

***Distribution Fitting (Arbitrarily Censored Data)***



Revised: 10/10/2017



Summary .....	1
Data Input.....	3
Analysis Options.....	5
Tables and Graphs.....	6
Analysis Summary .....	7
Scatterplot .....	8
Distribution Fitting.....	9
Plot of Fitted Distribution .....	10
Nonparametric Estimates .....	12
Box and Whisker Plot.....	13
Quantile-Quantile Plot .....	14
Cumulative Distribution.....	15
Survival Function.....	16
Save Results .....	17
Calculations.....	18
References.....	18

**Summary**

The **Distribution Fitting (Arbitrarily Censored Data)** procedure analyzes data in which one or more observations are not known exactly. In particular, observations may be:

1. **Left-censored:** known only to be less than a stated value.
2. **Right-censored:** known only to be greater than a stated value.
3. **Interval censored:** known only to fall within a stated interval.

The procedure calculates summary statistics, fits distributions, creates graphs, and calculates a nonparametric estimate of the survival function.

**Sample StatFolio:** *censored onevar.sgp*

## Sample Data

The file *bcos.sgd* file contains data from a study on breast cancer reported by Finkelstein and Wolfe (1985). It consists of data from 94 breast cancer patients who were randomly given either radiation therapy with chemotherapy or radiation therapy alone. The variable to be analyzed is the time between treatment and the onset of breast retraction. Times for patients observed with breast retraction at their first clinic visit are left-censored. Times for patients observed with no breast retraction at their last clinic visit are right-censored. Times for patients observed with no breast retraction at one clinic visit and with breast retraction at the next clinic visit are interval censored.

A portion of the file is shown below:

<i>Patient</i>	<i>Days</i>	<i>Left</i>	<i>Right</i>	<i>Treatment</i>
1	>45	45		Rad
2	[6,10]	6	10	Rad
3	<7	0	7	Rad
4	>46	46		Rad
5	>46	46		Rad
6	[7,16]	7	16	Rad
7	>17	17		Rad
8	[7,14]	7	14	Rad
9	[37,44]	37	44	Rad
10	<8	0	8	Rad

The length of time until breast retraction was observed is shown in 2 ways:

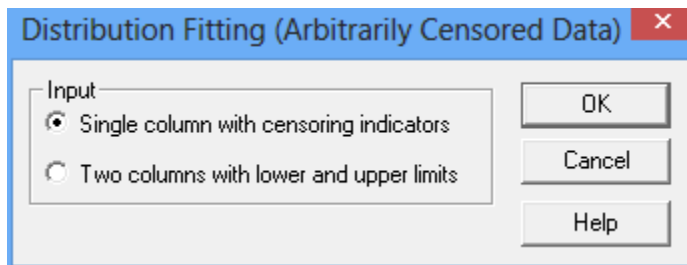
1. The column *Days* is a censored numeric column which represents the data using a notation such as:
  - a. >45 for right-censored data, meaning that retraction was not observed up to and including the last clinic visit at 45 days.
  - b. [6,10] for interval censored data, meaning that retraction was not observed during a visit at 6 days after treatment but was observed during a visit at 10 days.
  - c. <7 for left-censored data, meaning that retraction was observed at the first visit occurring 7 days after treatment.

- The columns *Left* and *Right* are the endpoints of the interval during which retraction occurred for each patient.

The data may be entered in either format.

## Data Input

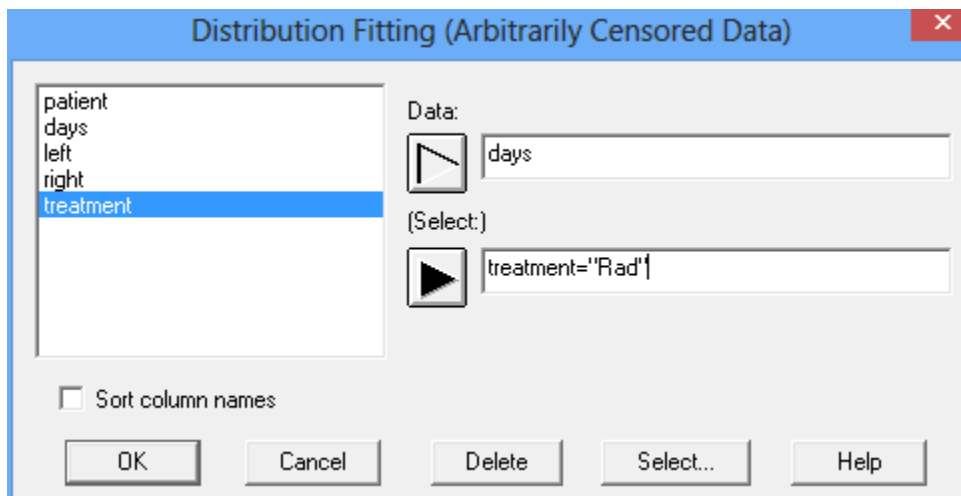
When the procedure is first selected, a dialog box is displayed requesting the format of the data to be analyzed:



- Single column with censoring indicators:** the data to be analyzed are in a single column of type “censored numeric”.
- Two columns with lower and upper limits:** two columns are supplied with the lower and upper limits for each observation. For left-censored data, the lower limit may be blank. For right-censored data, the upper limit may be blank.

### Single column with censoring indicators

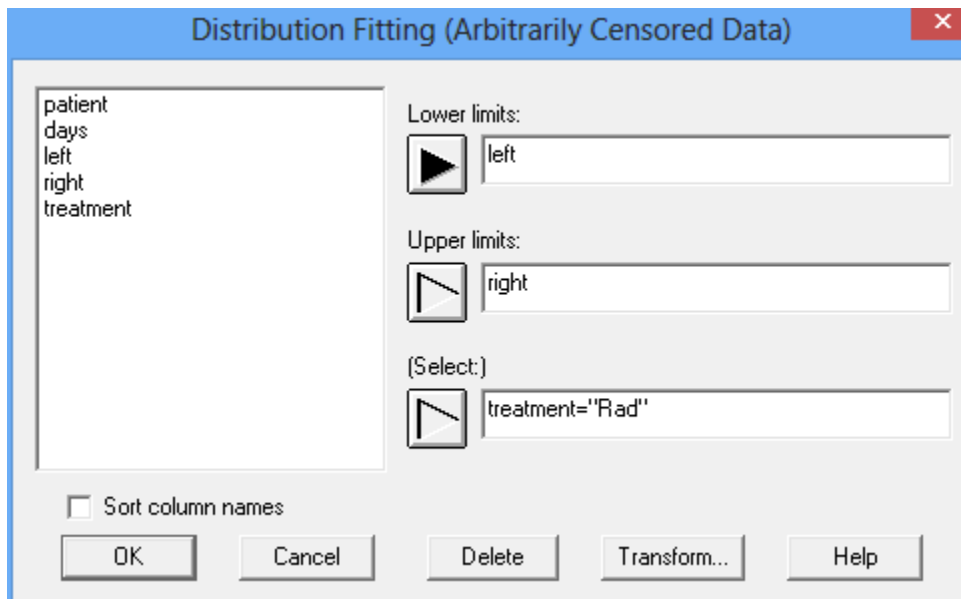
For a single column with censoring indicators, a second dialog box requests the name of the column:



The data column must have the “censored numeric” type.

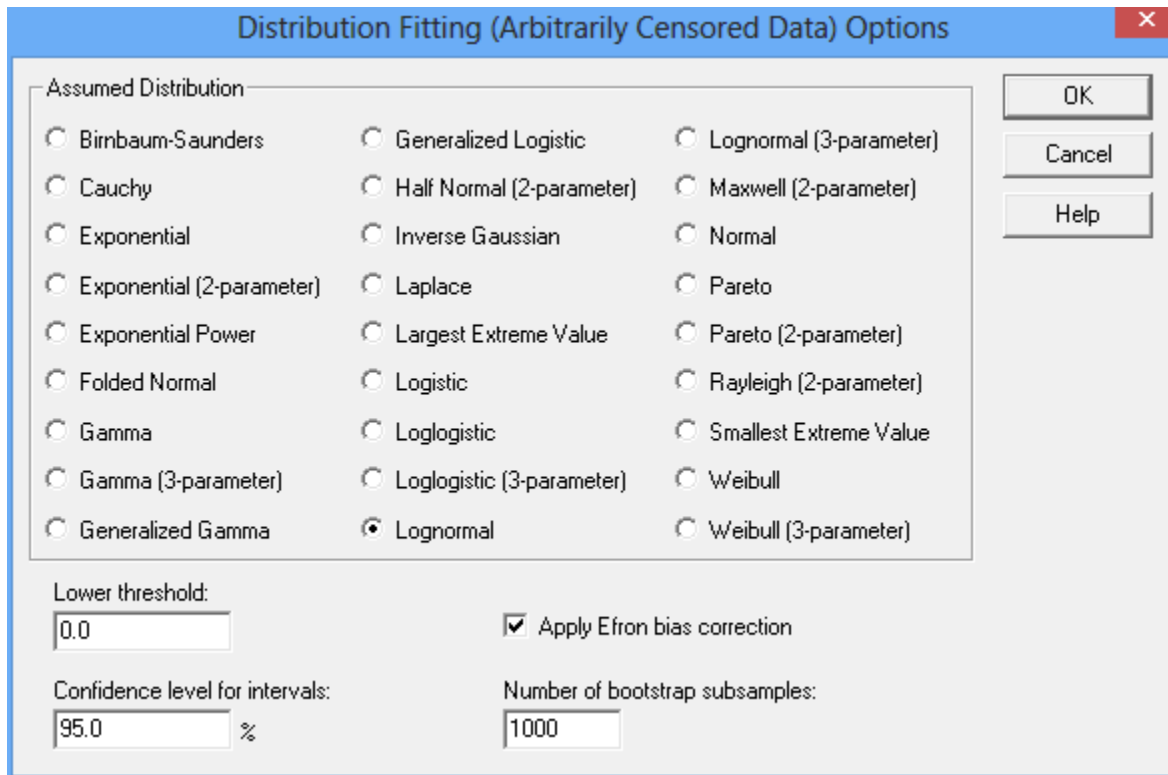
### Two columns with lower and upper limits

If the data consist of lower and upper limits, a second dialog box requests two columns containing the limits:



## Analysis Options

After the data have been specified, the *Analysis Options* dialog box is displayed:



**Distribution Fitting (Arbitrarily Censored Data) Options**

Assumed Distribution

- Birnbaum-Saunders
- Cauchy
- Exponential
- Exponential (2-parameter)
- Exponential Power
- Folded Normal
- Gamma
- Gamma (3-parameter)
- Generalized Gamma
- Generalized Logistic
- Half Normal (2-parameter)
- Inverse Gaussian
- Laplace
- Largest Extreme Value
- Logistic
- Loglogistic
- Loglogistic (3-parameter)
- Lognormal
- Lognormal (3-parameter)
- Maxwell (2-parameter)
- Normal
- Pareto
- Pareto (2-parameter)
- Rayleigh (2-parameter)
- Smallest Extreme Value
- Weibull
- Weibull (3-parameter)

Lower threshold:

Confidence level for intervals:  %

Number of bootstrap subsamples:

Apply Efron bias correction

OK  
Cancel  
Help

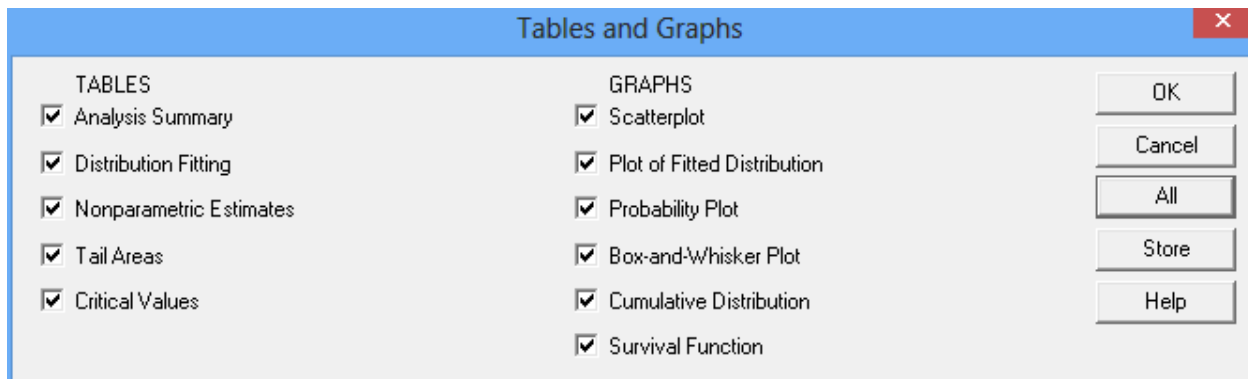
- **Assumed Distribution:** select a distribution to fit to the data.
- **Lower threshold:** when fitting distributions that have a lower threshold that is not estimated from the data, you may specify a threshold other than 0. The relevant distributions are:
  - exponential
  - gamma
  - half normal
  - loglogistic
  - lognormal
  - Maxwell
  - Pareto
  - Rayleigh
  - Weibull

In cases such as the gamma distribution where there are 2 forms (2-parameter and 3-parameter), selecting the higher parameter form will cause the lower threshold to be estimated from the data rather than specified by the user.

- **Apply Efron bias correction:** if selected and the smallest observation is left-censored, sets the KMT nonparametric CDF at that observation to 0 for purposes of calculating the mean and standard deviation. Otherwise, the CDF is assumed to go to 0 at the lower threshold linearly.
- **Number of bootstrap subsamples:** number of subsamples
- to be used when estimating confidence limits for the distribution parameters and other quantities. Setting a larger value will give more accurate estimates but may increase execution time significantly.
- **Confidence level for intervals:** confidence level used to create confidence limits for distribution parameters and other quantities.

## Tables and Graphs

The following tables and graphs may be created:



## Analysis Summary

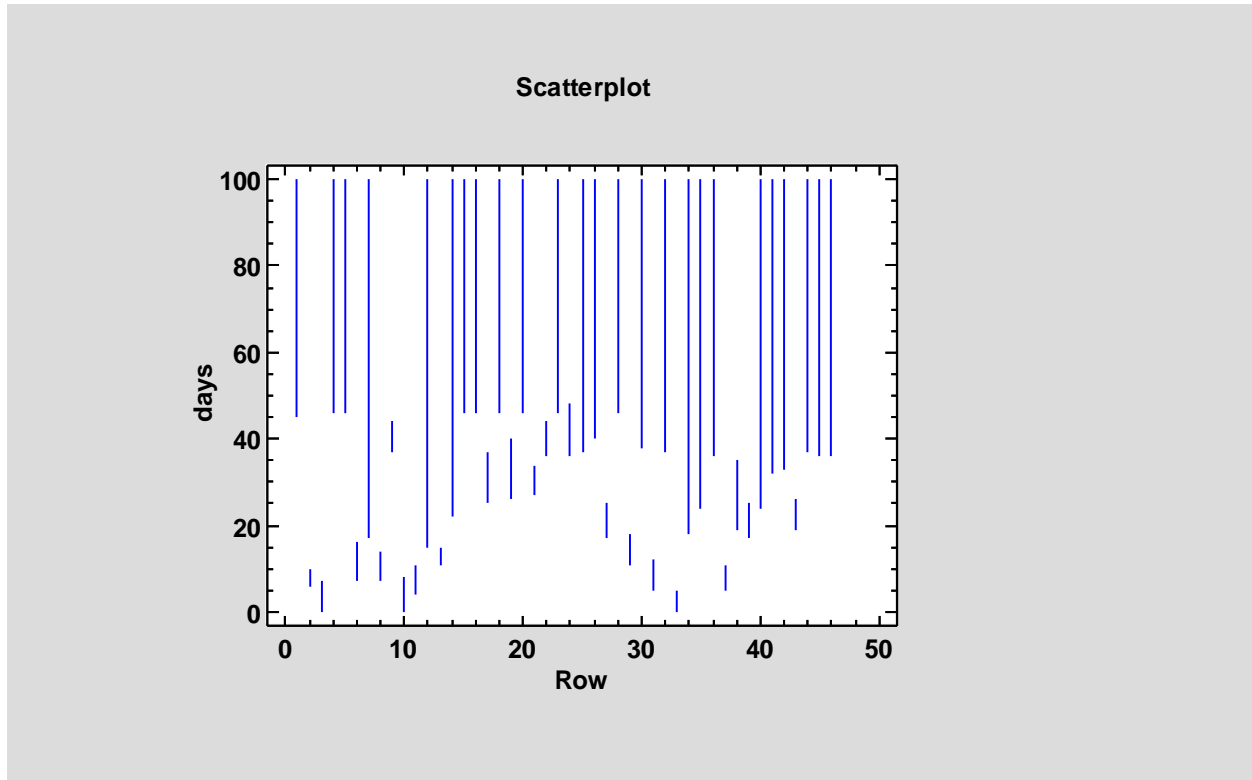
The *Analysis Summary* displays the data to be analyzed:

<u>Distribution Fitting (Arbitrarily Censored Data) - days (treatment="Rad")</u>	
Data variable: days	
Selection variable: treatment="Rad"	
Observations	
<i>Value</i>	<i>Frequency</i>
Uncensored	0
Left-censored	3
Interval-censored	18
Right-censored	25

It shows the number of observations with each type of censoring. For patients in the sample data set that were given radiation only, there are 3 left-censored observations, 18 interval-centered observations, and 25 right-censored observations.

## Scatterplot

This plot shows each of the observations. Exact observations are represented as points. Censored observations are represented by lines covering the range of possible values for that observation.



The 3 left-censored observations are shown as vertical lines extending upward from 0. The 18 interval-censored observations extend from their lower limits to their upper limits. The 25 right-censored observations extend from their lower limits to the top of the graph.



## Distribution Fitting

This table shows the results of fitting the selected distribution to the data:

<b>Distribution Fitting</b>			
Fitted distribution: Lognormal			
<i>Parameter</i>	<i>Estimate</i>	<i>95% LCL</i>	<i>95% UCL</i>
Mean	100.514	48.435	494.069
Std. Dev.	214.7	65.6497	3543.7
Distribution Properties			
	<i>Estimate</i>	<i>95% LCL</i>	<i>95% UCL</i>
Mean	100.514	48.435	494.069
Standard deviation	214.7	65.6497	3543.7
Median	42.6171	27.4519	81.4824
Lower quartile	17.6137	11.3238	28.4698
Upper quartile	103.114	56.6752	279.959
Interquartile range	85.5001	42.2701	257.812
Number of bootstrap subsamples: 1000			

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

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

where

$$l(x_i) = f(x_i) \text{ if the observation } x_i \text{ is uncensored} \quad (2)$$

$$l(x_i) = F(L_i) \text{ if the observation is left-censored at } L_i \quad (3)$$

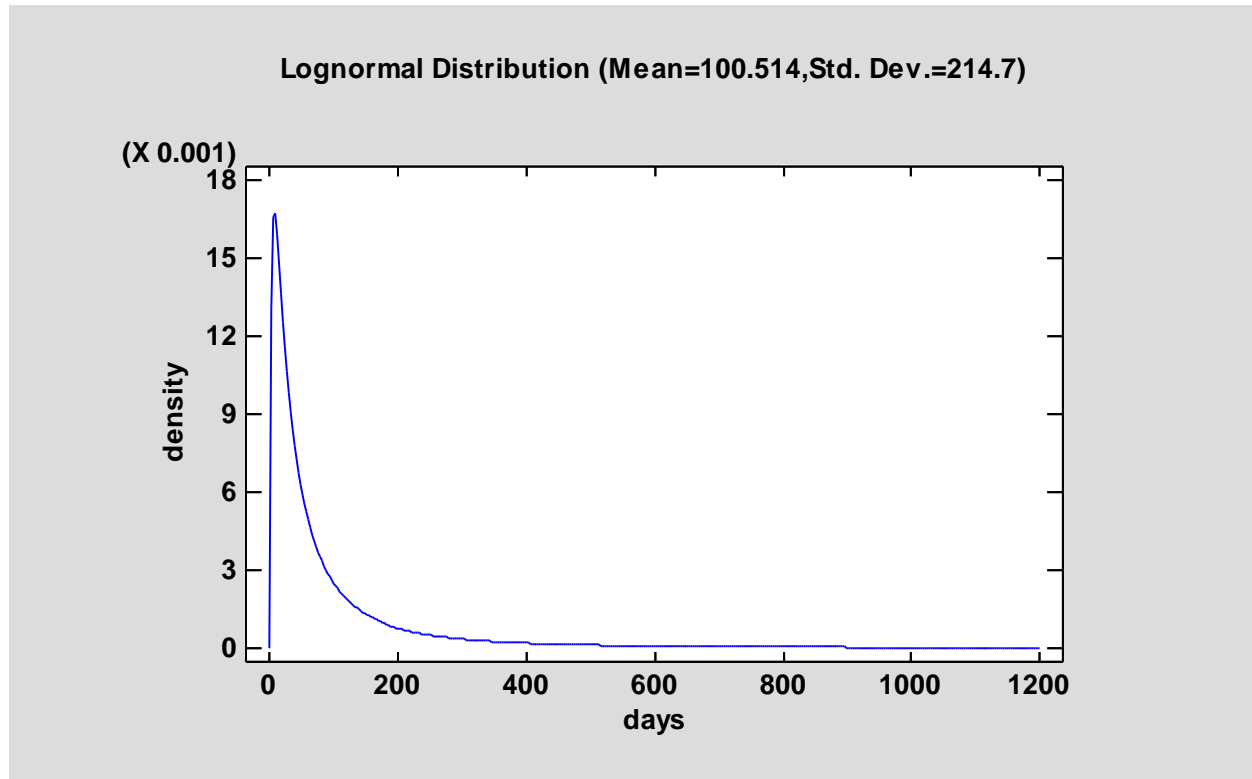
$$l(x_i) = 1 - F(U_i) \text{ if the observation is right-censored at } U_i \quad (4)$$

$$l(x_i) = F(U_i) - F(L_i) \text{ if the observation is interval-censored between } [L_i, U_i] \quad (5)$$

The lower and upper confidence limits for the parameters and related quantities are estimated using bootstrapping based on the number of subsamples shown.

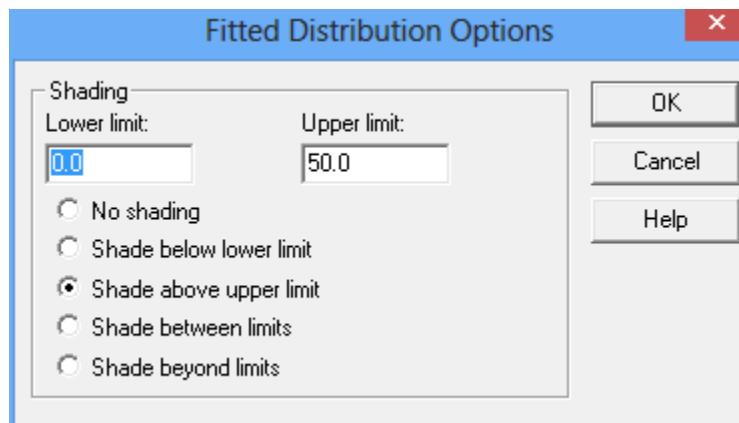
## Plot of Fitted Distribution

The fitted distribution may be plotted by selecting *Plot of Fitted Distribution* from the list of selected tables and graphs.



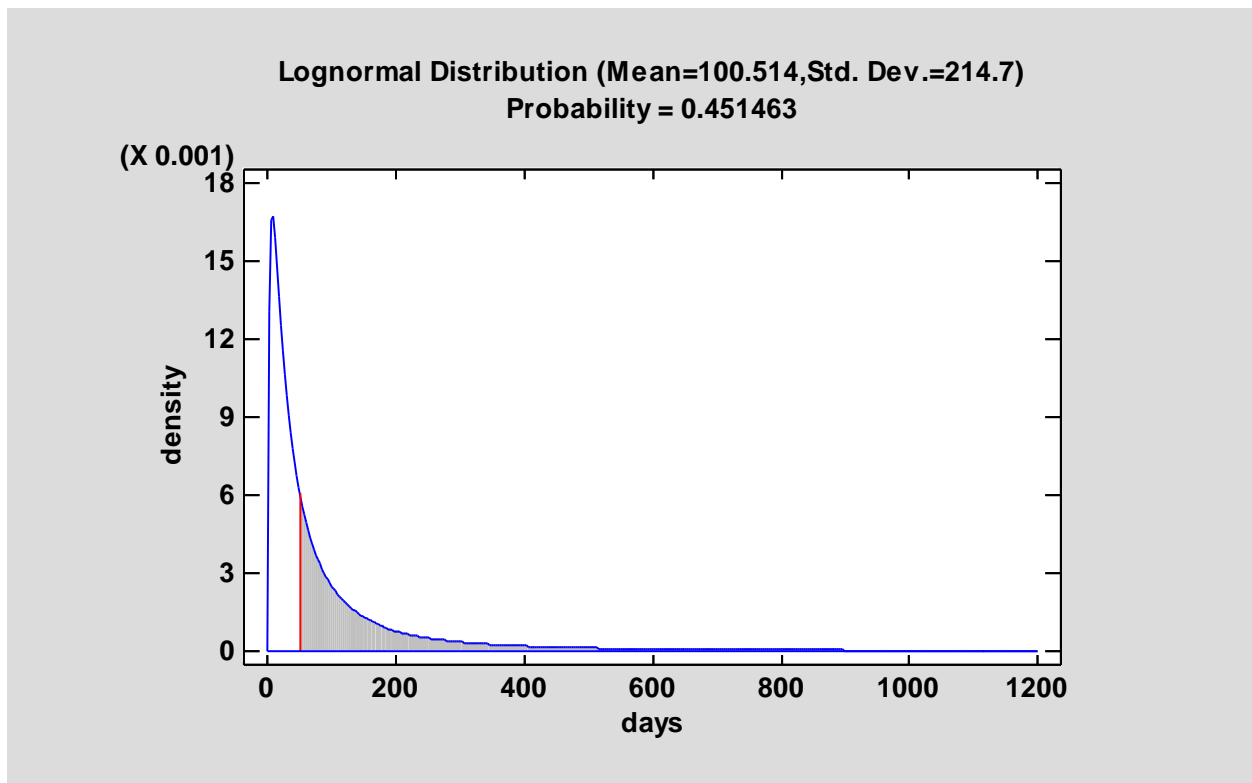
### Pane Options

Use this dialog box to shade areas under the density function:



Specify a lower limit, an upper limit, and the area to shade.

For example, the plot below shades the area above 50:



The area in the shaded region, which equals the probability of observing a value in that range, is displayed at the top of the graph.

## Nonparametric Estimates

An alternative nonparametric estimate is also calculated using the Kaplan-Meier-Turnbull (1976) procedure:

Nonparametric Estimates				
Kaplan-Meier-Turnbull Estimates				
<i>days</i>	<i>CDF</i>	<i>Survival</i>	<i>95% LCL</i>	<i>95% UCL</i>
4.0	0.0	1.0	0.891313	1.0
6.0	0.0463468	0.953653	0.826087	1.0
7.0	0.0797102	0.92029	0.782609	1.0
11.0	0.168378	0.831622	0.685386	1.0
24.0	0.23913	0.76087	0.599359	0.869565
33.0	0.331776	0.668224	0.470669	0.800272
38.0	0.413562	0.586438	0.332921	0.730274
46.0	0.534442	0.465558	0.283185	0.66923

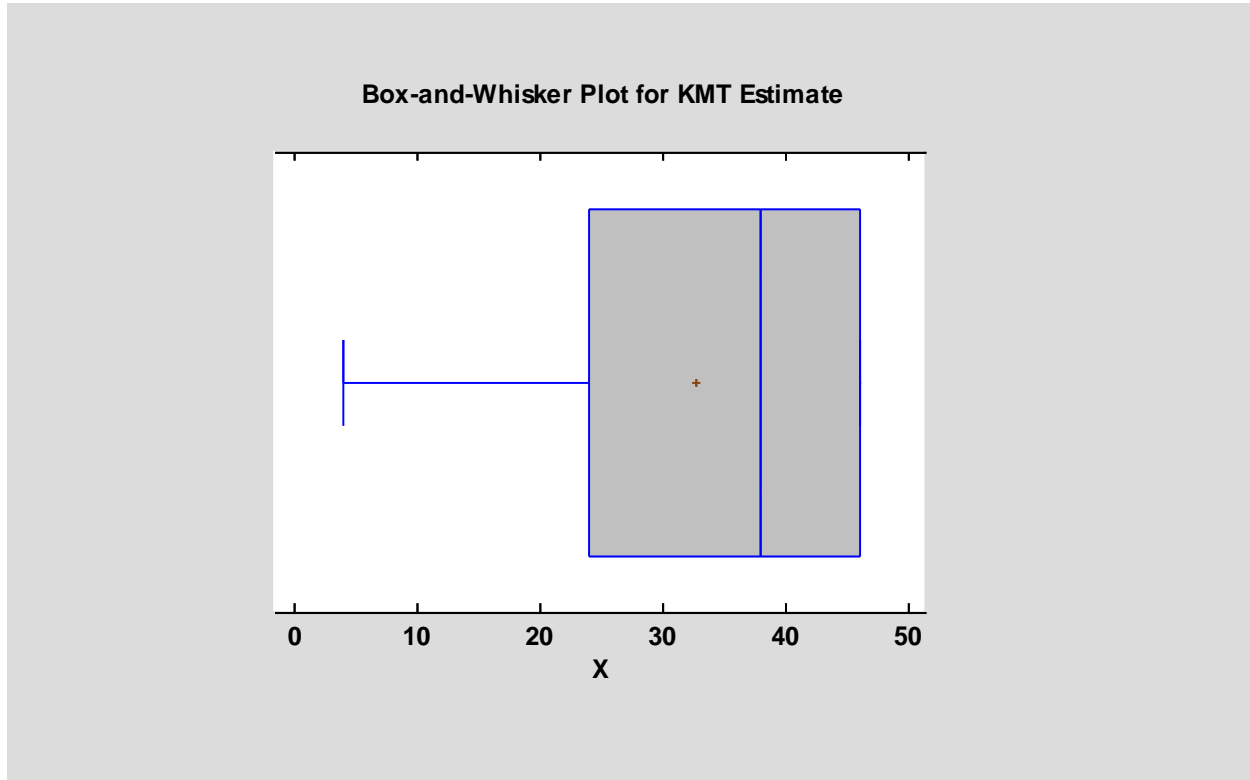
Statistics	
Mean	33.5093
Standard deviation	15.9287
Standard error	2.34855

If the data do not contain any interval-censored observations, the estimated mean, standard deviation and standard error of the mean are also displayed.

Note: The estimated mean and standard deviation are calculated from the estimated CDF by integration against the implied density function. If the largest observation is right censored, the CDF is assumed to equal 1 at that point. The standard error is obtained by dividing the estimated standard deviation by  $\sqrt{n}$  where  $n$  is the total number of observations in the data.

## Box and Whisker Plot

A box-and-whisker plot is created to display the percentiles calculated using the KMT procedure. It takes the form shown below:

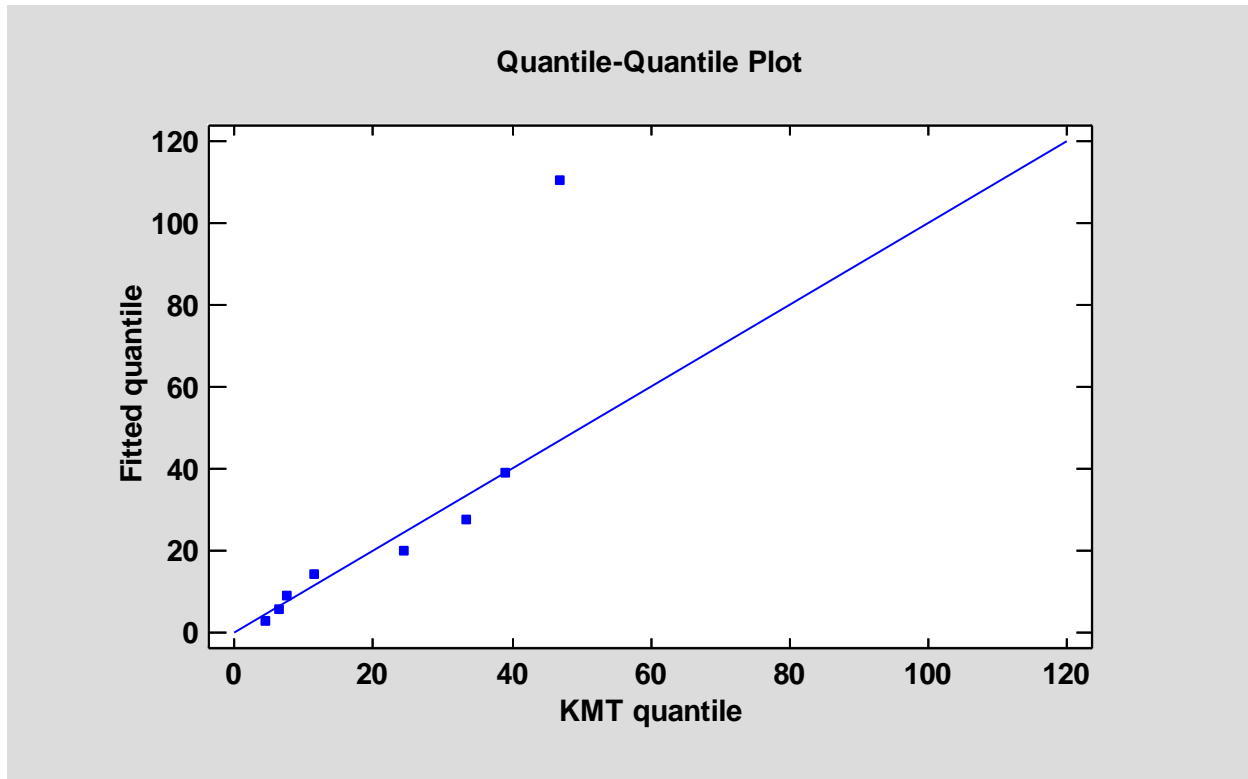


The central box extends from the lower quartile (25<sup>th</sup> percentile) to the upper quartile (75<sup>th</sup> percentile). The whiskers extend from the 1<sup>st</sup> percentile to the 99<sup>th</sup> percentile. There is a vertical line at the estimated median and a plus sign at the estimated mean. In the example above, the estimated 75<sup>th</sup> and 99<sup>th</sup> percentiles are the same, so no upper whisker appears.

To display a table of the estimated percentiles, select *Critical Values* from the list of tables and graphs.

## Quantile-Quantile Plot

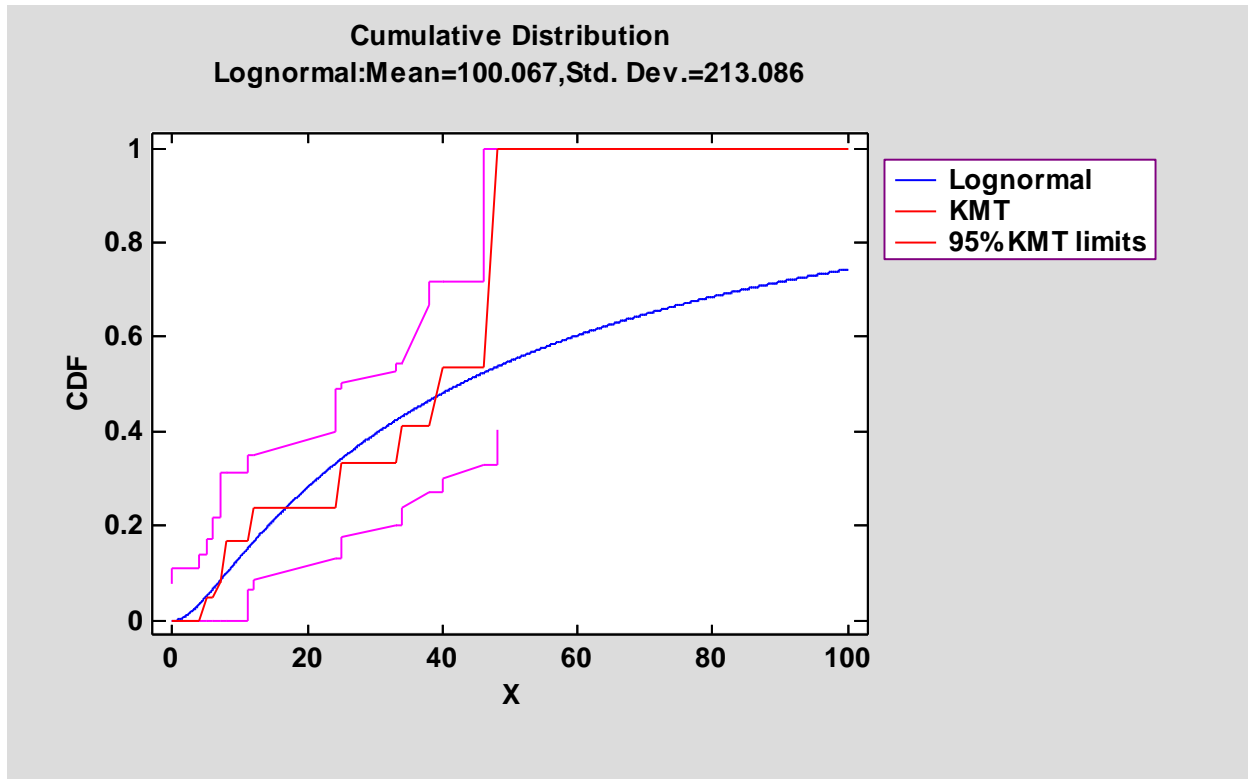
The quantile-quantile or Q-Q plot may be used to compare the fitted distribution with the nonparametric fit.



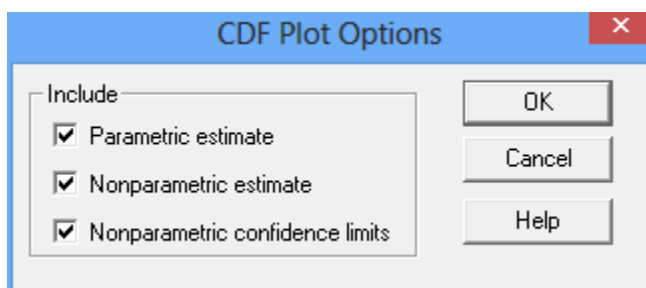
There is a point on the plot for each uncensored or interval-censored observation. The horizontal location of the point is the observed value if the point is not censored or the middle of the interval for interval-censored observations. The vertical location is the inverse probability distribution function for the fitted parametric distribution evaluated at the Kaplan-Meier-Turnbull CDF. (For interval-censored values, the KMT CDF is evaluated at the middle of the interval. For uncensored values, the average CDF before and after the step jump is used.) If the points fall close to the diagonal line, the 2 estimates are similar. In the example above, there is good correspondence between the fitted distribution and the KMT estimate except for the last observation, which is not as large as expected if the data come from a lognormal distribution.

## Cumulative Distribution

The fitted cumulative distribution  $F(X)$  is plotted using the fitted distribution and/or nonparametric estimate:



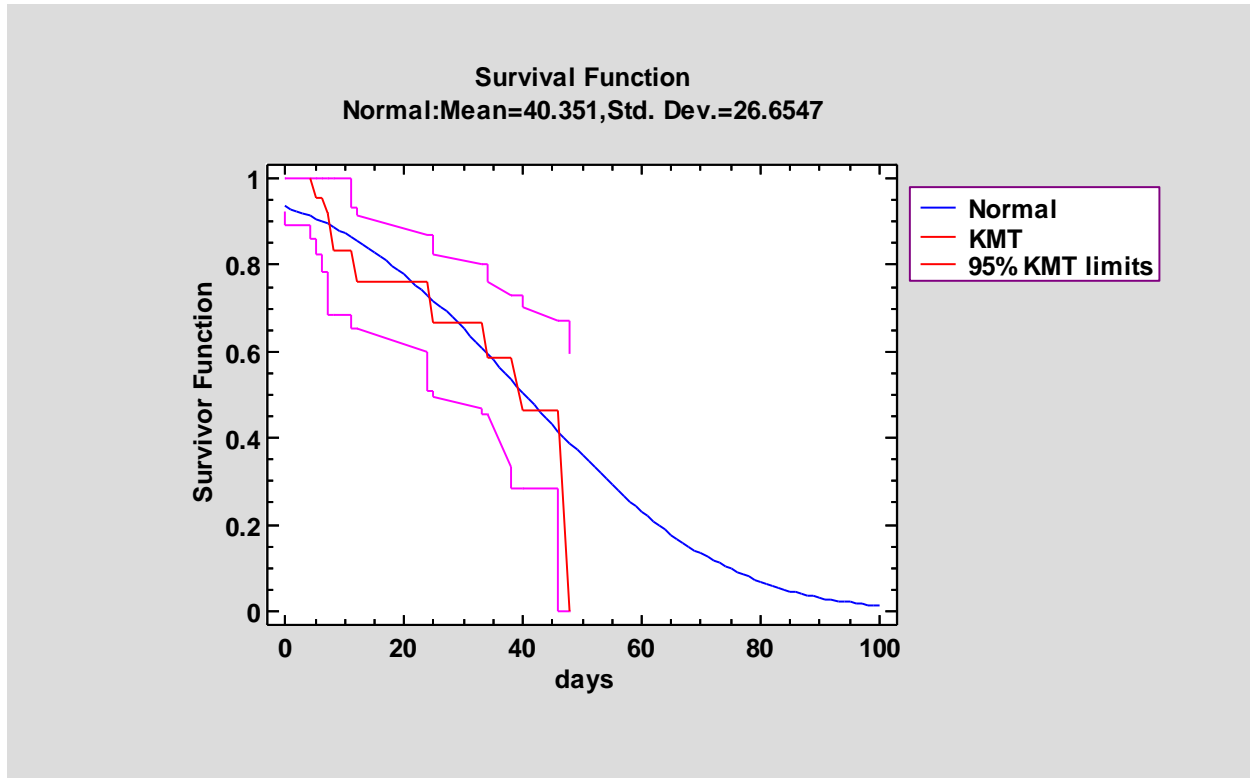
### Pane Options



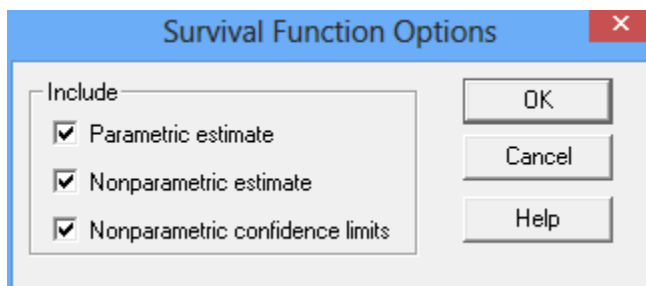
- **Parametric estimate:** if checked, the CDF of the fitted distribution will be plotted.
- **Nonparametric estimate:** if checked, the Kaplan-Meier-Turnbull estimate will be plotted.
- **Nonparametric confidence limits:** if checked, upper and lower confidence limits for the KMT estimate will be plotted. The confidence level is controlled by the *Analysis Options* dialog box.

## Survival Function

The fitted cumulative distribution  $1 - F(X)$  is plotted using the fitted distribution:



### Pane Options

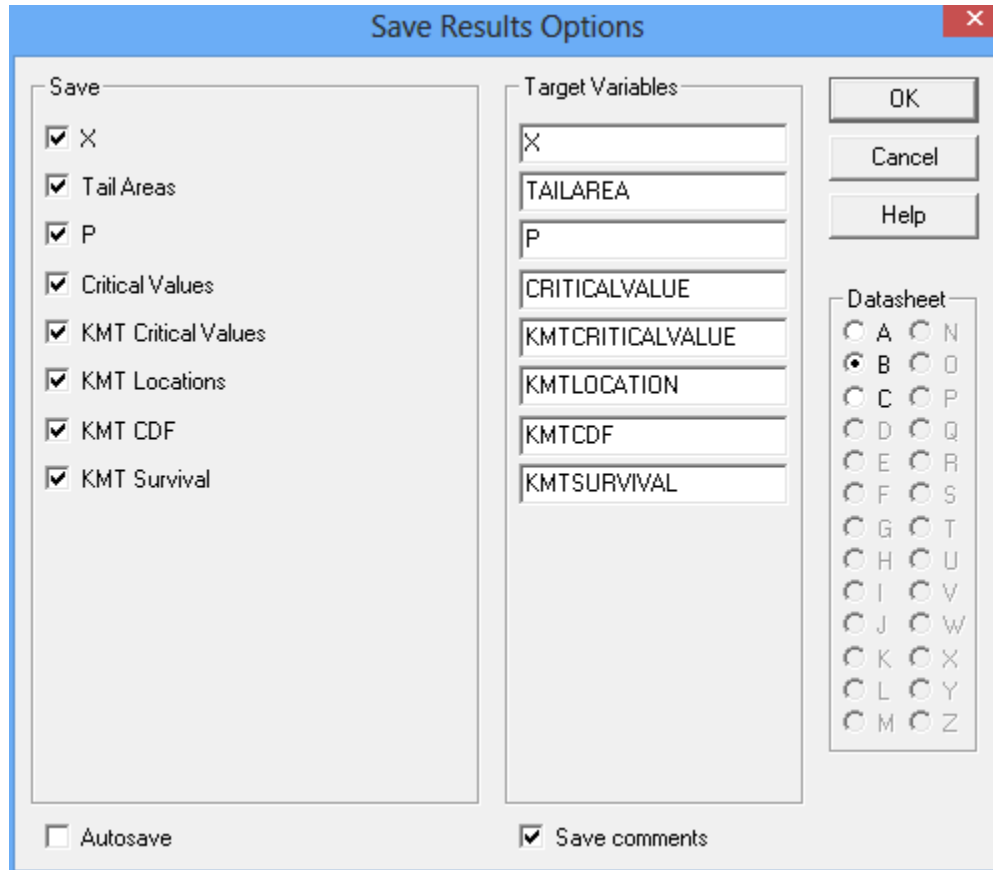


- **Parametric estimate:** if checked, the survivor function of the fitted distribution will be plotted.
- **Nonparametric estimate:** if checked, the Kaplan-Meier-Turnbull estimate will be plotted.
- **Nonparametric confidence limits:** if checked, upper and lower confidence limits for the KMT estimate will be plotted. The confidence level is controlled by the *Analysis Options* dialog box.



## Save Results

The results of selected calculations may be saved in a Statgraphics datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



The dialog box titled "Save Results Options" contains the following elements:

- Save:** A list of items to be saved, each with a checked checkbox:
  - X
  - Tail Areas
  - P
  - Critical Values
  - KMT Critical Values
  - KMT Locations
  - KMT CDF
  - KMT Survival
- Target Variables:** A list of text boxes containing the following labels:
  - X
  - TAILAREA
  - P
  - CRITICALVALUE
  - KMTCRITICALVALUE
  - KMTLOCATION
  - KMTCDF
  - KMTSURVIVAL
- Datasheet:** A grid of radio buttons for selecting a column letter:
 

<input type="radio"/> A	<input type="radio"/> N
<input checked="" type="radio"/> B	<input type="radio"/> O
<input type="radio"/> C	<input type="radio"/> P
<input type="radio"/> D	<input type="radio"/> Q
<input type="radio"/> E	<input type="radio"/> R
<input type="radio"/> F	<input type="radio"/> S
<input type="radio"/> G	<input type="radio"/> T
<input type="radio"/> H	<input type="radio"/> U
<input type="radio"/> I	<input type="radio"/> V
<input type="radio"/> J	<input type="radio"/> W
<input type="radio"/> K	<input type="radio"/> X
<input type="radio"/> L	<input type="radio"/> Y
<input type="radio"/> M	<input type="radio"/> Z
- Buttons:** OK, Cancel, and Help.
- Options:**
  - Autosave
  - Save comments

- **Save:** select the items to be saved.
  - **X and Tail Areas** – from the *Tail Areas* table.
  - **P, Critical Values and KMT Critical Values** – from the *Critical Values* table.
  - **KMT Locations, KMT CDF and KMT Survival** – from the *Nonparametric Estimates* table.
- **Target Variables:** enter names for the columns to be created.
- **Datasheet:** the datasheet into which the results will be saved.
- **Autosave:** if checked, the results will be saved automatically each time a saved StatFolio is loaded.
- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

## Calculations

The R *interval* package is used to calculate the non-parametric Kaplan-Meier-Turnbull estimate of the CDF and the survivor function. If interval-censored data is present, there are ranges of values within the intervals where the estimate is not exact. In those cases, the censored values are assumed to be equally likely to fall anywhere in the interval if the Efron bias correction is not applied.

## References

Finkelstein, D.M. and Wolfe, R.A. (1985). “A semiparametric model for regression analysis of interval-censored failure time data.” *Biometrics* **41**, 731-740.

Helsel, D.R. (2012). Statistics for Censored Environmental Data using Minitab and R, second edition. Wiley, Hoboken, N.J.

Lee, E.T. and Wang, J.W. (2003). Statistical Methods for Survival Data Analysis, 3<sup>rd</sup> edition. Wiley, New York.

R Package “interval” - <https://cran.r-project.org/web/packages/interval/interval.pdf>

Turnbull BW (1976). “The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data.” *Journal of the Royal Statistical Society. Series B*, **38**(3), 290–295.