statgraphics 18®
centurion

Distribution Fitting for Arbitrarily Censored Data

Dr. Neil Polhemus

Censored Data

- Censored data arise when the value of an observation is only partially known.

- For example, in a medical trial the survival time of a patient may only be known to greater than some value because the patient left the study.

- Or a measurement taken to study groundwater contamination may only be known to be less than some detection or quantitation limit.

Types of Censored Data

- Right-censored – known to be greater than.

  > 45

- Left-censored – known to be less than.

  < 7

- Interval-censored – known to be between.

  [6,10]

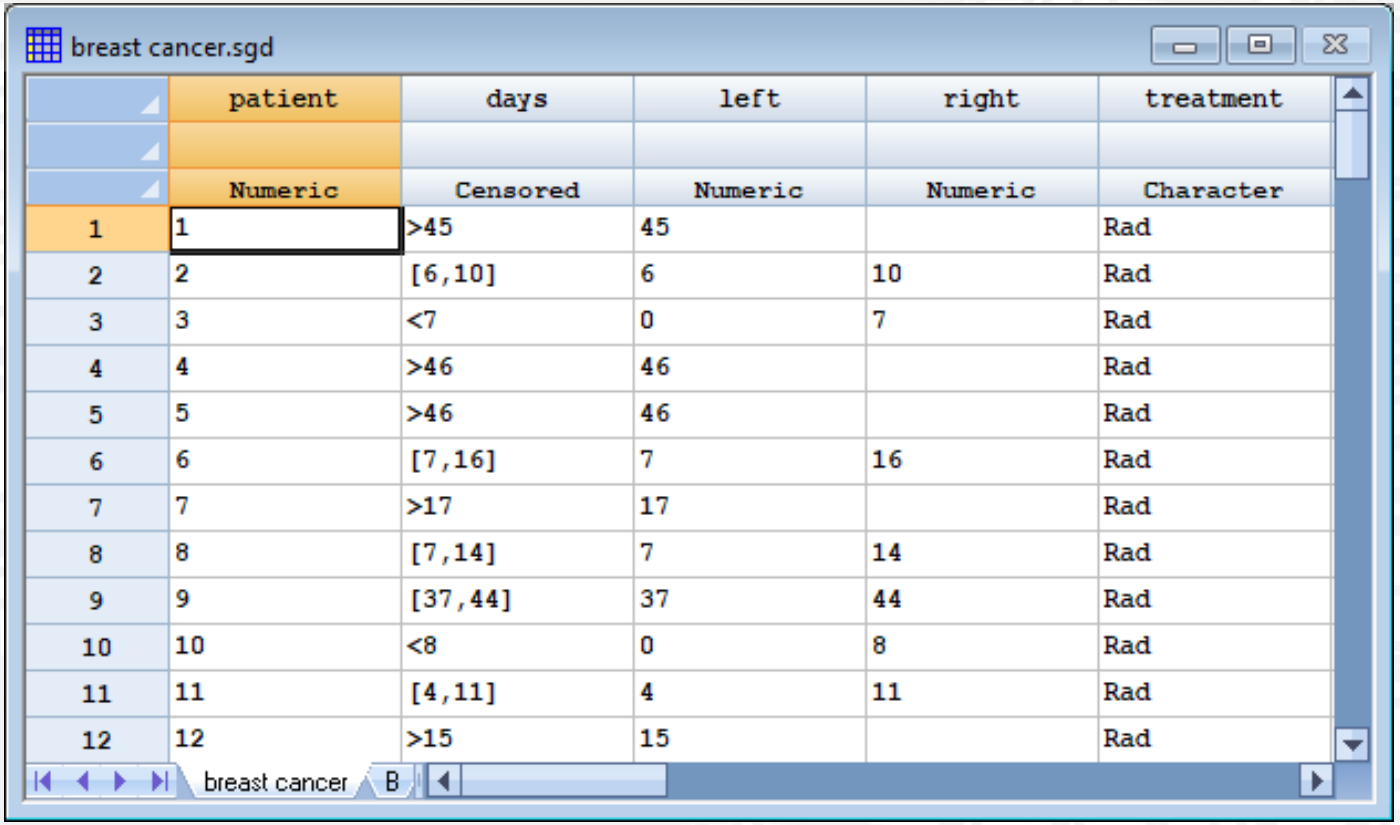
Example #1: Arsenic Concentrations

- Arsenic concentrations (ug/L) in an urban stream, Manoa Stream at Kanewai Field, on Oahu, Hawaii. (Tomlinson, 2003)

0.5	0.5	0.5	0.6	0.7	0.7	<0.9	0.9
<1.0	<1.0	<1.0	<1.0	1.5	1.7	<2.0	<2.0
<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	2.8	3.2

Example #2: Breast Cancer Study

- Days between treatment and breast retraction – Finkelstein and Wolfe (1985)



breast cancer.sgd

	patient	days	left	right	treatment
	Numeric	Censored	Numeric	Numeric	Character
1	1	>45	45		Rad
2	2	[6,10]	6	10	Rad
3	3	<7	0	7	Rad
4	4	>46	46		Rad
5	5	>46	46		Rad
6	6	[7,16]	7	16	Rad
7	7	>17	17		Rad
8	8	[7,14]	7	14	Rad
9	9	[37,44]	37	44	Rad
10	10	<8	0	8	Rad
11	11	[4,11]	4	11	Rad
12	12	>15	15		Rad

breast cancer | B

Statgraphics Procedures

Option #1: *Describe – Distribution Fitting – Fitting Censored Data.*

- Handles left-censored and right-censored data.
- Includes goodness-of-fit tests.

Option #2: *R Interface – Distribution Fitting (Arbitrarily Censored Data).*

- Also handles interval-censored data.
- Does not include goodness-of-fit tests.

Fitting the Arsenic Data

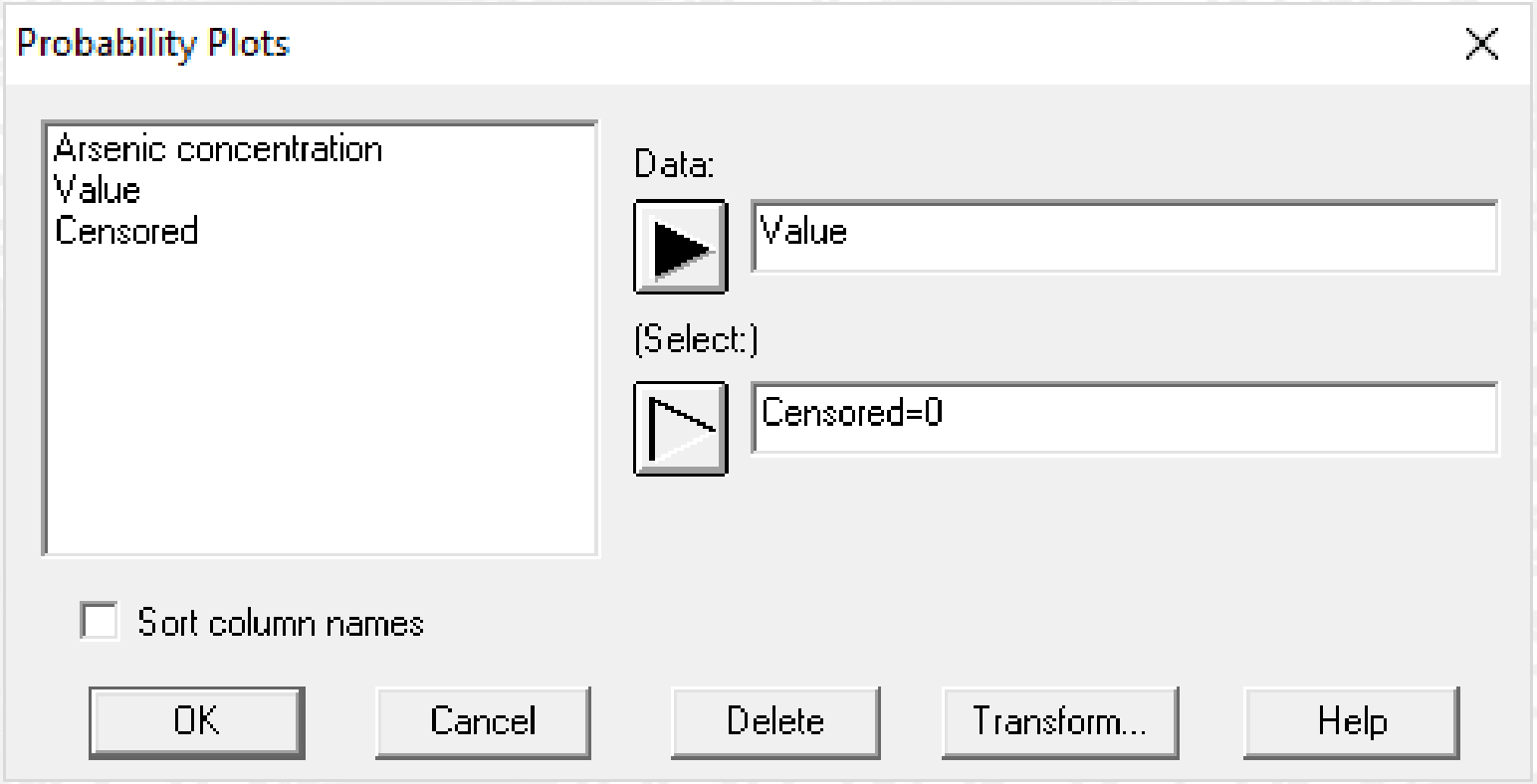
arsenic concentrations.sgd

	Arsenic concentration	Value	Censored
	mu-g/L		
	Censored	Numeric	Numeric
1	.5	.5	0
2	.5	.5	0
3	.5	.5	0
4	.6	.6	0
5	.7	.7	0
6	.7	.7	0
7	<.9	.9	-1
8	.9	.9	0
9	<1	1	-1
10	<1	1	-1
11	<1	1	-1
12	<1	1	-1
13	1.5	1.5	0
14	1.7	1.7	0

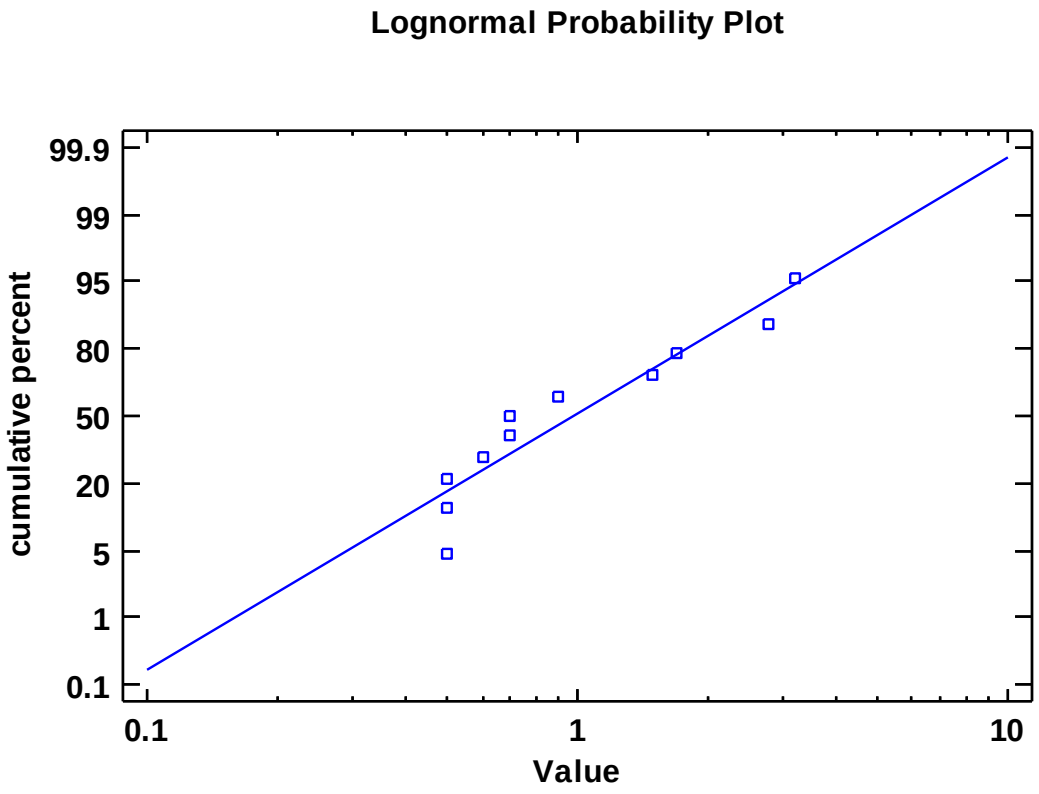
arsenic c

Note: may use *Edit – Replace Censored Values* to create the *Value* column.

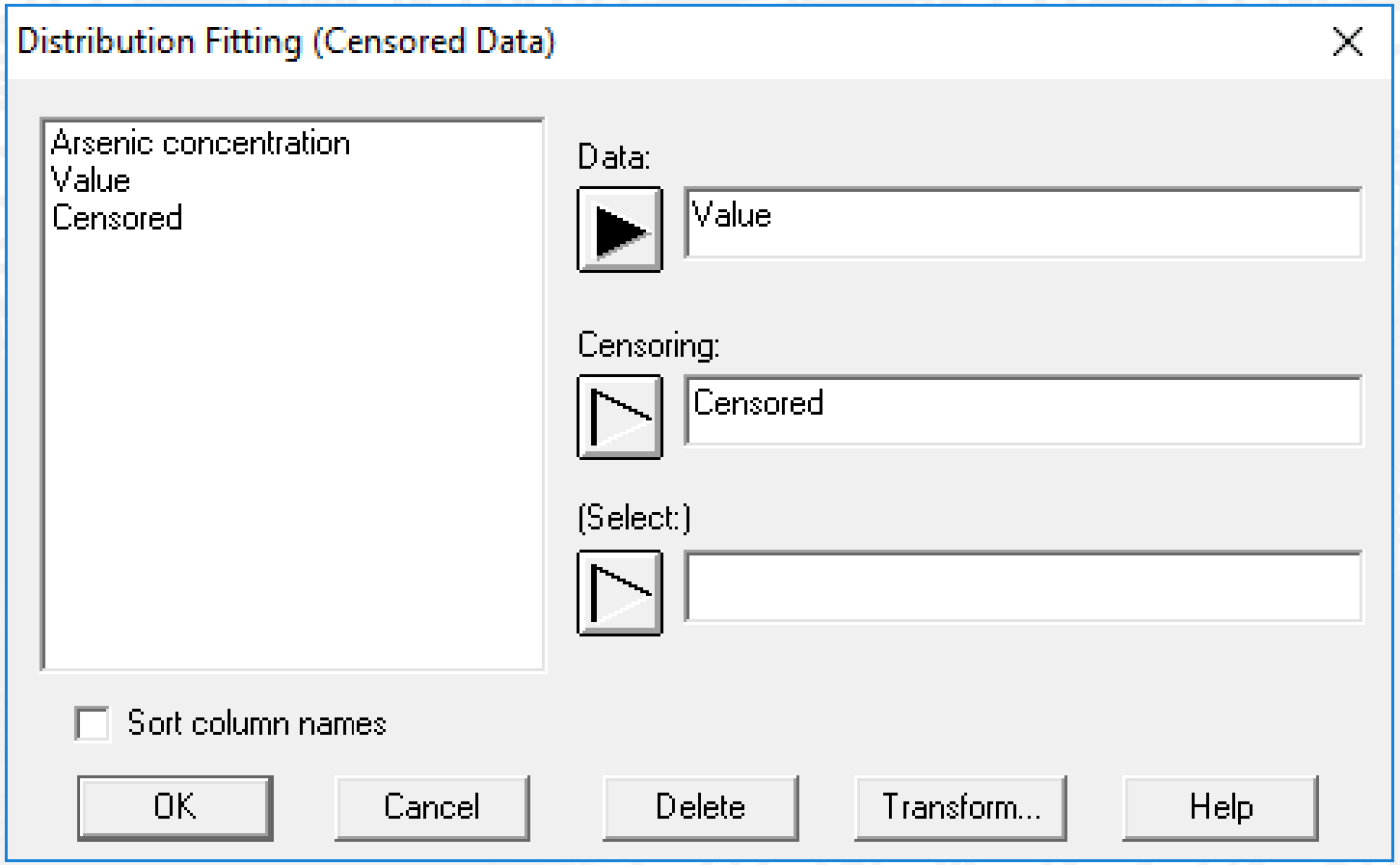
Probability Plots



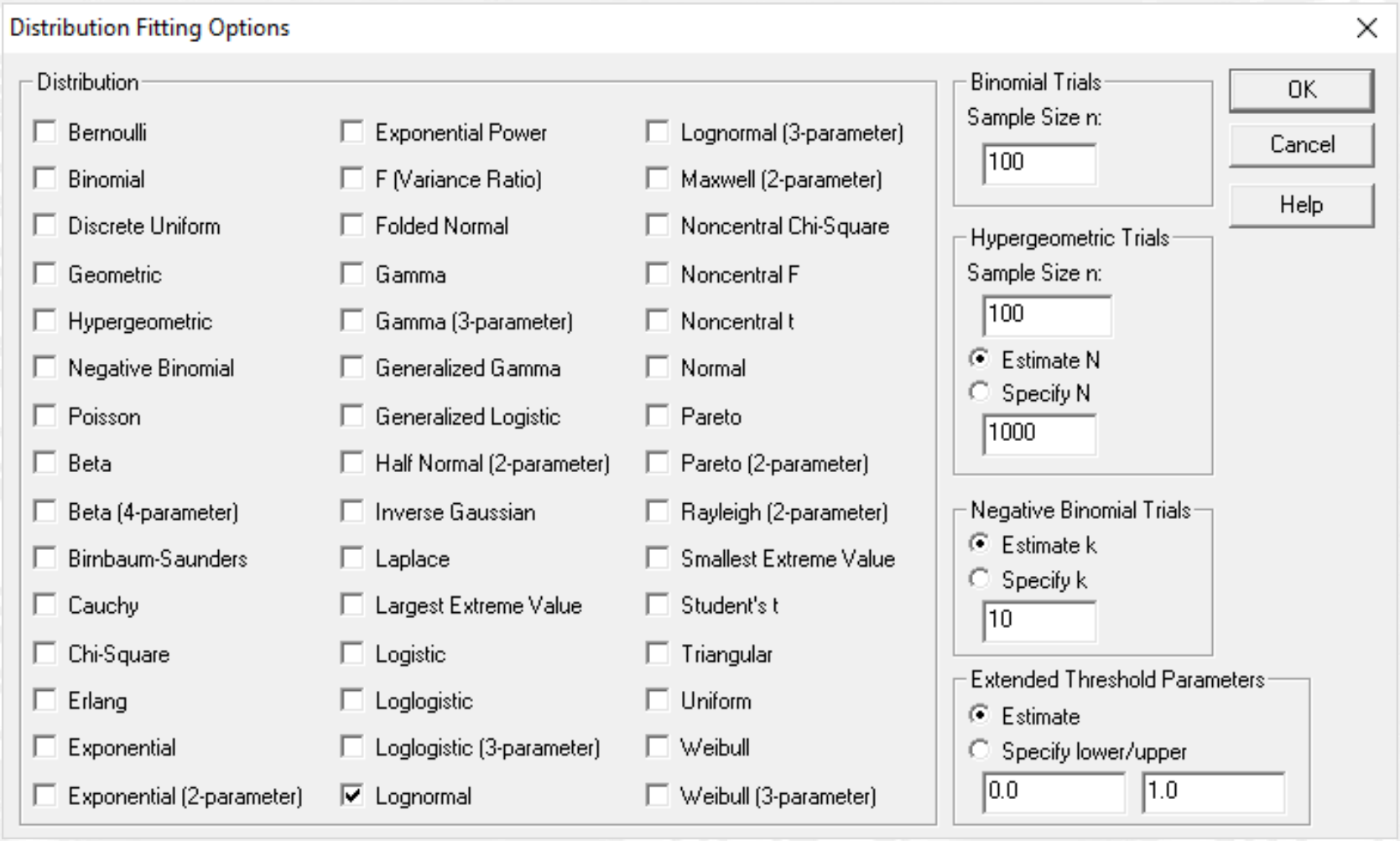
Probability Plots



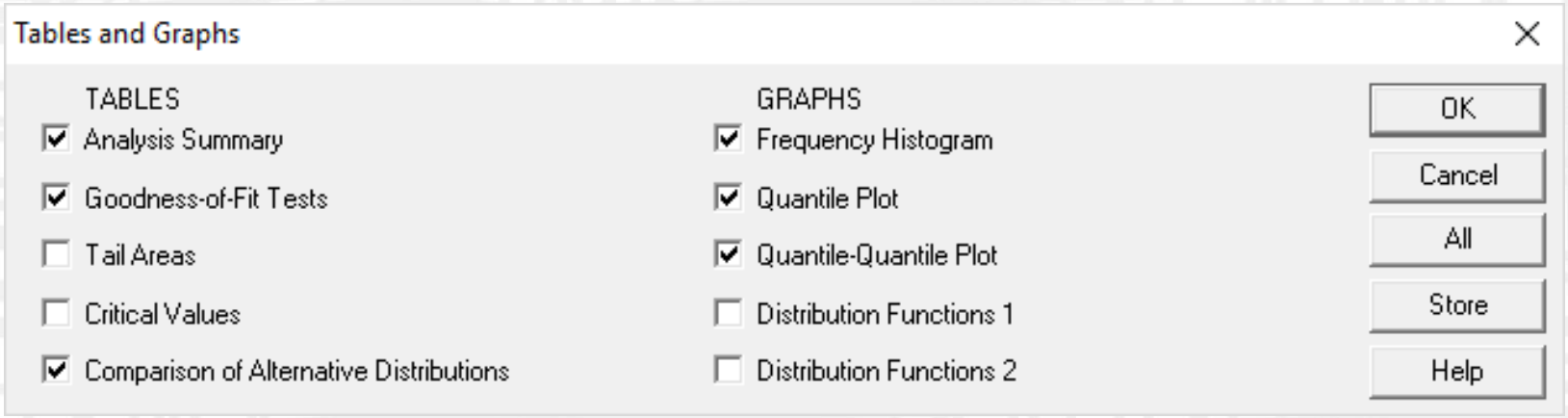
Distribution Fitting (Censored Data)



Analysis Options



Tables and Graphs



Analysis Summary

Distribution Fitting (Censored Data) - Value

Data variable: Value

Censoring: Censored

24 values ranging from 0.5 to 3.2

Number of left-censored observations: 13

Number of right-censored observations: 0

Fitted Distributions

Lognormal
mean = 0.94526
standard deviation = 0.655927
Log scale: mean = -0.252828
Log scale: std. dev. = 0.626949

Goodness-of-Fit Tests

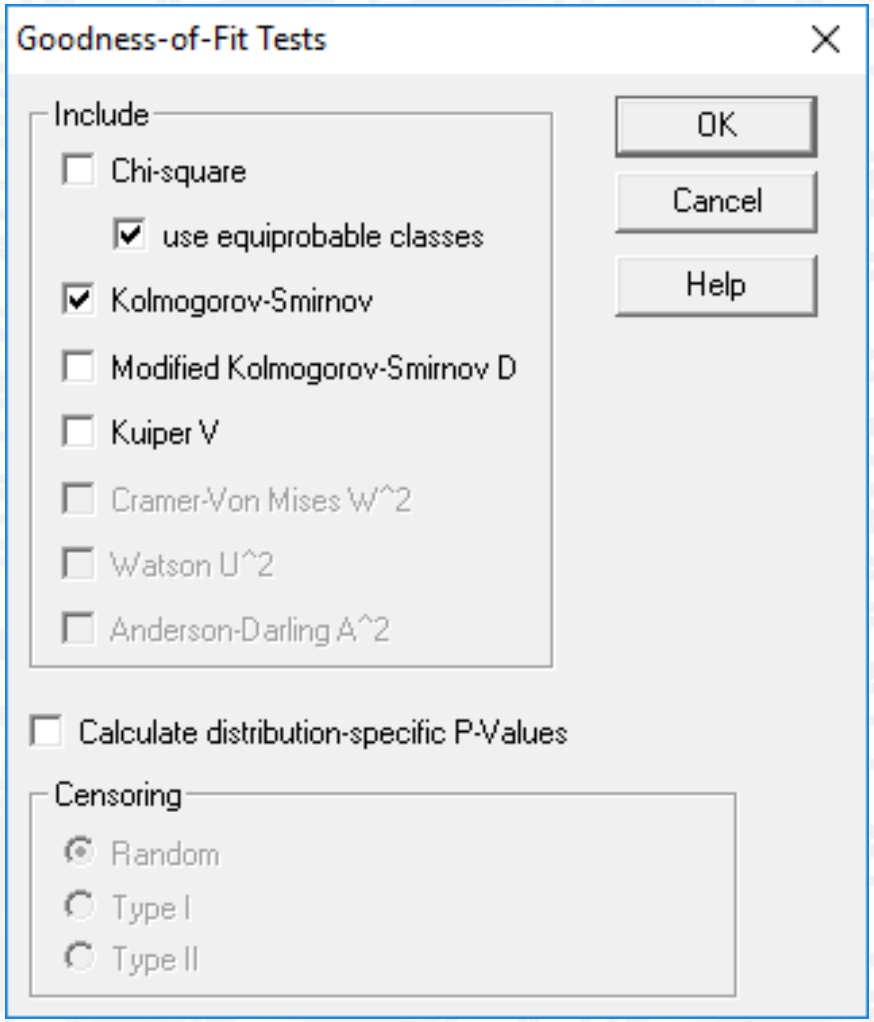
Goodness-of-Fit Tests for Value

Kolmogorov-Smirnov Test

	Lognormal
DPLUS	0.253282
DMINUS	0.168321
DN	0.253282
P-Value	0.091989

Type I censoring – items removed after prespecified times.

Type II censoring – test stopped after prespecified number of failures.



Comparison of Alternative Distributions

Comparison of Alternative Distributions

Distribution	*Est. Parameters*	*KS D*
Loglogistic	2	0.219342
Lognormal	2	0.253282
Weibull	2	0.255712
Gamma	2	0.257573
Inverse Gaussian	2	0.260837
Birnbaum-Saunders	2	0.268475
Laplace	2	0.276909
Logistic	2	0.283439
Largest Extreme Value	2	0.285595
Normal	2	0.287314
Exponential	1	0.366133
Smallest Extreme Value	2	0.374146
Uniform	2	0.409164
Pareto	<no fit>	

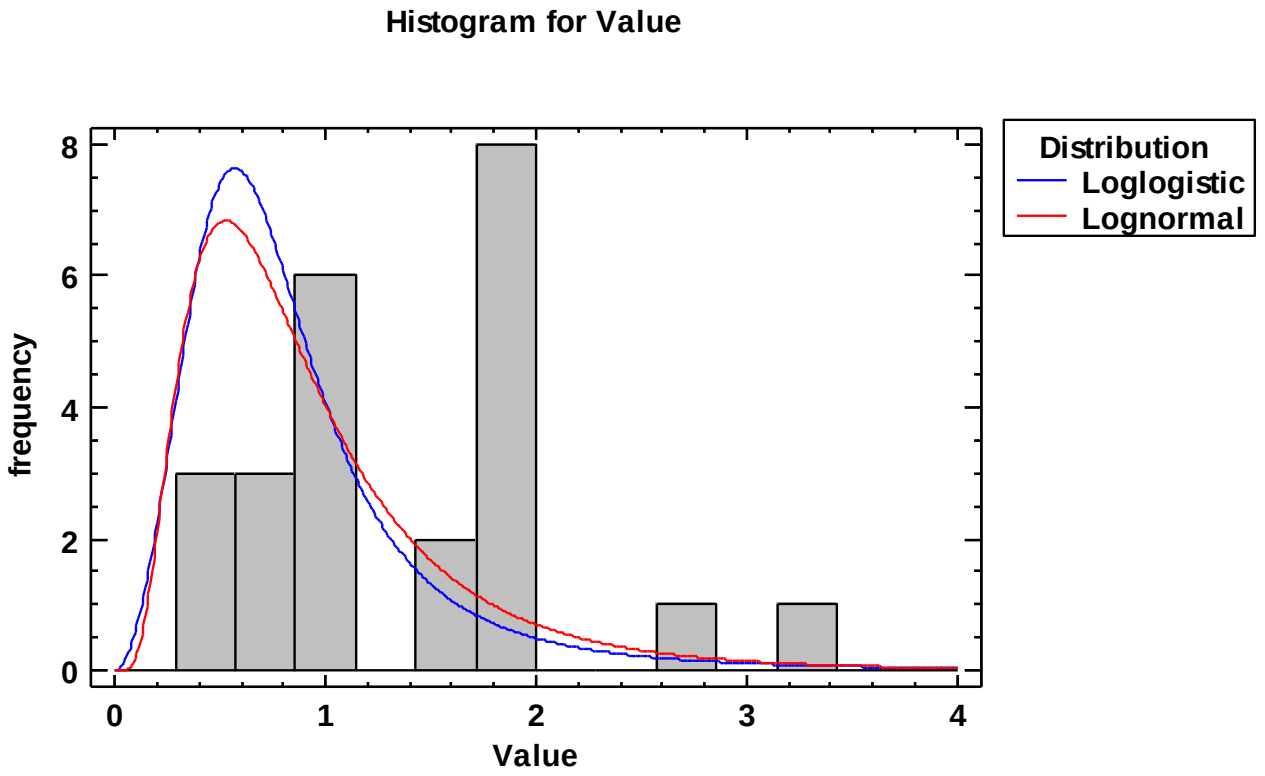
Goodness-of-Fit Tests

Goodness-of-Fit Tests for Value

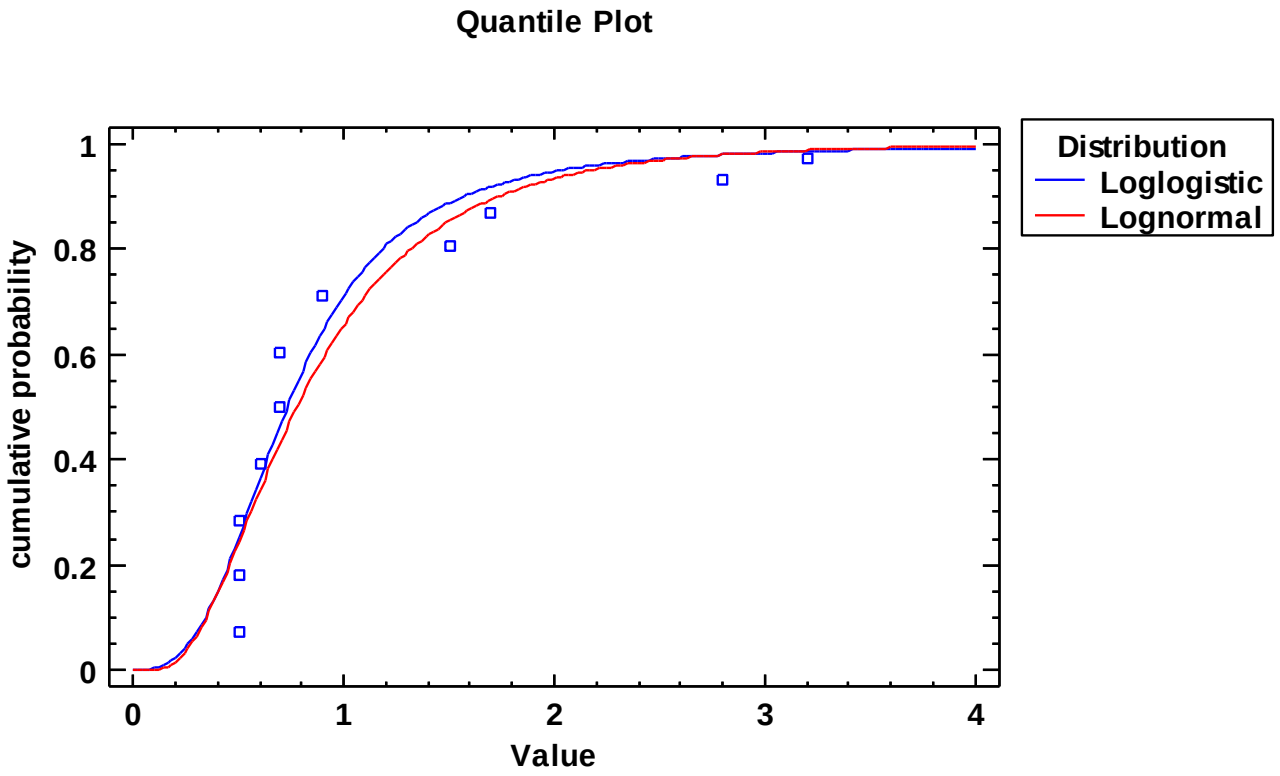
Kolmogorov-Smirnov Test

	Loglogistic	*Lognormal*
DPLUS	0.219342	0.253282
DMINUS	0.177452	0.168321
DN	0.219342	0.253282
P-Value	0.198755	0.091989

Histogram with Fits



Quantile Plot

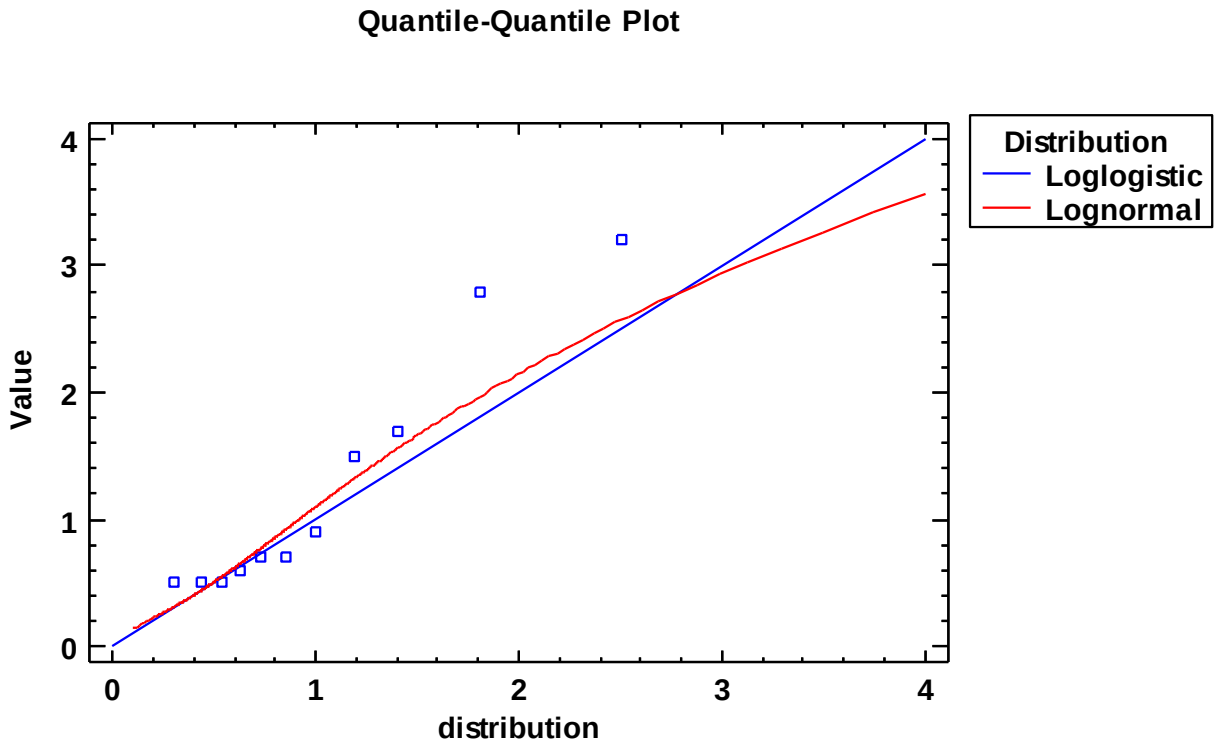


Critical Values

Critical Values for Value

Lower Tail Area (<=)	*Loglogistic*	*Lognormal*
0.01	0.148405	0.180625
0.1	0.341207	0.347741
0.5	0.731694	0.776602
0.9	1.56907	1.73436
0.99	3.60754	3.33901

Quantile-Quantile Plot

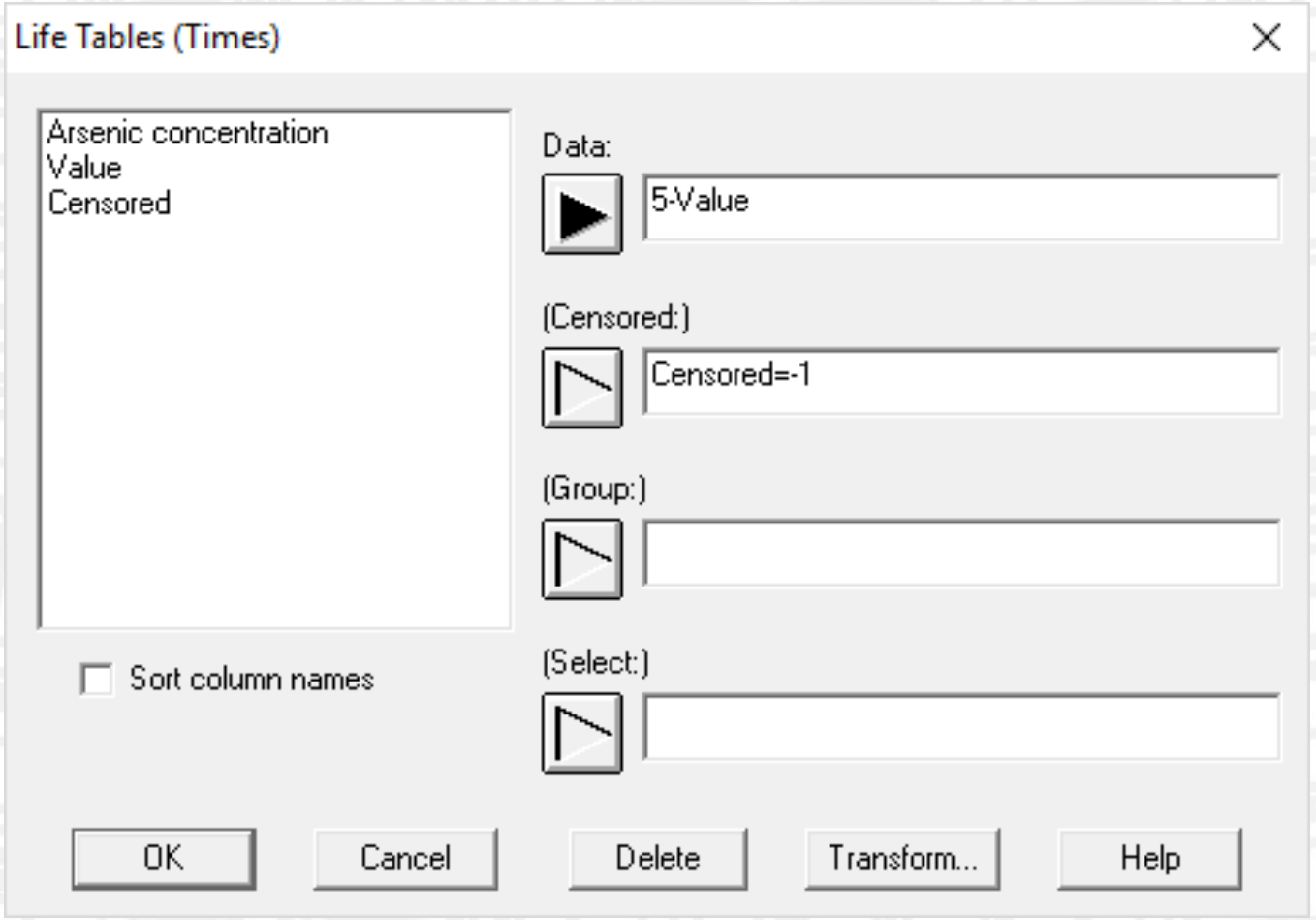


Nonparametric Estimates

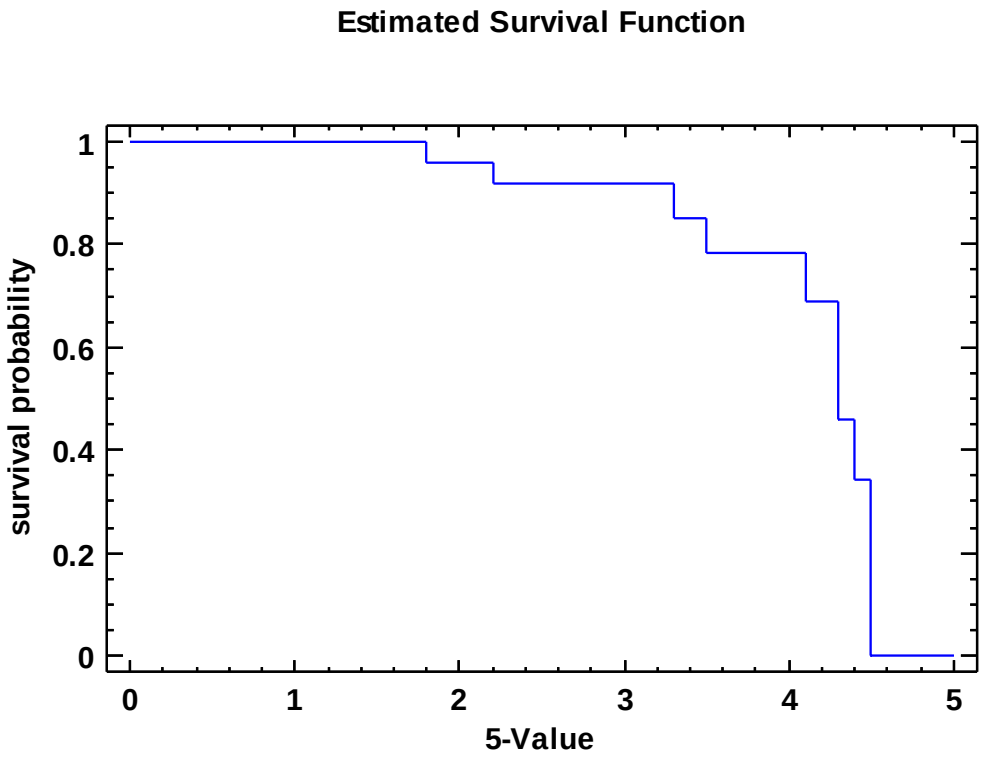
- Statgraphics contains a procedure under *Describe – Life Data – Life Tables (Times)* which will estimate a nonparametric survival function for right-censored data.

- Helsel (2012) shows how such a procedure may be used to analyze left-censored data by "flipping" the data values.

Data Input Dialog Box



Kaplan-Meier Estimate



Results

- Subtract each result from 5 to get what you want.

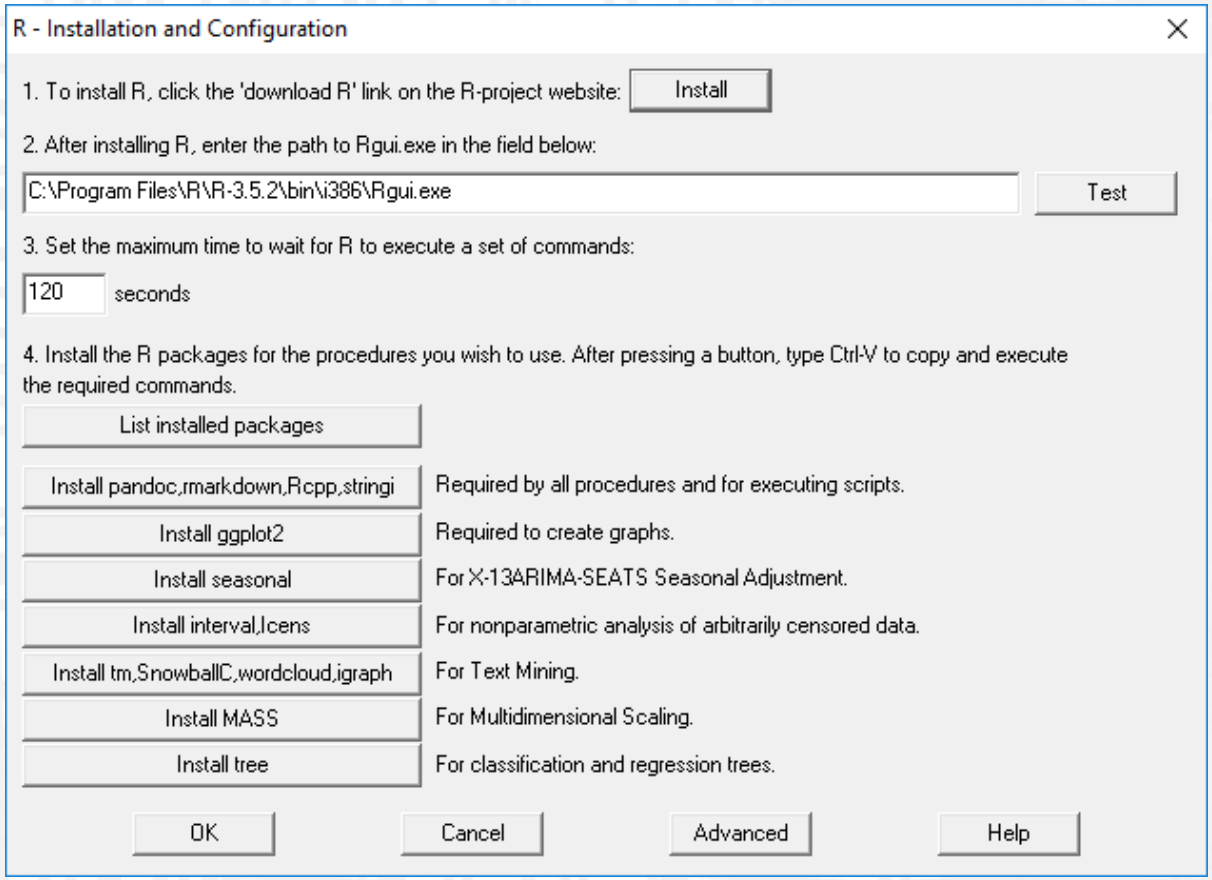
Percentile Table

		Standard
Percentile	*Estimate*	*Error*
75.0	4.1	0.615438
50.0	4.3	0.136753
25.0	4.5	

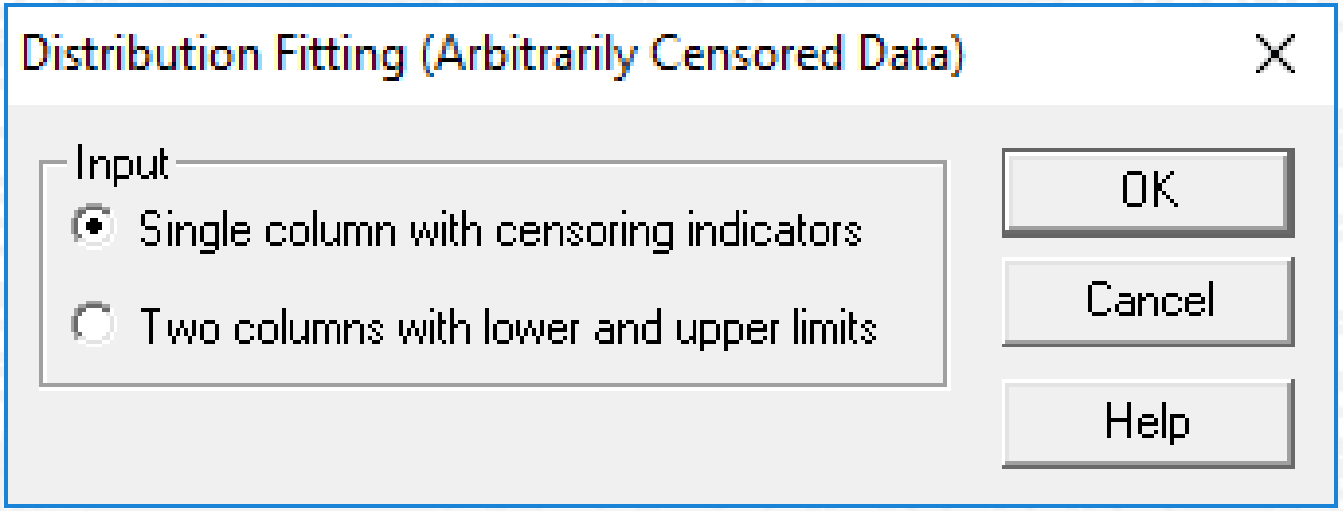
- Q1 = 0.5, Median = 0.7, Q3 = 0.9

Fitting the Breast Cancer Data

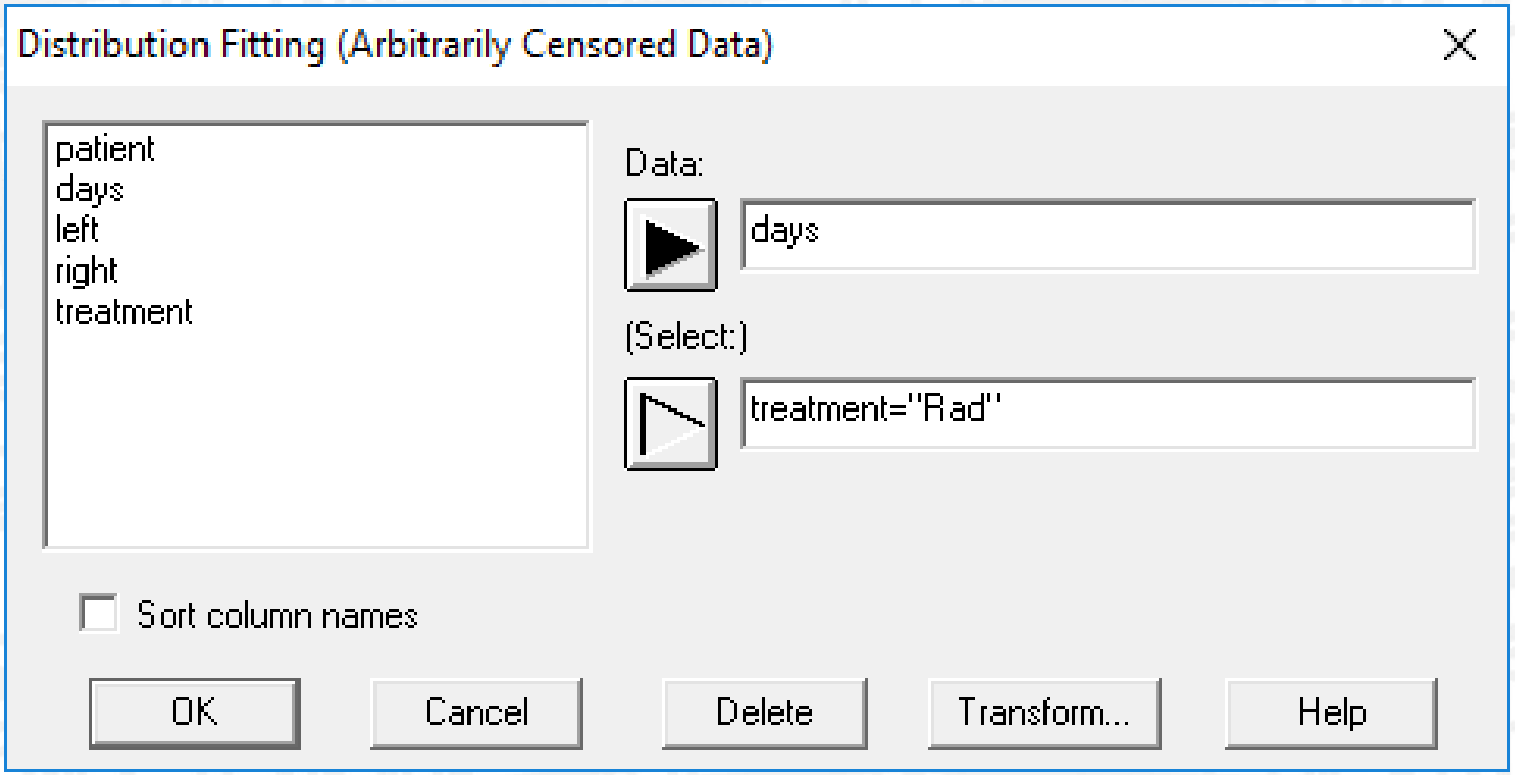
- Uses the “interval” and “Icens” packages written for R.



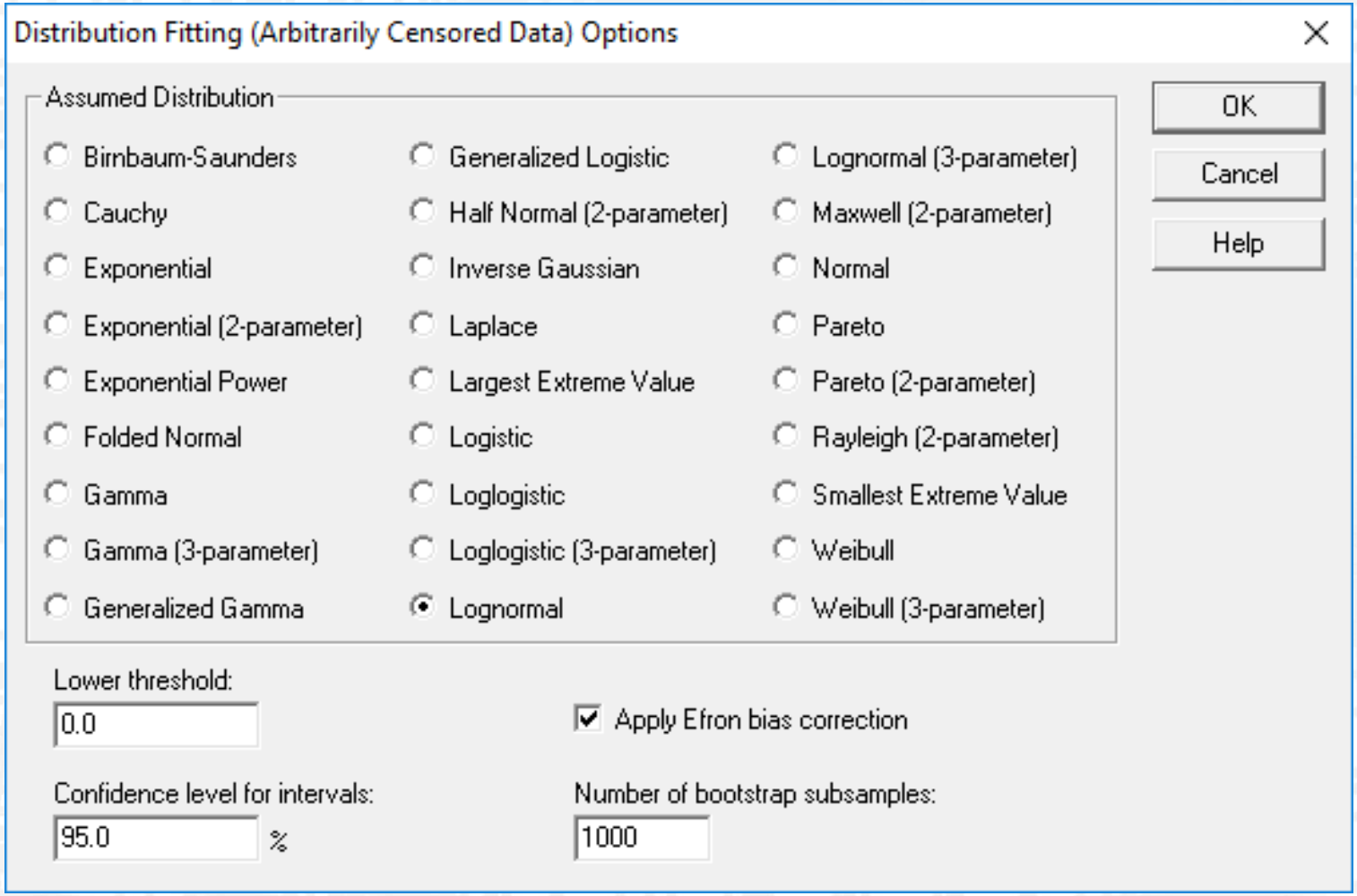
Data Input Dialog Boxes



Data Input Dialog Boxes



Analysis Options



Options

- **Assumed distribution** – will be fit to the data.
- **Lower threshold** – for distributions such as lognormal, the minimum possible value.
- **Apply Efron bias correction** – if smallest value is left-censored, sets KMT nonparametric CDF at that observation to 0 for purposes of calculating the mean and standard deviation. Otherwise, CDF is assumed to decay linearly to 0 at lower threshold.
- **Confidence level for intervals:** confidence level used to create confidence limits for distribution parameters and other quantities.
- **Number of bootstrap subsamples:** number of subsamples to be used when estimating confidence limits for the distribution parameters and other quantities.

Analysis Summary

Distribution Fitting (Arbitrarily Censored Data) (treatment="Rad")

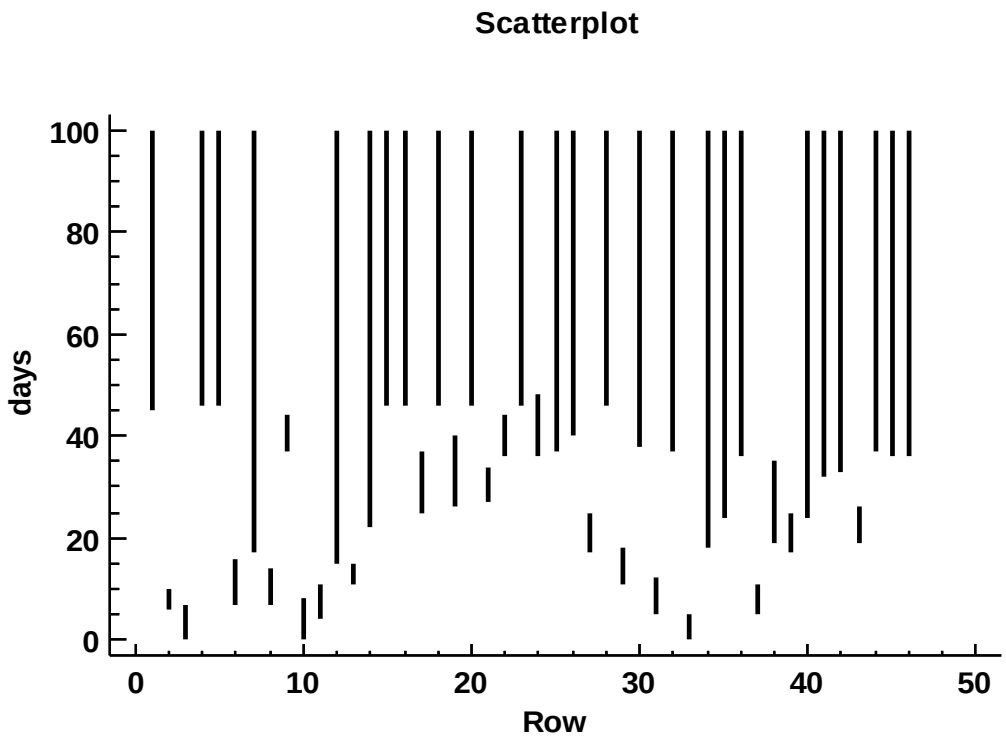
Data variable: days

Selection variable: treatment="Rad"

Observations

Value	*Frequency*
Uncensored	0
Left-censored	3
Interval-censored	18
Right-censored	25

Plotting the Data



Distribution Fitting

Parameter estimates are obtained by maximizing the likelihood function:

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

where

$l(x_i) = f(x_i)$ if the observation x_i is uncensored

$l(x_i) = F(L_i)$ if the observation is left-censored at L_i

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

$l(x_i) = F(U_i) - F(L_i)$ if the observation is interval-censored between $[L_i, U_i]$

Example

Distribution Fitting

Fitted distribution: Lognormal

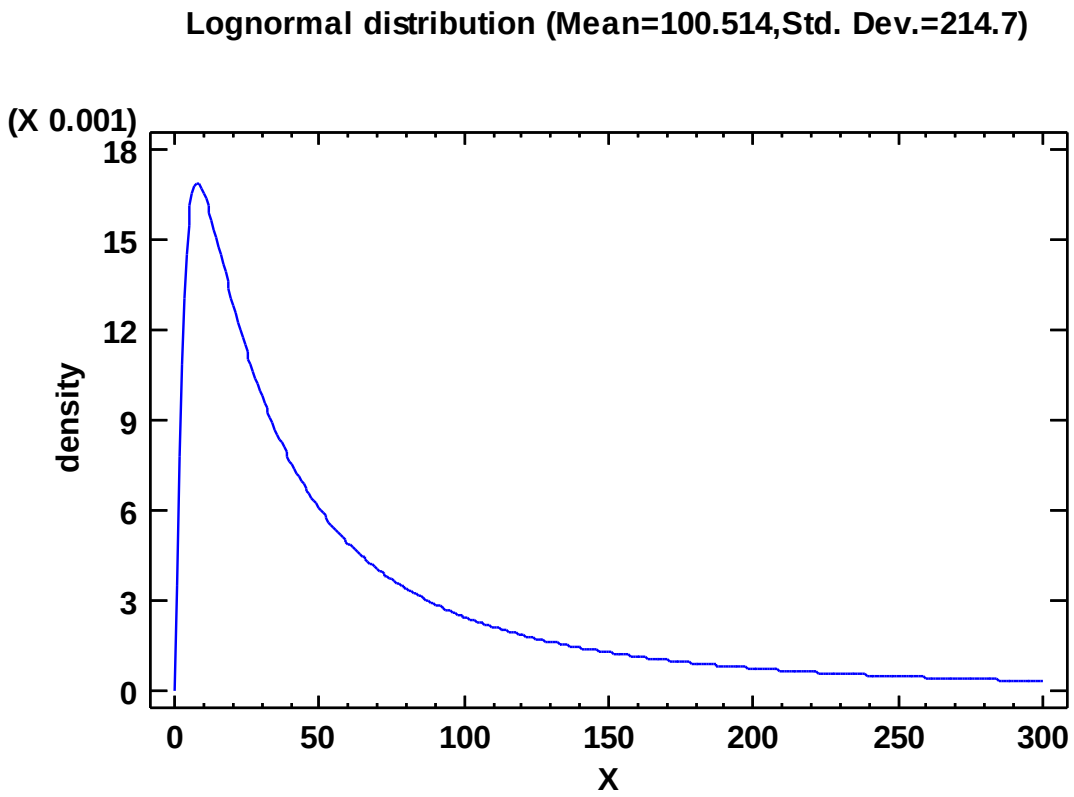
Parameter	*Estimate*	*95% LCL*	*95% UCL*
Mean	100.514	50.0203	447.337
Std. Dev.	214.7	63.2189	2986.17

Distribution Properties

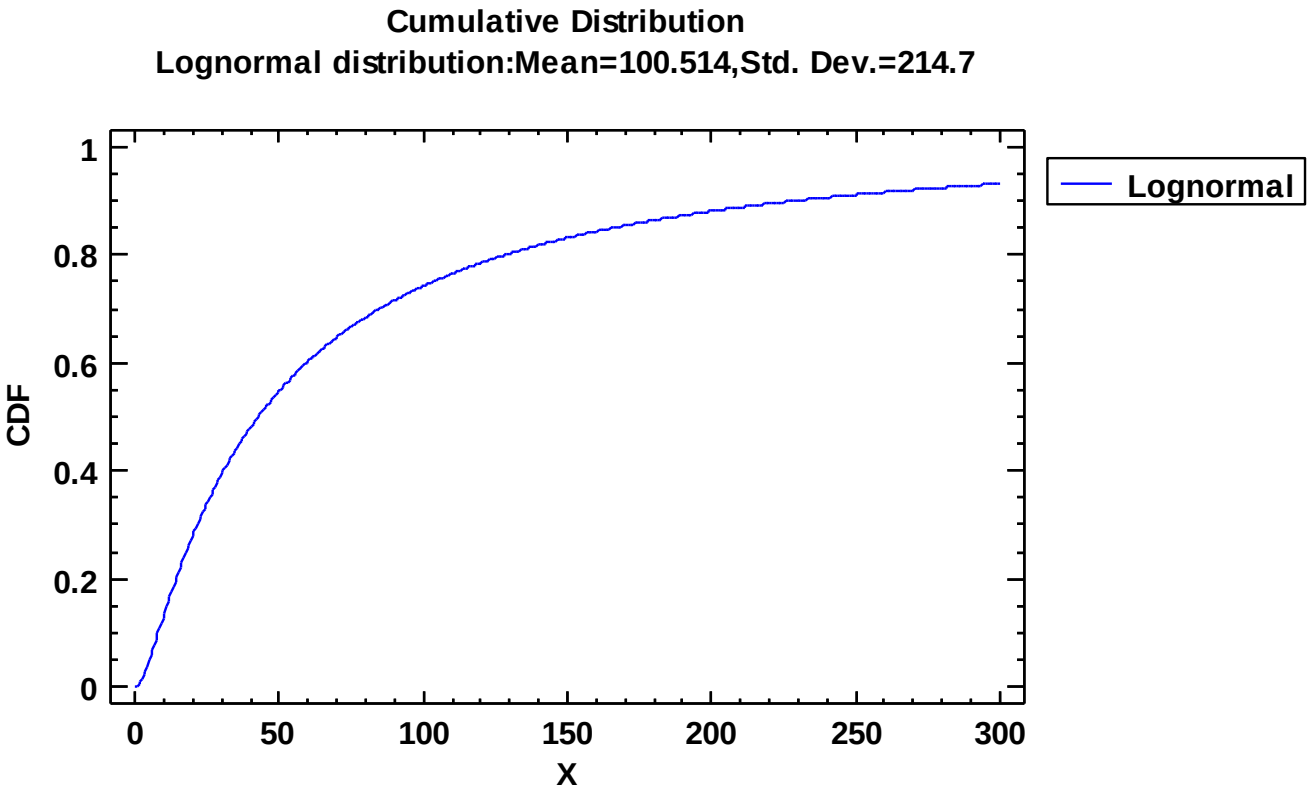
	Estimate	*95% LCL*	*95% UCL*
Mean	100.514	50.0203	447.337
Standard deviation	214.7	63.2189	2986.17
Median	42.6171	28.1219	77.6418
Lower quartile	17.6137	11.2882	27.5537
Upper quartile	103.114	59.1971	257.675
Interquartile range	85.5001	44.1705	238.614

Number of bootstrap subsamples: 1000

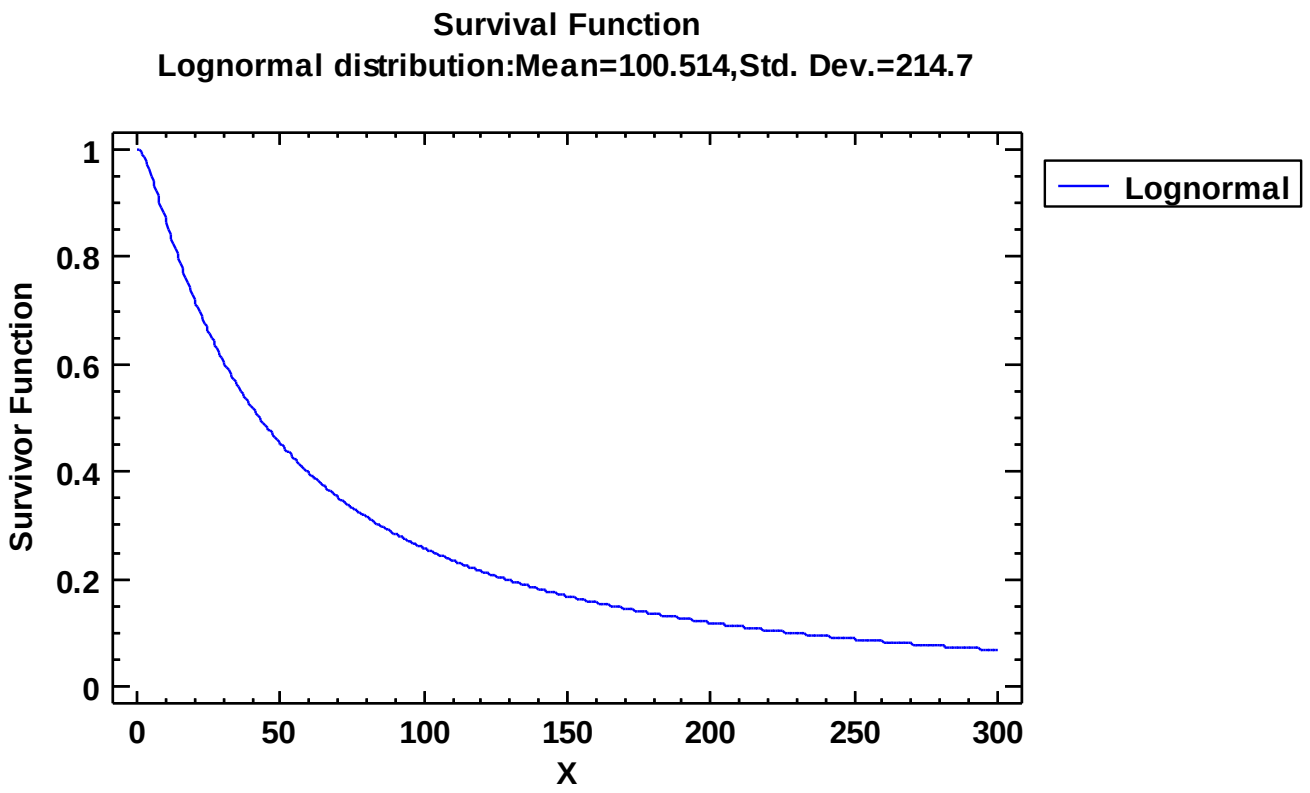
Plot of Fitted Distribution



Cumulative Distribution Function



Survival Function



Nonparametric Estimates

- A nonparametric estimate of the survival function may be obtained without assuming any particular distributional form.

- Estimated using the methods of Kaplan, Meier and Turnbull.

- Can also calculate confidence limits for the nonparametric survival function or CDF.

Example

Nonparametric Estimates

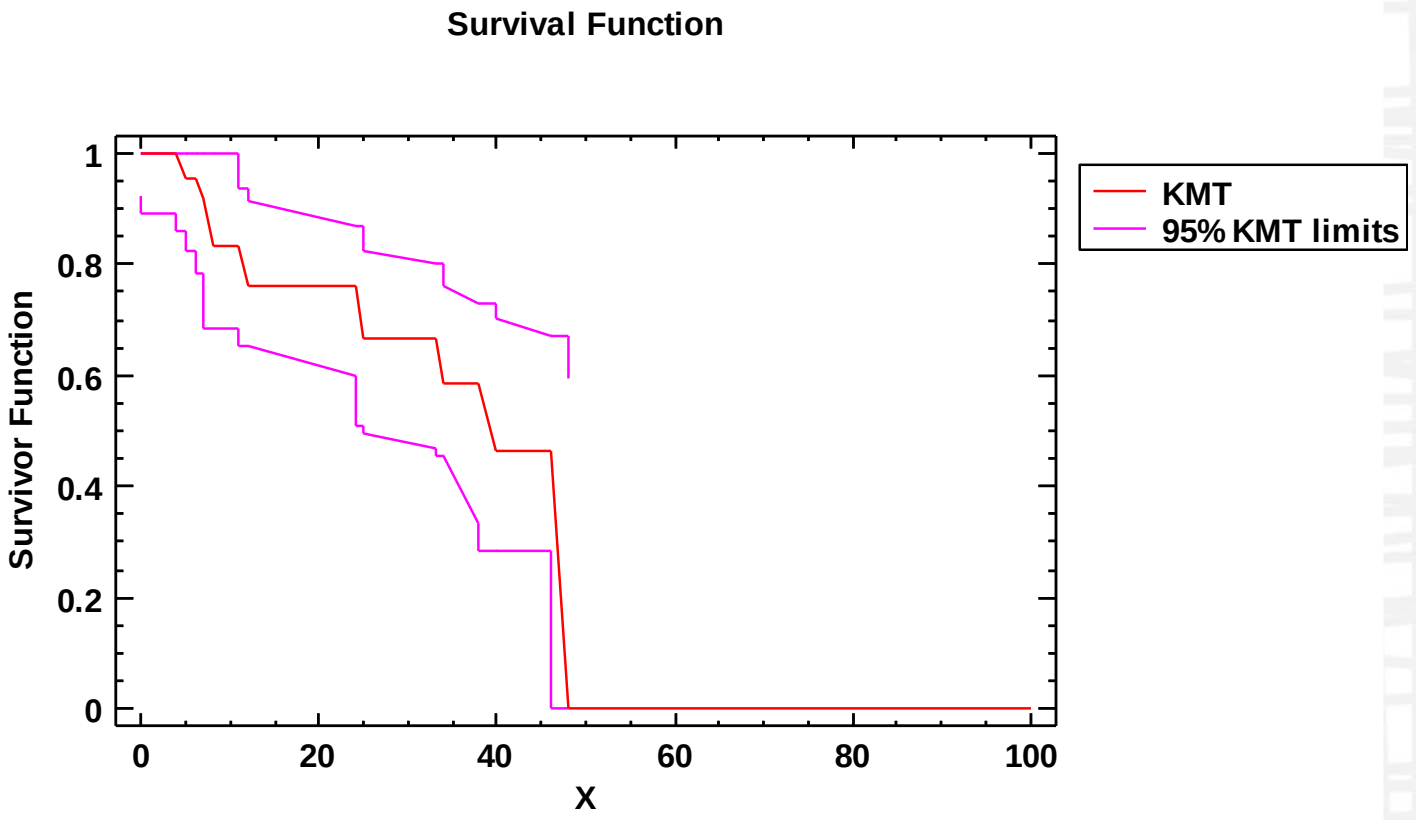
Kaplan-Meier-Turnbull Estimates

days	*CDF*	*Survival*	*95% LCL*	*95% UCL*
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

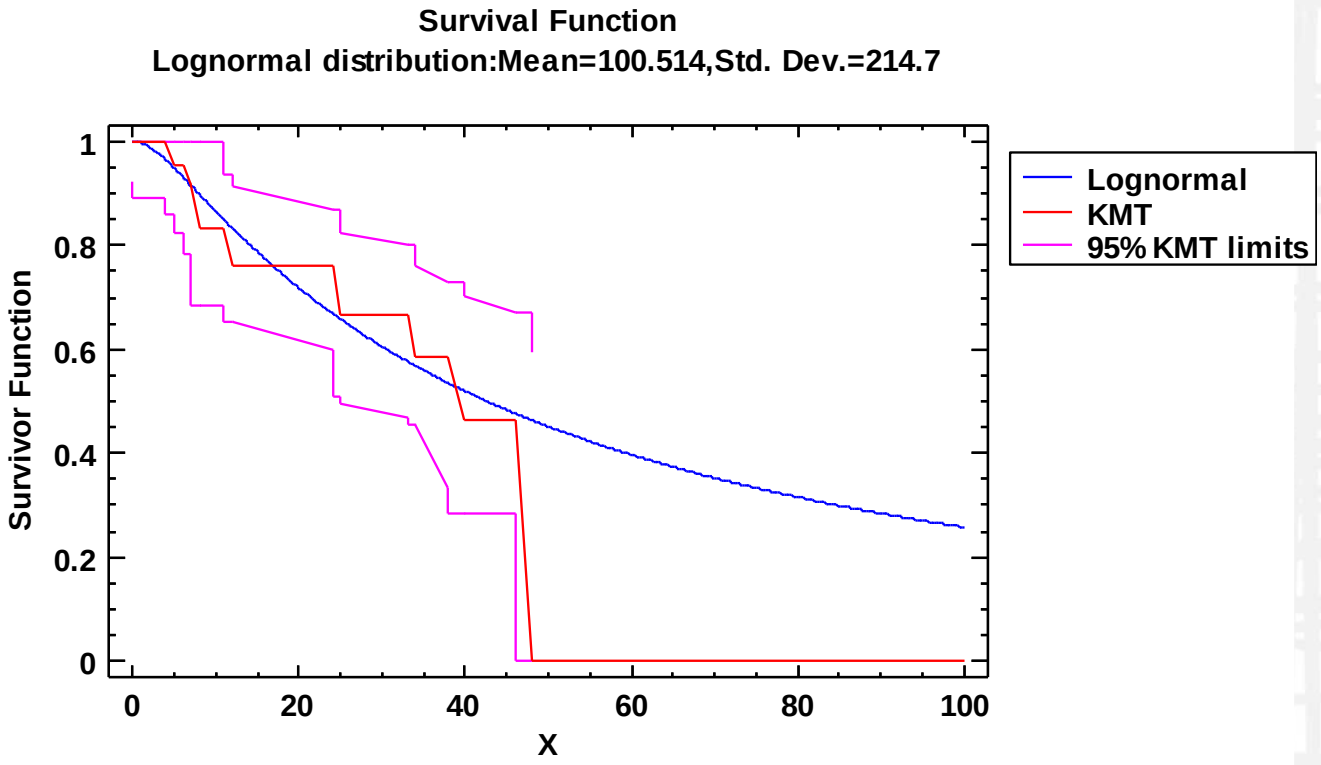
Statistics

Mean	33.5093
Standard deviation	15.9287
Standard error	2.34855

Example

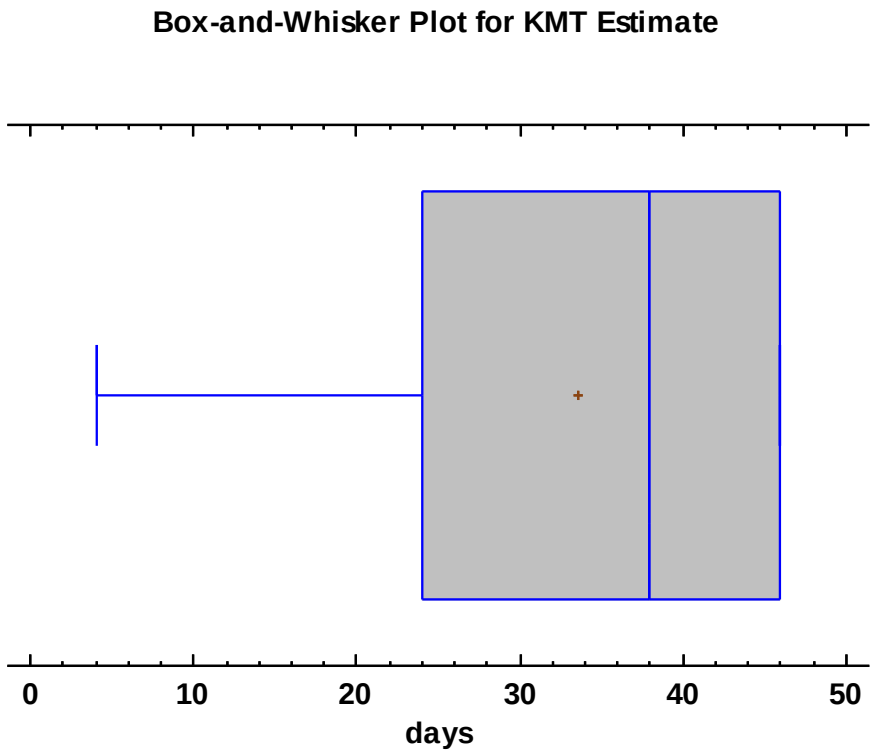

Survival Function
Survivor Function
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
X
KMT
95% KMT limits

Example

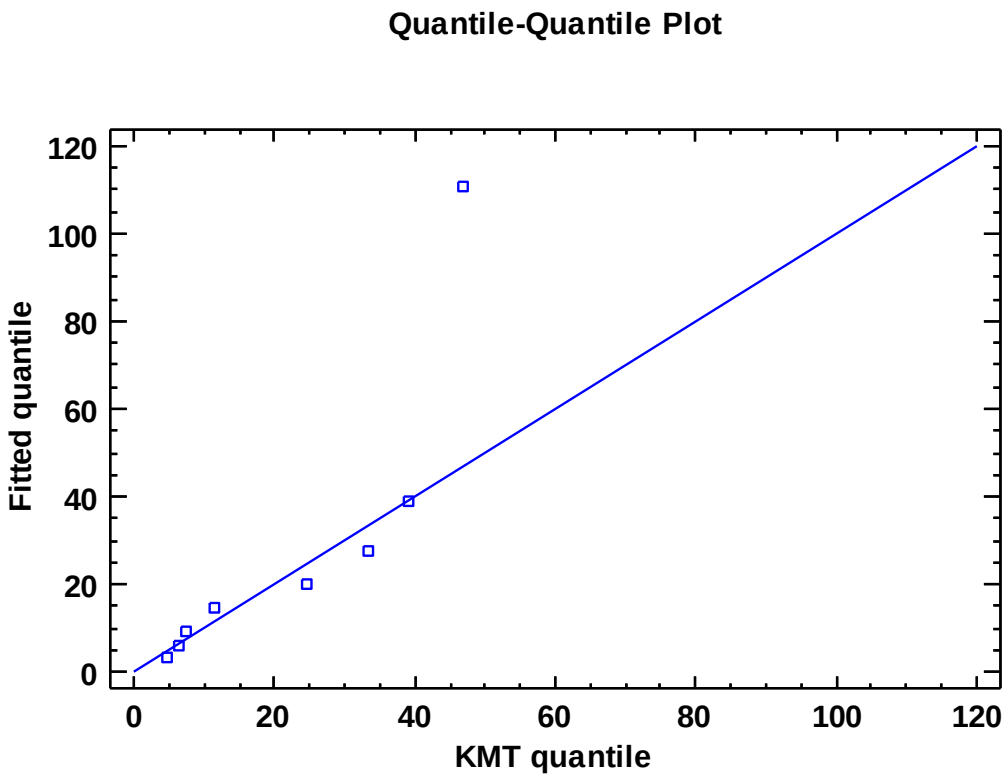

Survival Function
Lognormal distribution:Mean=100.514,Std. Dev.=214.7
Survivor Function
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
X
Lognormal
KMT
95% KMT limits

Box-and-Whisker Plot

- Based on 1st, 25th, 50th, 75th and 99th percentiles.



Quantile-Quantile Plot



References

- StatFolios and data files are at: *www.statgraphics.com/webinars*
- Finkelstein, D.M. and Wolfe, R.A. (1985). "A semiparametric model for regression analysis of interval-censored failure time data." *Biometrics* **41**, 731-740.
- Gentleman R, Vandal A (2018). *Icens: NPMLE for Censored and Truncated Data*. R package version 1.54.0.
- Helsel, D.R. (2005). Nondetects and Data Analysis: Statistics for Censored Environmental Data. Wiley, New York.
- 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, 3rd edition. Wiley, New York.
- R Package "interval" - https://cran.r-project.org/web/packages/interval/interval.pdf
- Tomlinson, M. S. (2003). "Effects of ground-water/surface-water interactions and land use on water quality." Written communication (draft USGS report).
- 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.