Distribution Fitting (Censored Data)

statgraphics 18®
centurion

Revised: 10/10/2017

Summary

The **Distribution Fitting (Censored Data)** procedure fits any of 45 probability distributions to a column of censored numeric data. Censoring occurs when some of the data values are not known exactly. For example, when measuring failure times, some items under study may not have failed when the study is stopped, resulting in only a lower bound on the failure times for those items.

Sample StatFolio: *distfit censored.sgp*

Sample Data:

The file *absorbers.sgd* contains $n = 38$ observations identifying the number of kilometers of use for a sample of vehicle shock absorbers, taken from Meeker and Escobar (1998). When inspected, some of the shock absorbers had failed, while others had not. The table below shows a partial list of the data from that file:

distance	*censored*
6700	0
6950	1
7820	1
8790	1
9120	0
9660	1
9820	1
11310	1
11690	1
11850	1
11880	1
12140	1

The file contains 11 observations corresponding to shock absorbers that had failed. The data for those absorbers are actual failure times. The file also contains 27 shock absorbers that had not failed. That data represents right-censored information on the failure time of those absorbers, since the true distance to failure is larger than the number recorded.

When analyzing censored data, STATGRAPHICS expects you to create a column with a censoring indicator, defined by:

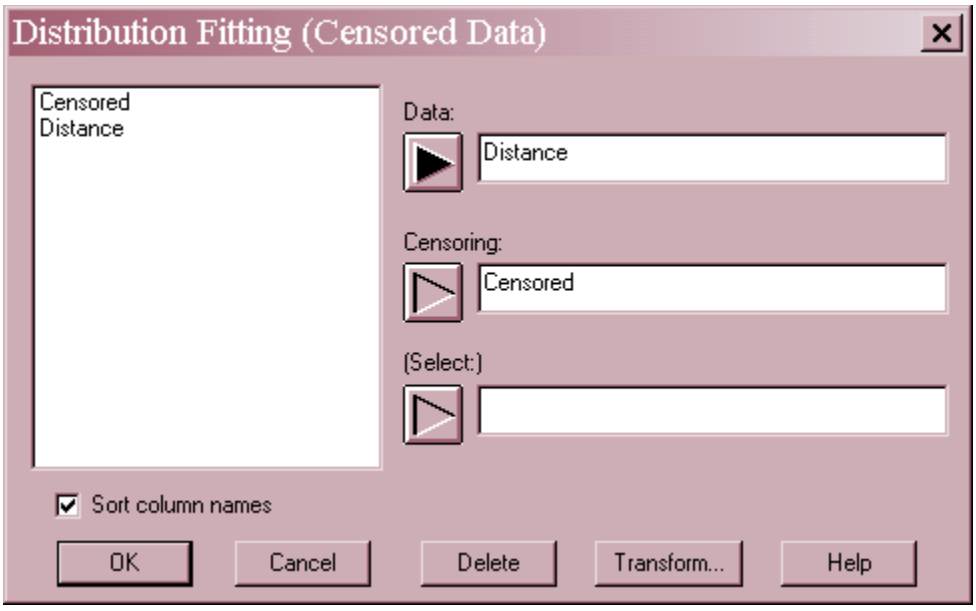
0 if the value has not been censored

1 if the value is right-censored (the actual value may be larger)

-1 if the value is left-censored (the actual value may be smaller)

Data Input

The data to be analyzed consist of a numeric column containing *n* observations and a second column containing censoring indicators.



- **Data**: column containing the *n* observations to be fit. The number of non-missing data values must be at least as large as the number of distribution parameters to be estimated.

- **Censoring**: column containing the censoring indicators. This column should contain a 0 for any row in which the data value is not censored, a 1 for any right-censored observation, and a –1 for any left-censored observation.

Analysis Summary

The *Analysis Summary* shows the number of observations, the range of the data, and the values of the estimated parameters for each distribution that is fit to the data.

Censored Data - Distance
Data variable: Distance
Censoring: Censored

38 values ranging from 6700.0 to 28100.0
Number of left-censored observations: 0
Number of right-censored observations: 27

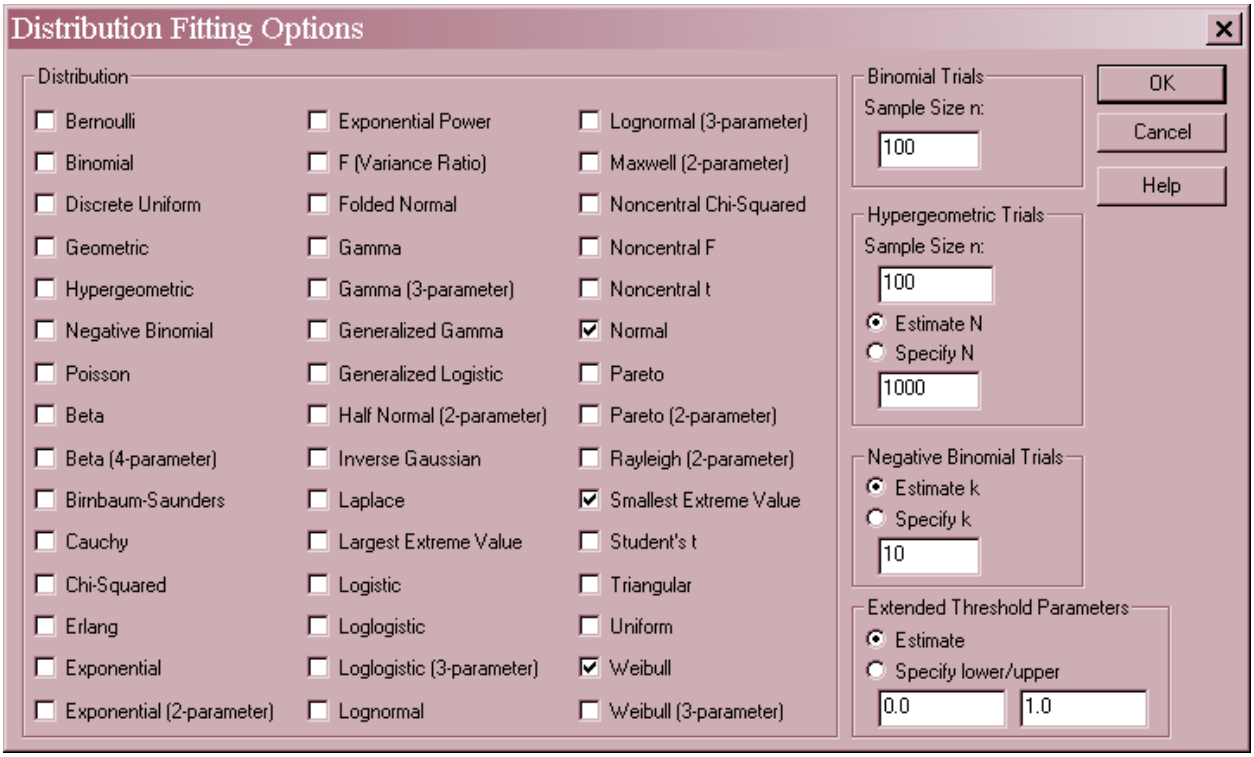
Fitted Distributions

Normal	*Smallest Extreme Value*	*Weibull*
mean = 24570.9	mode = 26896.4	shape = 3.16047
standard deviation = 8356.32	scale = 5668.58	scale = 27718.7

The parameters displayed depend upon the distributions selected (see the documentation for the *Probability Distributions* procedure). Estimates are obtained using Maximum Likelihood Estimation (MLE). You can fit between 1 and 5 distributions at the same time using *Analysis Options*.

In the above table, 3 distributions have been fit to the $n = 38$ distances. The *normal distribution* is defined by its mean and standard deviation. The *smallest extreme value distribution* is defined by its mode and scale parameter. The *Weibull distribution* is defined by a shape parameter and a scale parameter.

Analysis Options



- **Distribution**: select between 1 and 5 distributions to fit to the data. Each distribution is described in detail in the *Probability Distributions* documentation. To help determine which distributions to fit, the *Comparison of Alternative Distributions* pane described below can be extremely helpful. The following tables may also be helpful.

Discrete Distributions

Distribution	*Range of Data*	*Common Use*
Bernoulli	0 or 1	Model for event with 2 possible outcomes.
Binomial	0, 1, 2, …, *m*	Number of successes in *m* Bernoulli trials.
Discrete Uniform	a, a+1, a+2, …, b	Model for integers with fixed limits.
Geometric	0, 1, 2, …	Waiting time until first Bernoulli success.
Hypergeometric	0, 1, 2, …, *m*	Count when sampling from a finite population.
Negative Binomial	0, 1, 2, …	Waiting time until k-th Bernoulli success.
Poisson	0, 1, 2, …	Number of events in fixed interval.

Continuous Distributions

Distribution	*Range of Data*	*Common Use*
Beta	$0 \le X \le 1$	Distribution of a random proportion.
Beta (4-parameter)	$a \le X \le b$	Model for data with upper and lower thresholds.
Birnbaum-Saunders	$X > 0$	Failure times.
Cauchy	all real X	Measurement exhibiting long, flat tails.
Chi-Squared	$X \ge 0$	Reference distribution for sample variance.
Erlang	$X > 0$	Time between *k* arrivals in a Poisson process.
Exponential	$X > 0$	Time between consecutive Poisson events.
Exponential (2-parms)	$X > a$	Lifetimes with fixed lower threshold.
Exponential power	all real X	Symmetric data with variable kurtosis.
F	$X \ge 0$	Ratio of 2 independent variance estimates.
Folded Normal	$X \ge 0$	Absolute values of data from normal distribution
Gamma	$X \ge 0$	Model for positively skewed measurements.
Gamma (3-parameter)	$X \ge a$	Positively skewed data with lower threshold.
Generalized Gamma	$X > 0$	Includes several distributions as special cases.
Generalized Logistic	All real x	Used for the analysis of extreme values.
Half Normal	$X \ge \mu$	Normal data folded about its mean.
Inverse Gaussian	$X > 0$	First passage time in Brownian motion.
Laplace	all real X	Data with pronounced peak and long tails.
Largest Extreme Value	all real X	Largest value in a sample.
Logistic	all real X	Growth model; common alternative to normal.
Loglogistic	$X > 0$	Logs of data from logistic distribution.
Loglogistic (3-parms)	$X > a$	Logs of data with fixed lower threshold.
Lognormal	$X > 0$	Positively skewed data.
Lognormal (3-parameter)	$X > a$	Positively skewed data with lower threshold.
Maxwell	$X > a$	Speed of a molecule in an ideal gas.
Noncentral Chi-squared	$X \ge 0$	Calculating power of chi-squared test.
Noncentral F	$X \ge 0$	Calculating power of F test.
Noncentral t	all real X	Calculating power of t test.
Normal	all real X	Data with many sources of variability.
Pareto	$X \ge 1$	Socio-economic quantities with long upper tails.
Pareto (2-parameter)	$X \ge a$	Socio-economic quantities with lower threshold.
Rayleigh	$X > a$	Distance between neighboring items.
Smallest Extreme Value	all real X	Smallest value in a sample.
Student's t	all real X	Reference distribution for sample mean.
Triangular	$a \le X \le b$	Rough model in the absence of data.
Uniform	$a \le X \le b$	Data with equal probability over an interval.
Weibull	$X \ge 0$	Product lifetimes.
Weibull (3-parameter)	$X \ge a$	Product lifetimes with lower threshold.

- **Binomial Trials** – when fitting the binomial distribution, you must specify the sample size *n*.
- **Hypergeometric Trials** – when fitting the hypergeometric distribution, you must specify the sample size *n*. You may either specify the population size parameter *N* or estimate it from the data.

- **Negative Binomial Trials** – when fitting the negative binomial distribution, you may either specify the parameter *k* or estimate it from the data.
- **Extended Threshold Parameters** – when fitting distributions that have one or more threshold parameters, you may specify those parameters or estimate them from the data. The relevant distributions are:

  beta (4-parameter) – lower and upper
  exponential (2-parameter) – lower only
  half normal (2-parameter) – lower only
  gamma (3-parameter) – lower only
  loglogistic (3-parameter) – lower only
  lognormal (3-parameter) – lower only
  Maxwell (2-parameter) – lower only
  Pareto (2-parameter) – lower only
  Rayleigh (2-parameter) – lower only
  Weibull (2-parameter) – lower only

Goodness-of-Fit Tests

The *Goodness-of-Fit Tests* pane performs up to 7 different tests to determine whether or not the data could reasonably have come from each fitted distribution. For all tests, the hypotheses of interest are:

- Null hypothesis: data are independent samples from the specified distribution
- Alt. hypothesis: data are not independent samples from the specified distribution

The tests to be run are selected using *Pane Options*.

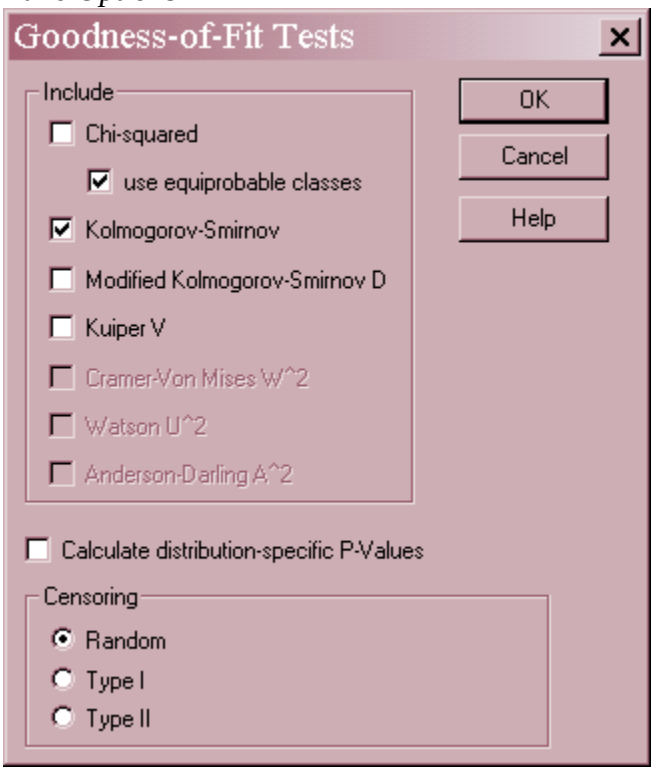
Goodness-of-Fit Tests for Distance
Modified Kolmogorov-Smirnov D

	Normal	*Smallest Extreme Value*	*Weibull*
D	0.0903629	0.122783	0.0901357
Modified Form	0.56949	0.77381	0.568059
P-Value	>=0.10	>=0.10	>=0.10

The goodness-of-fit tests are described in detail for uncensored data in the documentation for *Distribution Fitting (Uncensored Data).* For censored data, the tests are modified in a manner that depends on how the data are censored. Using *Pane Options*, you may select among 3 types of censoring: *Random*, *Type I*, or *Type II*, defined there. Modifications to the tests are described in the *Calculations* sections at the end of this document.

According to the tests displayed in the table above, any of the 3 distributions would provide a reasonable model for the data, since the P-Values all equal or exceed 0.10.

Pane Options



- **Include** – select the tests to be included. The available tests depend on the type of censoring. For the chi-squared test, select *use equiprobable classes* to group data into classes with equal expected frequencies. If this option is not checked, classes will be created that match the *Frequency Histogram*.

- **Calculate distribution-specific P-Values** – if checked, the P-Values will be based on tables or formulas specifically developed for the distribution being tested. Otherwise, the P-Values will be based on a general table or formula that applies to all distributions. The general approach is more conservative (will not reject a distribution as easily) but may be preferred when comparing P-Values amongst different distributions.

- **Censoring** – select the type of data censoring. The types are defined as:

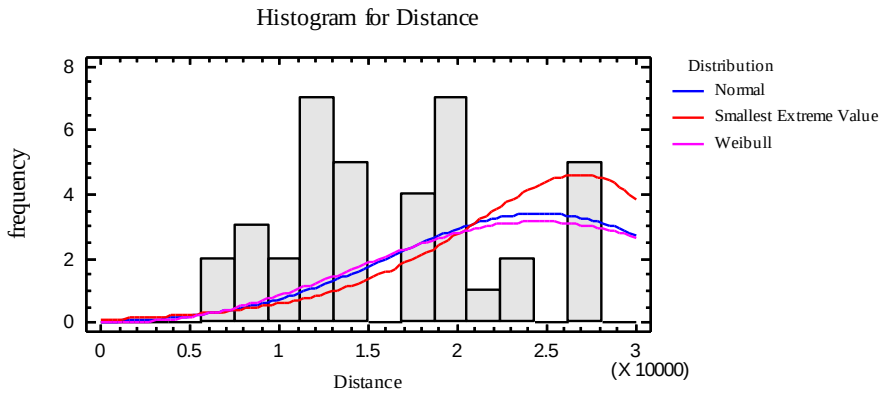
  Random - indicates that data values have been randomly censored. Random censoring occurs when values are censored for various reasons, not falling into either the Type I or Type II mechanism.

  Type I - indicates that the data are "time-censored", i.e., items have been removed from a test at a pre-specified time. If this type of censoring is selected, all of the censored values must be equal or an error message will be generated.

  Type II - indicates that the test was stopped after a predetermined number of failures had occurred. If this type of censoring is selected, all of the censored values must be equal or an error message will be generated.

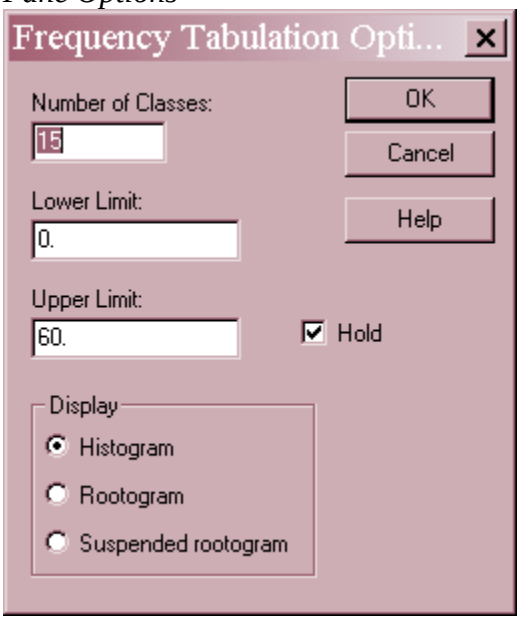
Frequency Histogram

The *Frequency Histogram* shows a histogram of the data as a set of vertical bars, together with the estimated probability density or mass functions.



If the data contain many censored observations, as in the plot above, the fitted distributions may not appear to match the bars.

Pane Options



- **Number of classes**: the number of intervals into which the data will be divided. Intervals are adjacent to each other and of equal width. The number of intervals into which the data is grouped by default is set by the rule specified on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu.

- **Lower Limit**: lower limit of the first interval.

- **Upper Limit**: upper limit of the last interval.

- **Hold**: maintains the selected number of intervals and limits even if the source data changes. By default, the number of classes and the limits are recalculated whenever the data changes. This is necessary so that all observations are displayed even if some of the updated data fall beyond the original limits.

- **Display**: the manner in which to display the frequencies. A *Histogram* scales the bars according to the number of observations in each class. A *Rootogram* scales the bars according to the square root of the number of observations. A *Suspended Rootogram* scales by the square roots and suspends the bars from the curve. The idea of using square roots is to equalize the variance of the deviations between the bars and the curve, which otherwise would increase with increasing frequency. The idea of suspending the bars from the curve is to allow an easier visual comparison with the horizontal line drawn at 0, since visual comparison with a curved line may be deceiving.

Comparison of Alternative Distributions

This pane automatically fits a collection of different distributions and displays them from top to bottom according to how well they fit the data.

Comparison of Alternative Distributions

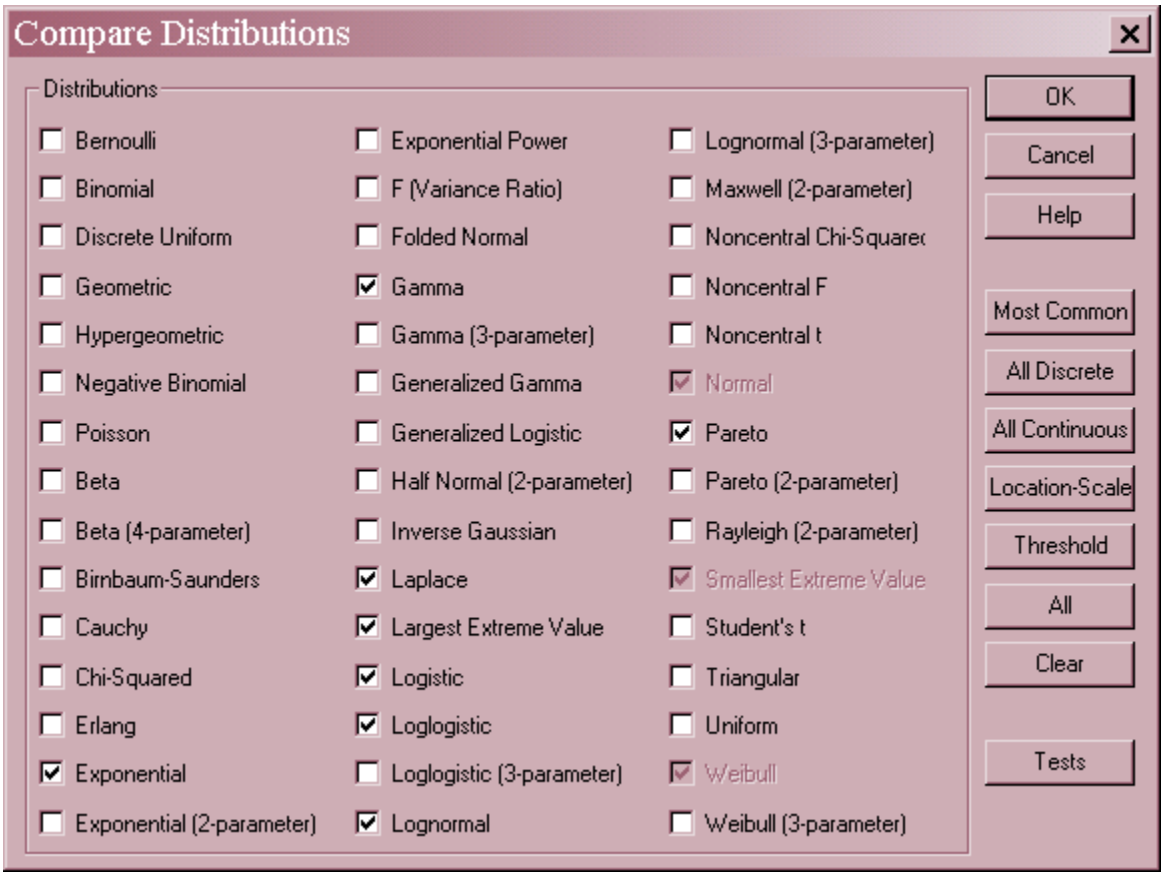
Distribution	*Est. Parameters*	*Log Likelihood*	*KS D*
Weibull	2	-404.991	0.0901357
Normal	2	-406.4	0.0903629
Logistic	2	-408.408	0.103344
Laplace	2	-413.516	0.108477
Smallest Extreme Value	2	-409.469	0.122783
Largest Extreme Value	2	-405.653	0.128409
Gamma	2	-404.845	0.128419
Loglogistic	2	-406.131	0.131113
Lognormal	2	-405.125	0.155015
Uniform	2	-400.338	0.159942
Exponential	1	-427.009	0.329046
Pareto	1	-510.249	0.448162

The table shows:

- **Distribution** – the name of the distribution fit. You may select additional distributions using *Pane Options*.
- **Est. Parameters** – the number of estimated parameters for that distribution.
- **Log Likelihood** – the natural logarithm of the likelihood function. Larger values tend to indicate better fitting distributions.
- **KS D, A^2, and other statistics** – values of various goodness-of-fit statistics, selected using the *Tests* button on the *Pane Options* dialog box. Smaller values tend to indicate better fitting distributions.

The distributions are sorted from best to worst according to one of the goodness-of-fit columns. That column is selected using the *Tests* button on the *Pane Options* dialog box. The above table shows the distributions sorted according to the value of the Kolmogorov-Smirnov D statistic. According to that statistic, the smallest extreme value distribution fits best.

Pane Options



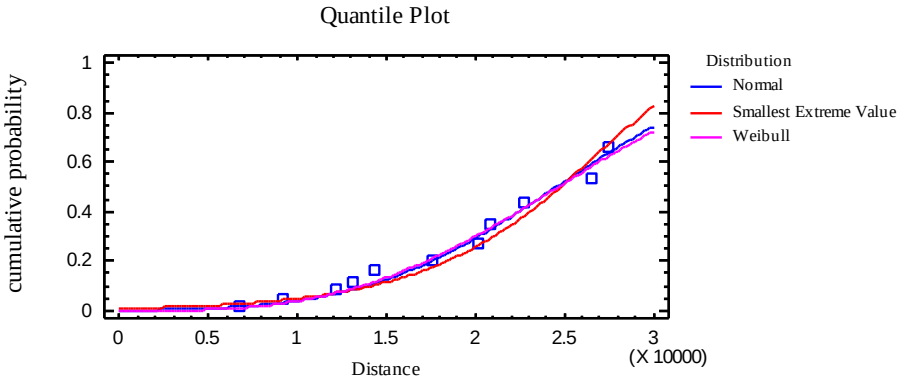
- **Distribution**: select the distributions to be fit to the data. The currently selected distributions are grayed out since they will be always included.
- **Most Common** – push this button to select the most commonly used distributions for variable (continuous) data.
- **All Discrete** – push this button to select all discrete distributions.
- **All Continuous** – push this button to select all continuous distributions.
- **Location-Scale** – push this button to select all distributions that are parameterized by a location parameter (such as a mean) and a scale parameter (such as a standard deviation).
- **Threshold** – push this button to select all distributions that contain a lower threshold parameter.
- **All** – push this button to select all distributions.
- **Clear** – push this button to deselect all distributions.
- **Tests** – push this button to display the dialog box used to specify the desired goodness-of-fit statistics:

- **Include** – select the goodness-of-fit statistics to be included in the table. The list includes the likelihood function and various statistics displayed on the *Goodness-of-Fit* pane.

- **Sort By** – select one of the included statistics to use to sort the distributions from best to worst.

Quantile Plot

The *Quantile Plot* shows the fraction of observations at or below each uncensored value of X, together with the cumulative distribution function of the fitted distributions.



To create the plot, the data are sorted from smallest to largest. The uncensored data values are then plotted at the coordinates

$$\left(\hat{F}(p_i), x_{(i)}\right) \tag{1}$$

where p_i are the Kaplan-Meier probabilities. The Kaplan-Meier probabilities are calculated according to

$$p_i = 1 - \frac{n-c+1}{n-2c+1} \prod_{\substack{j \in S_R \\ j \le i}} \left(\frac{n-j-c+1}{n-j-c+2} \right) \quad (2)$$

for all uncensored observations greater than the largest left-censored data value, where S_R is the set of all values which are not right-censored, and

$$p_i = \frac{n-c+1}{n-2c+1} \prod_{\substack{j \in S_L \\ j \ge i}} \left(\frac{j-c}{j-c+1} \right) \quad (3)$$

for all uncensored observations less than or equal to the largest left-censored data value, where S_L is the set of all values which are not left-censored, and c = 0.3175.

Ideally, the points will lie close to the line for the fitted distribution, as is the case in the plot above.

Tail Areas

This pane shows the value of the cumulative distribution at up to 5 values of X.

Tail Areas for Distance

Lower Tail Area (<=)

X	*Normal*	*Smallest Extreme Value*	*Weibull*
10000.0	0.040606	0.0494898	0.0390841
20000.0	0.29219	0.256386	0.299858
30000.0	0.74206	0.822526	0.723066
40000.0	0.967583	0.999959	0.958716
50000.0	0.998829	1.0	0.998423

Upper Tail Area (>)

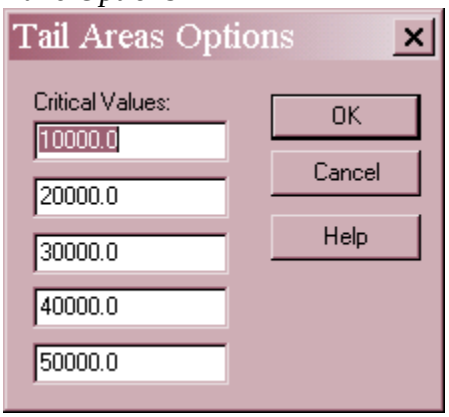
X	*Normal*	*Smallest Extreme Value*	*Weibull*
10000.0	0.959394	0.95051	0.960916
20000.0	0.70781	0.743614	0.700142
30000.0	0.25794	0.177474	0.276934
40000.0	0.0324166	0.000041464	0.0412835
50000.0	0.00117082	0.0	0.00157716

The table displays:

- **Lower Tail Area** – the probability that the random variable is less than or equal to X.
- **Upper Tail Area** – the probability that the random variable is greater than X.

For example, the probability of being less than or equal to *X* = 10,000 for the normal distribution is approximately 0.0406.

Pane Options



- **Critical Values**: values of X at which the cumulative probability is to be calculated.

Critical Values

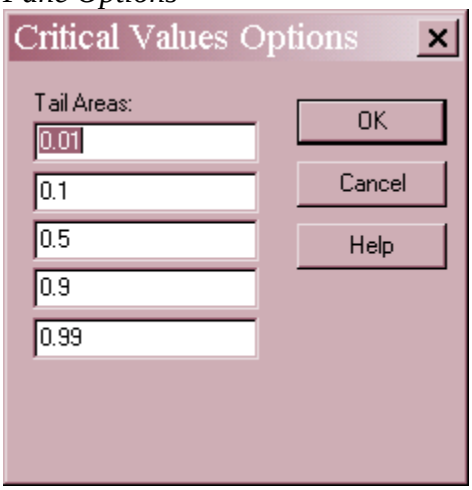
This pane calculates the value of the random variable X below which lies a specified probability.

Critical Values for Distance

Lower Tail Area (<=)	*Normal*	*Smallest Extreme Value*	*Weibull*
0.01	5131.13	820.116	6466.15
0.1	13861.8	14140.0	13600.0
0.5	24570.9	24818.8	24683.6
0.9	35279.9	31624.2	36089.5
0.99	44010.6	35553.4	44939.6

The table displays the smallest value of X such that the probability of being less than or equal to X is at least the tail area desired. The table above shows that the c.d.f. of the fitted normal distribution equals 0.01 at X = 5,131.13.

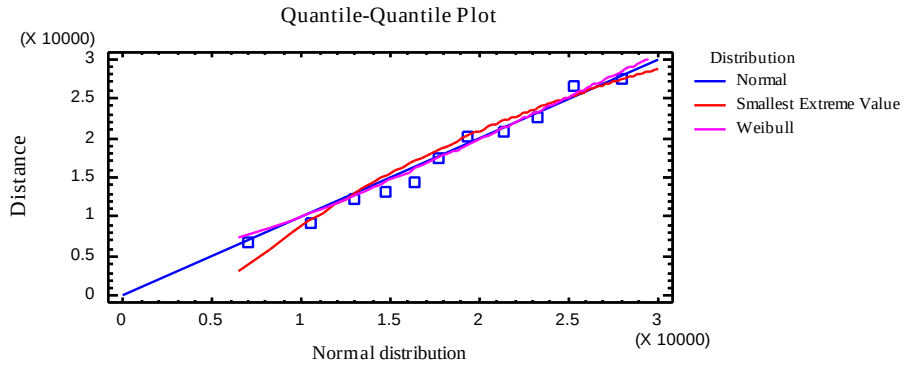
Pane Options



- **Tail Areas**: values of the c.d.f. at least to determine percentiles of the fitted distributions.

Quantile-Quantile Plot

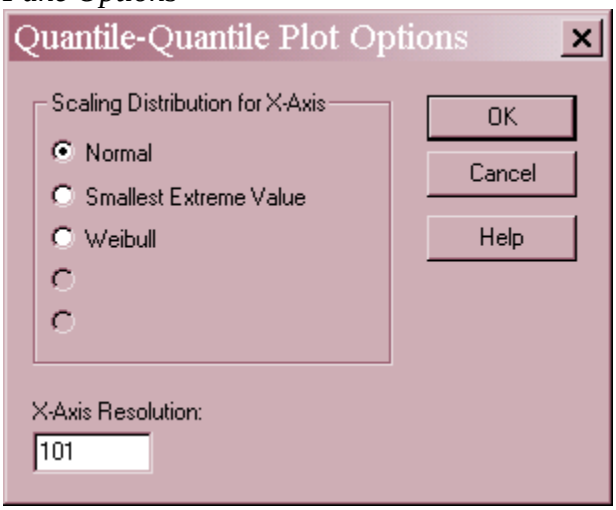
The *Quantile-Quantile Plot* shows the fraction of observations at or below X plotted versus the equivalent percentiles of the fitted distributions.



One distribution, selected using *Pane Options*, is used to define the X-axis and is represented by the diagonal line. The others are represented by curves.

In the above plot, the fitted normal distribution has been used to define the X-axis. With such a small sample, it is very hard to choose between the different distributions.

Pane Options

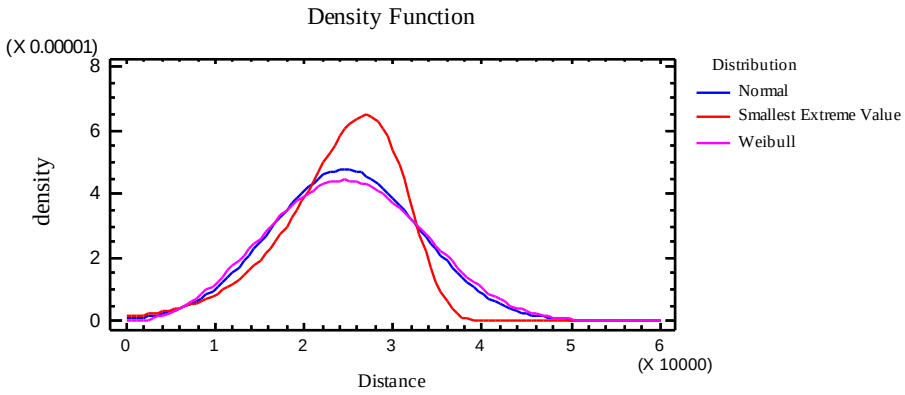


- **Scaling Distribution for X-Axis**: the distribution used to scale the horizontal axis, corresponding to the diagonal line.

- **X-Axis Resolution** – the number of X locations at which the functions are plotted. Increase this value if the lines are not smooth enough.

Distribution Functions 1 and 2

These two panes plot various functions for the fitted distributions.

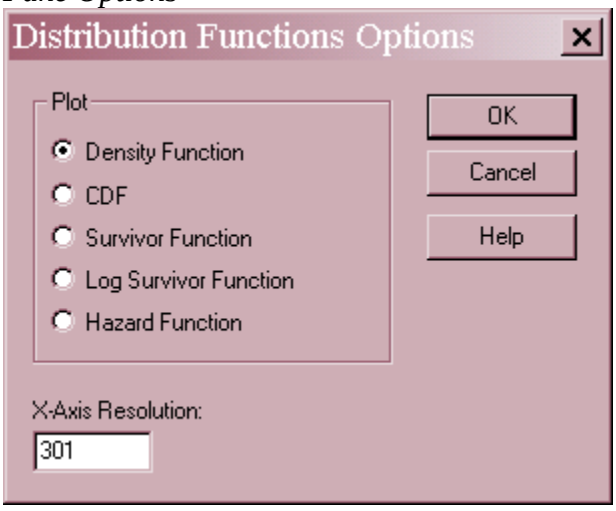


Using *Pane Options*, you may plot any of the following:

1. Probability density or mass function
2. Cumulative distribution function
3. Survivor function
4. Log survivor function
5. Hazard function

For definitions of these functions, see the documentation for *Probability Distributions*.

Pane Options



- **Plot**: the function to plot.
- **X-Axis Resolution** – the number of X locations at which the function is plotted. Increase this value if the lines are not smooth enough.

Calculations

Parameter Estimation
Parameter estimates are obtained numerically using Maximum Likelihood Estimation (MLE), where the likelihood function is given by

$$L = \prod_{i=1}^{n} l(x_i) \tag{4}$$

and

$$l(x_i) = \begin{cases} F(x_i) & left-censored \\ f(x_i) & \text{if } x_i \text{ is } uncensored \\ 1-F(x_i) & right-censored \end{cases} \tag{5}$$

Chi-Squared Test – When performing this test, after the initial intervals are constructed, all classes up to and including the class containing the largest left-censored observation are combined into a single lower class. In addition, all classes containing the smallest right-censored observation and higher values are combined into a single upper class. In some cases, this may not leave enough classes to perform the test.

EDF Tests – For the Kolmogorov-Smirnov and other EDF tests, the tests are performed by modifying the empirical c.d.f.. For *random censoring*, the Kolmogorov-Smirnov and Kuiper statistics are calculated by replacing the simple step function *i/n* by the Kaplan-Meier estimate

$$F_n(x) = 0, \qquad x < x_{(1)} \tag{6}$$

$$1 - \prod_{\substack{j \in S \\ x_{(j)} \le x}} \left(\frac{n-j}{n-j+1} \right), \qquad x_{(1)} \le x \le x_{(n)} \tag{7}$$

$$1 \qquad x > x_{(n)} \tag{8}$$

where *S* is the set of all uncensored observations. None of the other statistics are computed in this case. For *Type I and Type II censoring*, the sample of uncensored data values is converted to a complete sample over the uncensored region by modifying the fitted c.d.f. according to

$$\hat{F}^*(X_i) = \frac{\hat{F}(X_i) - A}{B - A} \tag{9}$$

For Type 1 censoring, A is the fitted c.d.f. evaluated at the lower censoring value (if any), while B is the fitted c.d.f. evaluated at the upper censoring value (if any). For Type II censoring, A is the fraction of the observations that are left-censored, and B is the fraction of the observations

that are right-censored. The usual e.d.f. formulas are then used, replacing *n* by the number of uncensored data values and letting

$$z_i = \hat{F}^*(x_i) \tag{10}$$

Save Results

The following results can be saved to the datasheet:

1. *X* – up to 5 values at which tail areas were calculated.
2. *Tail Areas* – the calculated tail areas.
3. *P* – up to 5 lower tail areas at which critical values were calculated.
4. *Critical Values* – the calculated critical values.