistical observations and calculations on the input data, and presents these in a simple view as shown above. Please note that as long as this window is open, the calculations will be updated when the input data is changed. In general, all open windows will be updated when the information upon which they depend changes. Therefore, it is a good idea, on slower machines, to close such calculation windows before changing the data.

Binned Data

The histogram / class data is available by selecting Statistics on the Menu bar and then Binned Data from the Submenu. 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. A typical output is shown below:

binned data

data points	100
precision	6

continuous relative frequency

intervals	6

end points	mid points	density	ascending cumulative	descending cumulative
3.25685e-003				1.
	6.88705e-002	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	4.e-002
	0.725007	4.e-002		
0.79062			1.	

continuous frequency

intervals	6

end points	mid points	density	ascending cumulative	descending cumulative
3.25685e-003				100.
	6.88705e-002	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.	

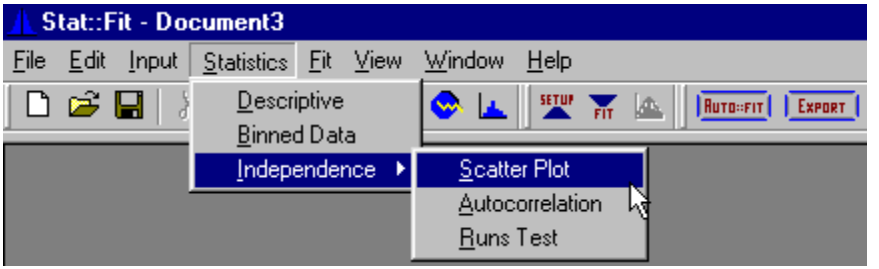
For convenience, frequency and relative frequency are given. If the data is sensed to be discrete (all integer), then the classes for the discrete representation are also given, at least up to 1000 classes. The availability of interval or class data can also be affected by forcing the distribution type to be either continuous or discrete.

Because the table can be large, it is viewed best expanded to full screen by selecting the up arrow box in the upper right corner of the screen. A scroll bar allows you to view the rest of the table. This grouping of the input data is used to produce representative graphs. For continuous data, the ascending and descending cumulative distributions match the appropriate endpoints. The density matches the appropriate midpoints. For discrete distributions, the data is grouped according to individual classes, with increments of one on the x-axis.

Independence Tests

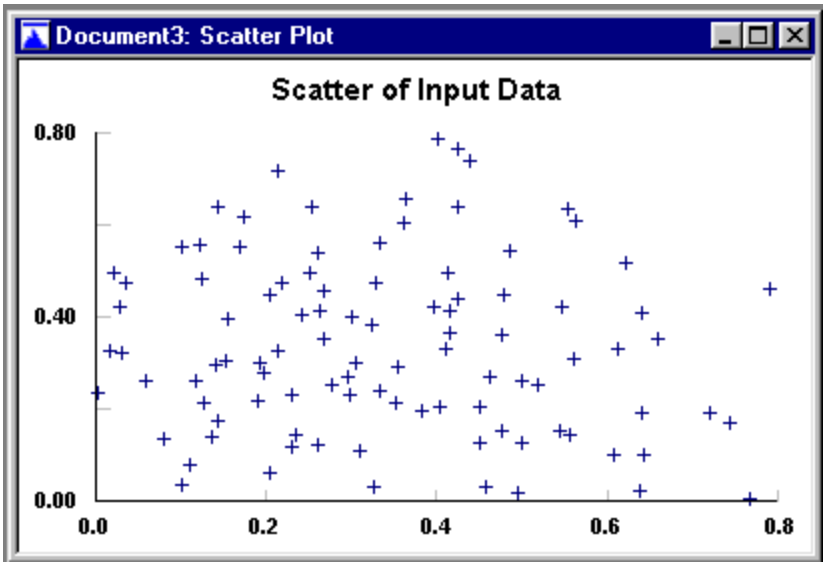
All of the fitting routines assume that your data are independent, identically distributed (IID), that is, each point is independent of all the other data points and all data points are drawn from identical distributions. **Stat::Fit** provides three types of tests for independence.

The Independence Tests are chosen by selecting Statistics on the Menu bar and then Independence from the Submenu. The following Submenu will be shown:



Scatter Plot:

This is a plot of 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 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. An example is shown below:



The structure of dependent data can be visualized graphically by starting with randomly generated data, choosing this plot, and then putting the data in ascending order with the Input /Operate commands. The position of each point is now dependent on the previous points and this plot would be close to a straight line.

Autocorrelation:

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 & Kelton[1]). 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[2]).

A graphical view of the autocorrelation can be displayed by plotting the scatter of related data points. The Scatter Plot, as previously described, 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:

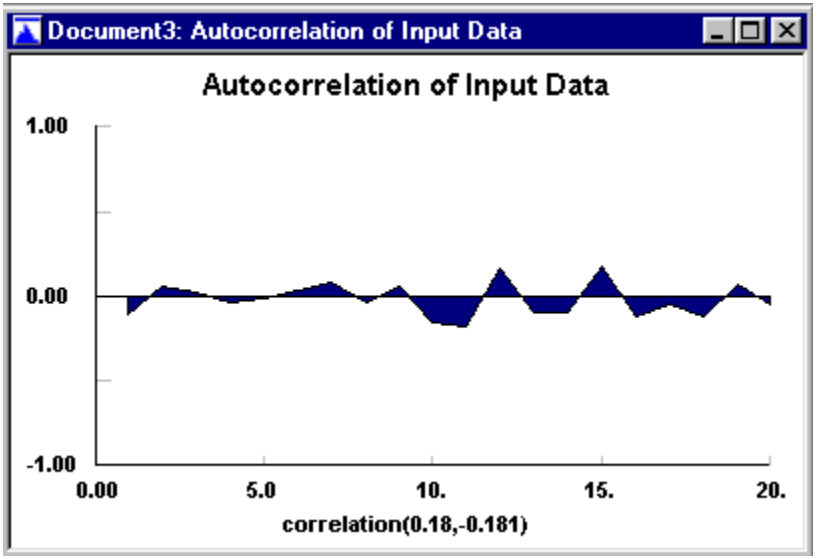
$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_{i+j} - \bar{x})}{\sigma^2 (n - j)}$$

where *j* is the lag between data points, σ 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

[1] "Simulation Modeling & Analysis" , Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p. 293
[2] "Time Series Analysis", George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, 1994, Prentice-Hall

to 1/5 of the length of the data set where diminishing pairs start to make the calculation unreliable.

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.



The numbers in parentheses along the x-axis are the maximum positive and negative correlations.

For large data sets, this plot can take a while to get to the screen. The overall screen redrawing can be improved by viewing this plot and closing it thereafter. The calculation is saved internally and need not be recalculated unless the input data changes.

Runs Tests:

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.

A run in a series of observations is the occurrence of an uninterrupted sequence of numbers with the same attribute. For instance, a consecutive set of increasing or decreasing numbers is said to provide runs 'up' or 'down' respectively. In particular, a single isolated occurrence is regarded as a run of one.

The number of runs in a series of observations indicates the randomness of those observations. Too few runs indicate strong correlation, point to point. Too many runs indicate cyclic behavior.

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 Brunk[3]). 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 above/below median runs test will not work if there are too few data points or for certain discrete distributions.

The second runs test is a turning point test which measures the number of times the series changes direction (see Johnson[4]). 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

[3] "An Introduction to Mathematical Statistics", H. D. Brunk, 1960, Ginn

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

total number of data points be greater than 12. An error message will be printed if this condition is not met.

While there are other runs tests for randomness, some of the most sensitive require larger data sets, in excess of 4000 numbers (see Knuth[5]).

Examples of the Runs Tests are shown below in the table. The length of the runs and their distribution is given.

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	5.e-002
runs statistic(2.5e-002)	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	5.e-002
turnings statistic(2.5e-002)	1.95996
p-value	0.264009
result	DO NOT REJECT

[5] "Seminumerical Algorithms", Donald E. Knuth, 1981, Addison-Wesley

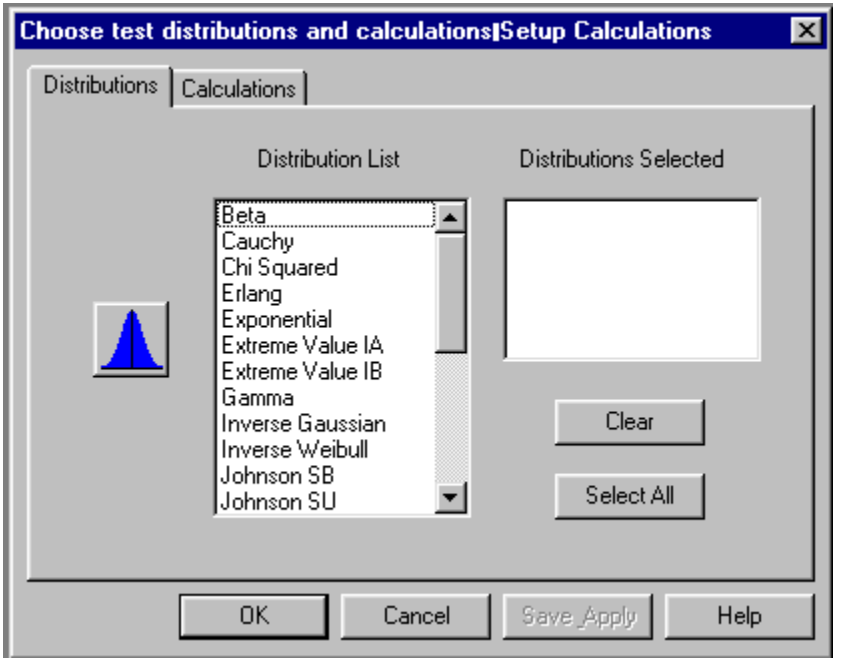
Distribution Fit

Automatic fitting of continuous and discrete distributions can be performed by using the **Auto::Fit** command. This command follows the same procedure as discussed below for 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 more details, see the section on **Auto::Fit** at the end of this chapter.

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 *moment* equations or the *maximum likelihood* equation. Finally the *goodness of fit tests* are calculated for each fitted distribution in order to ascertain the relative goodness of fit (see Breiman[6], Law & Kelton[7], Banks & Carson[8], Stuart & Ord[9]).

Begin the distribution fitting process by selecting Fit on the Menu bar and then Setup from the Submenu or by clicking on the Setup icon.

[6] "Statistics: With a View Toward Applications", Leo Breiman, 1973, Houghton Mifflin
[7] "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill
[8] "Discrete-Event System Simulation", Jerry Banks, John S. Carson II, 1984, Prentice-Hall
[9] "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart, J. Keith Ord, 1991, Oxford University Press



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 threre. 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 *distributions selected* box may be cleared by clicking the Clear button.

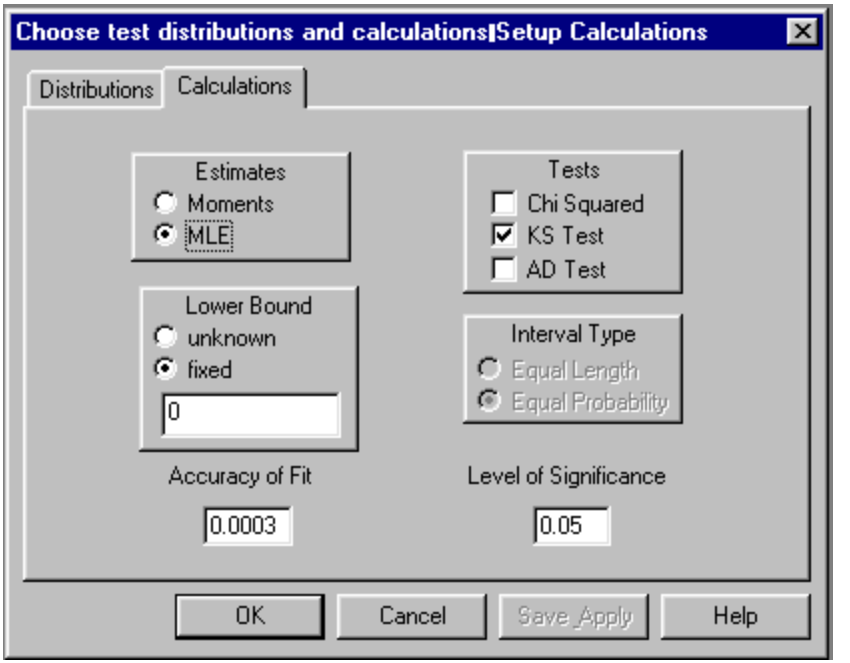
If the choice of distributions is uncertain or the data minimal, use the guides in the following Help directory guides:

Guide to Distribution Choices

Guide to No Data Representations

These guides should give some ideas on appropriate models for the input data. Also, each distribution is described separately in Appendix A and the Help files, along with examples.

After selecting the distribution(s), go to the next window of the dialog box to select the calculations to be performed.



Estimates can be obtained from either Moments or Maximum Likelihood Estimates (MLEs). The default setting for the calculation is MLE.

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 left unknown which causes all estimating procedures to vary the lower bound with the other parameters. If

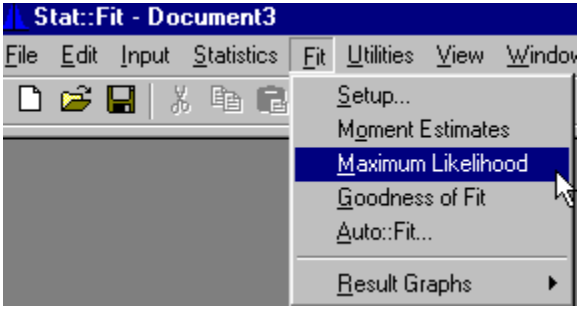
new data is added below a preset lower bound, the bound will be modified to assume the closest integer value below all input data.

The **Accuracy of Fit** describes the level of precision in iterative estimations. The default is 0.0003, but can be changed if greater accuracy is desired. Note that greater accuracy can mean much greater calculation time. Some distributions have either moments estimates and/or maximum likelihood estimates which do not require iterative estimation; in these cases, the accuracy will not make any difference in the estimation.

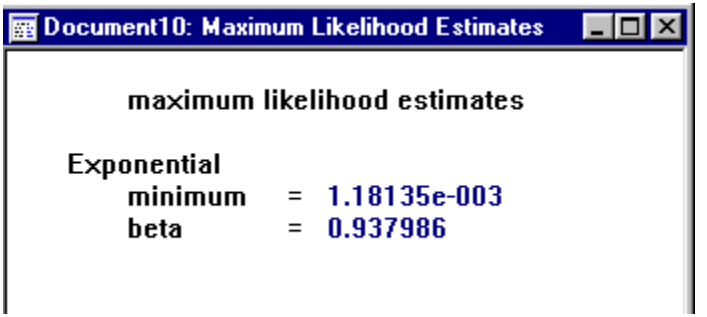
The **Level of Significance** refers to the level of significance of the test. The Chi-Squared, Kolmogorov-Smirnov and Anderson-Darling tests all ask to reject the fit to a given level of significance. The default setting is 5%, however this can be changed to 1% or 10% or any value you desire. This number is the likelihood that if the distribution is rejected, that it was the right distribution anyway. Stated in a different manner, it is the probability that you will make a mistake and reject when you should not. Therefore, the smaller this number, the less likely you are to reject when you should accept.

The **Goodness of Fit** tests described later in this chapter, may be chosen. Kolmogorov-Smirnow is the default test.

The maximum likelihood estimates and the moment estimates can be viewed independent of the goodness of fit tests. The MLE command is chosen by selecting Fit from the Menu and then Maximum Likelihood from the Submenu.

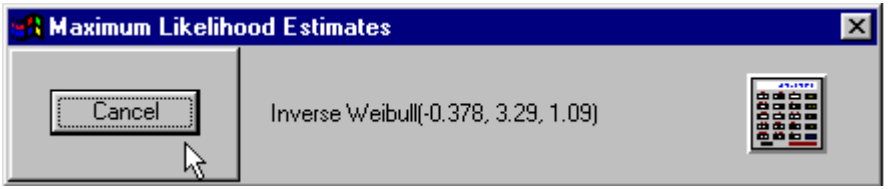


The maximum likelihood estimates of the parameters for all analytical distributions chosen in the fit setup dialog are calculated using the log likelihood equation and its derivatives for each choice. The parameters thus estimated are displayed in a new view as shown below:



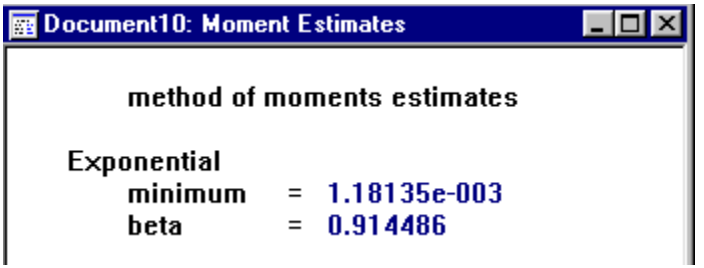
Some distributions do not have maximum likelihood estimates for given ranges of sample moments because initial estimates of the distribution's parameters are unreliable. This is especially evident for many of the bounded continuous distributions when the sample skewness is negative. When such situations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

Many of the MLEs require significant calculation, and therefore, significant time. Because of this, a Cancel dialog, shown below, will appear with each calculation.



Beside a Cancel button, it will display the values of the parameters in the current maximum likelihood calculation. If the Cancel button is clicked, the calculations will cease at the next iteration and an error message will be displayed in the Maximum Likelihood view next to the appropriate distribution.

The other choice for estimates is Moments. When the Moment Estimates command is chosen, the estimates of the parameters for all chosen analytical distributions 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 as shown below:



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.

Note that all chosen estimates (MLEs or Moments) must be finished before the Result Graphs can be displayed or the goodness of fit tests can be done. Any time the choice of estimates is changed, all

visible views of the Result Graphs and the goodness of fit tests will be redisplayed with the new calculated 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.

Goodness of Fit Tests

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 Smirnow test

Anderson Darling test

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

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, with that data appropriately separated into intervals (continuous data) or classes (discrete data). 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 oversmooth the data. Empirical rules can also be used. If none appear satisfactory, the number of intervals may be set manually. The intervals are set 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, the 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 the 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 & Kelton[10], Brunk[11], Stuart & Ord[12]).

The goodness of fit view also reports a REJECT or DO NOT REJECT decision for each Chi Squared test based on the comparison between the calculated test statistic and the *standard* statistic for the given level of significance. The level of significance can be changed in the Calculation page of the Fit Setup dialog.

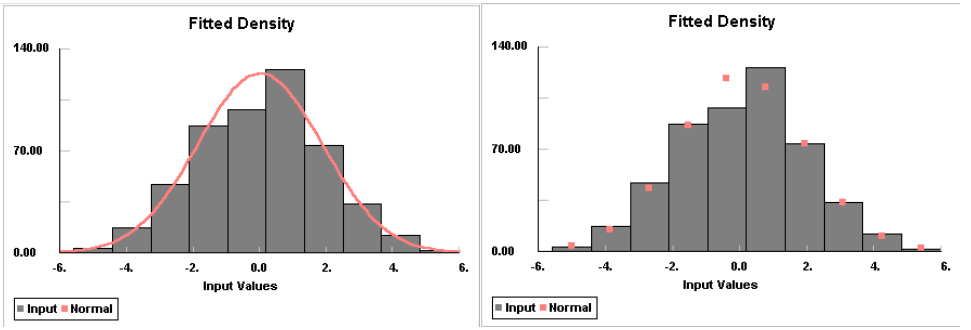
---

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

[11] "An Introduction to Mathematical Statistics", H. D. Brunk, 1960, Ginn & Co., p.261

[12] "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart & J. Keith Ord, 1991, Oxford University Press, p. 1159

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 choose 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

probability 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 probability choice is the default.

While the test statistic for the Chi Squared test 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 current 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 that the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-value is likely to be the better fit regardless of the level of significance.

Kolmogorov Smirnov Test

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

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

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

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

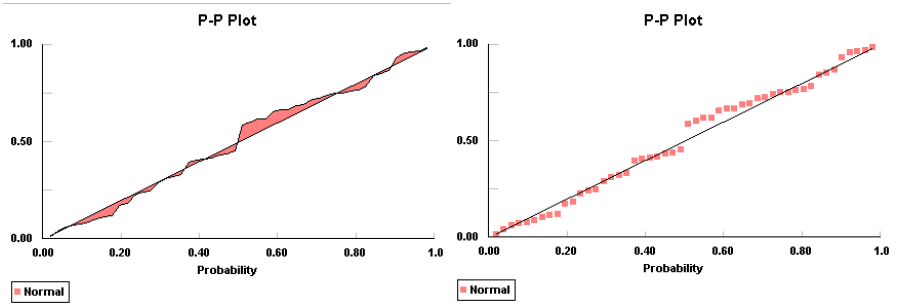
where D is the KS statistic, *x* is the value of the *ith* point out of *n* total data points, and F(x) is the fitted cumulative distribution. Note that the difference is determined separately for positive and negative discrepancies on a point by point basis.

The resulting test statistic is then compared to a standard value of the Kolmogorov Smirnov statistic with the appropriate number of data points and level of significance, usually labeled alpha. 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 because this is the most conservative test, that is, least likely to reject the fit in error. The KS test can be extended directly to some specific distributions, and these specific, more stringent, tests take the form of adjustment to the more general KS statistic. (See Law & Kelton[13], Brunk[14], Stuart & Ord[15].)

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

[13] "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p. 382
[14] "An Introduction to Mathematical Statistics", H. D. Brunk, 1960, Ginn & Co. p. 261
[15] "Kendall's Advanced Theory of Statistics, Volume 2", Alan Stuart & J. Keith Ord, 1991, Oxford University Press, p. 1159

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 a 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 as 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 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 Gleser[16]).

While the test statistic for the Kolmogorov Smirnov test 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

[16] "Exact Power of Goodness-of-Fit of Kolmogorov Type for Discontinuous Distributions", Leon Jay Gleser, J. Am. Stat. Assoc., 80 (1985) p. 954

will be as unusual as the current sample given that the fit is appropriate. A small p-value indicates that the current 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 that the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-value is likely to be the better fit regardless of the level of significance.

Anderson Darling Test

The Anderson Darling test (AD) 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 calculates 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:

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

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 μ_i is the value of the fitted cumulative distribution, $F(x_i)$, for the *ith* data point (see Law & Kelton[17], Anderson & Darling[18,19]).

[17] "Simulation Modeling & Analysis", Averill M. law, W. David Kelton, 1991, McGraw-Hill, p. 392
[18] "A Test of Goodness of Fit", T. W. Anderson, D. A. Darling, J.Am.Stat.Assoc., 1954, p. 765
[19] "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes", T. W. Anderson, D. A. Darling, Ann.Math.Stat., 1952, p. 193

The resulting test statistic is then compared to a *standard* value of the 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 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 a multiplicative adjustment to the general AD statistic.

The goodness of fit view also reports a REJECT or DO NOT REJECT decision for each AD test based on 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.

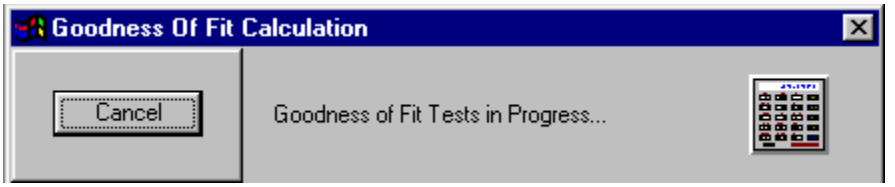
While the test statistic for the Anderson Darling test can be useful, the p-value is more useful in determining the goodness of fit. The p-value is defined as the probability that another sample will be as unusual as the current sample given that the fit is appropriate. A small p-value indicates that the current 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 that the fit is appropriate. When comparing two different fitted distributions, the distribution with the higher p-

value is likely to be the better fit regardless of the level of significance.

General

Each of these tests has its own regions of greater sensitivity, but they all have one criterion in common. The fit and the tests are totally insensitive for fewer than 10 data points (**Stat::Fit** will not respond to less data), and will not achieve much accuracy until 100 data points. On the order of 200 data points seems to be optimum. For large data sets, greater than 4000 data points, the tests can become too sensitive, occasionally rejecting a proposed distribution when it is actually a useful fit. This can be easily tested with the Generate command in the Input menu.

While the calculations are being performed, a window at the bottom of the screen shows its progress and allows for a Cancel option at any time.



The results are shown in a table. An example is given below:

goodness of fit

data points	1000
estimates	maximum likelihood estimates
accuracy of fit	3.e-004
level of significance	5.e-002

summary

distribution	Chi Squared
Exponential(1.18e-003, 0.935)	5.11 (16)

detail

Exponential

minimum	=	1.18135e-003
beta	=	0.9355

Chi Squared

total classes	17
interval type	equal probable
net bins	17
chi**2	5.11
degrees of freedom	16
alpha	5.e-002
chi**2(16,5.e-002)	26.3
p-value	0.995
result	DO NOT REJECT

In the *summary* section, the distributions you have selected for fitting are shown along with the results of the Goodness of Fit Test(s). The numbers in parentheses after the type of distribution are the parameters and they are shown explicitly in the detailed information, below the summary table. The above table shows results for the Chi Squared Test. The number in parenthesis is the

degrees of freedom. When you want to compare Chi Squared from different distributions, you can only make a comparison when they have the same degrees of freedom.

The detailed information, following the summary table, includes a section for each fitted distribution. This section includes:

Parameter values

Chi Squared Test

Kolmogorov Smirnov Test

Anderson Darling Test

If an error occurred in the calculations, the error message is displayed instead.

For the Chi Squared Test the details show:

total classes [intervals]

interval type [equal length, equal probable]

net bins [reduced intervals]

chi2 [the calculated statistic]**

degrees of freedom [net bins-1 here]

alpha [level of significance]

chi2(n, alpha) [the standard statistic]**

p-value

result

For both the Kolmogorov Smirnov and Anderson Darling tests, the details show:

data points

stat [the calculated statistic]

alpha [level of significance]

stat(n, alpha) [the standard statistic]

p-value

result

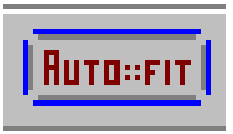
An example of the Kolmogorov Smirnov Test is shown below:

Kolmogorov-Smirnov	
data points	1000
ks stat	1.73e-002
alpha	5.e-002
ks stat(1000,5.e-002)	4.28e-002
p-value	0.922
result	DO NOT REJECT

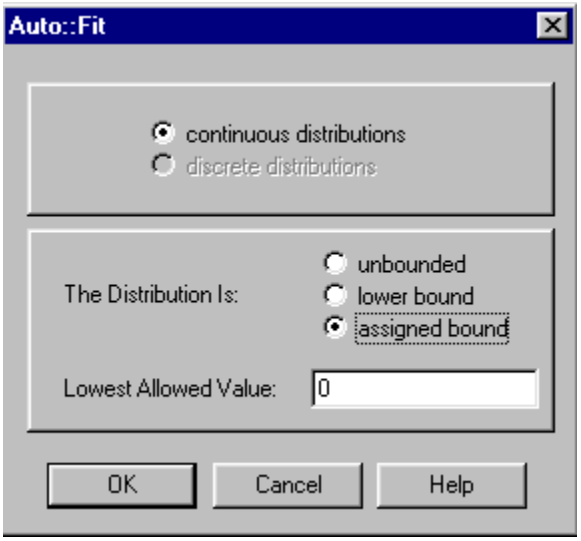
An example of the Anderson Darling Test is shown below:

Anderson-Darling	
data points	999
ad stat	2.36
alpha	5.e-002
ad stat(5.e-002)	2.49
p-value	5.9e-002
result	DO NOT REJECT

Distribution Fit – Auto::Fit



Automatic fitting of continuous or discrete distributions can be performed by clicking on the **Auto::Fit** icon or by selecting Fit from the Menu bar and then **Auto::Fit** from the Submenu.



This command follows the same procedure as previously discussed for manual fitting. **Auto::Fit** 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.

An example is shown below:

Auto::Fit of Distributions

distribution	rank	acceptance
Inverse Gaussian(-2.48304, 36.8058, 12.4311)	98.9	do not reject
Lognormal(-2.00183, 2.32172, 0.572285)	96.4	do not reject
Pearson 5(-5.40605, 6.02755, 77.6132)	78.7	do not reject
Gamma(0.270107, 1.88582, 5.13074)	66.4	do not reject
Inverse Weibull(-21.125, 6.05196, 3.62105e-002)	61.7	do not reject
Erlang(0.270107, 2., 4.83921)	47.2	do not reject
LogLogistic(-1.08811, 2.69717, 9.25314)	45.7	do not reject
Pearson 6(0.44382, 98.9852, 1.80798, 19.7105)	36.3	do not reject
Weibull(0.410585, 1.41541, 10.4908)	21.	do not reject
Beta(0.44382, 59.3738, 1.52761, 7.89056)	8.75	do not reject
Chi Squared(-7.5199, 17.2907)	1.95e-003	reject
Rayleigh(-1.37887, 9.37822)	4.02e-007	reject
Exponential(0.44382, 9.50422)	0.	reject
Pareto(0.44382, 0.351087)	0.	reject
Power Function(0.443199, 38.2082, 0.588391)	0.	reject
Triangular(0.316432, 38.2564, 1.2507)	0.	reject
Uniform(0.44382, 38.1042)	0.	reject
Johnson SB	no fit	reject

For continuous distributions, the **Auto::Fit** dialog limits the number of distributions by choosing only those distributions with a lower bound or by forcing a lower bound to a specific value as in Fit Setup. Also, the number of distributions will be limited if the skewness of the input data is negative. Many continuous distributions with lower bounds do not have good parameter estimates in this situation.

For discrete distributions, the **Auto::Fit** dialog limits the distributions by choosing only those distributions that can be fit to the data. The discrete distributions must have a lower bound.

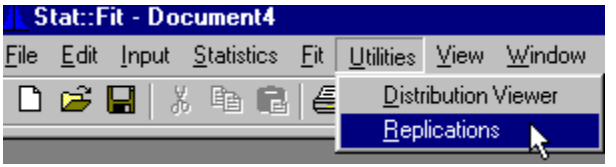
The acceptance of fit usually reflects the results of the goodness of fit tests at the level of significance chosen by the user. However, the acceptance may be modified if the fitted distribution would

generate significantly more data points in the tails of the distribution than are indicated by the input data.

Replication and Confidence Level Calculator

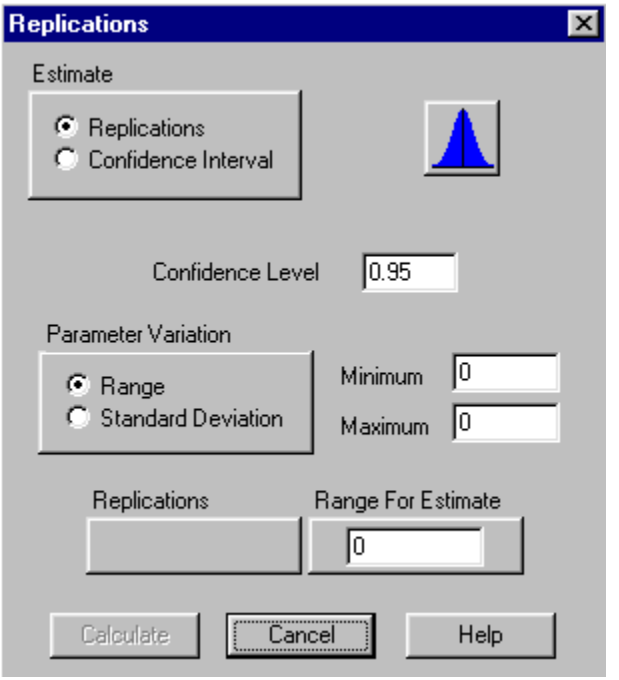
The Replications command allows the user to calculate the number of independent data points, or replications, of an experiment that are necessary to provide a given range, or confidence interval, for the estimate of a parameter. The confidence interval is given for the confidence level specified, with a default of 0.95. The resulting number of replications is calculated using the *t* distribution[20].

To use the Replications calculator, select Utilities from the Menu bar and then Replications.



The following dialog will be displayed.

[20] "Discrete-Event System Simulation, Second Edition", Jerry Banks, John S. Carson II, Barry L. Nelson, 1966, Prentice-Hall, p. 447



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.

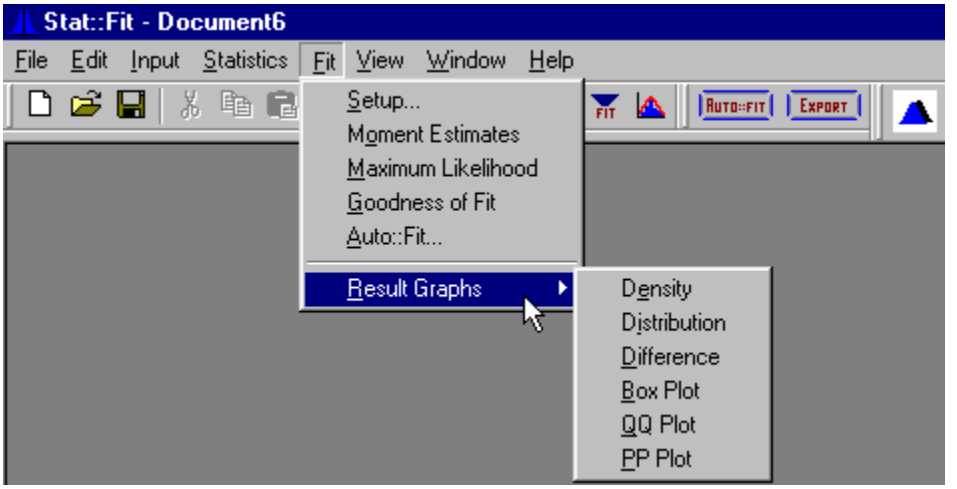
Chapter 6 – Graphs

This chapter describes the types of graphs and the Graphics Style options. Graphical analysis and output is an important part of **Stat::Fit**. The input data in the Data Table may be graphed as a histogram or line chart and analyzed by a scatter plot or autocorrelation graph. The resulting fit of a distribution may be compared to the input via a direct comparison of density or distribution, a difference plot, a box plot, a Q-Q plot, and a P-P plot for each analytical distribution chosen. The analytical distributions can be displayed for any set of parameters.

The resulting graphs can be modified in a variety of ways using the Graphics Style dialog in the Graphics menu, which becomes active when a graph is the currently active window.

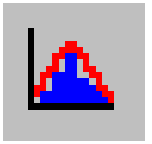
Result Graphs

A density graph of your input data and the fitted density can be viewed by choosing Fit from the Menu bar and then Result Graphs from the Submenu.



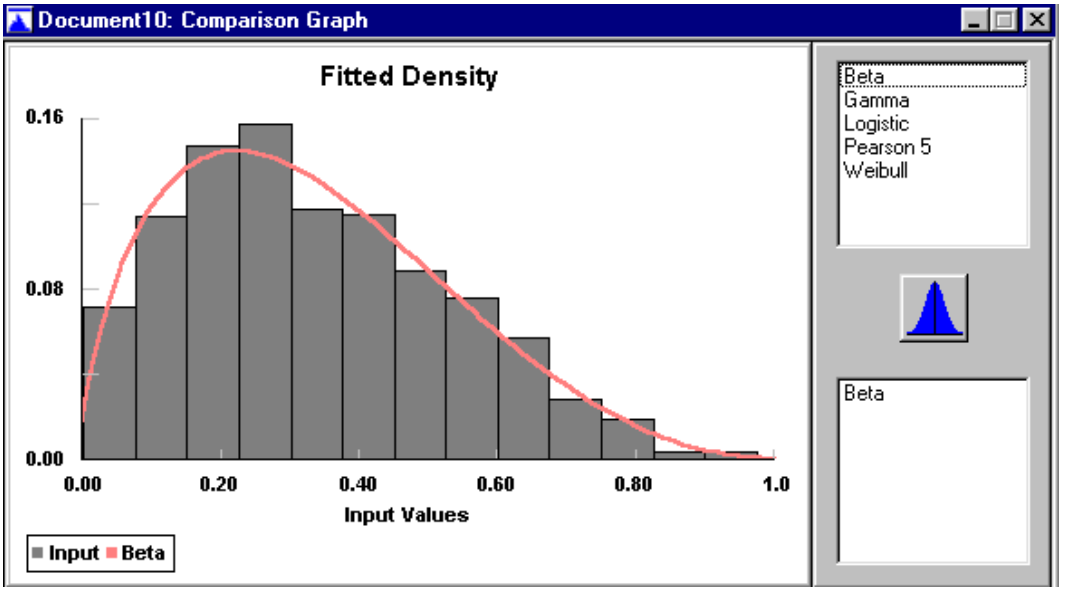
This graph displays a histogram of the input data overlaid with the fitted densities for specific distributions.

From the next menu that appears (see above), choose Density.



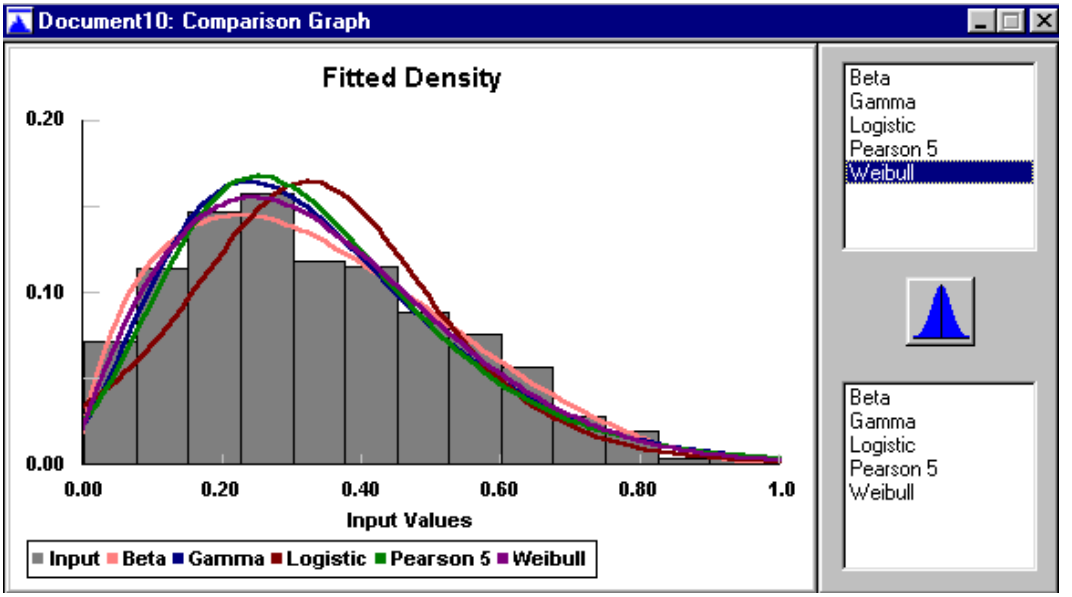
Quicker access to this graph is accomplished by clicking on the Graph Fit icon on the Control Bar.

The graph will appear with the default settings of the input data in a blue histogram and the fitted data in a red polygon, as shown below.



The distribution being fit is listed in the lower box on the right. If you have selected more than one distribution to be fit, a list of the distributions is given in the upper box on the right. Select additional distributions to be displayed, as comparisons, by clicking on the distribution name(s) in the upper box. The additional fit(s) will be added to the graph and the name of the

distribution(s) added to the box on the lower right. There will be a Legend at the bottom of the graph, as shown below:



To remove distributions from the graph, click on the distribution name in the box on the lower right side and it will be removed from the graphic display.

Stat::Fit provides many options for graphs in the Graphics Style dialog, including changes in the graph character, the graph scales, the title texts, the graph fonts and the graph colors.

This dialog can be activated by selecting Graphics from the menu bar and then Graphics Style from the Submenu.

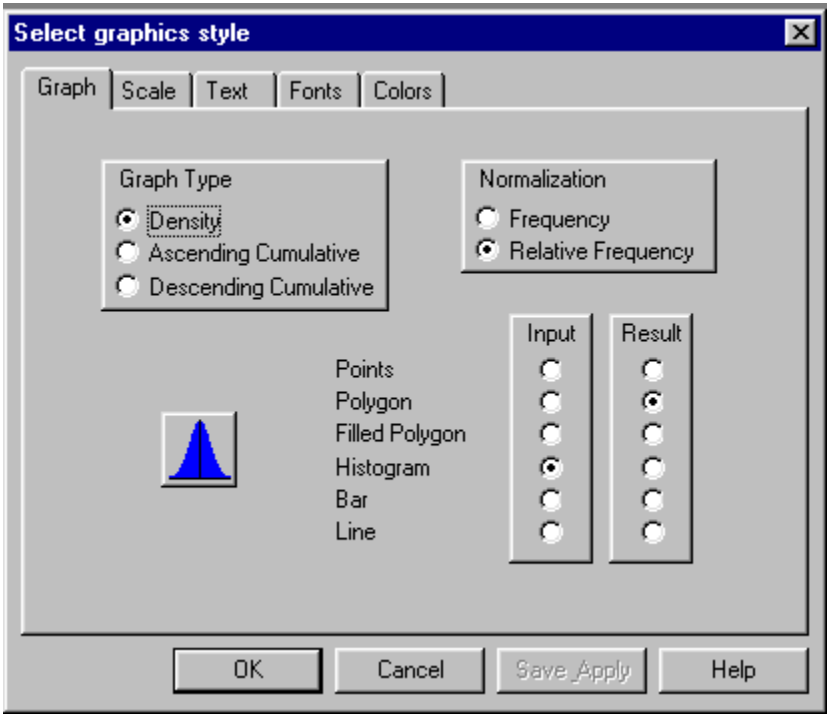
The graph remains modified as long as the document is open, even if the graph itself is closed and reopened. It will also be saved with the project as modified. Note that any changes are singular to that particular graph; they do not apply to any other graph in that document or any other document.

If a special style is always desired, the default values may be changed by changing any graph to suit, checking the Save Apply button at the bottom of the dialog. The resulting style becomes the default for all new graphs in **Stat::Fit** with the exception of some specialized titles, styles and legends (such as the autocorrelation style and x-axis title).

Graphics Style – Graph

The Graphics Style dialog box has 5 tabs (or pages). When you select a tab, the dialog box changes to display the options and default settings for that tab. You determine the settings for any tab by selecting or clearing the check boxes on the tab. The new settings take effect when you close the dialog box. If you want your new settings to be permanent, select Save Apply and they will remain in effect until you wish to change them again.

The dialog box for the graph type options is shown below:



The **Graph Type** chooses between three types of distribution functions:

Density indicates the probability density function, f(x), for continuous random variables and the probability mass function, p(j), for discrete random variables. Quite frequently, f(x) is substituted for p(j) with the understanding that x then takes on only integer values.

Ascending cumulative indicates that the cumulative distribution function, F(x), where x can be either a continuous random variable or a discrete random variable. F(x) is continuous or discrete accordingly. F(x) varies from 0 to 1.

Descending cumulative indicates the survival function (1-F(x)).

Graph Type is not available for some graphs.

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).

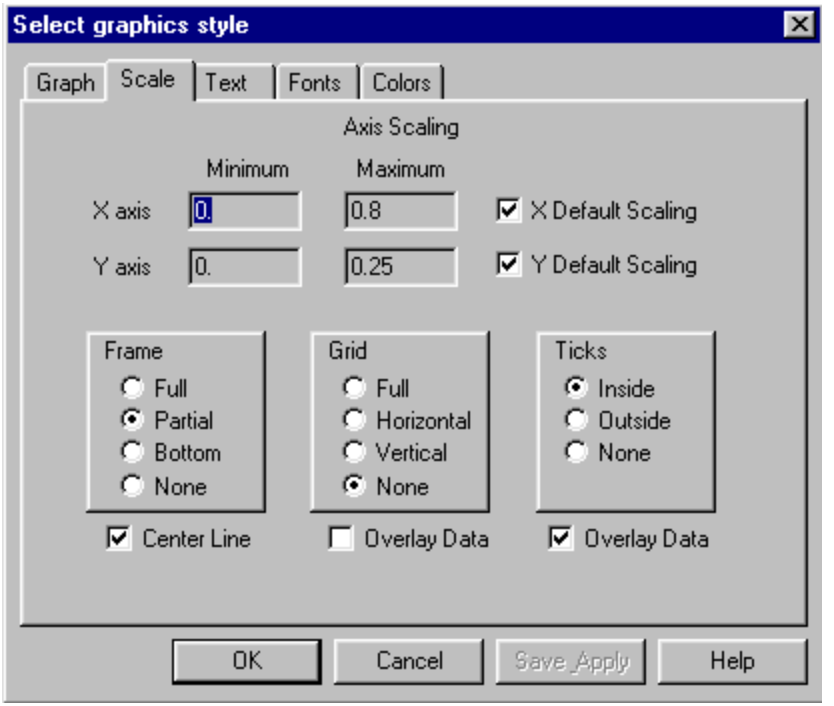
Relative Frequency represents the relative fraction of the total counts for each interval (continuous random variable) or class (discrete random variable).

Normalization is not available for some graphs.

The graph style can be modified for both the input data and the fitted distribution. Choices include points, line, bar, polygon, filled polygon and histogram. For Scatter Plots, the choices are modified and limited to: points, cross, dots.

Graphics Style – Scale

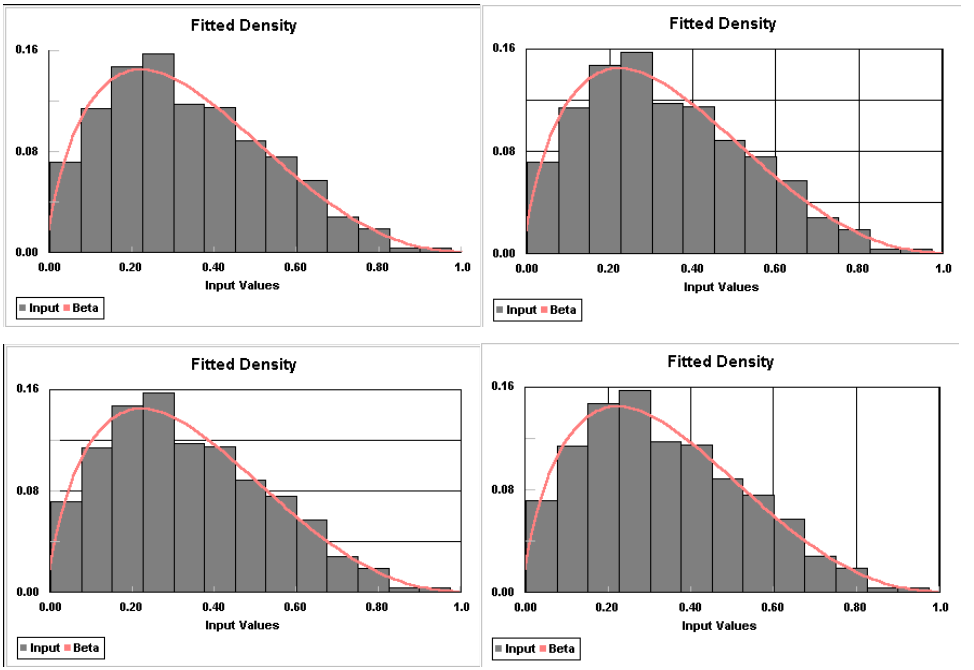
The dialog box for Scale is shown below:



The scale page 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 default settings for Scale allow the data and fitted distribution to be displayed. These settings can be changed by deselecting the default and adding Min and Max values.

Moreover, the printed graph will maintain that aspect ratio as will the bitmap that can be saved to file or copied to the Clipboard.

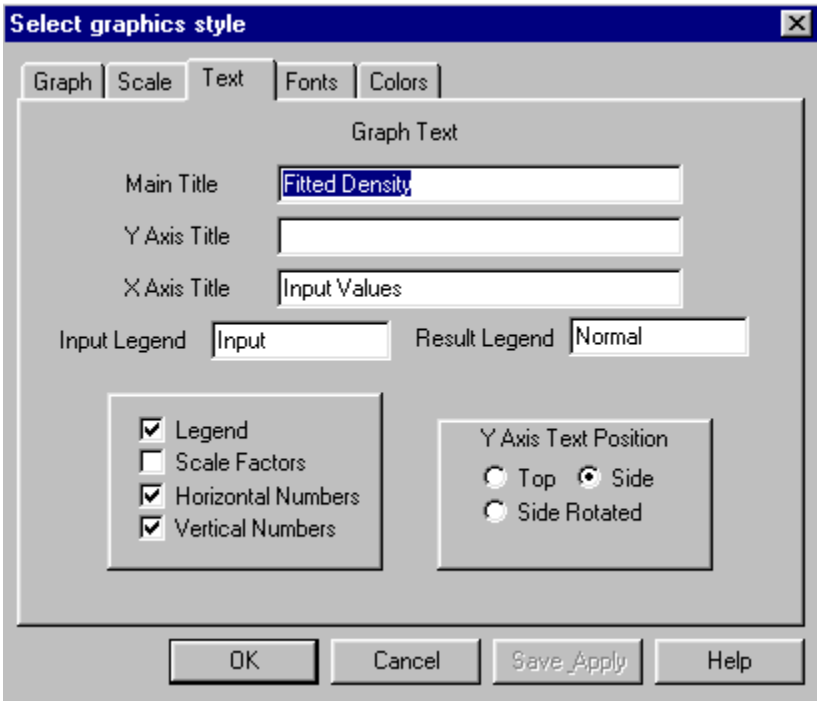
Frame allows you to have a full, partial or no frame around your graph. A grid can be added to your graph in both x and y, or just a horizontal or vertical grid can be displayed, as shown below.



Tick marks can be selected to be inside, outside, or absent. Both ticks and the grid can overlay the data.

Graphics Style – Text

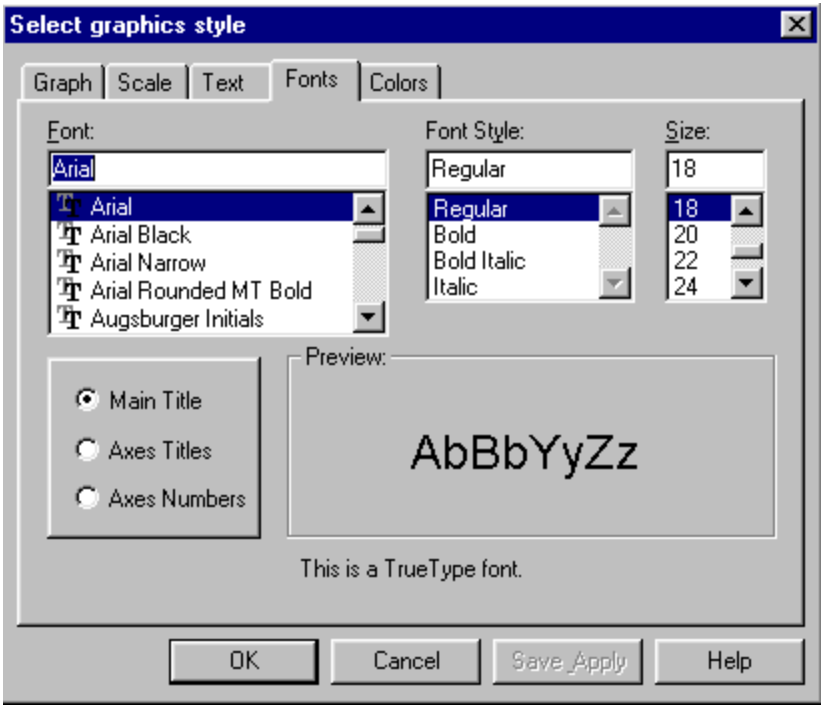
The dialog box for Text is shown below:



The Text function allows you to add text to your graph. A Main Title, x-axis and y-axis titles, and legends can be included. Scale factors can be added. The layout of the y-axis title can be modified to be at the top, on the side or rotated along the side of the y-axis. Some graphs load default titles initially.

Graphics Style – Fonts

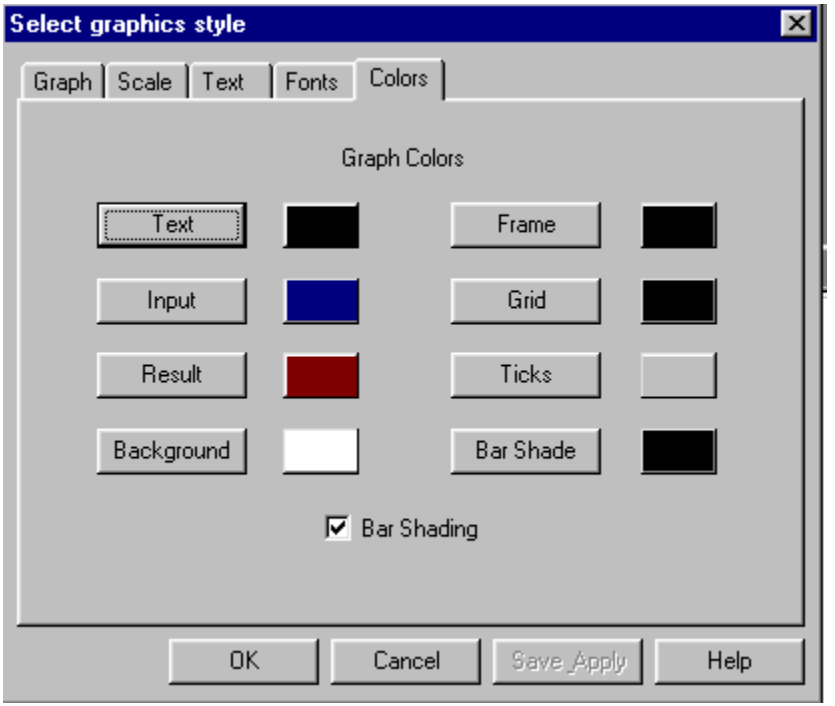
The dialog box for Fonts is shown below:



The Fonts page of the dialog provides font selection for the text titles and scales in 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 Style – Color

The dialog box for Color is shown below:



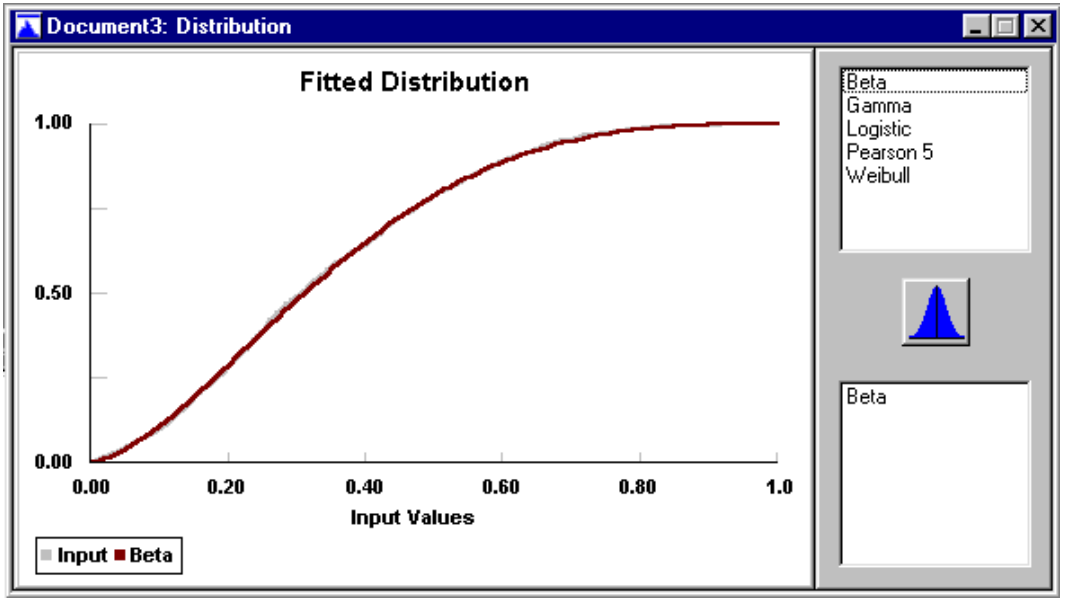
The Colors page of the dialog provides color options for all the fields of the currently active graph. For each object in the graph, a button to call the color dialog is located to the left and a color patch is located on the right. *Text* refers to all text including scales. *Input* refers to the first displayed graph, the input data in comparison graphs. *Result* refers to the first fitted data. *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 the screen. 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.

Other Graphs

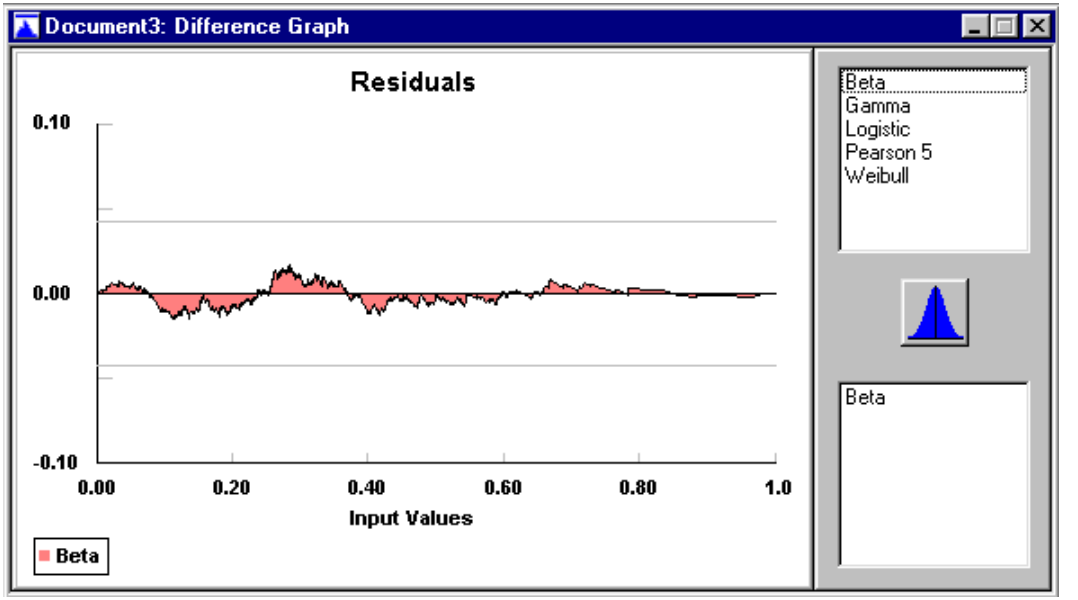
Stat::Fit provides additional Result Graphs for visualizing the fit of your data to a distribution. Above we have described the Density Graph. The other choices on the Submenu are Distribution, Difference, Box Plot, Q-Q Plot and P-P Plot. All of these graph types allow the comparison of multiple distributions. If you have selected more than one distribution to be fit, a list of the distributions is given in the upper box on the right. Select additional distributions to be displayed, as comparisons, by clicking on the distribution name(s) in the upper box. The additional fit(s) will be added to the graph and the name of the distribution(s) added to the box on the lower right. There will be a Legend at the bottom of the graph. To remove distributions from the graph, click on the distribution name in the box on the lower right side and it will be removed from the graphic display.

Distribution Graph:



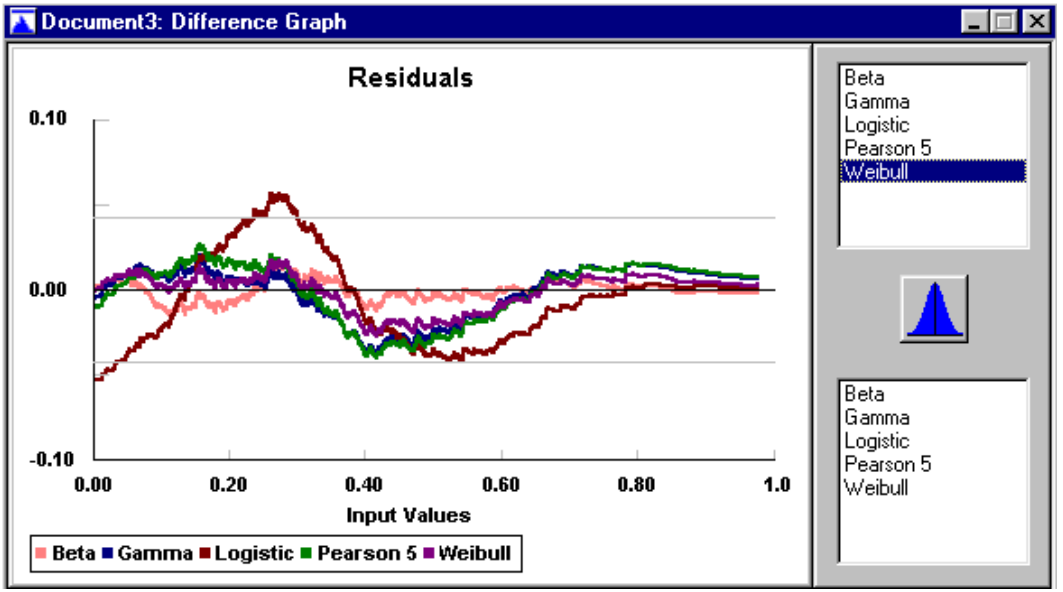
The Distribution graph displays the cumulative distribution of the input data overlaid with the fitted cumulative distributions for specific distributions.

Difference Graph:



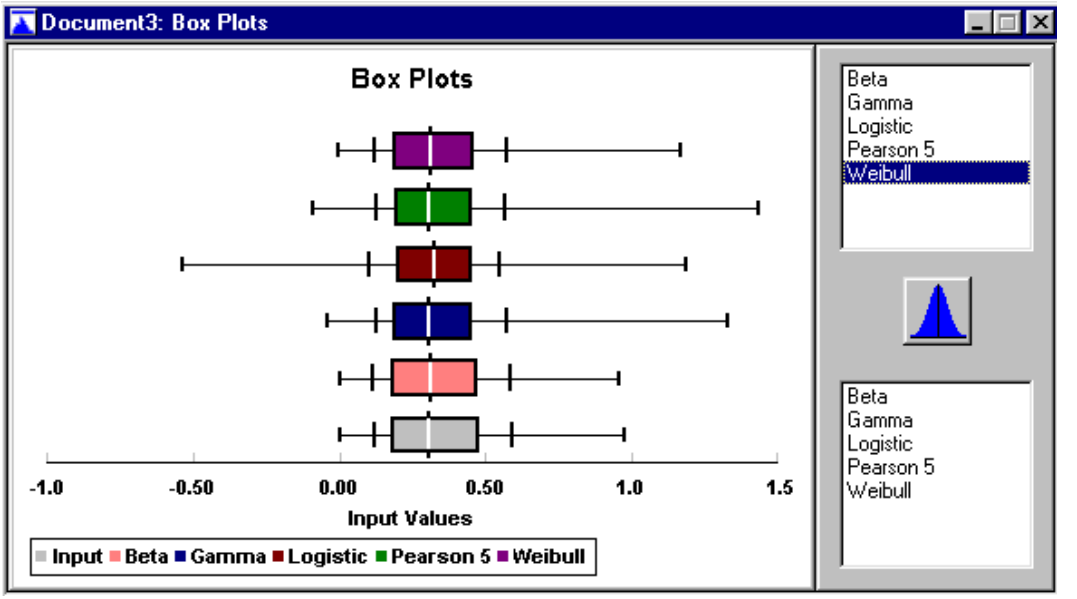
The Difference Graph is a plot of the difference between the cumulative input data in the Data Table minus the fitted cumulative distribution. Note that conservative error bars shown in the graph are not a function of the number of intervals for continuous data. Although the graph may be modified to plot the difference between the input density and the fitted density, the error bars derived from the conservative Kolmogorov Smirnov calculations, are only applicable for the cumulative distribution.

Multiple distributions can be compared with respect to their difference plots as shown below:


Document3: Difference Graph
Residuals
0.10
0.00
-0.10
0.00
0.20
0.40
0.60
0.80
1.0
Input Values
Beta
Gamma
Logistic
Pearson 5
Weibull
Beta
Gamma
Logistic
Pearson 5
Weibull
Beta
Gamma
Logistic
Pearson 5
Weibull

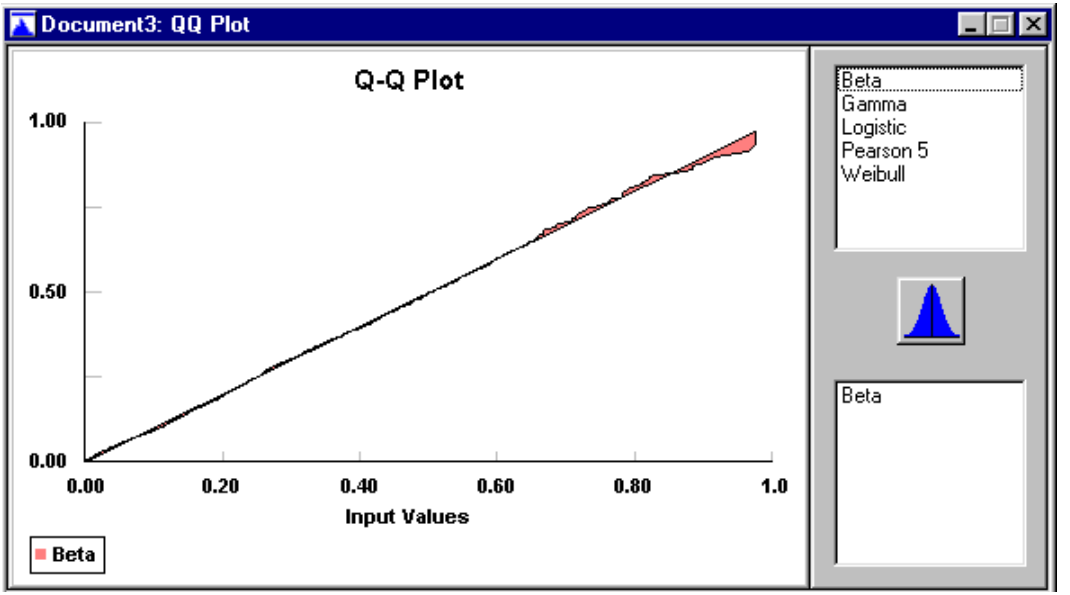
Box Plot:

Box Plots are another way to compare the data with the distributions. It is particularly good for looking at the extremes. The center line is the median. The box represents the quartiles (25% and 75%). The next set of lines are the octiles and the outer lines are the extremes of the data or distribution. A box plot gives a quick indication of potential skewness in a data set by relating the location of the box to the median. If one side of the box is further away from the median than the other side, the distribution tends to be skewed in the direction furthest from the median.



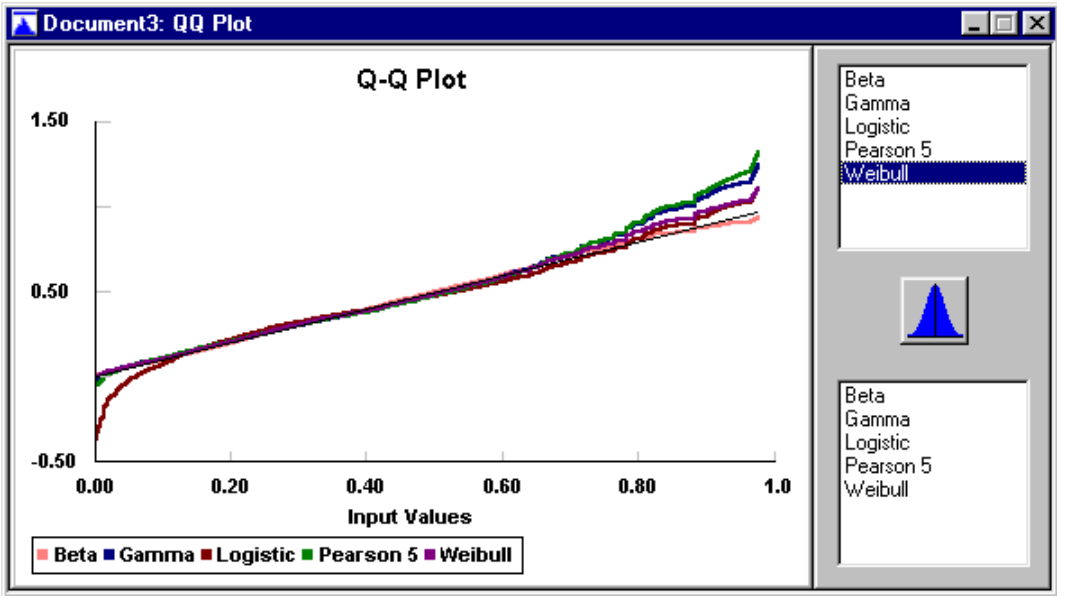
Q-Q Plot:

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]).



Multiple distributions can be added to the graph for comparison.

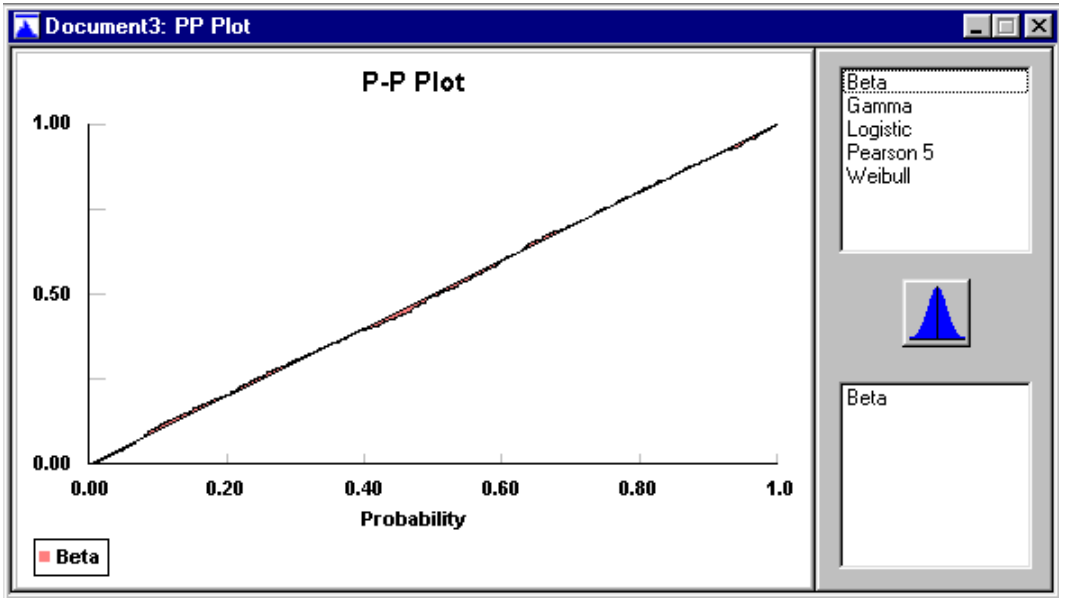
[1] "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p. 374



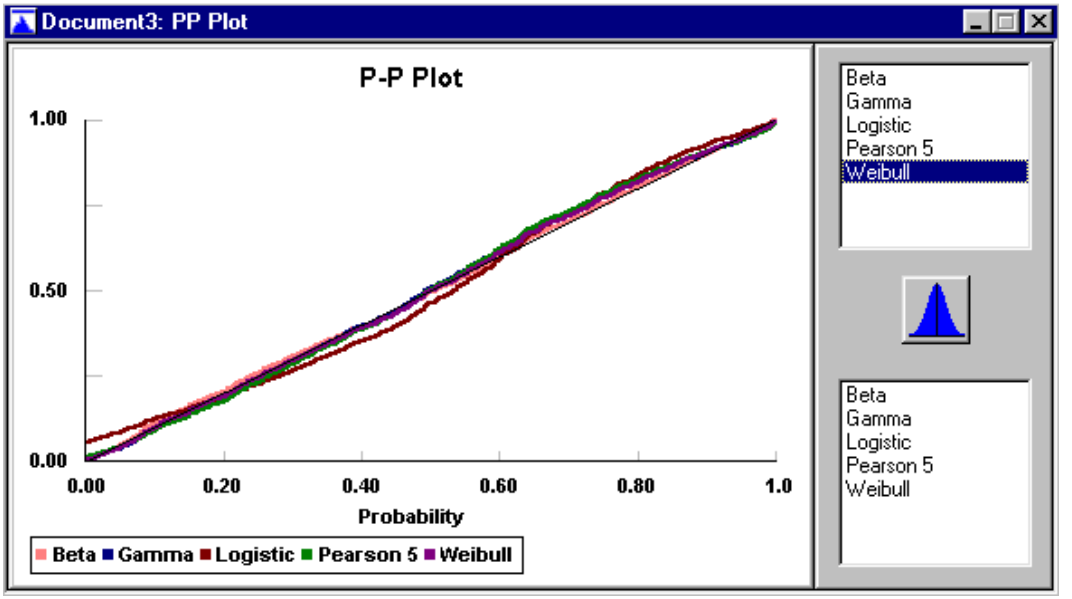
P-P Plot:

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 & Kelton[2]).

[2] ibid, p. 339



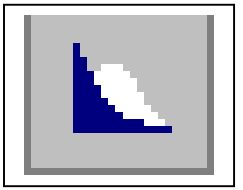
Multiple distributions can be added to the graph for comparison.



Distribution Viewer

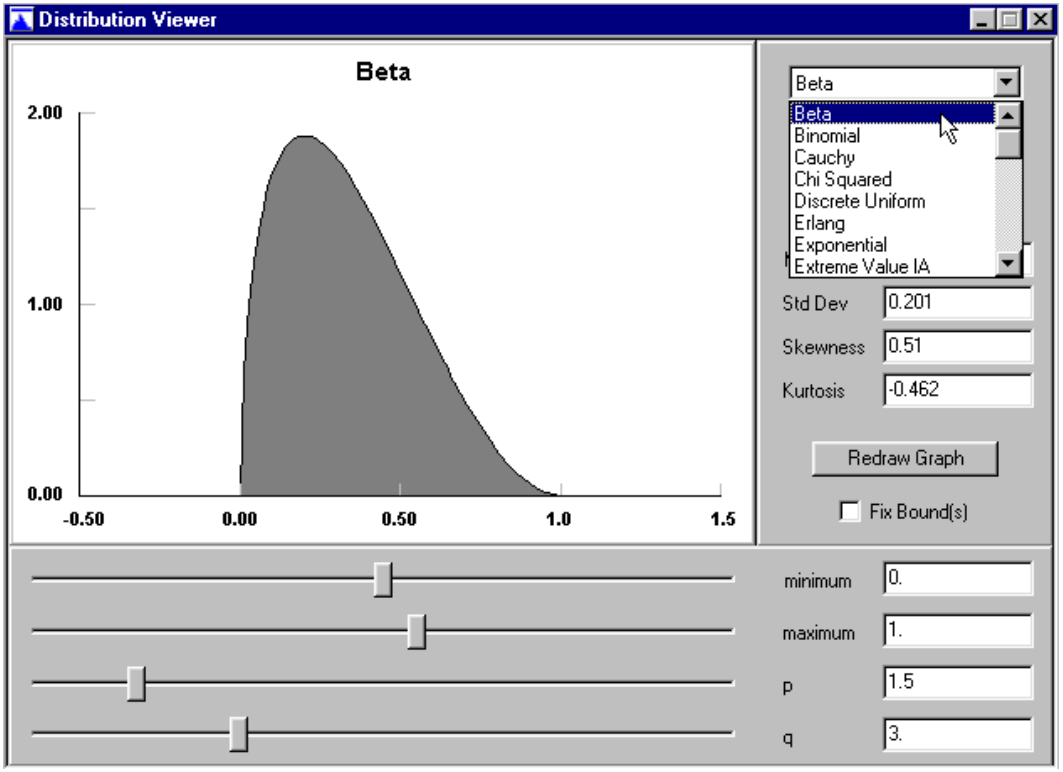
The Distribution Viewer option allows you to display the functional form of any distribution with specified parameters totally independent of data. This is just a picture of what that analytical distribution would look like with the parameters you have selected. It can be used to visualize the functional form of distributions and can be useful in selecting a distribution for fitting. The Distribution Viewer allows active viewing of a distribution while the parameters or moments are changed.

To display the Distribution Viewer, select Utilities from the Menu bar and then Distribution Viewer from the Submenu.



Quicker access to the Distribution Viewer by clicking on the Distribution Viewer icon.

A dialog box will allow you to select the type of distribution and its parameters. An example of the distribution graph viewer is shown below:



The Distribution Viewer uses the distribution and parameters provided in this dialog to create a graph of any analytical distribution supported by **Stat::Fit**. This graph is not connected to any input data or document. This graph of the distribution may then be modified using the *sliders* for each parameter, specifying the 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**, the 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 a graphic file [.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.

Copy and Save As

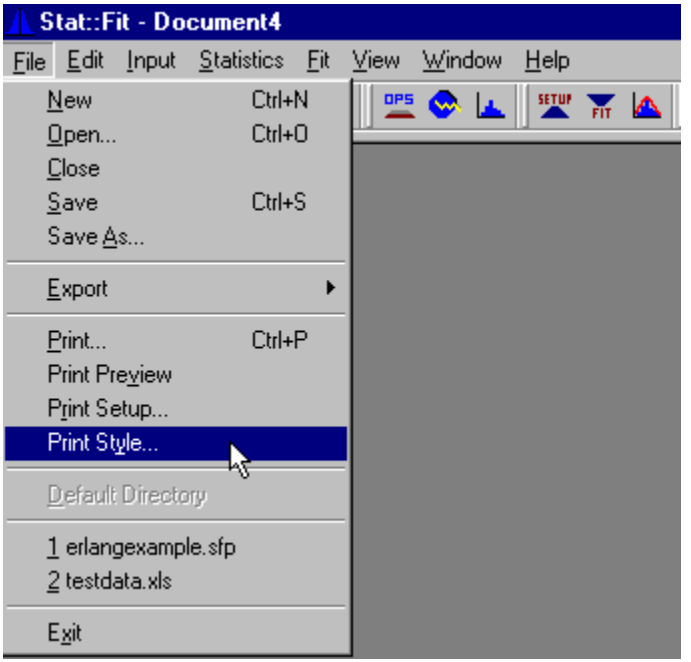
Any graph can be copied onto the Clipboard, so that you can transfer it to another program, by selecting Graphics from the Menu bar and then Copy from the Submenu. The Graphics Copy command places a copy of the current graph in the Clipboard as a bitmap.

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. It can also be loaded into **Stat::Fit** but will no longer be connected with a document. Note that the copy can no longer be modified with the Graphics Style dialog.

Chapter 7 – Print and Output Files

This chapter provides details on how to Print graphs and reports tailored to meet your needs. Information on exporting files will also be given.

The printing process is started by selecting File from the Menu bar. The following Submenu is displayed.

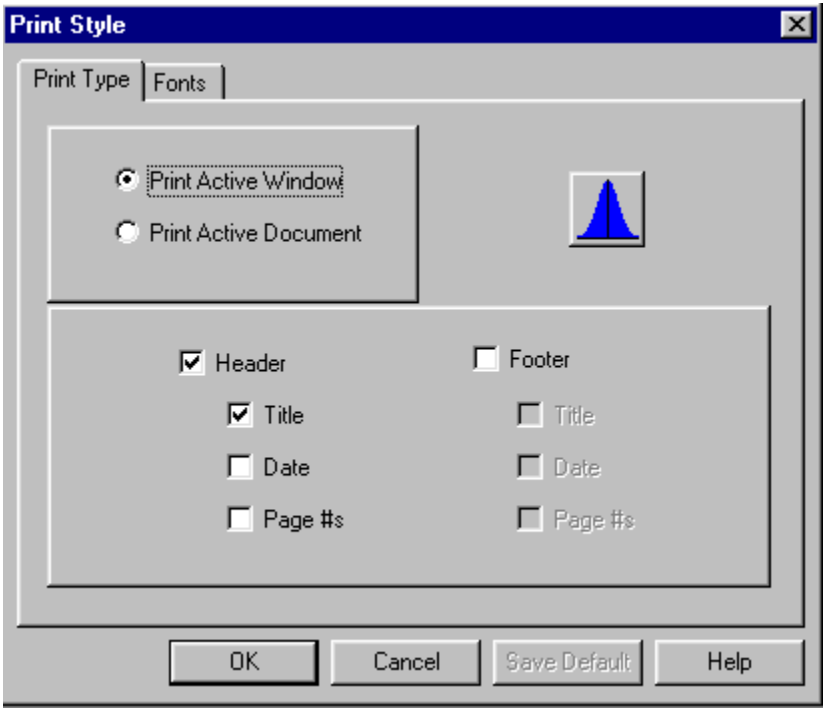


Print Style

Select Print Style from the Submenu. All print output is a copy of selected windows of a **Stat::Fit** document or of a stand alone distribution graph. The type of print output, whether a single active window or all active windows in the document, is chosen in the Print Style dialog. Unless changed by the user, the default style

is to print the active window (colored titled bar). Other options, such as fonts and labels, may be chosen in the Print Style dialog as well.

The Print Style dialog box has 2 tabs as shown below:



Print Type

Print Type is the first tab in the Print Style dialog box. It allows you to select the items you want to have printed. Your choices include:

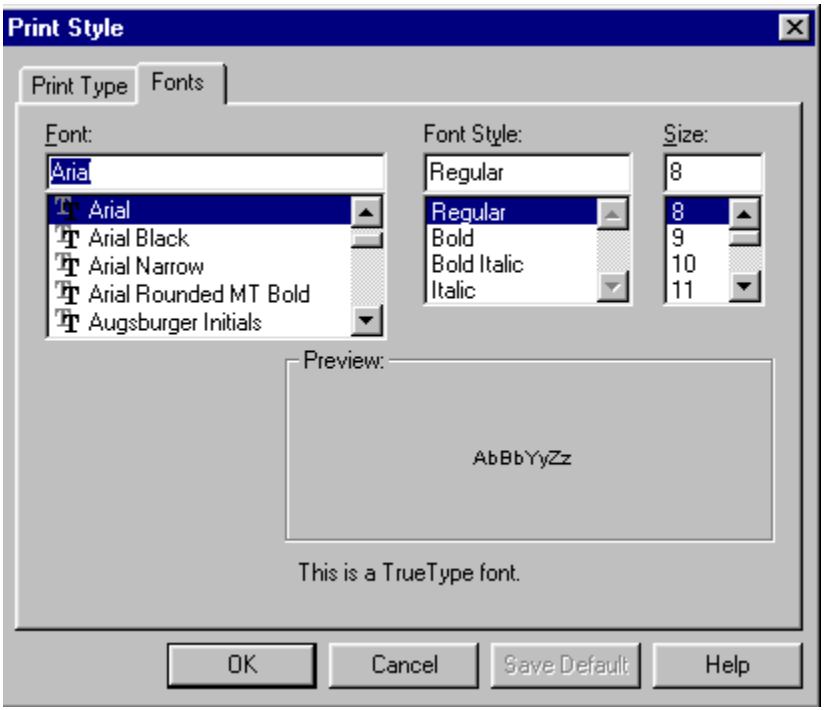
Active Window: Print only the active window.

Active Document: Print all of the open windows in a single document.

Other options available for printing are the inclusion of a Header or Footer. In either you can put the title of the document, a date and page numbers.

Fonts

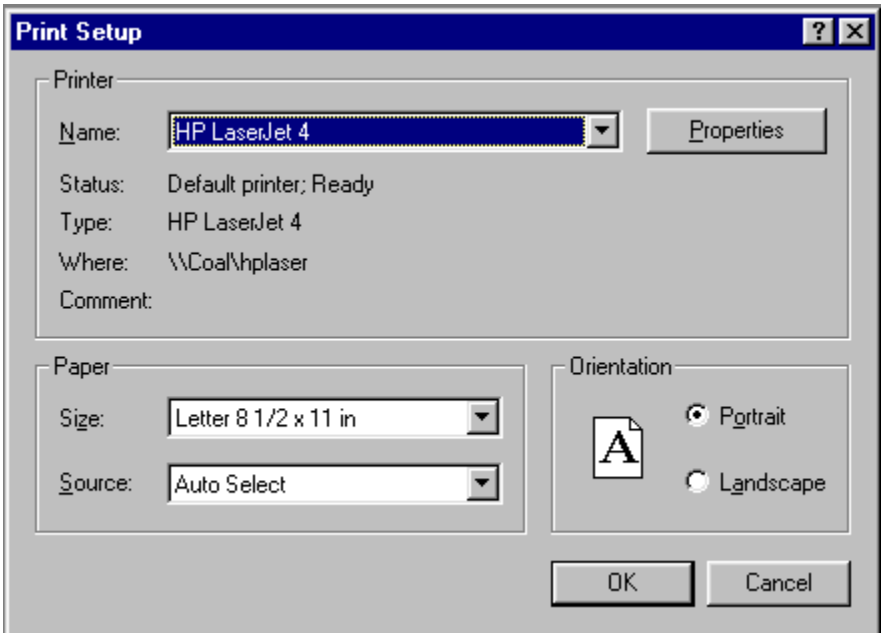
The dialog box for Fonts is shown below:



The Font page of the 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. This allows you to select the font type and size for the written text in the report. The fonts for the graphs should have been previously selected in the Graphics Style/Fonts dialog box.

Printer Setup

Return to the Submenu under File shown at the beginning of this chapter, and select Printer Setup. The standard windows dialog box for printer setup will be displayed, as shown below:



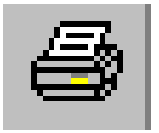
This standard Print Setup dialog will allow specification of the printer, paper size, and orientation of printed output. It will also allow access to the Properties dialog of the chosen printer. This setup will subsequently be used by the Print command.

Print Preview

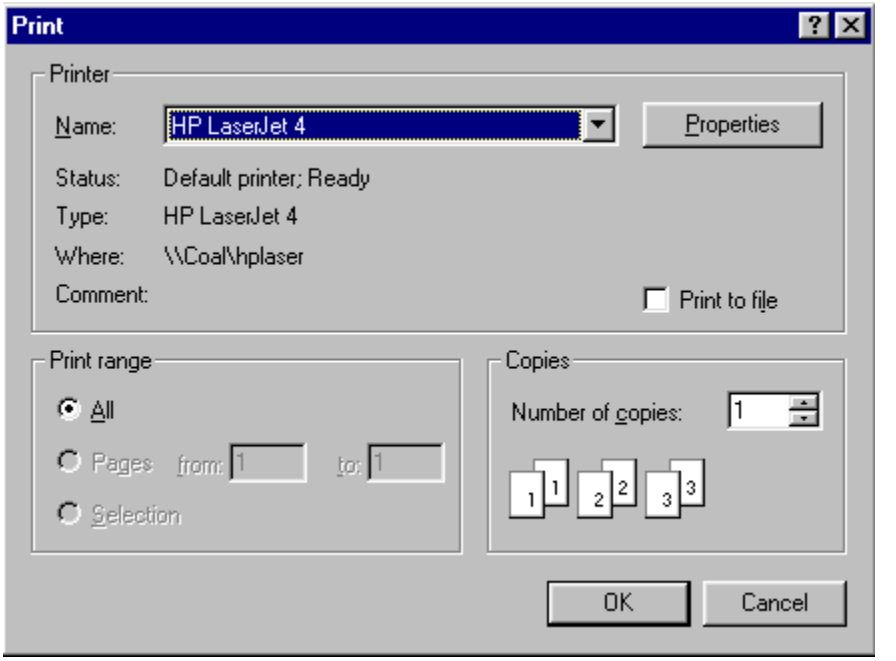
The Print Preview command opens a separate set of windows to display each expected page of the print output, using the options specified in the Print Style command. These windows can be closed by double clicking on the upper right Close button.

The Print Preview windows give a scaled version of the output to be generated by the Print command. Each graph will maintain the aspect ratio of the on screen window while maximizing the graph size to fit the printed page.

Print



Selecting the Print icon or clicking on Print on the Submenu under File, will display the following dialog box:



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 Properties 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 Output

When input data is entered into **Stat::Fit**, whether through manual entry in a new document, opening a data file, pasting data from the Clipboard, or reopening a **Stat::Fit** project file, a **Stat::Fit** document is created which contains the data and all subsequent calculations and graphs. If the document is initiated from an existing file, it assumes the name of that file and the document can be saved automatically as a **Stat::Fit** project (.SFP extension) with the Save command.

If the document is unnamed, it can be saved as either a **Stat::Fit** project or a text data file with the Save As command. In either situation, the on-screen document assumes the name of the file used.

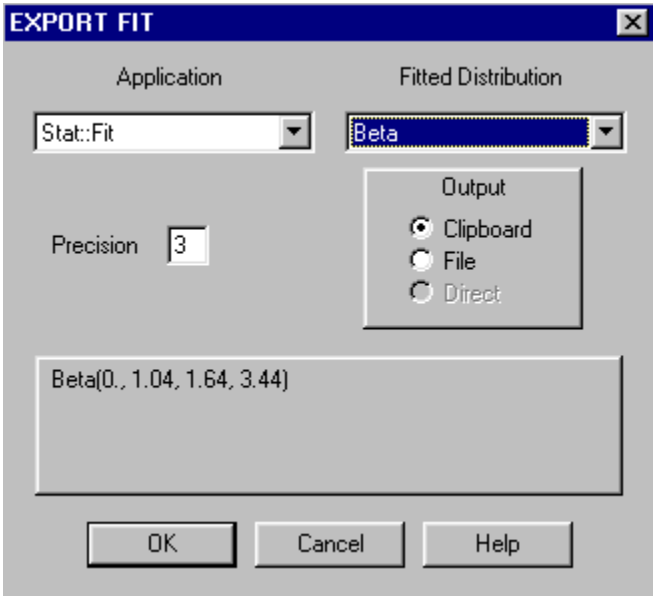
A text file of the input data can be saved with the Save Input command which will prompt if the file already exists. Unless specifically changed, the file will be saved with the *.txt* extension.

All graphs can be saved independently as bitmap files (.bmp format), by using the Save As command in the Graphics menu when the graph is the active window.

Export Fit

The fitted distributions, ready for inclusion in a specific Application, may be saved in a text file by using the Export command in the File menu or the Export icon on the Control Bar.

The Export 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, as shown below, 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 Export command requires that the appropriate estimates have been calculated, either manually or automatically with the **Auto::Fit** command.

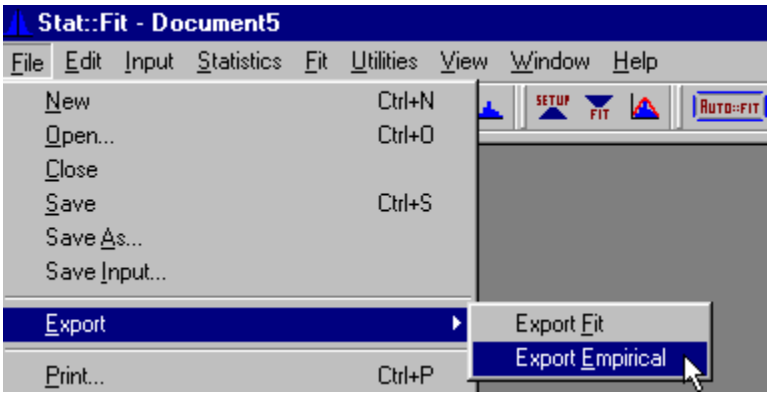
For some applications, the requested distribution may not be supported. If the generator for the unsupported distribution can be formed from a known distribution, the analytical form of this generator is given instead. If the generator is not straightforward, no output will be provided.

The output can be directed to either the Windows clipboard or a text file.

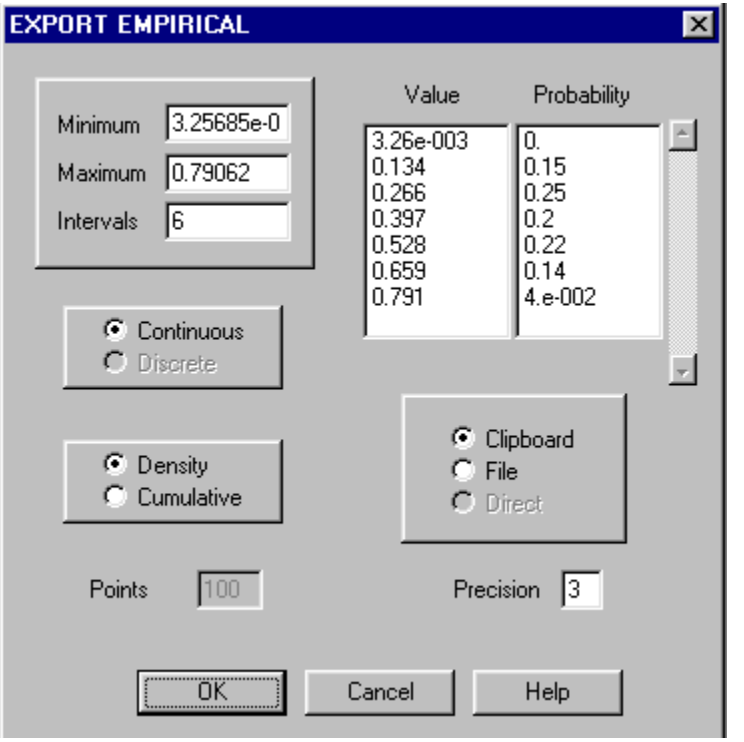
Export of Empirical Distributions

The Export Empirical command allows the export of an empirical distribution for the input data to either the Windows clipboard or a text file. 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.

To export an empirical distribution, select File from the Menu bar, Export from the Submenu, and then Export Empirical.



The following dialog will be displayed:



If 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. 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 oversmoothing. The default interval value is the same as the *auto* value set by **Stat::Fit** and will avoid oversmoothing. The distribution may be either a cumulative or density distribution.

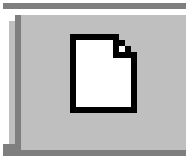
If the data is integer, then the exported distribution may be either a continuous or discrete distribution with the minimum and

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 indicated in the Points box. The distribution may be either a cumulative or density distribution.

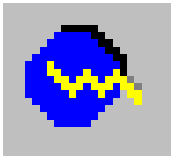
The Precision of the numbers in the output is set by the Precision box whose default value is 3.

Chapter 8 – Tutorial

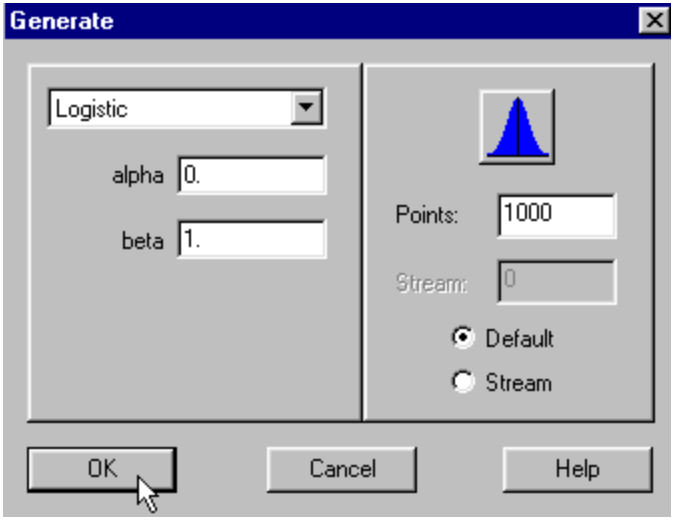
This chapter provides an example to illustrate some of the functions of **Stat::Fit**. It will also give some insights for fitting your data and understanding the results of the tests and graphical outputs.



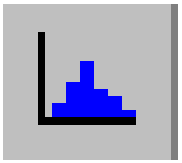
To begin, start a New File for data by clicking on the New File icon. An input data table will appear on the screen.



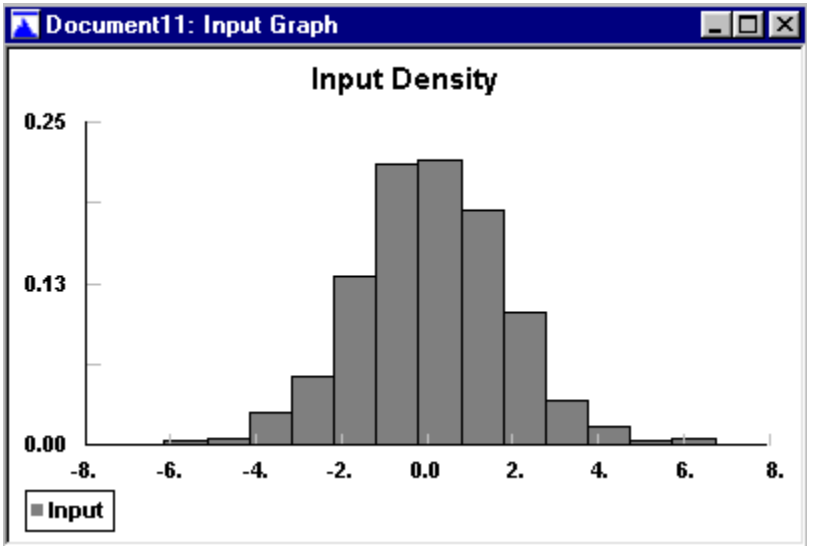
An easy way to provide some input data is to use the built-in number generator provided with **Stat::Fit**. Click on the Generate icon.



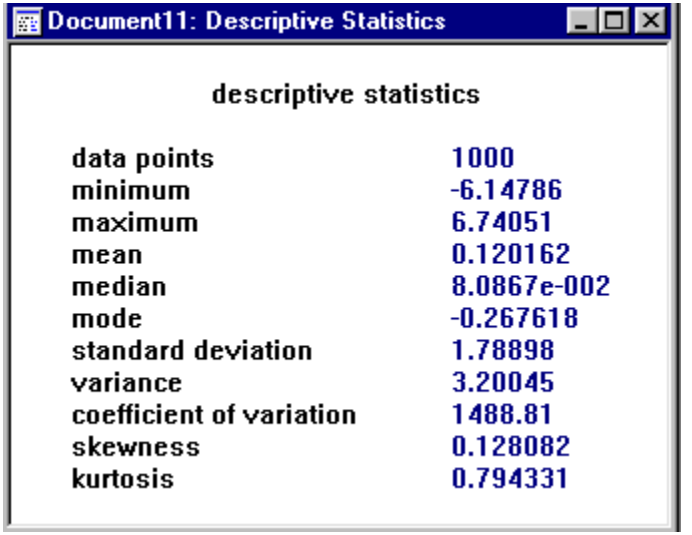
Select the Logistic distribution, enter 1000 for the number of points and use the default settings. Click on OK. The data table will be filled with 1000 points of generated data.



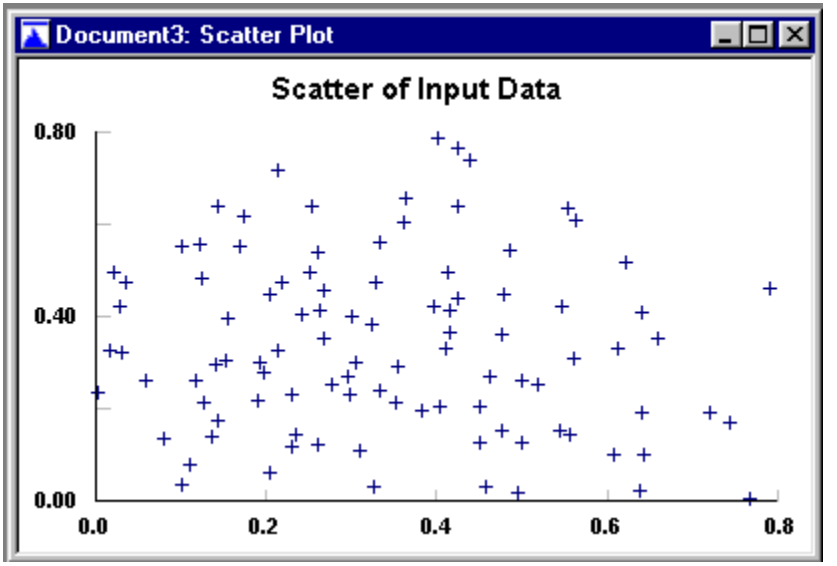
A histogram of your new data can be displayed by clicking on the Input Graph icon. A graph very similar to the following will be shown:



The Descriptive Statistics of the input data can be viewed by selecting Statistics from the menu bar and then Descriptive from the Submenu. The following table will be shown:

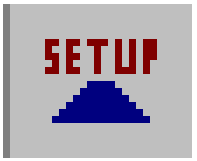


The independence of your data can be checked by selecting Statistics from the Menu bar and then Independence from the Submenu. From the next Submenu which appears, select Scatter Plot.



If the data is independent, it will scatter all over the graph.

The distribution fitting process is started by clicking on the Setup icon. A dialog box for Fit Setup will appear. Looking at the graph of the input data, the shape resembles a Normal distribution. Therefore, let us try to fit the data to both a Logistic distribution and a Normal distribution.



Click on Logistic and then click on Normal.

After selecting the distributions, go to the next window of the dialog box, for Calculations. In addition to the Chi Squared Test, also select the Kolmogorov Smirnow and Anderson Darling Tests. Then click on OK.

The tests will be performed by clicking on the Fit icon. The calculations for the goodness of fit tests will start. A Summary of the results will be presented in a table as shown below:

goodness of fit

data points	1000
estimates	maximum likelihood estimates
accuracy of fit	3.e-004
level of significance	5.e-002

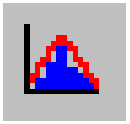
summary

distribution	Chi Squared	Kolmogorov Smirnov	Anderson Darling
Logistic(0.12, 0.986)	6.07 (12)	1.77e-002	0.291
Normal(0.12, 1.79)	8.77 (12)	2.57e-002	1.01

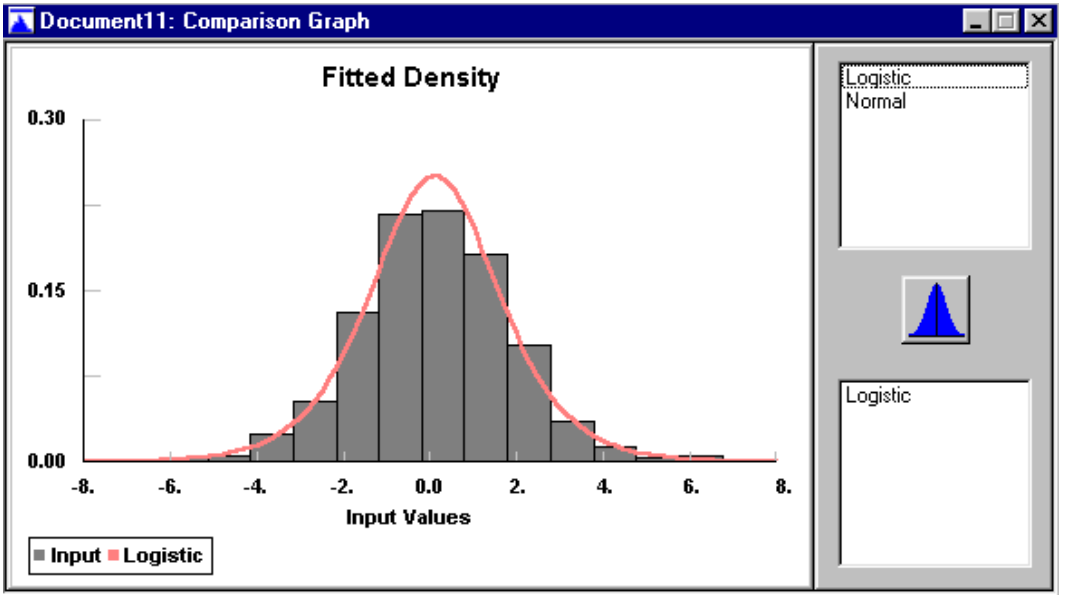
```
detail

Logistic
    alpha       =           0.120162
    beta        =           0.986316
    Chi Squared
        total classes                   13
        interval type                   equal probable
        net bins                        13
        chi**2                          6.07
        degrees of freedom              12
        alpha                           5.e-002
        chi**2(12,5.e-002)              21.
        p-value                         0.913
        result                          DO NOT REJECT
    Kolmogorov-Smirnov
        data points                     1000
        ks stat                         1.77e-002
        alpha                           5.e-002
        ks stat(1000,5.e-002)           4.28e-002
        p-value                         0.908
        result                          DO NOT REJECT
    Anderson-Darling
        data points                     1000
        ad stat                         0.291
        alpha                           5.e-002
        ad stat(5.e-002)                2.49
        p-value                         0.945
        result                          DO NOT REJECT
```

Normal		
mean =	0.120162	
sigma =	1.78808	
Chi Squared		
total classes		13
interval type		equal probable
net bins		13
chi**2		8.77
degrees of freedom		12
alpha		5.e-002
chi**2(12,5.e-002)		21.
p-value		0.722
result		DO NOT REJECT
Kolmogorov-Smirnov		
data points		1000
ks stat		2.57e-002
alpha		5.e-002
ks stat(1000,5.e-002)		4.28e-002
p-value		0.515
result		DO NOT REJECT
Anderson-Darling		
data points		1000
ad stat		1.01
alpha		5.e-002
ad stat(5.e-002)		2.49
p-value		0.354
result		DO NOT REJECT

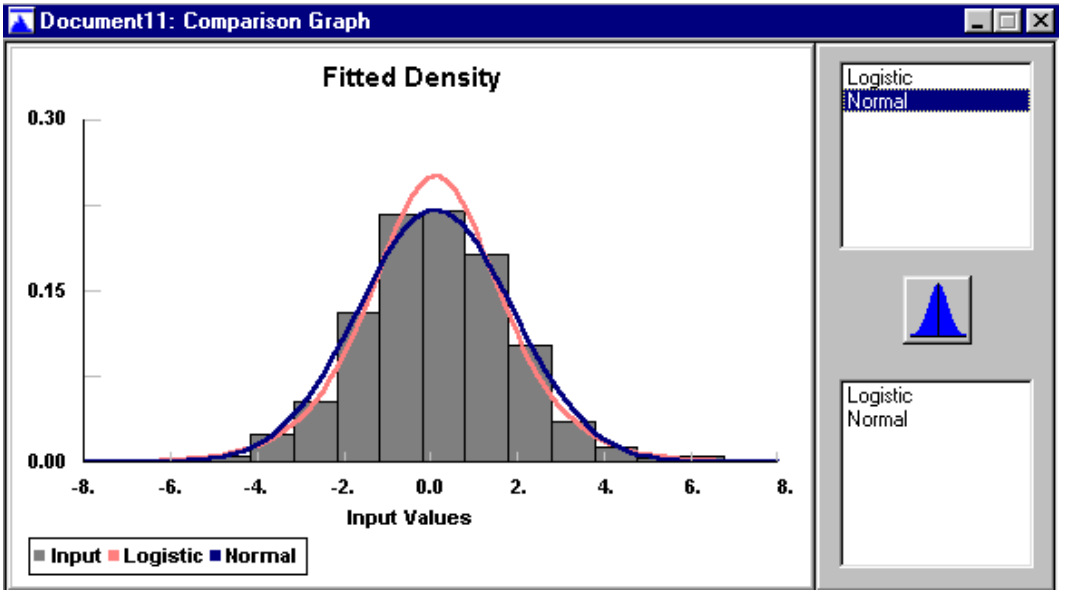
The results of the tests indicate that neither distribution should be rejected. However you will notice that the p-values for the Logistic distribution are higher than for the Normal distribution, indicating a better fit.



Graphical results can be viewed by clicking on the Graph Fit icon. The following graph will be shown:

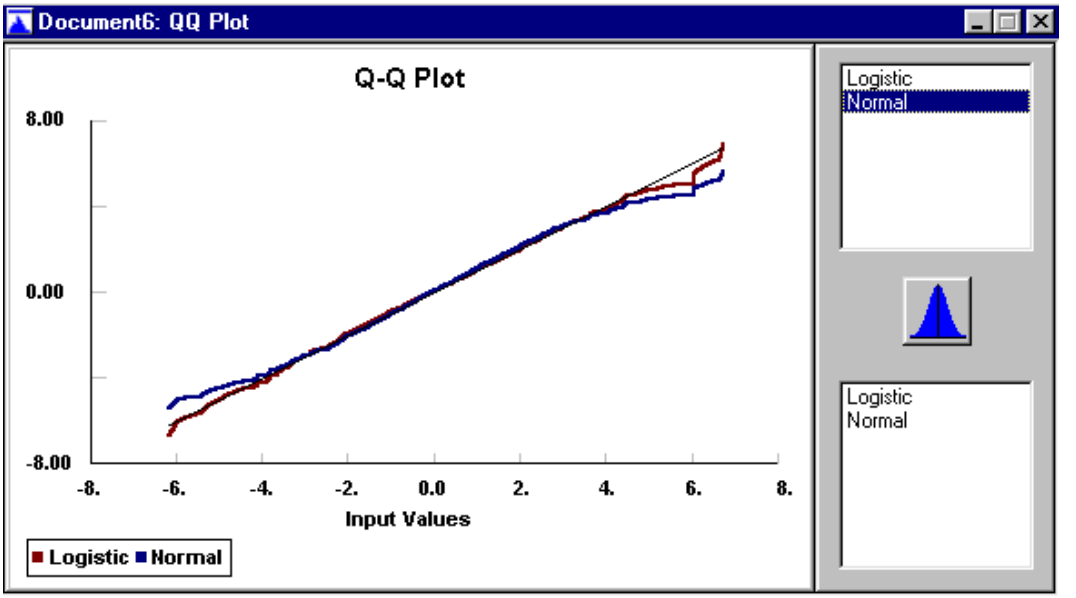


To compare this to the graph of the Normal distribution, click on Normal in the upper right box. It will be added to the graph as shown below:



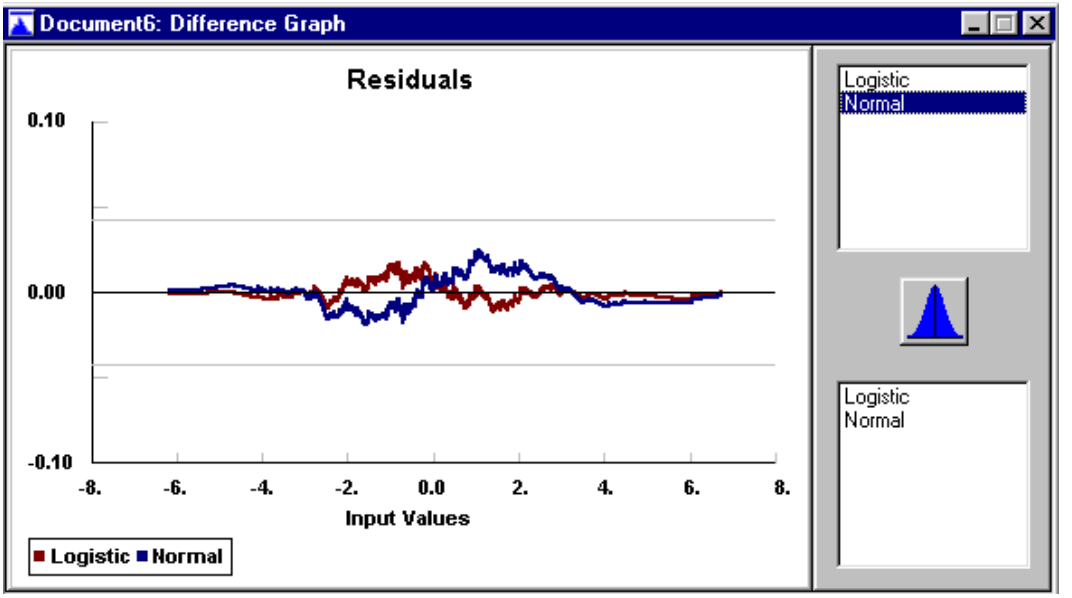
Graphically, both distributions appear to fit the data. Remember that the goodness of fit tests indicated that the Logistic distribution was a better fit. To get a better visualization of this, let us try some of the other Result Graphs.

Select Fit from the Menu bar and then Result Graphs from the Submenu. From the next Submenu choose Q-Q Plot.



The Q-Q Plot is sensitive to the tails of a distribution and you can see in the above graph that the Normal distribution does not provide as good a fit for the tails of the data as does the Logistic distribution.

Another useful graph for displaying the fit is the Difference Graph. Select Fit from the menu bar, Result Graphs from the Submenu and then Difference Graph.

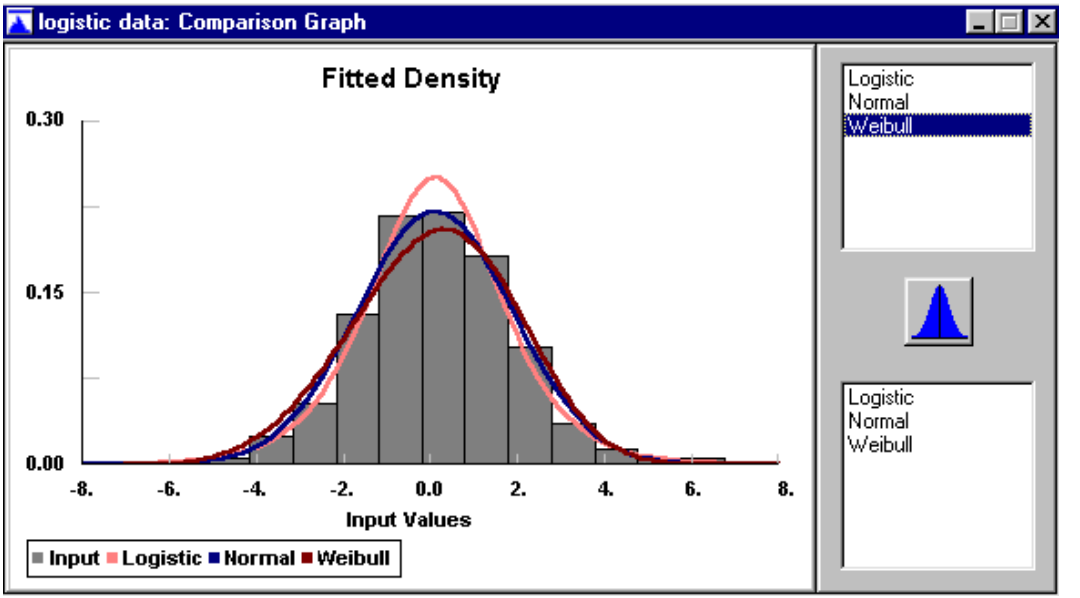


The Difference Graph illustrates the difference of the cumulative distribution you have fit compared to the data.

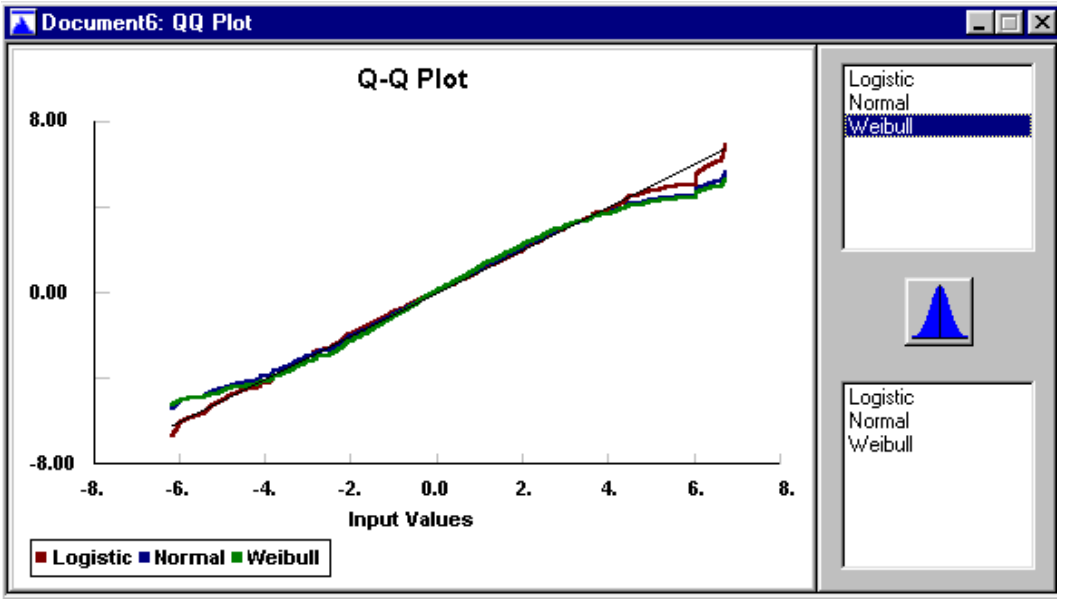
It is instructive to look at another distribution which might fit. Let us try the Weibull distribution. Repeat the previous procedures: click on the Setup icon and select the Weibull distribution for fitting. Hit OK and then click on the Test icon.

The goodness of fit summary table is shown below:

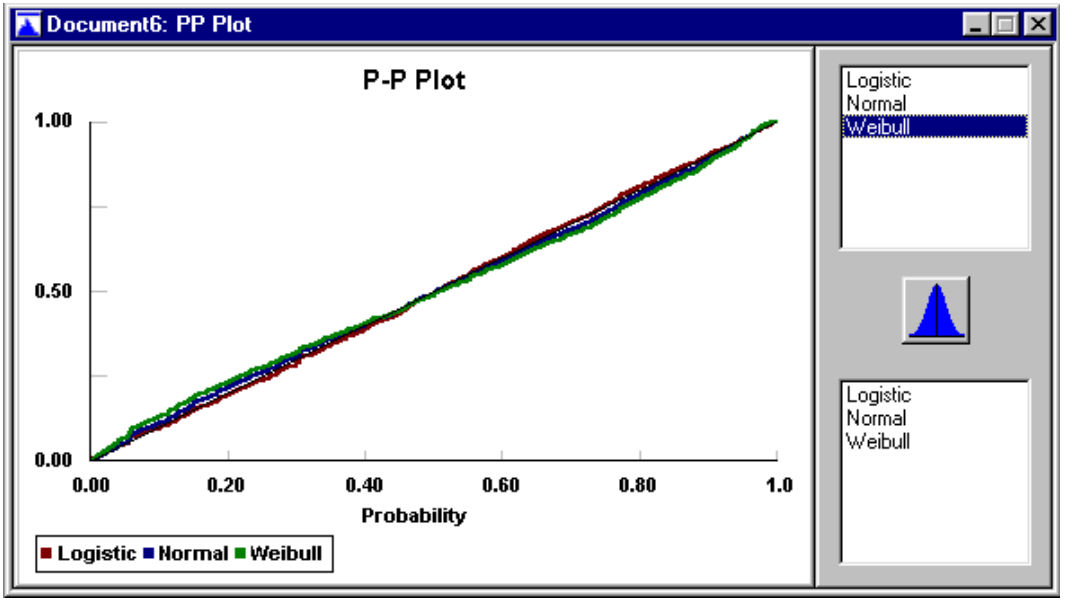
The results of these tests indicate that you should reject this distribution. Let us look at the Comparison Graph to see if it is obvious from its visual appearance. Click on the Graph Fit icon and choose Weibull.



For a better visual display, try the Q-Q Plot for the Weibull distribution.

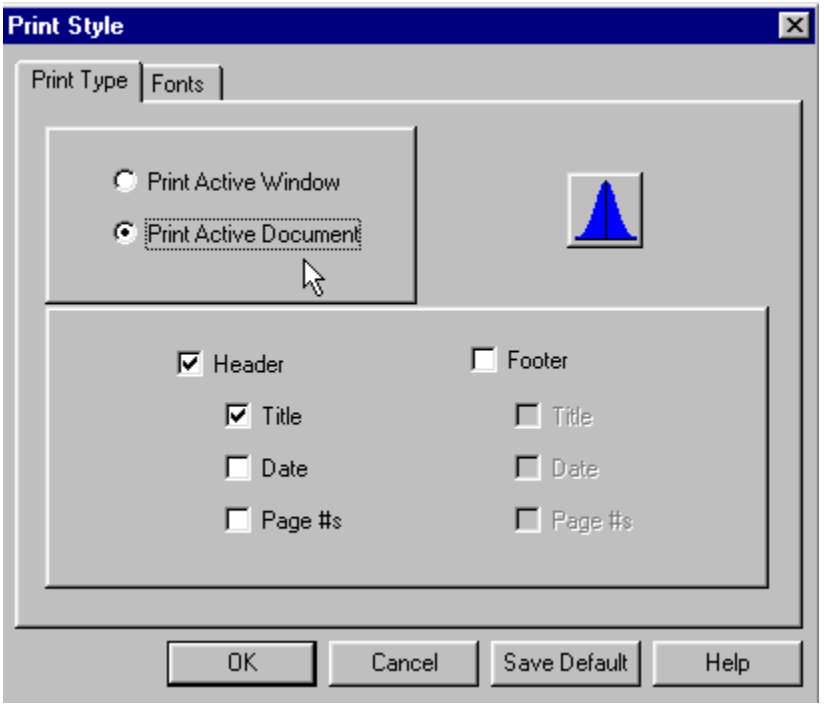


Here we see the tails of the distribution do not fit well. Let us look at the P-P Plot for the center of the distribution.



The combination of these plots provides a good visualization of how well the data is fit to a particular distribution.

Let us now print a Report with our results. Select File from the menu bar and then Print Style from the Submenu. In the first page of the dialog box, select Print Active document in order to print all of the test results and graphs.



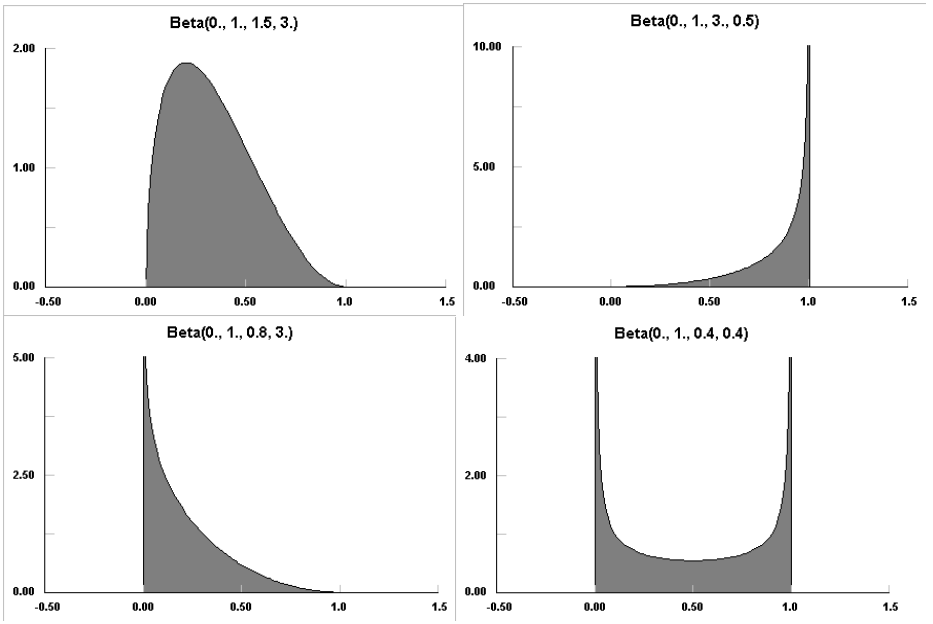
Select Print Setup from the Submenu under File in order to specify your printer, paper and orientation. Select Print Preview to display your report before printing. If you are pleased with its contents, Select Print from the Submenu and hit OK.

Appendix A – Distributions

Beta Distribution (*min, max, p, q*)

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

$\min \le x \le \max$
min = minimum value of x
max = maximum value of x
p = lower shape parameter > 0
q = upper shape parameter > 0
B(p,q) Beta Function



Description:

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 $p,q = 1$.

As can be seen in the examples above, 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 $p,q<1$ cause the Beta distribution to approach infinity at that bound. Values of $p,q>1$ cause the Beta distribution to be finite at that bound.

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, steady-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, p. 236-237

Binomial Distribution (*n, p*)

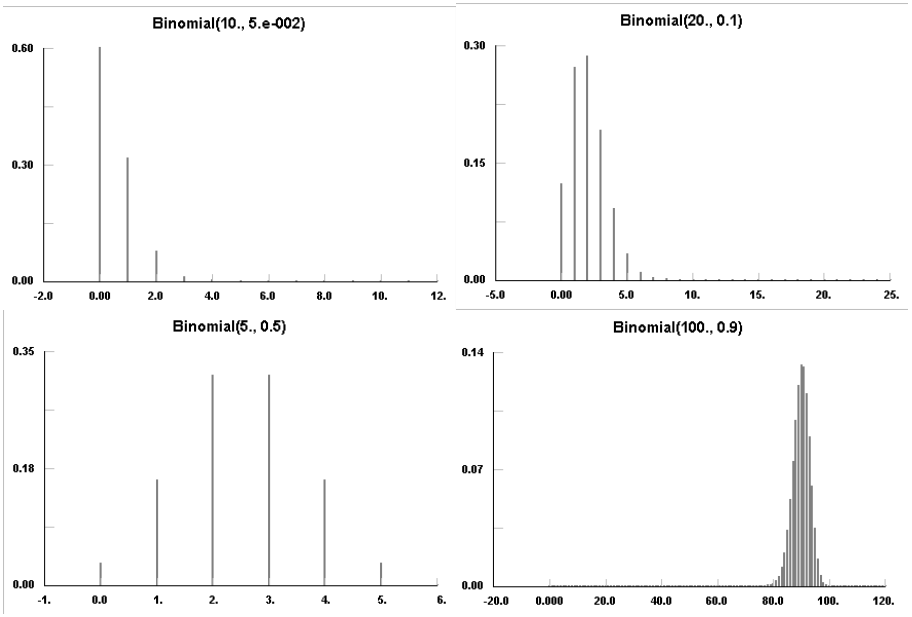
$$p(x) = \left(\frac{n}{x}\right) p^x (1-p)^{n-x}$$

x = 0, 1, …n

n = number of trials

p = probability of the event occurring

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



Description:

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 above, 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.

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]

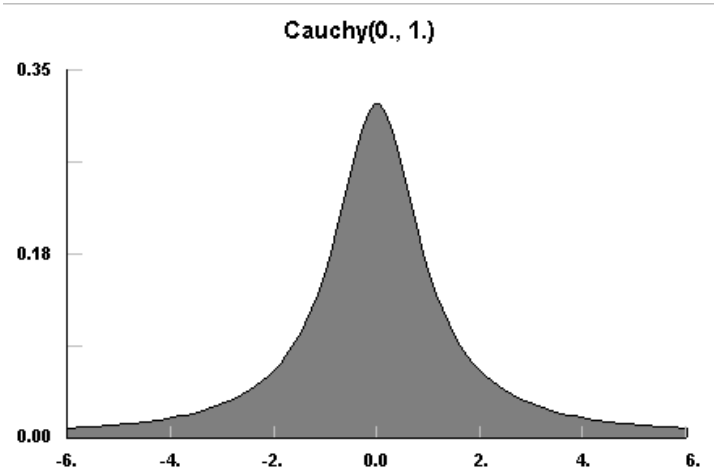
[2] "Univariate Discrete Distribution" Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp., 1992, John Wiley & Sons, p. 134

Cauchy Distribution (*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



Description:

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 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 a long as the data has a central tendency. (see Johnson et. al.[2])

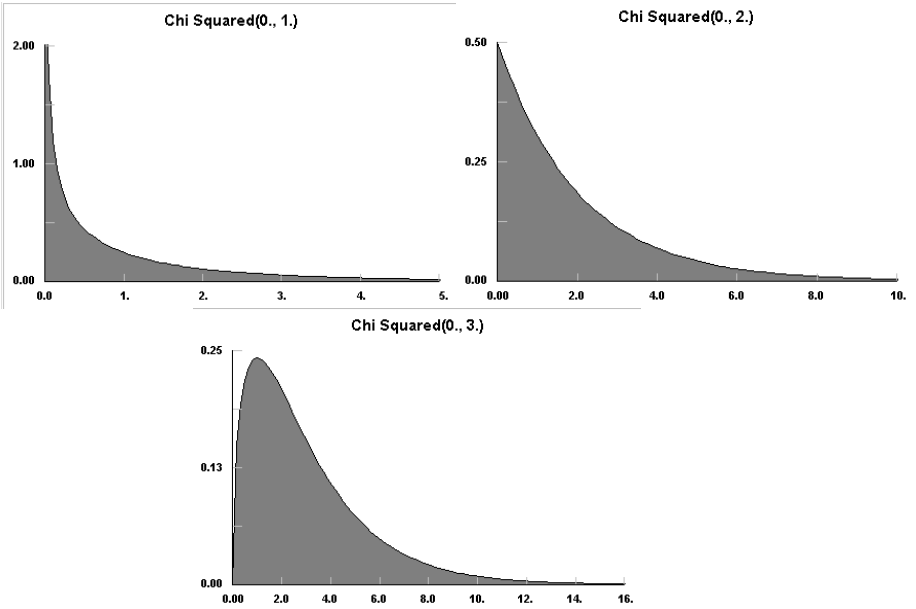
The Cauchy Distribution, as shown above, has a distinct peaked shape. It is unchanged in shape with changes in theta or lambda. More examples can be viewed by using the Distribution viewer capability.

[2] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & sons, p. 298

Chi Squared Distribution (*min, nu*)

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

min = minimum x value
nu = shape parameter



Description:

The Chi Squared is a bounded continuous distribution bounded on the lower side. Note that the Chi Squared distribution is a subset of the Gamma distribution with beta=2 and alpha=nμ/2. Like the Gamma distribution, it has three distinct regions. For nμ=2, the Chi Squared distribution reduces to the Exponential distribution, starting at a finite value at minimum x and decreasing

monotonically thereafter. For $n\mu<2$, the Chi Squared distribution tends to infinity at minimum x and decreases monotonically for increasing x. For $n\mu>2$, the Chi Squared distribution is 0 at minimum x, peaks at a value that depends on $n\mu$, decreasing monotonically thereafter.

Because the Chi Squared distribution does not have a scaling parameter, its utilization is somewhat limited. Frequently, this distribution will try to represent data with a clustered distribution with $n\mu$ less than 2. However, it can be viewed as the distribution of the sum of squares of independent unit normal variables with $n\mu$ degrees of freedom and is used in many statistical tests. (see Johnson et al.[3])

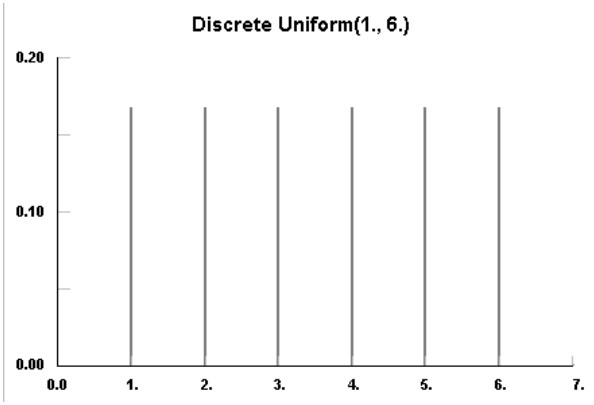
Examples of each of the regions of the Chi Squared distribution are shown above. Note that the peak of the distribution moves away from the minimum value for increasing $n\mu$, but with a much broader distribution. More examples can be viewed by using the Distribution Viewer.

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

Discrete Uniform Distribution (*min, max*)

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

x = min, min+1, …, max
min = minimum x
max = maximum x



Description:

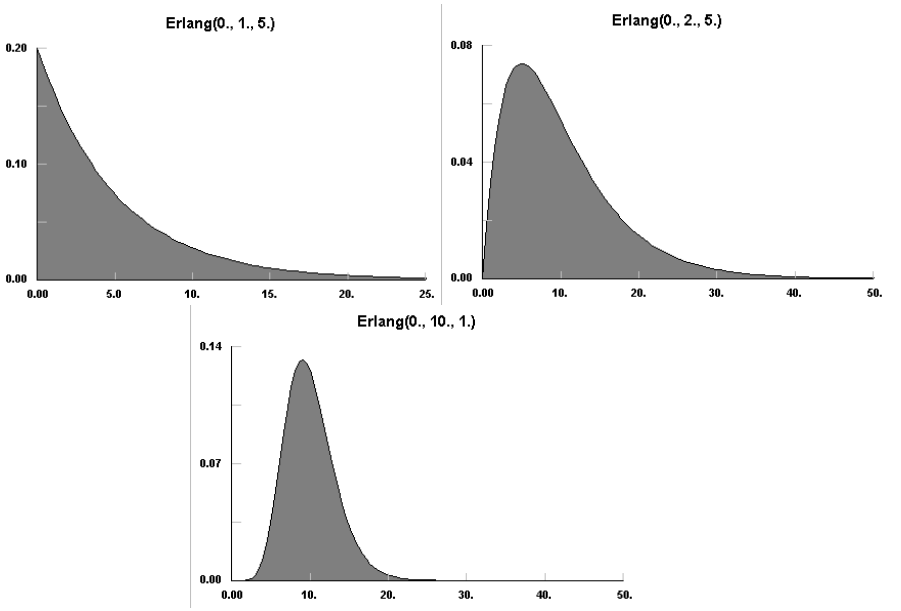
The Discrete Uniform distribution is a discrete distribution bounded on [min, max] with constant probability at every value on or between the bounds. Sometimes called the discrete rectangular distribution, it arises when an event can have a finite and equally probable number of outcomes. (See Johnson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

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

Erlang Distribution (*min, m, beta*)

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

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



Description:

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[4], Banks & Carson[5]).

As can be seen in the examples above, 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.

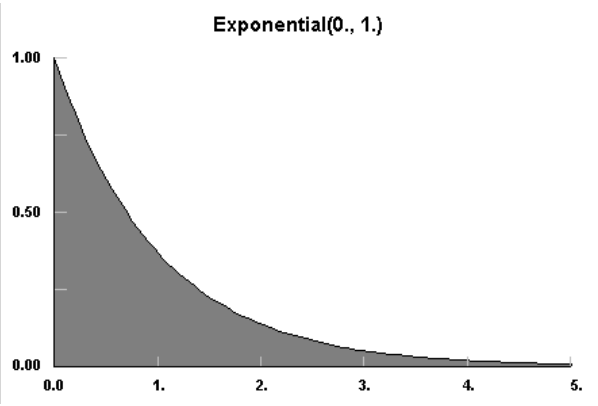
[4] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons

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

Exponential Distribution (*min, beta*)

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

min = minimum x value
β = scale parameter = mean



Description:

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. As shown in the example above, the Exponential distribution decreases rapidly for increasing 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[6], Johnson et al.[7])

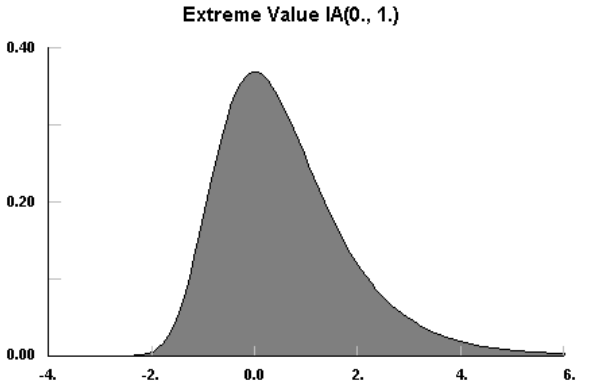
[6] "Simulation Modeling & Analysis", Averill M. Law, W. David Kelton, 1991, McGraw-Hill, p. 330
[7] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 499

Extreme Value Type 1A Distribution (*tau, beta*)

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

τ = threshold / shift parameter

β = scale parameter



Description:

The Extreme Value 1A 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 1A 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.

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[8]).

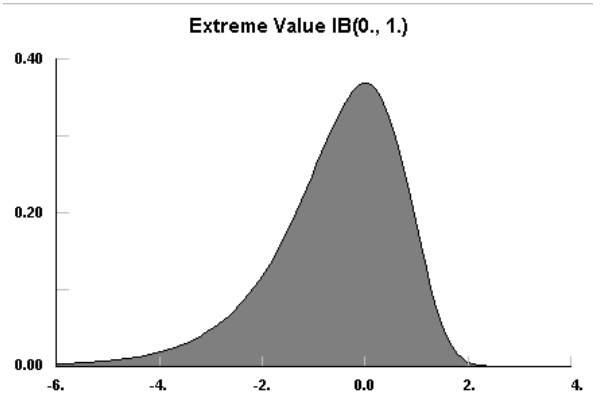
The Extreme Value 1A distribution starts below τ, is skewed in the positive direction peaking at τ, then decreasing monotonically thereafter. β determines the breadth of the distribution.

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

Extreme Value Type 1B Distribution (*tau, beta*)

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

τ = threshold/shift parameter
β = scale parameter



Description:

The Extreme Value 1B distribution is an unbounded continuous distribution. It's shape is always the same but may be shifted or scaled to need.

The Extreme Value 1B 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.[4])

The Extreme Value 1B distribution starts below τ, is skewed in the negative direction peaking at τ, then decreasing monotonically thereafter. β determines the breadth of the distribution.

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

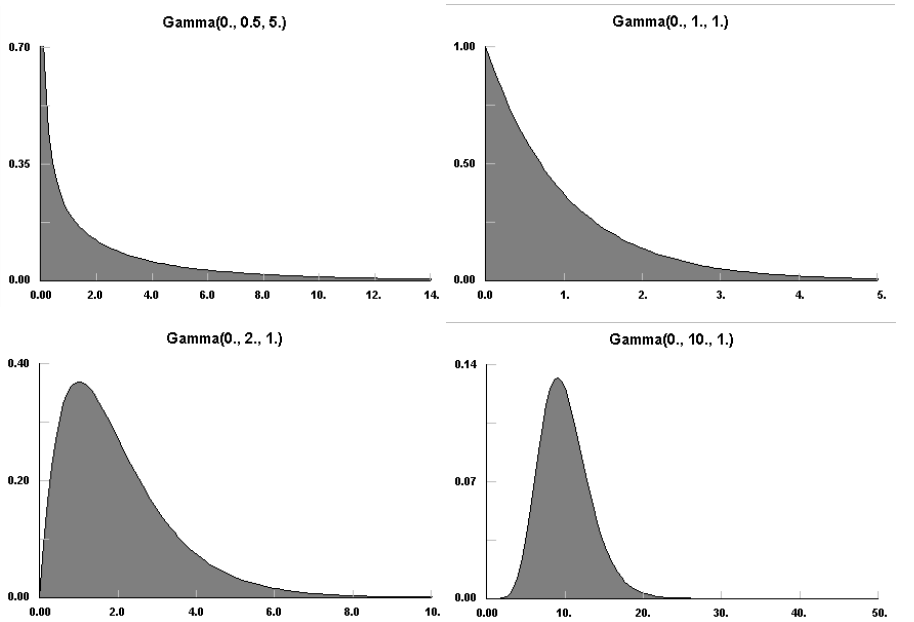
Gamma Distribution (*min, alpha, beta*)

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

min = minimum x

α = shape parameter > 0

β = scale parameter > 0



Description:

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 $\alpha=n\mu/2$. It can then be viewed as the distribution of the sum of squares of independent unit normal variables, with $n\mu$ 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, interarrival times, and service times. In particular, it can represent lifetime with redundancy (see Johnson[9], Shooman[10]).

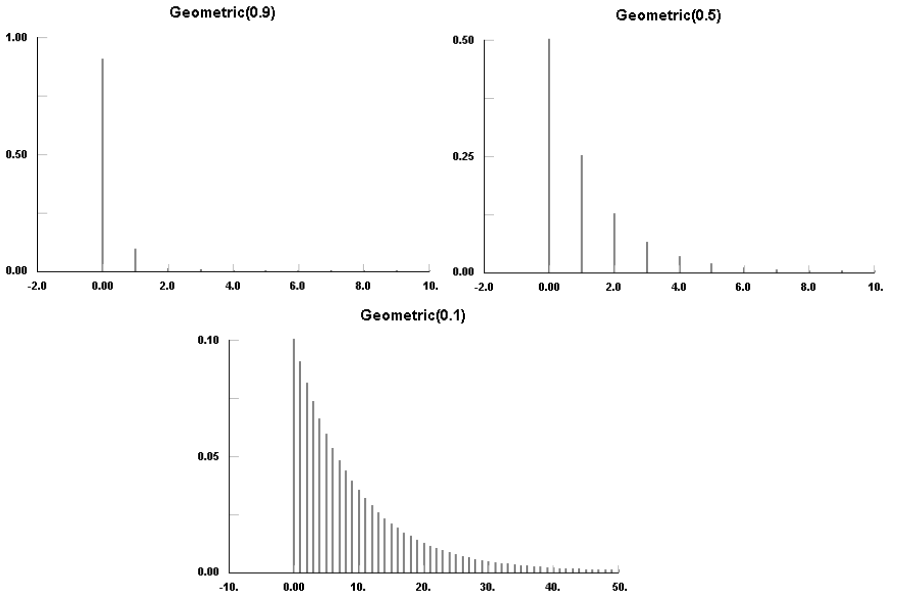
Examples of each of the regions of the Gamma distribution are shown above. Note the peak of the distribution moving away from the minimum value for increasing alpha, but with a much broader distribution.

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

Geometric Distribution (*p*)

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

p = probability of occurrence



Description:

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

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

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[11], Law & Kelton[12])

Several examples with decreasing probability are shown above.

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

[12] "Simulation Modeling & Analysis", Averill M. Law, w. David Kelton, 1991, McGraw-Hill, p. 366

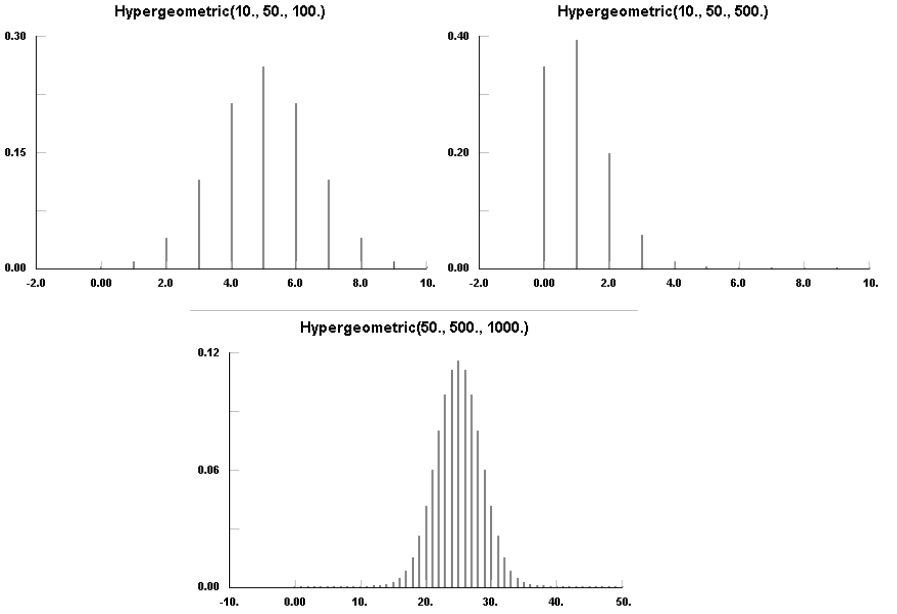
Hypergeometric Distribution (*s, m, M*)

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

s = sample size

m = number of defects in the population

M = size of the population



Description:

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 $m/N=p$ 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.[5])

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.

[5] "Univariate Discrete Distribution", Norman L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p. 237

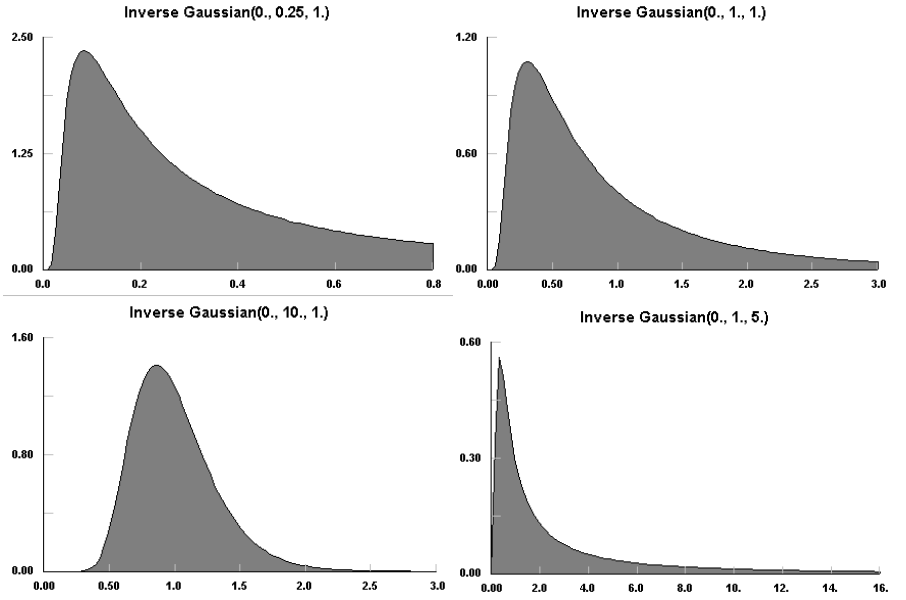
Inverse Gaussian Distribution (*min, alpha, beta*)

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

min = minimum x

α = shape parameter > 0

β = mixture of shape and scale > 0



Description:

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[13]).

Examples of Inverse Gaussian distributions are shown above. In particular, notice the drastically increased upper tail for increasing β.

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

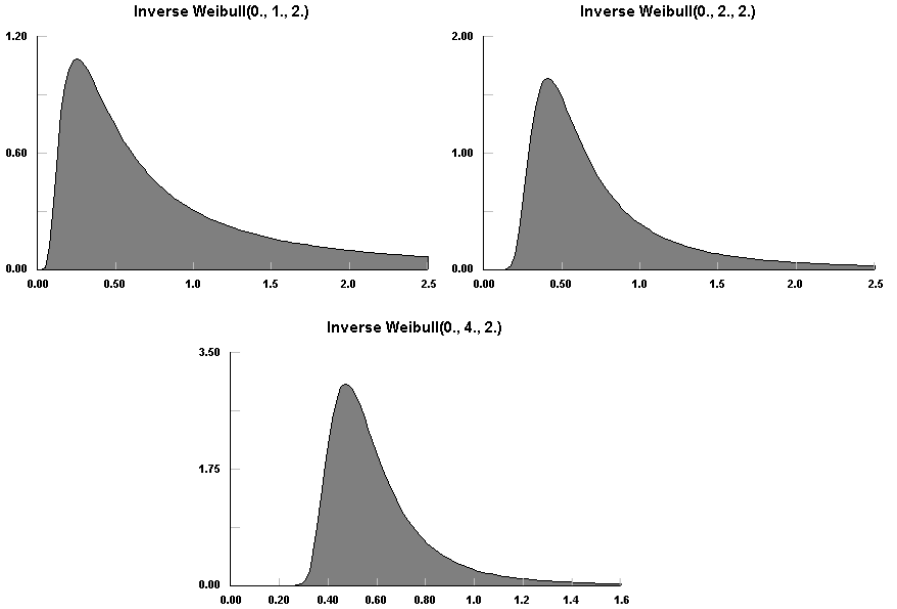
Inverse Weibull Distribution (*min, alpha, beta*)

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

min = minimum x

α = shape parameter > 0

β = mixture of shape and scale > 0



Description:

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 & Pulcini[6]) 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 above. In particular, notice the increased peakedness and movement from the minimum for increasing α

[6] R. Calabria, G. Pulcini, "On the maximum likelihood and least-squares estimation in the Inverse Weibull Distribution", Statistica Applicata, Vol. 2, n.1, 1990, p.53

Johnson SB Distribution (*min, lambda, gamma, delta*)

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

where

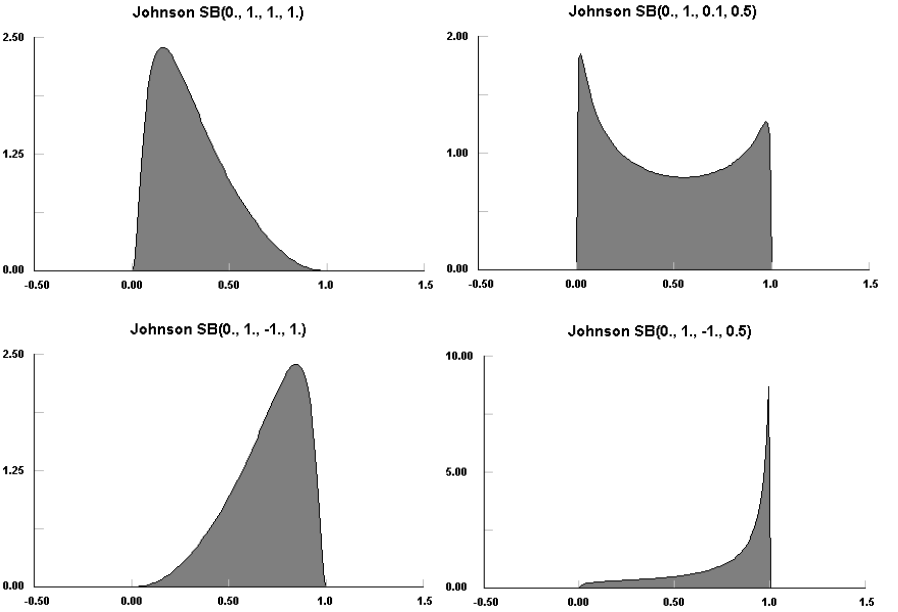
$$y = \frac{x-\min}{\lambda}$$

min = minimum value of x

λ **= range of x above the minimum**

γ **= skewness parameter**

δ **= shape parameter > 0**



Description:

The Johnson SB distribution is a continuous distribution 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 γ controlling the skewness and δ controlling the shape. The distribution can be either unimodal or bimodal. (see Johnson et al. [7] and N. L. Johnson[8])

[7] "Continuous Univariate distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, Johns Wiley & Sons, p. 34

[8] N. L. Johnson, "Systems of frequency curves generated by methods of translation", Biometrika, Vol. 36, 1949, p. 149

Johnson SU Distribution (*xi, lambda, gamma, delta*)

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

where

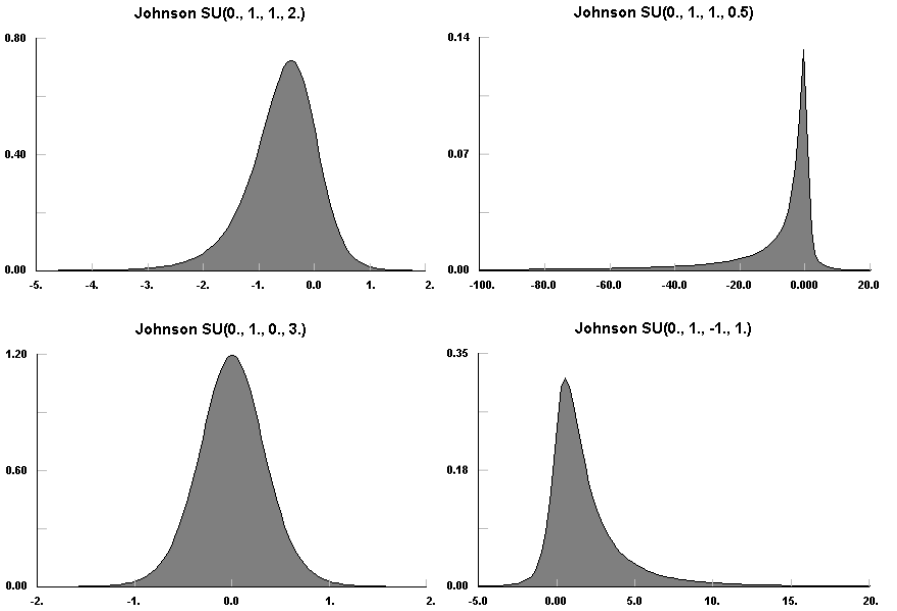
$$y = \frac{x-\xi}{\lambda}$$

x_i = minimum value of x

λ =range of x above the minimum

γ= skewness parameter

δ = shape parameter > 0



Description:

The Johnson SU distribution is an unbounded continuous distribution. The Johnson SU distribution, together with the Lognormal and the Johnson SB distributions, 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 of the notoriously unstable Pearson IV distribution, with reasonably good fidelity over the most probable range of values.

As can be see in the examples above, the Johnson SU distribution is one of the few unbounded distributions that can vary its shape, with γ controlling the skewness and δ controlling the shape. The scale is controlled by γ, δ, and λ. (see Johnson et al.[9] and N. L. Johnson[10])

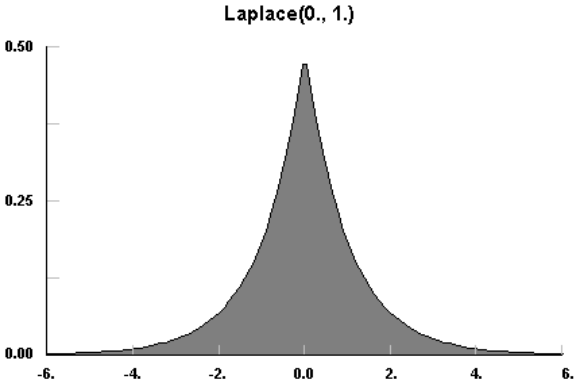
[9] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 34

[10] N. L. Johnson, "Systems of frequency curves generated by methods of translation", Biometrika, Vol. 36, 1949, p. 149

Laplace Distribution (*theta, phi*)

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

θ = mode or central peak position
φ = scaling parameter



Description:

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

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.[11])

[11] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balkrishnan, 1994, John Wiley & Sons, p. 164

The Laplace distribution, as shown above, has a distinct spike at its mode. It is unchanged in shape with changes in θ or φ.

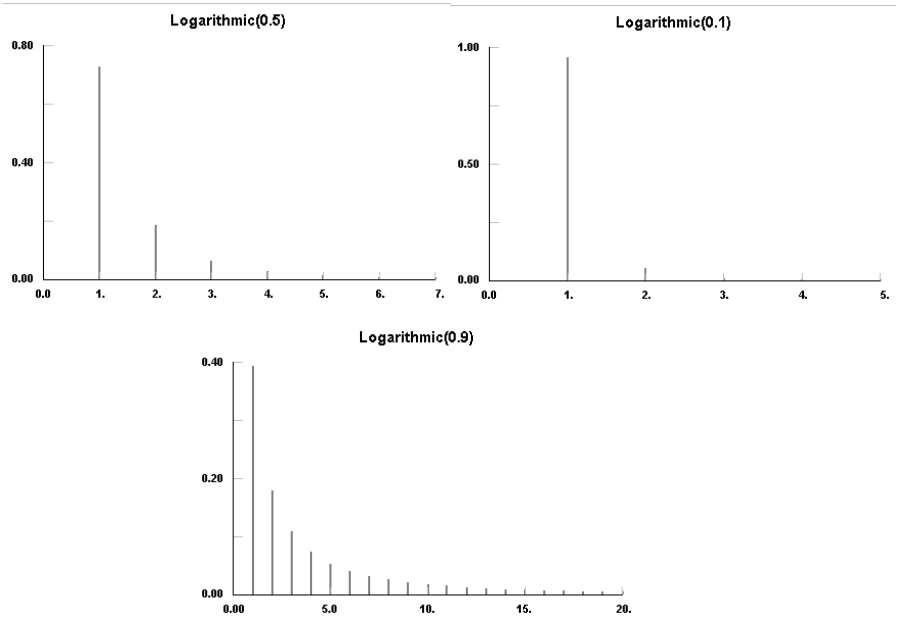
Logarithmic Distribution (*theta*)

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

where

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

Θ = shape/scale parameter $0 < \text{theta} < 1$



Description:

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.[12])

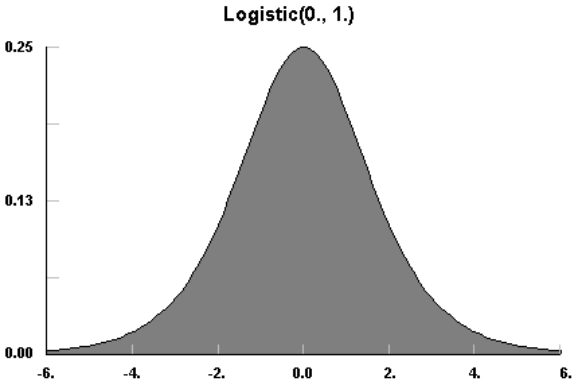
As shown in the examples, low values of theta give high probabilities for low values of x with the distribution expanding as theta nears 1.

[12] "Univariate Discrete Distributions", Normn L. Johnson, Samuel Kotz, Adrienne W. Kemp, 1992, John Wiley & Sons, p. 285

Logistic Distribution (*alpha, beta*)

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

α = shift parameter
β = scale parameter > 0



Description:

The Logistic distribution is an unbounded continuous distribution which is symmetrical about its mean (and shift parameter), α. As shown in the example above, 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 as a growth model; for populations, for weight gain, for business failure, etc.. The Logistic

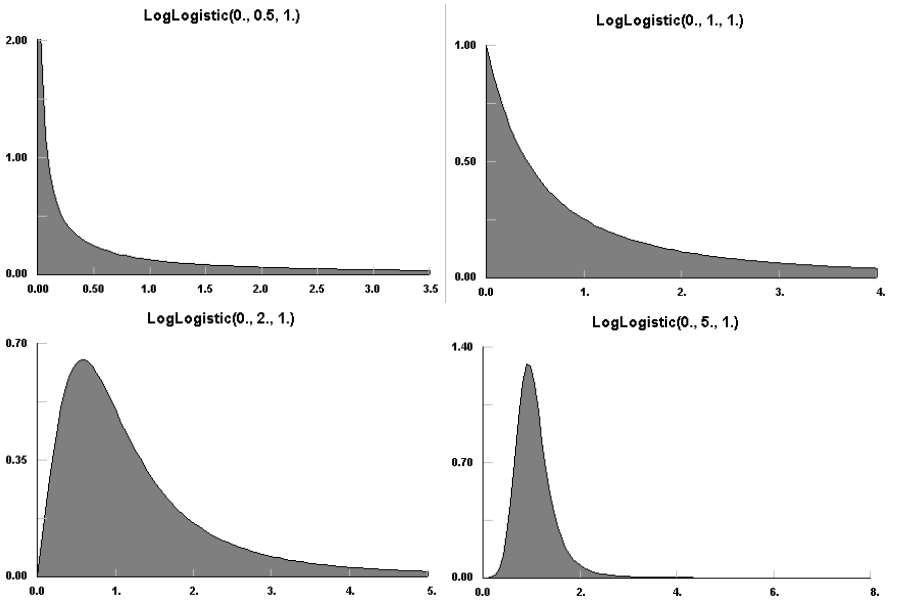
distribution can be used to test for the suitability of such a model, with transformation to get back to the minimum and maximum values for the Logistic function. Occasionally, the Logistic function is used in place of the Normal function where exceptional cases play a larger role (see Johnson[14]).

[14] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p. 113

Log-Logistic Distribution (*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
p = shape parameter > 0
β = scale parameter > 0



Description:

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

By definition, the natural logarithm of a Log-Logistic 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 Log-Logistic parameters, LLp and LLβ, by

$$\text{Lapha} = \ln(\text{LL}\beta)$$

$$\text{Lbeta} = 1/\text{LLp}$$

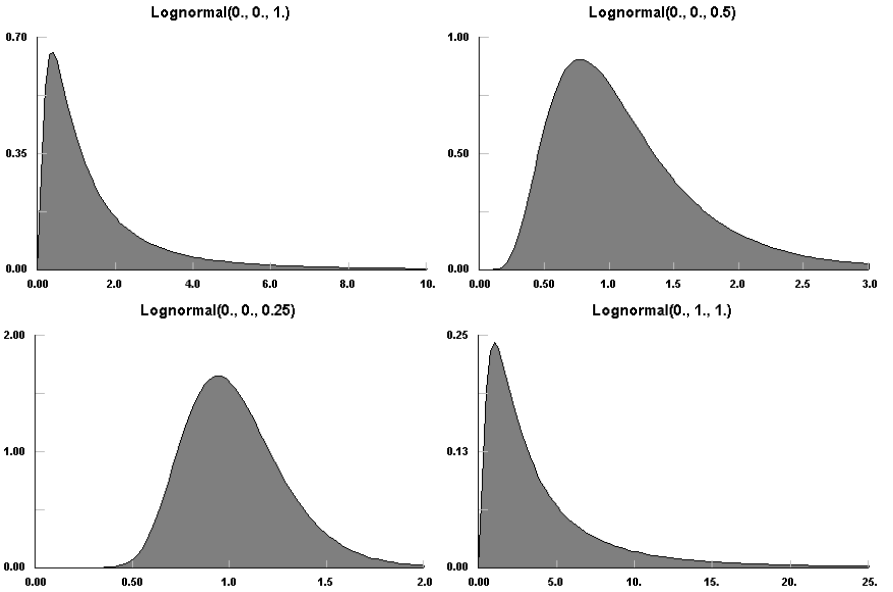
The Log-Logistic distribution is used to model the *output* of complex processes such as business failure, product cycle time, etc. (see Johnson[15]).

[15] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p. 151

Lognormal Distribution (*min, mu, sigma*)

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

min = minimum x
μ = mean of the included Normal
σ = sandard deviation of the included Normal



Description:

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 μ and σ, 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 σ, 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[16]).

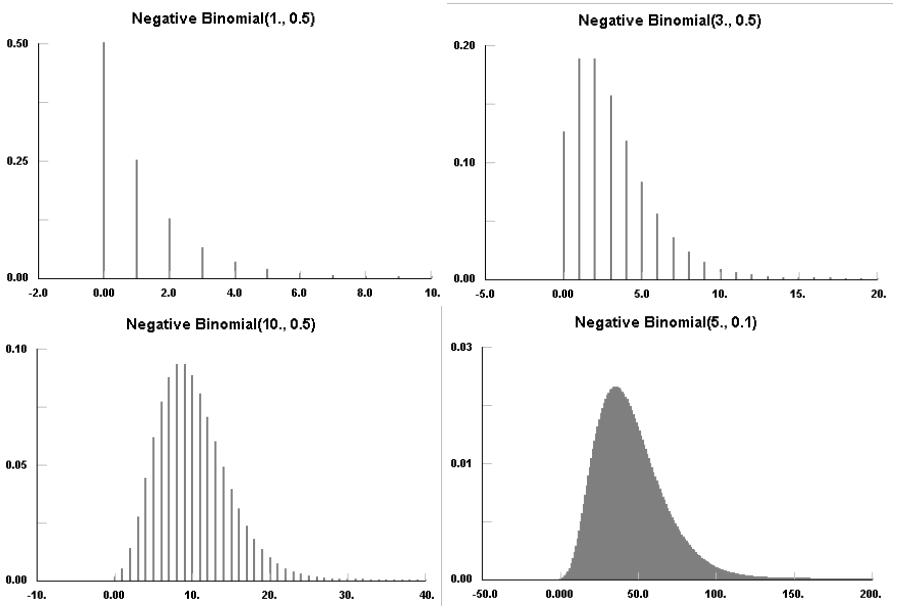
The Lognormal distribution can provide very peaked distributions for increasing σ, indeed, far more peaked than can be easily represented in graphical form.

[16] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 207

Negative Binomial Distribution (*p*, *k*)

$$p(x) = \left(\frac{k+x-1}{x} \right) p^k (1-p)^x$$

x = number of trials to get k events
p = probability of event = [0,1]
k = number of desired events = positive integer



Description:

The Negative Binomial distribution is a 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, biometrics data, and many others (see Johnson[17]).

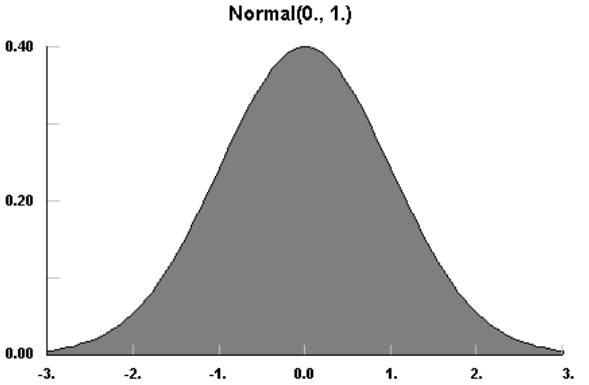
Several examples with increasing *k* are shown above. With smaller probability, *p*, the number of classes is so large that the distribution is best plotted as a filled polygon.

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

Normal Distribution (*mu, sigma*)

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

μ = shift parameter
σ = scale parameter = standard deviation



Description:

The Normal distribution is an unbounded continuous distribution. It is 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[18]).

The Normal distribution is used as an approximation for the Binomial distribution when the values of *n, p* are in the appropriate

[18] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 80

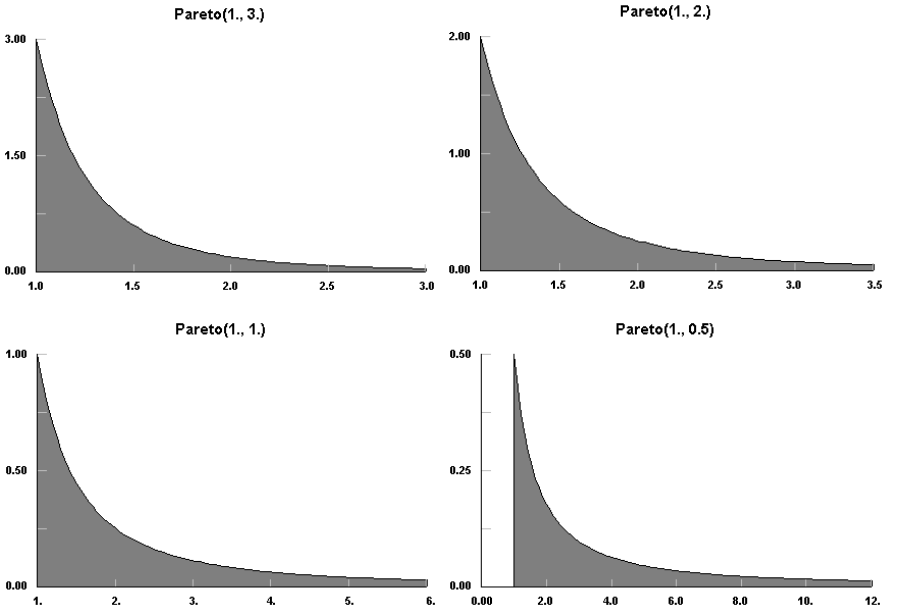
range. The Normal distribution is frequently used to represent symmetrical data, but suffers from being unbounded in both directions. If the data is known to have a lower bound, it may be better represented by suitable parameterization of the Lognormal, Weibull or Gamma distributions. If the data is 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, shown above, has the familiar bell shape. It is unchanged in shape with changes in μ or σ.

Pareto Distribution (*min, alpha*)

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

min = minimum x
α = scale parameter >0



Description:

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[19]).

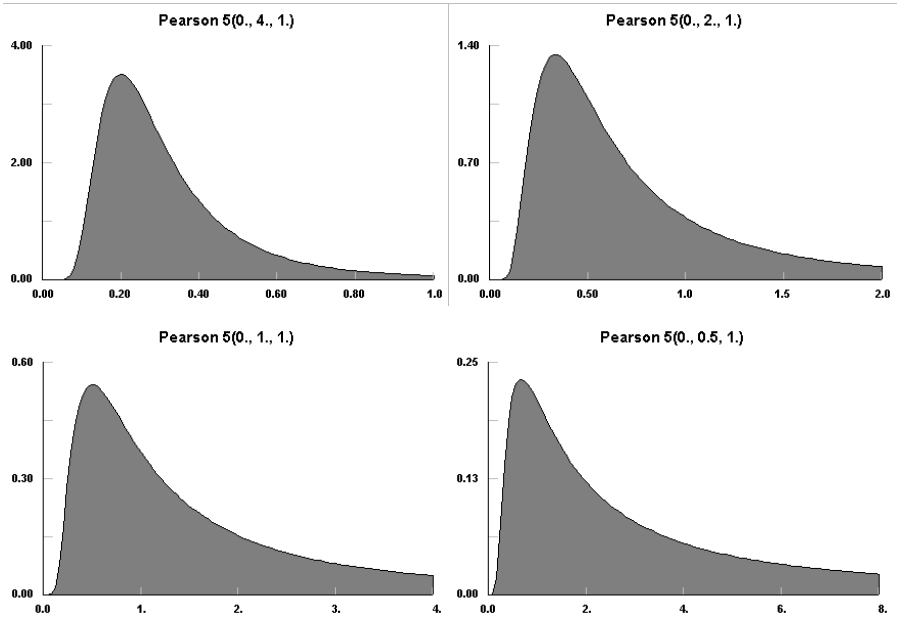
The shape of the Pareto curve changes slowly with α, but the tail of the distribution increases dramatically with decreasing α.

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

Pearson 5 Distribution (*min, alpha, beta*)

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

min = minimum x
α = shape parameter > 0
β = scale parameter > 0



Description:

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 & Kelton[20]).

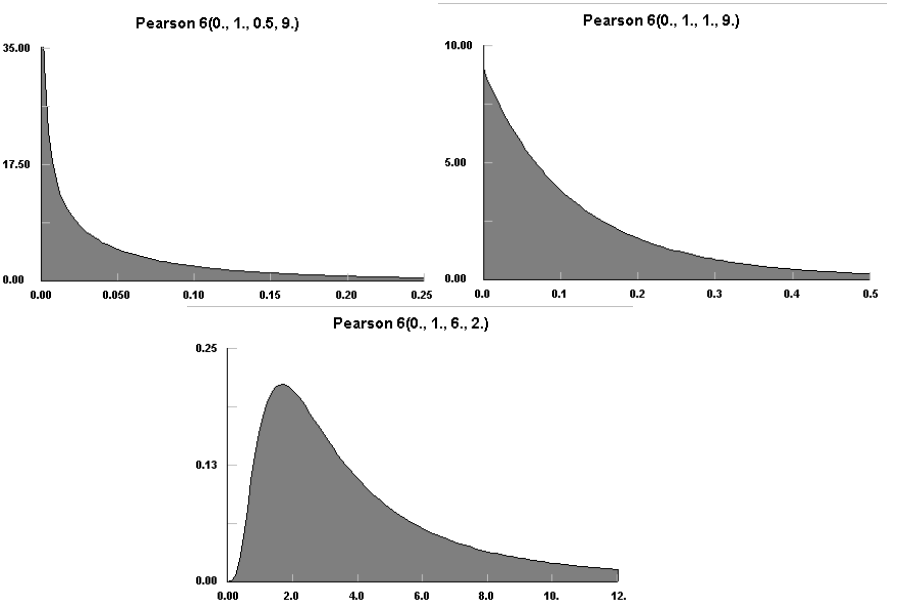
The Pearson 5 distribution starts slowly near its minimum and has a peak slightly removed from it, as shown above. With decreasing α, the peak gets flatter (see vertical scale) and the tail gets much broader.

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

Pearson 6 Distribution (*min, beta, p, q*)

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

$x > \min$
$\min \in (-\infty, \infty)$
$\beta > 0$
$p > 0$
$q > 0$



Description:

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, β=1, $p=n\mu_1/2$, $q=n\mu_2/2$, the Pearson 6 distribution reduces to the F distribution of $n\mu_1$, $n\mu_2$ which is used for many statistical tests of goodness of fit (see Johnson[21]).

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 is serves as the distribution of the ratio of independent estimators of variance and provides the final test for the analysis of variance.

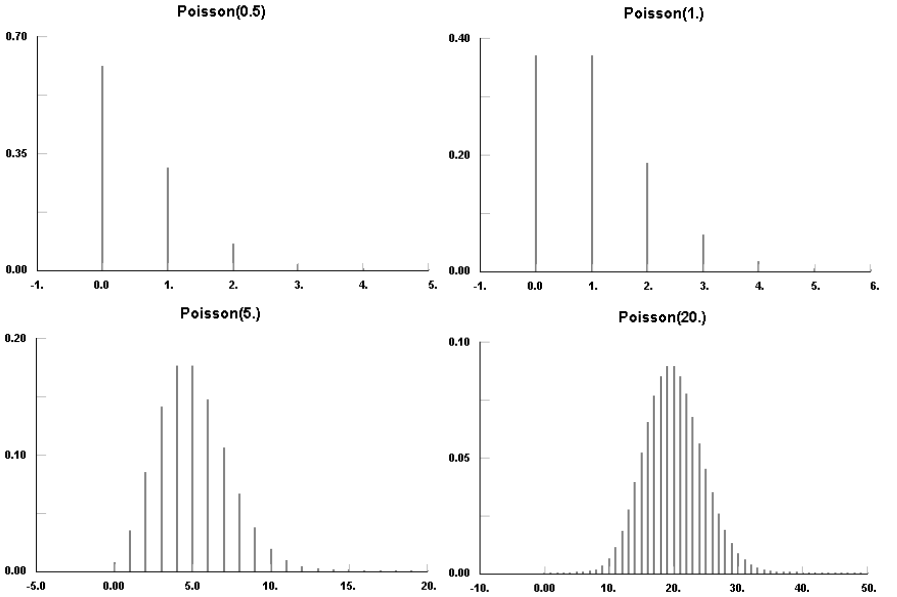
The three regions of the Pearson 6 distribution are shown in the examples above. Also note that the distribution becomes sharply peaked just off the minimum for increasing q.

[21] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p. 322

Poisson Distribution (*lambda*)

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

λ = rate of occurrence



Description:

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 semiconductor 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[22]. 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.

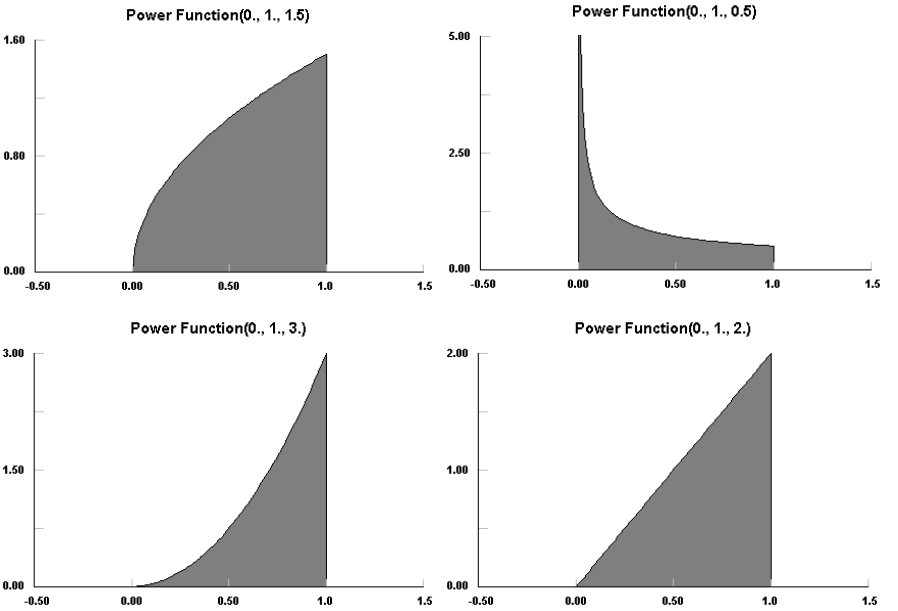
The Poisson distribution peaks near λ and falls off rapidly on either side.

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

Power Function Distribution (*min, max, alpha*)

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

min = minimum value of x
max = maximum value of x
α = shape parameter > 0



Description:

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.[13]) The Uniform distribution is a special case of the Power Function distribution with p=1.

As can be seen from the examples above, 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.

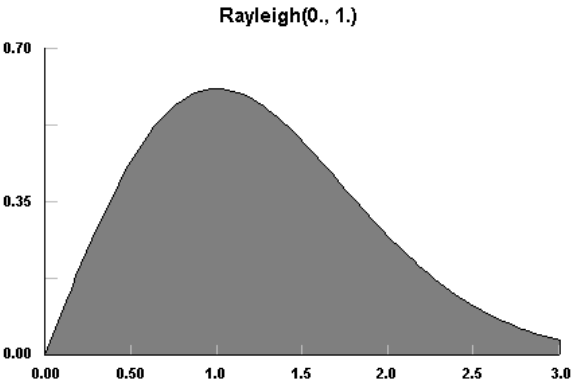
[13] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p. 210

Rayleigh Distribution (*min, sigma*)

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

$\min$ = minimum x

σ = scale parameter > 0



Description:

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.

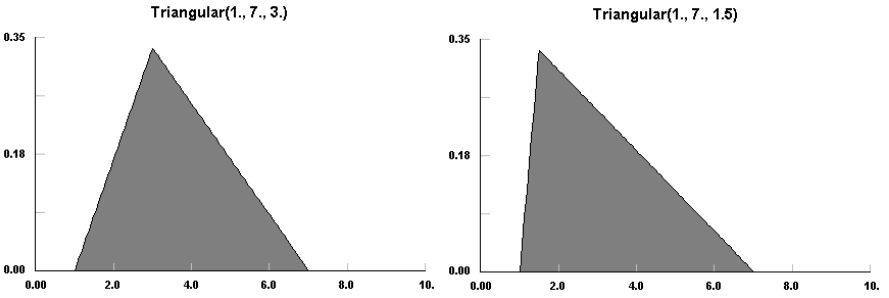
The Rayleigh distribution is frequently used to represent lifetimes because its hazard rate increases linearly with time, e.g. the lifetime of vacuum tubes. This distribution also finds application in noise problems in communications. (see Johnson et al.[14] and Shooman[15])

[14] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 456

Triangular Distribution (*min, max, mode*)

$$f(x) = \begin{cases} \dfrac{2(x - \min}{(\max - \min)(\mathrm{mod}\, e - \min)} & \min < x \le \mathrm{mode} \\ \dfrac{2(\max - x)}{(\max - \min)(\max - \mathrm{mod}\, e)} & \mathrm{mode} < x \le \max \end{cases}$$

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



Description:

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

[15] "Probabilistic Reliability: An Engineering Approach", Martin L. Shooman, 1990, RobertE. Krieger, p. 48

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 & Kelton[23]). However, it is employed as the functional form of regions for fuzzy logic due to its ease of use.

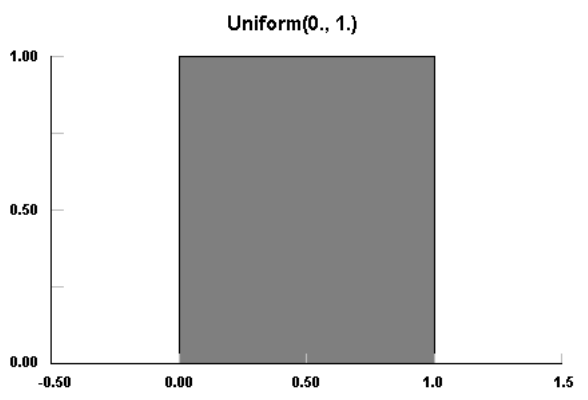
The Triangular distribution can take on very skewed forms, as shown above, including negative skewness. For the exceptional cases where the mode is either the min or max, the Triangular distribution becomes a right triangle.

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

Uniform Distribution (*min, max*)

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

min = minimum x
max = maximum x



Description:

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 rectangular distribution (see Johnson[24]). Most random number generators provide samples from the Uniform distribution on (0,1) and then convert these samples to random variates from other distributions.

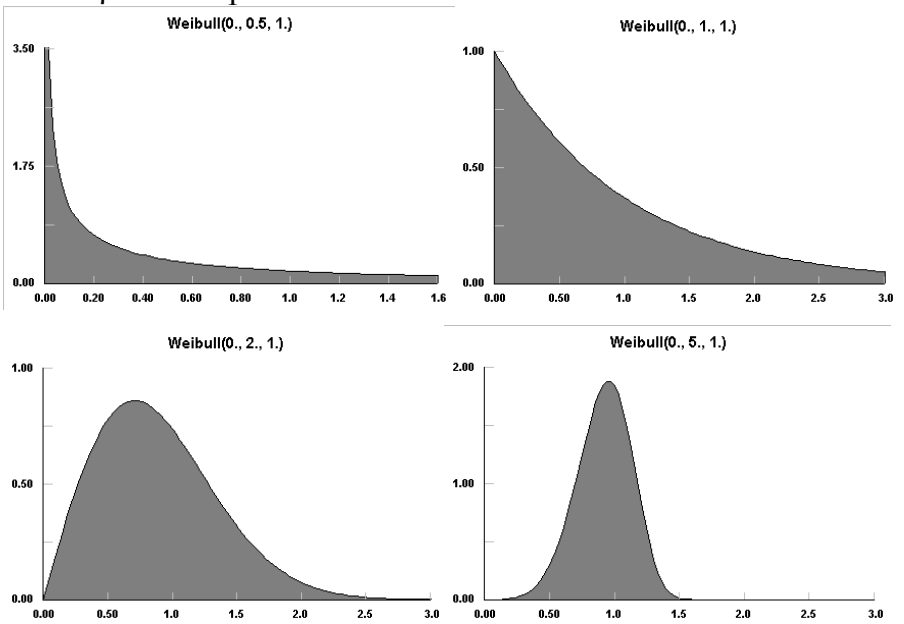
[24] "Continuous Univariate Distributions, Volume 2", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1995, John Wiley & Sons, p. 276

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.

Weibull Distribution (*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
α = shape parameter > 0
β = scale parameter > 0



Description:

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[25].

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 α and β, 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 wearout lifetimes in reliability, wind speed, rainfall intensity, health related issues, germination, duration of industrial stoppages, migratory systems, and thunderstorm data (see Johnson[26] and Shooman[27]).

[25] "Continuous Univariate Distributions, Volume 1", Norman L. Johnson, Samuel Kotz, N. Balakrishnan, 1994, John Wiley & Sons, p. 628

[26] ibid

[27] "Probabilistic Reliability: An Engineering Approach", Martin L. Shooman, 1990, Robert E. Krieger, p. 190

Appendix B – Reference Books

"An Introduction in Mathematical Statistics" H.D. Brunk, 1960, Ginn & Co.

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

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

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

"Introductory Statistical Analysis", Donald L. Harnett, James L. Murphy, 1975, Addison-Wesley

"Kendall's Advanced Theory of Statistics, Volume 1 – Distribution Theory", Alan Stuart & J. Keith Ord, 1994, Edward Arnold

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

"Seminumerical Algorithms, Volume 2", Donald E. Knuth, 1981, Addison-Wesley

"Simulation Modeling & Analysis – Third Edition", Averill M. Law, W. David Kelton, 2000, McGraw-Hill

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

"Statistical Distributions – Second Edition", Merran Evans, Nicholas Hastings, Brian Peacock, 1993, John Wiley & Sons

O

P

Q

R

S

T

U

W