Weibull Analysis

Summary

The **Weibull Analysis** procedure is designed to fit a Weibull distribution to a set of *n* observations. It is commonly used to analyze data representing lifetimes or times until failure. The data may include censoring, in which some failure times are not known exactly due to removals of items from the test. The distribution is plotted and estimated percentiles are displayed.

If desired, the data for more than one group may be specified. In such cases, a separate estimate of the distribution for each group will be derived.

Sample StatFolio: *Weibull analysis.sgp*

Sample Data:

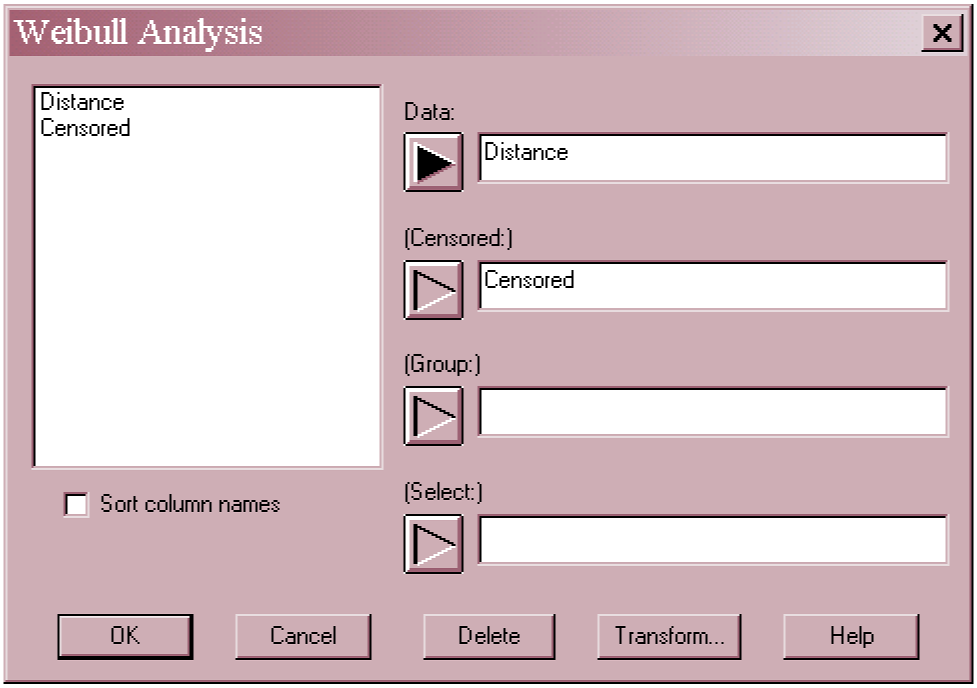
The file *absorbers.sgd* contains the data from a life test on $n = 38$ shock absorbers, reported by Meeker and Escobar (1998). A portion of the data is shown below:

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
…	…
28100	1

The *Distance* column represents the number of kilometers of use for each shock absorber when it was inspected. The *Censored* column contains a 0 for each absorber that had failed at the time of inspection and a 1 for each absorber that had not failed. All of the data contain information about the time until failure of the shock absorbers. For those that had failed, the observation is a true failure time. For those that had not failed, the observation is a right-censored failure time, with the actual time until failure known to be greater than the value indicated.

Data Input

The data input dialog box requests information about the failure times and their status:



- **Data:** numeric column with the *n* observed times.

- **Censored:** numeric column of 0's and 1's. A 0 indicates that the observation is not censored and therefore represents a true time until failure. A 1 indicates that the observation is right-censored, with the failure time known only to be greater than that indicated.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* displays a table showing the fitted Weibull distribution:

Weibull Analysis - Distance
Data variable: Distance
Censoring: Censored
Estimation method: maximum likelihood

Sample size = 38
Number of failures = 11
Estimated shape = 3.16047
Estimated scale = 27718.7
Specified threshold = 0.0

The 3-parameter Weibull distribution has a probability density function defined by:

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

It has 3 parameters:

1. *Shape* parameter $\alpha > 0$
2. *Scale* parameter $\beta > 0$
3. *Threshold* parameter θ

The range of values for the random variable $X \geq \theta$. The mean and variance of the Weibull distribution are:

$$E(X) = \theta + \frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right) \tag{2}$$

$$V(X) = \frac{\beta^2}{\alpha}\left[2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)^2\right] \tag{3}$$

Often the threshold parameter θ is set to 0, resulting in the 2-parameter Weibull distribution.

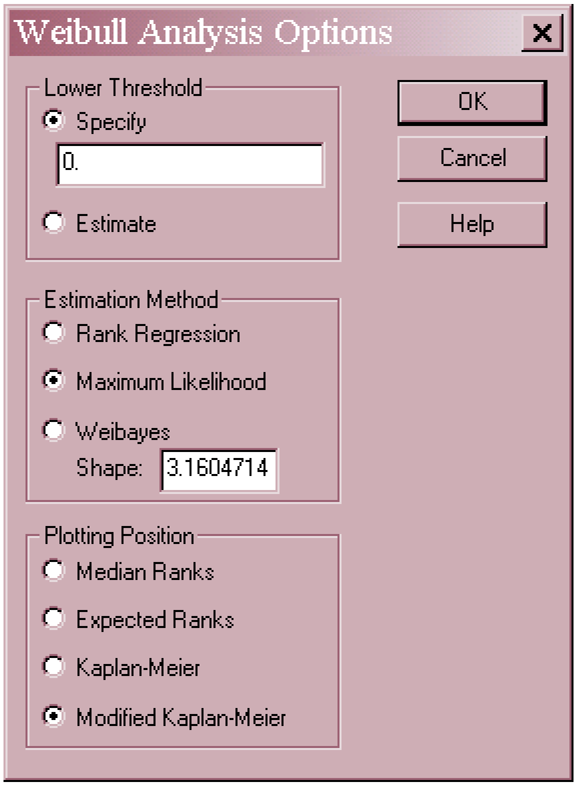
The *Analysis Summary* table shows:

- **Sample size** - the total number of observations *n*.
- **Number of failures** - the number of observations corresponding to uncensored failure times.
- **Estimated shape and scale** – the estimates of the shape and scale parameters, $\hat{\alpha}$ and $\hat{\beta}$.
- **Specified or estimated threshold** - the value of the threshold parameter. Depending upon the settings on the *Analysis Options* dialog box, this parameter may either be estimated from the data or specified by the user.

- **Method of estimation** – the method used to estimate the parameters. The default method is maximum likelihood, which is the same method used by the *Distribution Fitting* procedures, although other methods may be requested on the *Analysis Options* dialog box.

Analysis Options

The *Analysis Options* dialog box controls how the parameters of the Weibull distribution are estimated:



- **Lower Threshold:** Specify a value for the threshold parameter θ, or request that it be estimated from the data. If estimating θ, you must select the maximum likelihood method.

- **Estimation Method:** the method used to estimate the parameters. Three choices are available:

  1. *Rank Regression* – Fits a line to the data on the *Weibull Plot* by regressing n values $ln(X_i-\theta)$ against the plotting positions specified on the dialog box.

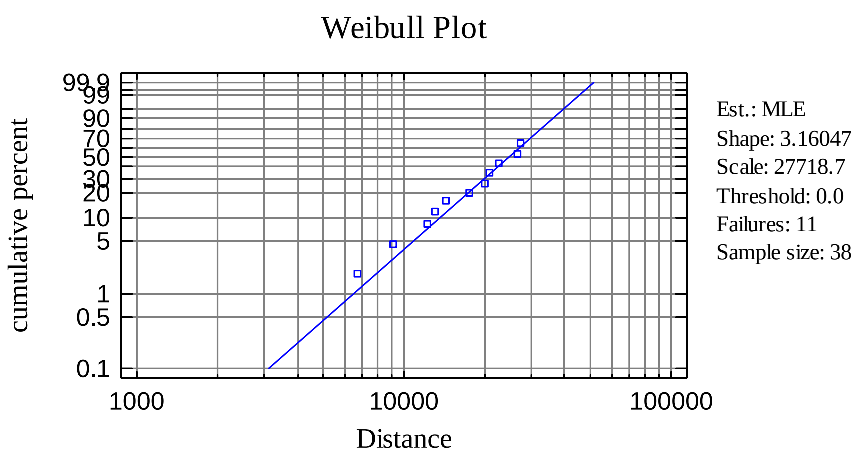
  2. *Maximum likelihood* - Estimates the parameters by maximizing the likelihood function.

  3. *Weibayes* - Estimates the scale parameter assuming that both the threshold and shape parameters are known and equal to the values indicated on the dialog box.

- **Plotting Position** - defines the vertical plotting positions on the Weibull plot. This also affects the estimated parameters if the *Rank Regression* method has been selected. The default method is *Modified Kaplan-Meier*. See the *Calculations* section for definitions of the different options.

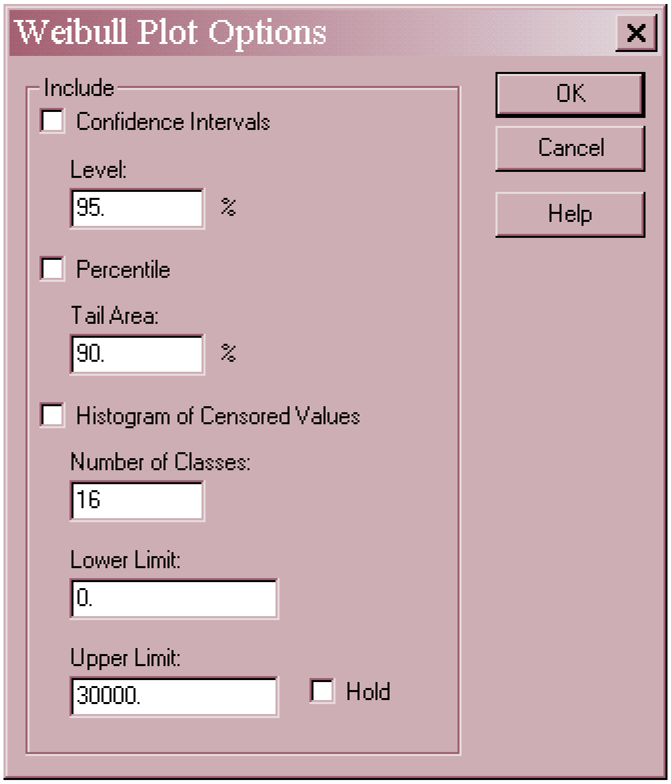
Weibull Plot

The *Weibull Plot* shows the uncensored failure times plotted on a logarithmically scaled horizontal X axis, with the plotting positions on the vertical axis dependent on *Analysis Options*.



If the data come from a Weibull distribution, the points should fall approximately along a straight line on this plot, which corresponds to the fitted Weibull distribution.

Pane Options

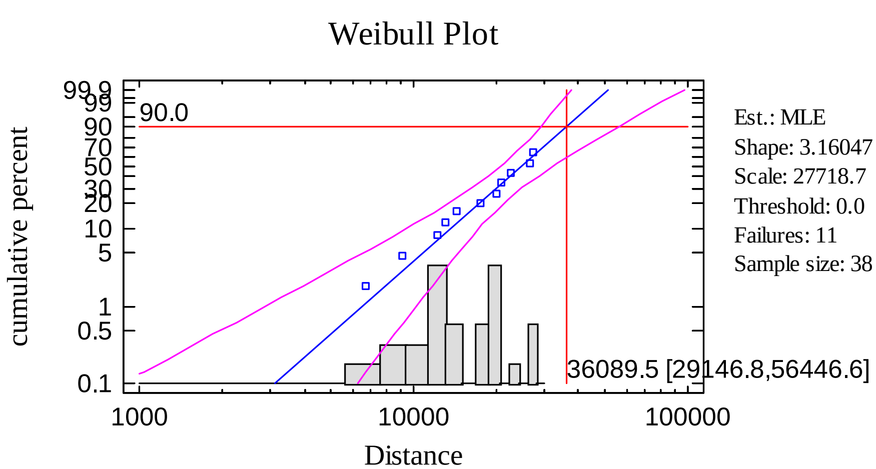


- **Confidence Intervals** - adds likelihood ratio confidence intervals for X_p, the p-th percentile of the Weibull distribution.

- **Percentile** - calculates and displays the selected percentile.

- **Histogram of Censored Values** – adds a histogram of censored data values.

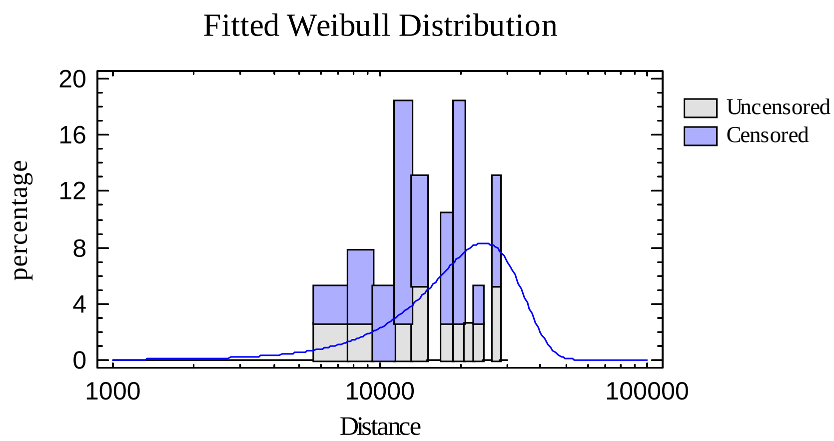
Example: Plot Showing Confidence Bands, 90th Percentile, and Censored Histogram



The 90th percentile of *Distance* is estimated to be approximately 36,090 km. The 95% confidence limits range from 29,147 km to 56,447 km, corresponding to the bands around the fitted line.

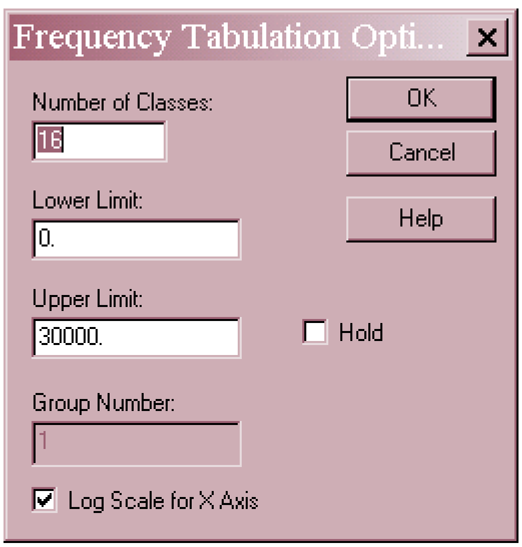
Frequency Histogram

The *Frequency Histogram* shows the distribution of the observations, together with the estimated density function:



The presence of many censored values causes the curve to lie to the right of the observed histogram.

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 (before the logarithm is calculated). 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.

- **Group Number:** if more than one group of data has been fit, the number of the group to be displayed.

- **Log Scale for X Axis**: whether the horizontal axis should be displayed using logarithmic scaling.

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 a Weibull distribution. For all tests, the hypotheses of interest are:

- **Null hypothesis**: data are independent samples from the estimated Weibull distribution

- **Alt. hypothesis**: data are not independent samples from the estimated Weibull distribution

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

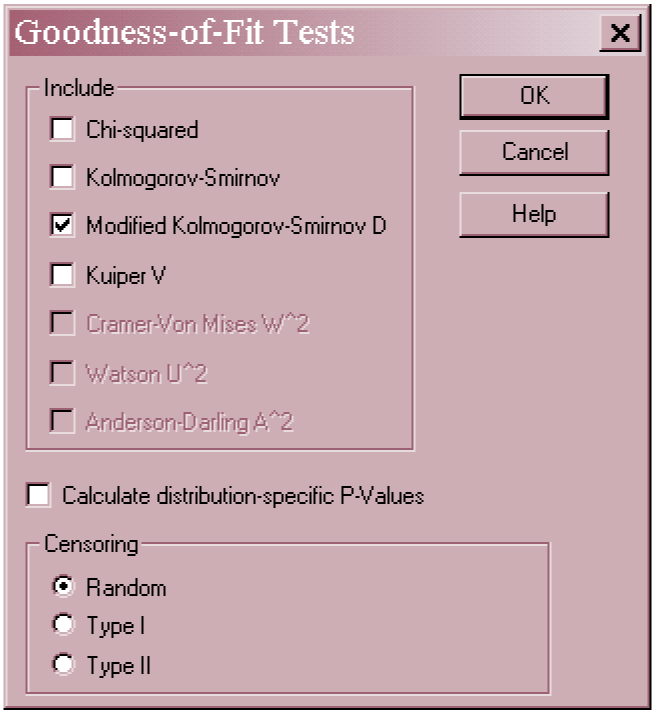
Goodness-of-Fit Tests for Distance
Modified Kolmogorov-Smirnov D

	Weibull
D	0.0901357
Modified Form	0.568059
P-Value	>=0.10

Small P-Values (less than 0.05 if operating at the 5% significance level) lead to a rejection of the Weibull distribution. In the current example, the P-Value is large, suggesting that the Weibull distribution is a reasonable model for the data.

The goodness-of-fit tests are described in detail for uncensored in the documentation for *Distribution Fitting (Uncensored Data)* and for censored data in *Distribution Fitting (Censored Data*.

Pane Options



- **Include** – select the tests to be included. The available tests depend on the type of censoring.

- **Calculate distribution-specific P-Values** – if checked, the P-Values will be based on tables or formulas specifically developed for the Weibull distribution. 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:

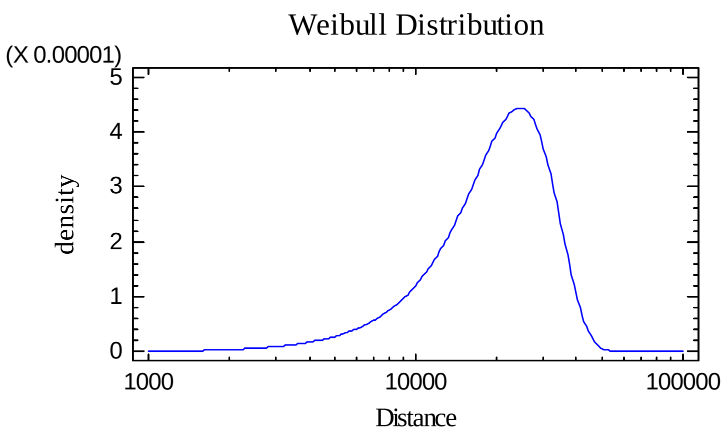
  Random - indicates that data values have been randomly censored. Random censoring occurs when values are censored for various reasons, not falling into either the Type I or Type II mechanism.

  Type I - indicates that the data are "time-censored", i.e., items have been removed from the 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.

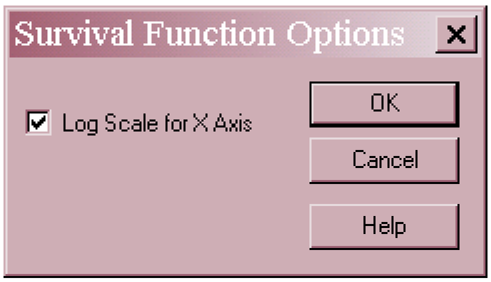
Density Function

This plot shows the estimated probability density function f(x):



The density function can be used to determine the probability that an item's failure time will lie within a specified interval.

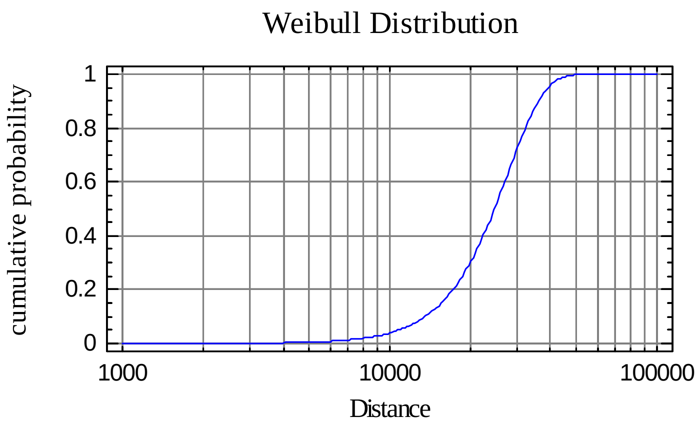
Pane Options



- **Log Scale for X-Axis**: if selected, a log scale will be used for the horizontal axis.

CDF

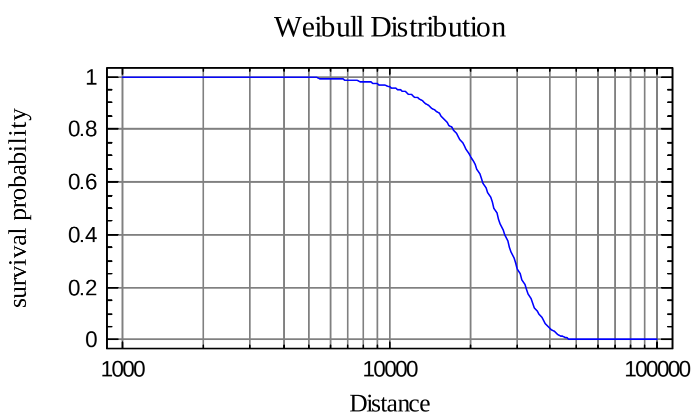
The *Cumulative Distribution Function* (CDF) shows the estimated probability that an item will have failed by time *t*:



It increases from 0.0 at θ to 1.0 at large values of X.

Survival Function

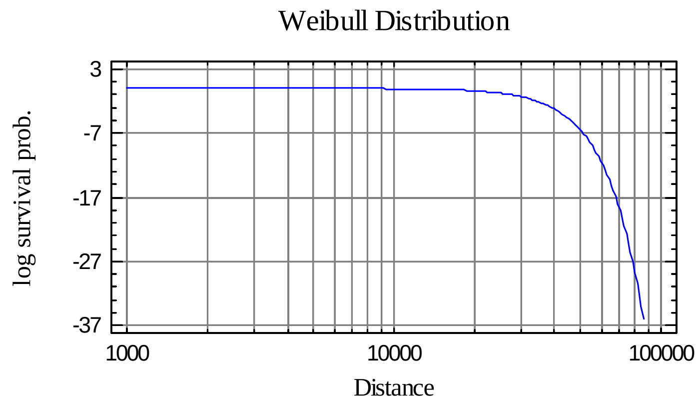
The *Survival Function* plots the estimated probability that an item will survive until time *t*:



It decreases from 1.0 at θ to 0.0 at large values of X.

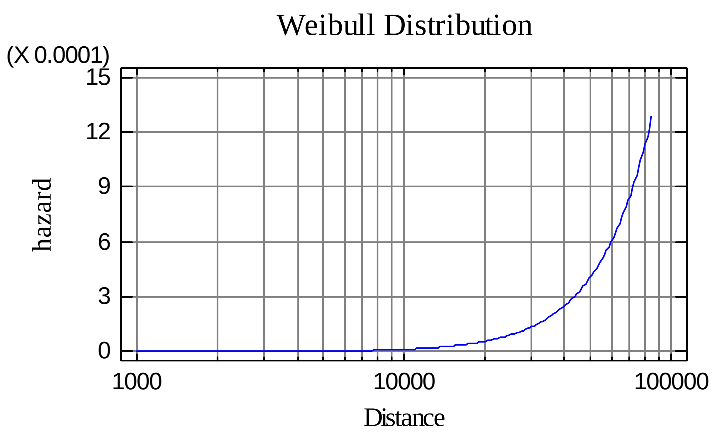
Log Survival Function

The *Log Survival Function* is the natural logarithm of the survival function:



Hazard Function

The *Hazard Function* is an estimate of the instantaneous rate of failure:



The units of the hazard function are the fraction of items failed per unit time. If the estimated shape parameter $\hat{\beta} < 1$, the hazard function will be monotonically decreasing. If $\hat{\beta} > 1$, the hazard function will be monotonically increasing. If $\hat{\beta} = 1$, the hazard function will be constant.

Tail Areas

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

Tail Areas for Distance

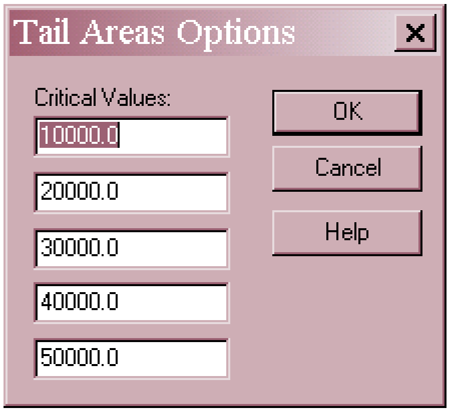
X	*Lower Tail Area (<)*	*Upper Tail Area (>)*
10000.0	0.0390841	0.960916
20000.0	0.299858	0.700142
30000.0	0.723066	0.276934
40000.0	0.958716	0.0412835
50000.0	0.998423	0.00157716

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* = 30,000 is approximately 72.3%.

Pane Options



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

Critical Values

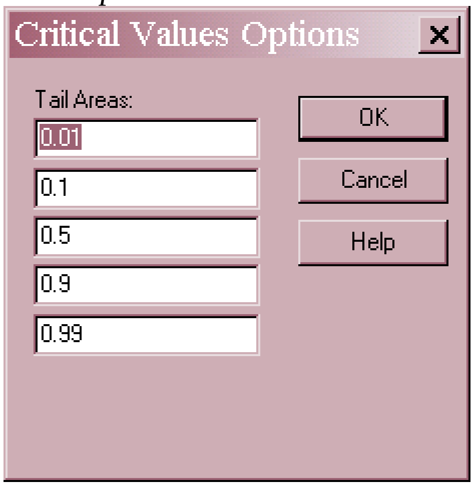
This pane calculates the value of the random variable X below which lies a specified probability.

Critical Values for Distance

X	*Lower Tail Area (<)*	*Upper Tail Area (>)*
6466.15	0.01	0.99
13600.0	0.1	0.9
24683.6	0.5	0.5
36089.5	0.9	0.1
44939.6	0.99	0.01

The table displays the value of X such that the probability of being less than or equal to X equals the tail area desired. The table above shows that the c.d.f. of the fitted Weibull distribution equals 50% at *X* = 24,683.6.

Pane Options



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

Calculations

Estimation Using Rank Regression Method

Regress $ln(X_i\text{-}\theta)$ against the plotting positions specified by *Analysis Options*. The shape and scale parameters are estimated from the intercept and slope of the fitted line according to:

$$\hat{\alpha} = \frac{1}{slope} \tag{4}$$

$$\hat{\beta} = exp(intercept) \tag{5}$$

Estimation Using Maximum Likelihood Method

The estimates are obtained by numerical maximization of the likelihood function

$$L = \prod_{i=1}^{n} l(X_i) \tag{6}$$

where

$$l(X_i) = \begin{cases} f(X_i) & \text{if } X_i \text{ is } \quad uncensored \\ 1 - F(X_i) & \phantom{\text{if } X_i \text{ is }} \quad right-censored \end{cases} \tag{7}$$

Estimation Using Weibayes Method
If there are no failures, the scale parameter is estimated by

$$\hat{\beta} = \left\{ \frac{\sum_{i=1}^{n} (X_i - \theta)^{\alpha}}{-\ln(0.05)} \right\}^{1/\alpha} \tag{8}$$

which is a lower 95% confidence bound for β. Otherwise,

$$\hat{\beta} = \left[\frac{\sum_{i=1}^{n} (X_i - \theta)^{\alpha}}{d} \right]^{1/\alpha} \tag{9}$$

where d is the number of uncensored failure times.

Plotting Positions: Median ranks
The data are ranked from smallest to largest and assigned an adjusted rank $j_i = j_{i-1} + \Delta$, where Δ = 1 initially and is modified at each censoring time according to

$$\Delta = \frac{\text{(n+1) - adjusted rank of previous failure}}{\text{1 + number of items beyond previous censored item}} \tag{10}$$

The plotting positions are given by

$$F_i = \frac{j_i - 0.3}{n + 0.4} \tag{11}$$

Plotting Positions: Expected ranks
The data are first ranked in reverse order, i.e., $r_1 = n, \ldots, r_n = 1$. Reliabilities are computed from

$$R_i = \left[\frac{r_i}{r_i + 1} \right] R_{i-1} \tag{12}$$

where R_0=1, and the plotting positions are $F_i = (1-R_i)$.

Plotting Positions: Kaplan-Meier
Similar to expected ranks, except that

$$R_i = \left[\frac{r_i - 1}{r_i}\right] R_{i-1} \tag{13}$$

Plotting Positions: Modified Kaplan-Meier
Using the Kaplan-Meier reliabilities,

$$R_{i_i}^{'} = \frac{R_i + R_{i-1}}{2} \tag{14}$$

where R_0=1, and the plotting positions are $F_i = 1 - R_{i_i}^{'}$.

Confidence Limits
The confidence limits are computed on a pointwise basis by determining all values $X = Q + \theta$ for which

$$\Lambda = -2\log L(\tilde{\alpha}, \tilde{\beta}) + 2\log L(\hat{\alpha}, \hat{\beta}) \le \chi^2_{1-p,1} \tag{15}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators, while $\tilde{\alpha}$ and $\tilde{\beta}$ satisfy the equations

$$\frac{r}{\tilde{\beta}} - r\ln(Q) + \sum_{i \in S} \ln(X_i - \theta) + \ln(1-p) \sum_{i=1}^{n} \left(\frac{(X_i - \theta)}{Q}\right)^{\tilde{\beta}} \ln\left(\frac{(X_i - \theta)}{Q}\right) = 0 \tag{16}$$

$$\tilde{\alpha} = \frac{Q}{\left[-\ln(1-p)\right]^{1/\tilde{\beta}}} \tag{17}$$

and S is the set of the *d* failure times.

