Chapter 551

Beta Distribution Fitting

Introduction

This module fits the beta probability distributions to a complete set of individual or grouped data values. It outputs various statistics and graphs that are useful in reliability and survival analysis.

The beta distribution is useful for fitting data which have an absolute maximum (and minimum). It finds some application as a lifetime distribution.

Technical Details

The four-parameter beta distribution is indexed by two shape parameters (*P* and *Q*) and two parameters representing the minimum (*A*) and maximum (*B*). We will not estimate *A* and *B*, but rather assume that they are known parameters.

Using these symbols, the beta density function may be written as

$$f(t|P,Q,A,B) = \frac{1}{B(P,Q)} \frac{(t-A)^{P-1}(B-t)^{Q-1}}{(B-A)^{P+Q-1}}, \quad P>0,\ Q>0,\ A<t<B$$

where

$$B(P,Q) = \frac{\Gamma(P)\Gamma(Q)}{\Gamma(P+Q)}$$

Making the transformation

$$X = \frac{(t-A)}{(B-A)}$$

results in the two-parameter beta distribution. This is also known as the standardized form of the beta distribution. In this case, the density function is

$$f(x|P,Q) = \frac{1}{B(P,Q)} x^{P-1}(1-x)^{Q-1}, \quad P>0,\ Q>0,\ 0<x<1$$

Reliability Function

The reliability (or survivorship) function, *R(t)*, gives the probability of surviving beyond time *t*. For the beta distribution, the reliability function is

$$R(T) = 1 - \int_A^T f\left(t|P,Q,A,B\right)dt$$

where the integral is known as the *incomplete beta function ratio*.

The conditional reliability function, *R(t,T)*, may also be of interest. The is the reliability of an item given that it has not failed by time *T*. The formula for the conditional reliability is

$$R(t) = \frac{R(T+t)}{R(T)}$$

Hazard Function

The hazard function represents the instantaneous failure rate. For this distribution, the hazard function is

$$h(t) = \frac{f(t)}{R(t)}$$

Kaplan-Meier Product-Limit Estimator

The production limit estimator is covered in the Distribution Fitting chapter and will not be repeated here.

Data Structure

Beta datasets require only a failure time variable. Censored data may not be fit with this program. An optional count variable which gives the number of items occurring at that time period. If the count variable is omitted, all counts are assumed to be one.

The table below shows the results of a study to test failure rate of a particular machine which has a maximum life of 100 hours. This particular experiment began with 10 items under test. After all items had failed, the experiment was stopped. These data are contained on the Beta dataset.

Beta dataset

Time
23.5
50.1
65.3
68.9
70.4
77.3
81.6
85.7
89.9
95.3

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Variable

Time Variable

This variable contains the failure times. Note that negative time values and time values less than the minimum parameter are treated as missing values. Zero time values are replaced by the value in the Zero Time Replacement option.

These time values represent elapsed times. If your data has dates (such as the failure date), you must subtract the starting date so that you can analyze the elapsed time.

Zero Time Replacement

Under normal conditions, a respondent beginning the study is “alive” and cannot “die” until after some small period of time has elapsed. Hence, a time value of zero is not defined and is ignored (treated as a missing value). If a zero time value does occur on the database, it is replaced by this positive amount. If you do not want zero time values replaced, enter a “0.0” here.

This option would be used when a “zero” on the database does not actually mean zero time. Instead, it means that the response occurred before the first reading was made and so the actual survival time is only known to be less.

Frequency Variable

Frequency Variable

This variable gives the number of individuals (the count or frequency) at a given failure (or censor) time. When omitted, each row receives a frequency of one. Frequency values should be positive integers.

Group Variable

Group Variable

An optional categorical (grouping) variable may be specified. If it is used, a separate analysis is conducted for each unique value of this variable.

Estimation Options

Survival and Hazard Confidence Limits Method

The standard nonparametric estimator of the reliability function is the Product Limit estimator. This option controls the method used to estimate the confidence limits of the estimated reliability. The options are Linear, Log Hazard, Arcsine Square Root, and Nelson-Aalen. The formulas used by these options were presented in the Technical Details section of the Distribution Fitting chapter. Although the Linear (Greenwood) is the most commonly used, recent studies have shown either the Log Hazard or the Arsine Square Root Hazard are better in the sense that they require a smaller sample size to be accurate.

Beta Minimum

This option sets the value of the minimum. Usually, this value is zero. All data values used must be greater than this value.

Beta Maximum

This option sets the value of the maximum. Often, this value is one. All data values used must be less than this value.

Options – Search

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of about 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained.

Minimum Relative Change

This value is used to control the iterative algorithms used in parameter estimation. When the relative change in any of the parameters is less than this amount, the iterative procedure is terminated.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary – Beta Percentiles

These options indicate whether to display the corresponding report.

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want to automatically attach labels to the values of the group variable (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Survival and Haz Rt Calculation Values

Percentiles

This option specifies a list of percentiles (range 1 to 99) at which the reliability (survivorship) is reported. The values should be separated by commas.

Specify sequences with a colon, putting the increment inside parentheses after the maximum in the sequence. For example: 5:25(5) means 5,10,15,20,25 and 1:5(2),10:20(2) means 1,3,5,10,12,14,16,18,20.

Times

This option specifies a list of times at which the percent surviving is reported. Individual values are separated by commas. You can specify a sequence by specifying the minimum and maximum separate by a colon and putting the increment inside parentheses. For example: 5:25(5) means 5,10,15,20,25. Avoid 0 and negative numbers. Use '(10)' alone to specify ten values between zero and the maximum value found in the data.

Report Options – Decimal Places

Time Decimals

This option specifies the number of decimal places shown on reported time values.

Plots Tab

These options control the attributes of the survival curves and the hazard curves.

Select Plots

Survivorship Plot - Probability Plot

These options indicate whether to display the corresponding plot. Click the plot format button to change the plot settings.

Line Resolution

This option specifies the number of points along the time axis at which calculations are made. This controls the resolution of the plots. Usually, a value between 50 and 100 is sufficient.

F(t) Calculation Method

This option specifies the method used to determine F(t), which is used to calculate the vertical plotting positions of points in the probability plot (the probability plot shows time (t) on the vertical axis and the distribution (normal, beta, Weibull, etc.) quantile on the horizontal axis).

The five calculation options are

- **Median (Approximate)**

  The most popular method is to calculate the median rank for each sorted data value. This is the median rank of the j^{th} sorted time value out of n values. Since the median rank requires extensive calculations, this approximation to the median rank is often used.

$$F(t_j) = \frac{j - 0.3}{n + 0.4}$$

- **Median (Exact)**

  The most popular method is to calculate the median rank for each sorted data value. This is the median rank of the j^{th} sorted time value out of n values. The exact value of the median rank is calculated using the formula

$$F(t_j) = \frac{1}{1 + \left(\frac{n-j+1}{j}\right) F_{0.5,2(n-j+1),2j}}$$

- **Mean**

  The mean rank is sometimes recommended. In this case, the formula is

$$F(t_j) = \frac{j}{n+1}$$

- **White's Formula**

  A formula proposed by White is sometimes recommended. The formula is

$$F(t_j) = \frac{j - 3/8}{n + 1/4}$$

- **$F(t_j)$ = [j - 0.5]/n**

  The following formula is sometimes used

$$F(t_j) = \frac{j - 0.5}{n}$$

Example 1 – Fitting a Beta Distribution

This section presents an example of how to fit a beta distribution. The data used were shown above and are found in the Beta dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab of the Beta Distribution Fitting window.

1 Open the Beta dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Beta.NCSS**.
- Click **Open**.

2 Open the Beta Distribution Fitting window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Beta Distribution Fitting** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Beta Distribution Fitting window, select the **Variables tab**.
- Double-click in the **Time Variable** box. This will bring up the variable selection window.
- Select **Time** from the list of variables and then click **Ok**.
- Click in the **Beta Maximum** box. Enter **100** for the maximum value.

4 Specify the plots.

- On the Beta Distribution Fitting window, select the **Plots tab**.
- Click on the **Beta Reliability Plot Format** button.
- Click on the **At-Risk Table** tab on the left and check **Show At-Risk Table**.
- Click **OK** to save the plot settings.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Data Summary

Data Summary

Type of Observation	Rows	Count	Minimum	Maximum	Average	Sigma
Failed	10	10	23.5	95.3	70.8	21.2021

This report displays a summary of the data that were analyzed. Scan this report to determine if there were any obvious data errors by double-checking the counts and the minimum and maximum.

Parameter Estimation Section

Beta Parameter Estimation

Parameter	Method of Moments Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
A (Minimum)	0	0			
B (Maximum)	100	100			
P (Shape 1)	2.548055	3.301583	1.485834	0.3894027	6.213764
Q (Shape 2)	1.050893	1.414615	0.577846	0.2820573	2.547172
Log-Likelihood		-3.403845			
Mean	70.8	70.00519			
Median	74.91825	73.002			
Mode	96.81711	84.73547			
Sigma	21.2021	19.16614			

This report displays parameter estimates along with standard errors and confidence limits in the maximum likelihood case.

Method of Moments Estimate

By equating the theoretical moments with the data moments, the following estimates are obtained.

$$\tilde{P} = \frac{\left[\frac{m_1 - A}{B-A}\right]^2\left[1 - \frac{m_1 - A}{B-A}\right]}{\left[\frac{m_2}{(B-A)^2}\right]} - \left[\frac{m_1 - A}{B-A}\right]$$

$$\tilde{Q} = \frac{\left[\frac{m_1 - A}{B-A}\right]\left[1 - \frac{m_1 - A}{B-A}\right]}{\left[\frac{m_2}{(B-A)^2}\right]} - \tilde{P}$$

where m_1 is the usual estimator of the mean and m_2 is the usual estimate of the standard deviation.

Maximum Likelihood Estimates of A, C, and D

These estimates maximize the likelihood function. The maximum likelihood equations are

$$\psi\left(\hat{P}\right) - \psi\left(\hat{P} + \hat{Q}\right) = \frac{1}{n}\sum_{j=1}^{n}\log\left(\frac{t_j - A}{B-A}\right)$$

$$\psi\left(\hat{Q}\right) - \psi\left(\hat{P} + \hat{Q}\right) = \frac{1}{n}\sum_{j=1}^{n}\log\left(\frac{B - t_j}{B-A}\right)$$

where $\psi(x)$ is the digamma function.

The formulas for the standard errors and confidence limits come from the inverse of the Fisher information matrix, *{f(i,j)}*. The standard errors are given as the square roots of the diagonal elements *f(1,1)* and *f(2,2)*. The confidence limits for *P* are

$$\hat{P}_{lower,1-\alpha/2} = \hat{P} - z_{1-\alpha/2}\sqrt{f(1,1)}$$

$$\hat{P}_{upper,1-\alpha/2} = \hat{P} + z_{1-\alpha/2}\sqrt{f(1,1)}$$

The confidence limits for Q are

$$\hat{Q}_{lower,1-\alpha/2} = \hat{Q} - z_{1-\alpha/2}\sqrt{f(2,2)}$$

$$\hat{Q}_{upper,1-\alpha/2} = \hat{Q} + z_{1-\alpha/2}\sqrt{f(2,2)}$$

Log-Likelihood

This is the value of the log-likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log-likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Mean

This is the mean time to failure (MTTF). It is the mean of the random variable (failure time) being studied given that the beta distribution provides a reasonable approximation to your data's actual distribution.

The formula for the mean is

$$Mean = A + \frac{P(B-A)}{P+Q}$$

Median

The median of the beta distribution is the value of t where $F(t)=0.5$.

$$Median = A + I(0.5, P, Q)$$

where $I(0.5,P,C)$ is the incomplete beta function.

Mode

The mode of the beta distribution is given by

$$Mode = A + \frac{(P-1)(B-A)}{P+Q-2}$$

when $A > 1$ and D otherwise.

Sigma

This is the standard deviation of the failure time. The formula for the standard deviation (sigma) of a beta random variable is

$$\sigma = \sqrt{\frac{PQ(B-A)^2}{(P+Q)^2(P+Q+1)}}$$

Inverse of Fisher Information Matrix

Inverse of Fisher Information Matrix

Parameter	P	Q
P	2.207702	0.6725335
Q	0.6725335	0.333906

This table gives the inverse of the Fisher information matrix for the two-parameter beta. These values are used in creating the standard errors and confidence limits of the parameters and reliability statistics. The approximate Fisher information matrix is given by the 2-by-2 matrix whose elements are

$$f(1,1) = \frac{\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

$$f(1,2) = f(2,1) = \frac{\psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

$$f(2,2) = \frac{\psi'\left(\hat{P}\right) - \psi'\left(\hat{P}+\hat{Q}\right)}{n\left(\psi'\left(\hat{P}\right)\psi'\left(\hat{Q}\right) - \psi'\left(\hat{P}+\hat{Q}\right)\left\{\psi'\left(\hat{P}\right)+\psi'\left(\hat{Q}\right)\right\}\right)}$$

where $\psi'(z)$ is the trigamma function and n represents the sample size.

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution
Confidence Limits Method: Linear (Greenwood)

Failure Time	Estimated Survival	Lower 95% C.L. Survival	Upper 95% C.L. Survival	Estimated Hazard	Lower 95% C.L. Hazard	Upper 95% C.L. Hazard	Sample Size
23.5	0.9000	0.7141	1.0000	0.1054	0.0000	0.3368	10
50.1	0.8000	0.5521	1.0000	0.2231	0.0000	0.5941	9
65.3	0.7000	0.4160	0.9840	0.3567	0.0161	0.8771	8
68.9	0.6000	0.2964	0.9036	0.5108	0.1013	1.2162	7
70.4	0.5000	0.1901	0.8099	0.6931	0.2108	1.6602	6
77.3	0.4000	0.0964	0.7036	0.9163	0.3515	2.3396	5
81.6	0.3000	0.0160	0.5840	1.2040	0.5378	4.1368	4
85.7	0.2000	0.0000	0.4479	1.6094	0.8031		3
89.9	0.1000	0.0000	0.2859	2.3026	1.2520		2
95.3							1

This report displays the Kaplan-Meier product-limit survival distribution and hazard function along with confidence limits. The formulas used were presented in the Technical Details section earlier in this chapter. Note that these estimates do not use the beta distribution in any way. They are the nonparametric estimates and are completely independent of the distribution that is being fit. We include them for reference.

Note that the Sample Size is given for each time period. As time progresses, participants are removed from the study, reducing the sample size. Hence, the survival results near the end of the study are based on only a few participants and are therefore less reliable. This shows up in a widening of the confidence limits.

Beta Reliability

Beta Reliability

Fail Time	Method of Moments Estimated Reliability	MLE Estimated Reliability
5.0	0.9995	0.9999
10.0	0.9969	0.9990
15.0	0.9914	0.9964
20.0	0.9821	0.9908
25.0	0.9685	0.9811
30.0	0.9500	0.9662
35.0	0.9261	0.9450
40.0	0.8964	0.9164
45.0	0.8606	0.8796
50.0	0.8182	0.8338
55.0	0.7689	0.7786
60.0	0.7126	0.7135
65.0	0.6488	0.6387
70.0	0.5776	0.5546
75.0	0.4986	0.4621
80.0	0.4120	0.3629
85.0	0.3178	0.2598
90.0	0.2164	0.1572
95.0	0.1088	0.0632
100.0	0.0000	0.0000

This report displays the estimated reliability (survivorship) at the time values that were specified in the Times option of the Reports Tab. Reliability may be thought of as the probability that failure occurs after the given failure time. Thus, (using the ML estimates) the probability is 0.944961 that failure will not occur until after 35 hours.

Two reliability estimates are provided. The first uses the method of moments estimates and the second uses the maximum likelihood estimates. Confidence limits are not available. The formulas used are as follows.

Estimated Reliability

The reliability (survivorship) is calculated using the beta distribution as

$$\hat{R}(t) = \hat{S}(t) = 1 - I\left(\frac{t-A}{B-A}; P, Q\right)$$

Beta Percentiles

Beta Percentiles

Percentile	Method of Moments Failure Time	MLE Failure Time
5.0	30.0	33.9
10.0	39.5	42.4
15.0	46.3	48.3
20.0	51.9	53.2
25.0	56.8	57.3
30.0	61.0	61.0
35.0	64.9	64.3
40.0	68.5	67.4
45.0	71.8	70.3
.	.	.
.	.	.
.	.	.

This report displays failure time percentiles using the method of moments and the maximum likelihood estimates. No confidence limit formulas are available.

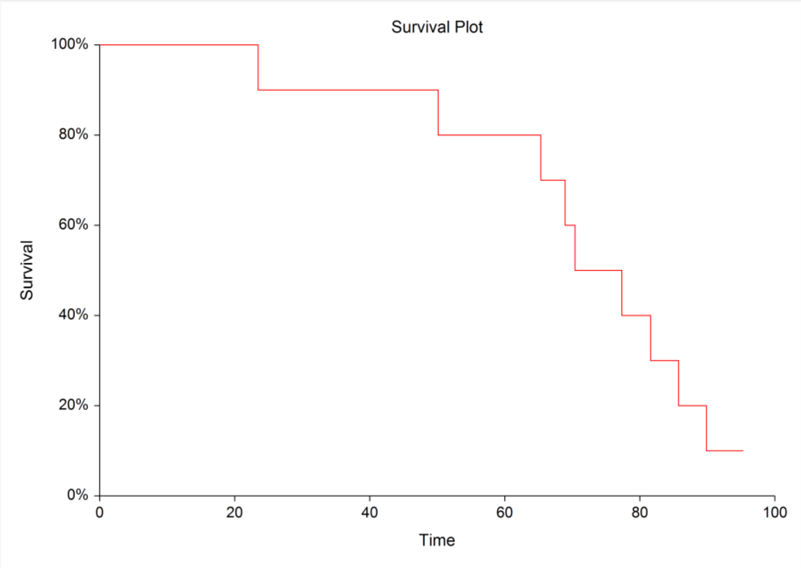
The formulas used are

Estimated Percentile

The time percentile at P (which ranges between zero and one hundred) is calculated using

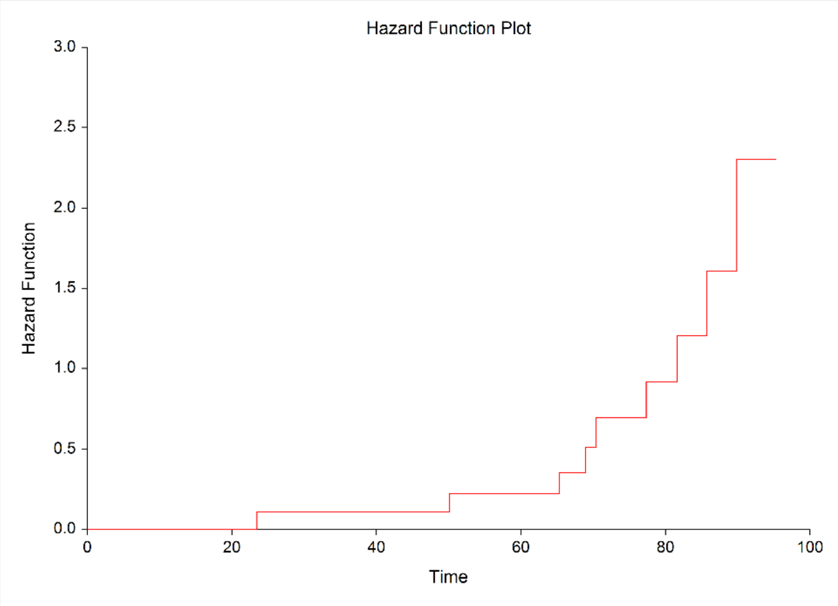
$$\hat{t}_p = \left[A + I(p; A, C)(B - A)\right] \times 100$$

Product-Limit Survivorship Plot



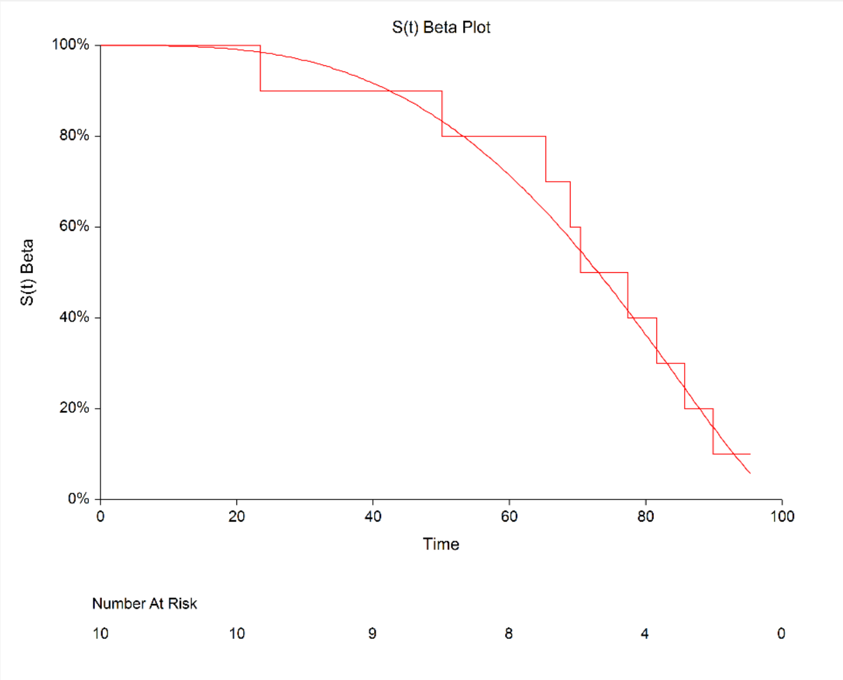
This plot shows the product-limit survivorship function for the data analyzed. If you have several groups, a separate line is drawn for each group. The step nature of the plot reflects the nonparametric product-limit survival curve.

Hazard Function Plot



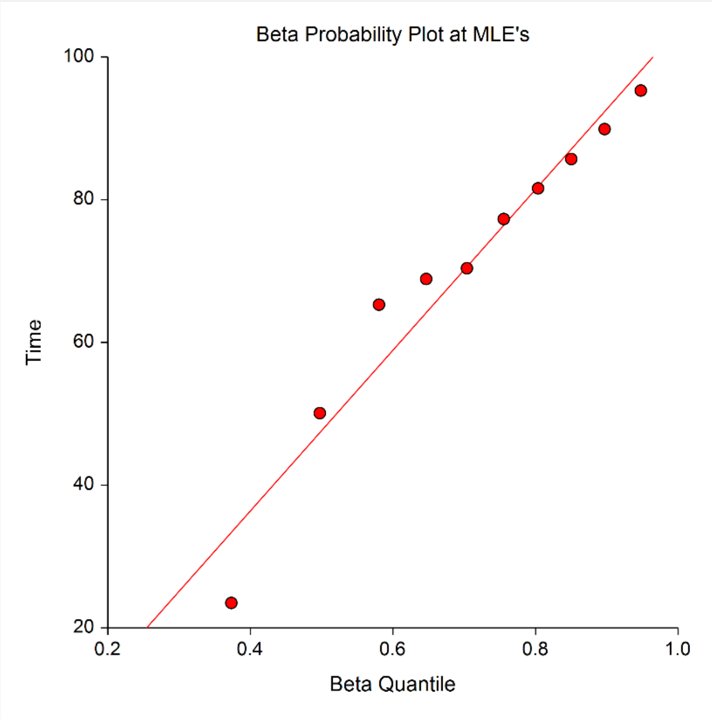
This plot shows the cumulative hazard function for the data analyzed. If you have several groups, then a separate line is drawn for each group. The shape of the hazard function is often used to determine an appropriate survival distribution.

Beta Reliability Plot



This plot shows the product-limit survival function (the step function) and the beta distribution overlaid. If you have several groups, a separate line is drawn for each group. The plot includes the number at risk at several times.

Beta Probability Plot



This is a beta probability plot for these data. The expected quantile of the theoretical distribution is plotted on the horizontal axis. The time value is plotted on the vertical axis. Also note that for grouped data, only one point is shown for each group.

This plot lets you investigate the goodness of fit of the beta distribution to your data. If the points seem to fall along a straight line, the beta probability model may be useful. You have to decide whether the beta distribution is a good fit to your data by looking at this plot and by comparing the value of the log-likelihood to that of other distributions.

Grouped Data

The case of grouped data causes special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the beta distribution for each (sorted) failure time. In the regular case, we used the rank of the observation in the overall dataset. However, in case of grouped data, we most use a modified rank. This modified rank, O_j , is computed as follows

$$O_j = O_p + I_j$$

where

$$I_j = \frac{(n+1) - O_p}{1+c}$$

were I_j is the increment for the jth failure; n is the total number of data points; O_p is the order of the previous failure; and c is the number of data points remaining in the data set, including the current data. Implementation details of this procedure may be found in Dodson (1994).