Chapter 552

Gamma Distribution Fitting

Introduction

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

The gamma distribution competes with the Weibull distribution as a model for lifetime. Since it is more complicated to deal with mathematically, it has been used less. While the Weibull is a purely heuristic model (approximating the data well), the gamma distribution does arise as a physical model since the sum of exponential random variables results in a gamma random variable.

At times, you may find that the distribution of log lifetime follows the gamma distribution.

The Three-Parameter Gamma Distribution

The three-parameter gamma distribution is indexed by a shape , a scale, and a threshold parameter. Many symbols have been used to represent these parameters in the statistical literature. We have selected the symbols *A*, *C*, and *D* for the shape, scale, and threshold. Our choice of symbols was made to make remembering their meanings easier. That is, just remember shApe, sCale, and thresholD and you will remember the general meaning of each symbol. Using these symbols, the three parameter gamma density function may be written as

$$f(t|A,C,D) = \frac{1}{C\Gamma(A)}\left(\frac{t-D}{C}\right)^{A-1} e^{-\frac{t-D}{C}}, \quad A>0, C>0, -\infty<D<\infty, t>D$$

Shape Parameter - A

This parameter controls the shape of the distribution. When $A = 1$, the gamma distribution is identical to the exponential distribution. When $C = 2$ and $A = \nu/2$, where ν is an integer, the gamma becomes the chi-square distribution with ν degrees of freedom. When *A* is restricted to integers, the gamma distribution is referred to as the Erlang distribution used in queueing theory.

Scale Parameter - C

This parameter controls the scale of the data. When *C* becomes large, the gamma distribution approaches the normal distribution.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. When *D* is set to zero, we obtain the two parameter gamma distribution. In *NCSS*, the threshold is not an estimated quantity but rather a fixed constant. Care should be used in using the threshold parameter because it forces the probability of failure to be zero between 0 and *D*.

Reliability Function

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

$$R(t) = 1 - I(t)$$

where *I(t)* in this case represents the incomplete gamma function.

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

Most gamma datasets require two (and often three) variables: the failure time variable, an optional censor variable formed by entering a zero for a censored observation or a one for a failed observation, and an optional count variable which gives the number of items occurring at that time period. If the censor variable is omitted, all time values represent observations from failed items. 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. This particular experiment began with 30 items under test. After the twelfth item failed at 152.7 hours, the experiment was stopped. The remaining eighteen observations were censored. That is, we know that they will fail at some time in the future. These data are contained on the Weibull dataset.

Weibull dataset

Time	Censor	Count
12.5	1	1
24.4	1	1
58.2	1	1
68.0	1	1
69.1	1	1
95.5	1	1
96.6	1	1
97.0	1	1
114.2	1	1
123.2	1	1
125.6	1	1
152.7	1	1
152.7	0	18

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 threshold 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.

Censor Variable

Censor Variable

This optional variable contains the censor indicator variable. The value is set to zero for censored observations and one for failed observations.

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.

Options

Threshold Value

This option controls the setting of the threshold parameter. When this value is set to zero (which is the default) the two-parameter gamma distribution is fit. You can put in a fixed, nonzero value for *D* here.

A cautionary note is needed. The maximum value that *D* can have is the minimum time value. If the minimum time is a censored observation you may be artificially constraining *D* to an inappropriately low value. It may make more sense to ignore these censored observations or to fit the two-parameter gamma.

Product Limit and Hazard Conf. 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.

Options – Probability Plot

Least Squares Model

When a probability plot is used to estimate the parameters of the gamma model, this option designates which variable (time or frequency) is used as the dependent variable.

- **F=A+B(Time)**

  On the probability plot, *F* is regressed on Time and the resulting intercept and slope are used to estimate the gamma parameters. See the discussion of probability plots below for more information.

- **Time=A+B(F)**

  On the probability plot, Time is regressed on *F* and the resulting intercept and slope are used to estimate the gamma parameters.

Shape Values

This options specifies values for the shape parameter, *A*, at which probability plots are to be generated. You can use a list of numbers separated by blanks or commas. Or, you can use the special list format: e.g. 0.5:2.0(0.5) which means 0.5 1.0 1.5 2.0. All values must be greater than zero.

Final Shape Value

This option specifies the shape parameter value that is used in the reports. The value in the list above that is closest to this value is used.

Use of this option usually requires two runs. In the first run, the probability plots of several trial *A* values are considered. The value of *A* for which the probability plot appears the straightest (in which all points fall along an imaginary straight line) is determined and used in a second run. Or, you may decide to use a value near the maximum likelihood estimate of *A*.

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 Report - Percentiles Report

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.

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 report or plot. Click the plot format button to change the plot settings.

Example 1 – Fitting a Gamma Distribution

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

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Gamma Distribution Fitting window.

1 Open the Weibull dataset.

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

2 Open the Gamma Distribution Fitting window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Gamma 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 Gamma 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**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select **Count** from the list of variables and then click **Ok**.
- Double-click in the **Censor Variable** box. This will bring up the variable selection window.
- Select **Censor** from the list of variables and then click **Ok**.

4 Specify the plots.

- On the Gamma Distribution Fitting window, select the **Plots tab**.
- Check the **Confidence Limits** box after clicking the format button for the survival plots.

5 Run the procedure.

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

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Minimum	Maximum	Average	Sigma
Failed	12	12	12.5	152.7	86.41666	41.66633
Censored	1	18	152.7	152.7		
Total	13	30	12.5	152.7		

Type of Censoring: Singly

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

Parameter Estimation Section

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Shape	2	2.407362	0.9228407	0.598627	4.216096
Scale	107.21	85.21822	36.31201	36.96837	196.4422
Threshold	0	0			
Log Likelihood		-80.6078			
Mean	214.42	205.1511			
Median	179.9356	177.551			
Mode	107.21	119.9329			
Sigma	151.6178	132.2218			

This report displays parameter estimates along with standard errors and confidence limits in the maximum likelihood case. In this example, we have set the threshold parameter to zero so we are fitting the two-parameter gamma distribution.

Probability Plot Estimate

This estimation procedure uses the data from the gamma probability plot to estimate the parameters. The estimation formula depends on which option was selected for the Least Squares Model. Note that the value of *A* is given—only *C* is estimated from the plot.

Least Squares Model: F=A+B(Time)

Using simple linear regression through the origin, we obtain the estimate of *C* as

$$\tilde{C} = \text{slope}$$

Least Squares Model: Time=A+B(F)

Using simple linear regression through the origin, we obtain the estimate of *C* as

$$\tilde{C} = \frac{1}{slope}$$

Maximum Likelihood Estimates of A, C, and D

These estimates maximize the likelihood 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 *A* are

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

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

The confidence limits for *C* are

$$\hat{C}_{lower,1-\alpha/2} = \frac{\hat{C}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{f(2,2)}}{\hat{C}}\right\}}$$

$$\hat{C}_{upper,1-\alpha/2} = \hat{C}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{f(2,2)}}{\hat{C}}\right\}$$

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 gamma distribution provides a reasonable approximation to your data's actual distribution.

The formula for the mean is

$$Mean = D + AC$$

Median

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

$$Median = D + I(0.5, A, C)$$

where $I(0.5,A,C)$ is the incomplete gamma function.

Mode

The mode of the gamma distribution is given by

$$Mode = D + C(A - 1)$$

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 gamma random variable is

$$\sigma = C\sqrt{A}$$

Inverse of Fisher Information Matrix

Inverse of Fisher Information Matrix

Parameter	Shape	Scale
Shape	0.8516349	-30.14704
Scale	-30.14704	1318.562

This table gives the inverse of the Fisher information matrix for the two-parameter gamma. These values are used in creating the standard errors and confidence limits of the parameters and reliability statistics. These statistics are very difficult to calculate directly for the gamma distribution when censored data are present. We use a large sample approximation that has been suggested by some authors. These results are only accurate when the shape parameter is greater than two.

The approximate Fisher information matrix is given by the 2-by-2 matrix whose elements are

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

$$f(1,2) = f(2,1) = \frac{-\hat{C}}{n\left(\hat{A}\psi'\left(\hat{A}\right) - 1\right)}$$

$$f(2,2) = \frac{\hat{C}^2 \psi'\left(\hat{A}\right)}{n\left(A\psi'\left(\hat{A}\right) - 1\right)}$$

where $\psi'(z)$ is the trigamma function and *n* represents the number of failed items (does not include censored items).

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution

Failure Time	Lower 95% C.L. Survival	Estimated Survival	Upper 95% C.L. Survival	Lower 95% C.L. Hazard	Estimated Hazard	Upper 95% C.L. Hazard	Sample Size
12.5	0.902433	0.966667	1.000000	0.000000	0.033902	0.102661	30
24.4	0.844073	0.933333	1.000000	0.000000	0.068993	0.169517	29
58.2	0.792648	0.900000	1.000000	0.000000	0.105361	0.232376	28
68.0	0.745025	0.866667	0.988308	0.011760	0.143101	0.294338	27
69.1	0.699975	0.833333	0.966692	0.033875	0.182322	0.356711	26
95.5	0.656864	0.800000	0.943136	0.058545	0.223144	0.420278	25
96.6	0.615318	0.766667	0.918016	0.085541	0.265703	0.485616	24
97.0	0.575091	0.733333	0.891576	0.114765	0.310155	0.553227	23
114.2	0.536018	0.700000	0.863982	0.146203	0.356675	0.623588	22
123.2	0.497980	0.666667	0.835354	0.179900	0.405465	0.697196	21
125.6	0.460893	0.633333	0.805774	0.215952	0.456758	0.774590	20
152.7	0.424695	0.600000	0.775305	0.254499	0.510826	0.856383	19
152.7+							18

Confidence Limits Method: Linear (Greenwood)

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 gamma 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 censored observations are marked with a plus sign on their time value. The survival and hazard functions are not calculated for censored observations.

Also 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.

Reliability Section

Reliability Section

Fail Time	ProbPlot Estimated Reliability	MLE Estimated Reliability	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
8.0	0.997351	0.998953	0.929195	1.000000
16.0	0.989912	0.994798	0.921475	1.000000
24.0	0.978387	0.987065	0.910223	1.000000
32.0	0.963401	0.975768	0.895440	1.000000
40.0	0.945512	0.961121	0.877317	1.000000
48.0	0.925214	0.943434	0.856131	1.000000
56.0	0.902947	0.923062	0.832187	1.000000
64.0	0.879099	0.900376	0.805796	0.994957
72.0	0.854012	0.875743	0.777251	0.974235
80.0	0.827987	0.849515	0.746827	0.952203
88.0	0.801290	0.822024	0.714773	0.929274
96.0	0.774151	0.793576	0.681318	0.905834
104.0	0.746772	0.764451	0.646672	0.882229
112.0	0.719328	0.734901	0.611036	0.858766
120.0	0.691969	0.705152	0.574602	0.835701
128.0	0.664826	0.675402	0.537563	0.813240
136.0	0.638009	0.645826	0.500114	0.791538
144.0	0.611611	0.616574	0.462450	0.770698
152.0	0.585711	0.587777	0.424773	0.750781
160.0	0.560373	0.559543	0.387282	0.731804

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 0.975768 that failure will not occur until after 32 hours. The 95% confidence for this estimated probability is 0.895440 to 1.000000.

Two reliability estimates are provided. The first uses the parameters estimated from the probability plot and the second uses the maximum likelihood estimates. Confidence limits are calculated for the maximum likelihood estimates. The formulas used are as follows.

Estimated Reliability

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

$$\hat{R}(t) = \hat{S}(t) = 1 - I\left(t - D; A, C\right)$$

Confidence Limits for Reliability

The confidence limits for this estimate are computed using the following formulas. Note that these estimates lack accuracy when A is less than 2.0.

$$\hat{R}_{upper}(t) = \hat{R}(t) - z_{1-\alpha/2}\sqrt{Var\left(\hat{R}(t)\right)}$$

$$\hat{R}_{upper}(t) = \hat{R}(t) + z_{1-\alpha/2}\sqrt{Var\left(\hat{R}(t)\right)}$$

where

$$Var\left(\hat{R}(t)\right) \cong \frac{\phi^2\left(\hat{\beta}\right)}{n}\left[\frac{2\left(t-D\right)^2}{\hat{C}\hat{A}^2} - \left(2\hat{C}-1\right)\left(1+\frac{\hat{\beta}}{2\sqrt{\hat{C}}}\right)\left(1+\frac{3\hat{\beta}}{2\sqrt{\hat{C}}}\right)\right]$$

where $\phi\left(z\right)$ is the standard normal density and

$$\hat{\beta} = \frac{\left(t-D\right)/\hat{A} - \hat{C}}{\sqrt{\hat{C}}}$$

Percentile Section

Percentile Section

Percentile	MLE Failure Time
5.00	45.2
10.00	64.1
15.00	79.9
20.00	94.2
25.00	107.9
30.00	121.4
35.00	134.9
40.00	148.6
45.00	162.7
50.00	177.6
55.00	193.2
60.00	210.1
65.00	228.5
70.00	249.1
75.00	272.6
80.00	300.4
85.00	335.1
90.00	382.2
95.00	459.4

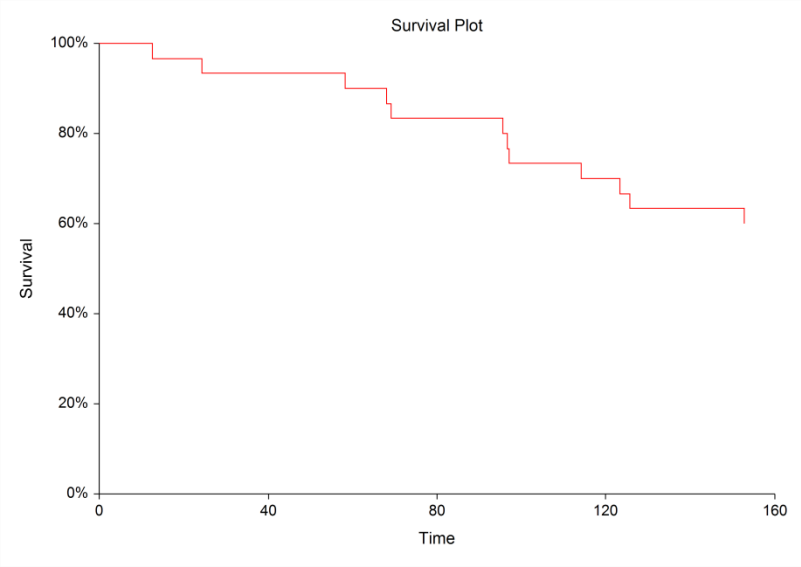
This report displays failure time percentiles using the maximum likelihood estimates. No confidence limit formulas are available.

Estimated Percentile

The time percentile at P (which ranges between 0 and 100) is calculated using

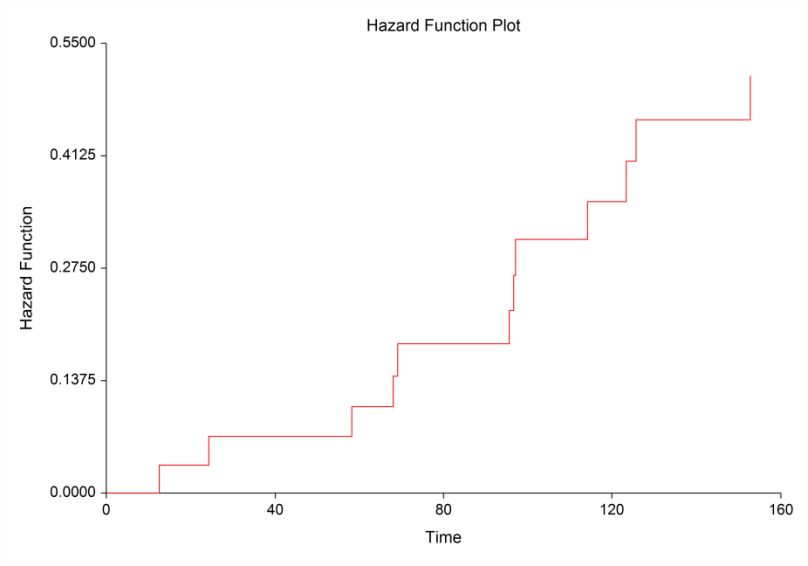
$$\hat{t}_p = \left[D + I(p; A, C)\right] \times 100$$

Product-Limit Survivorship Plot



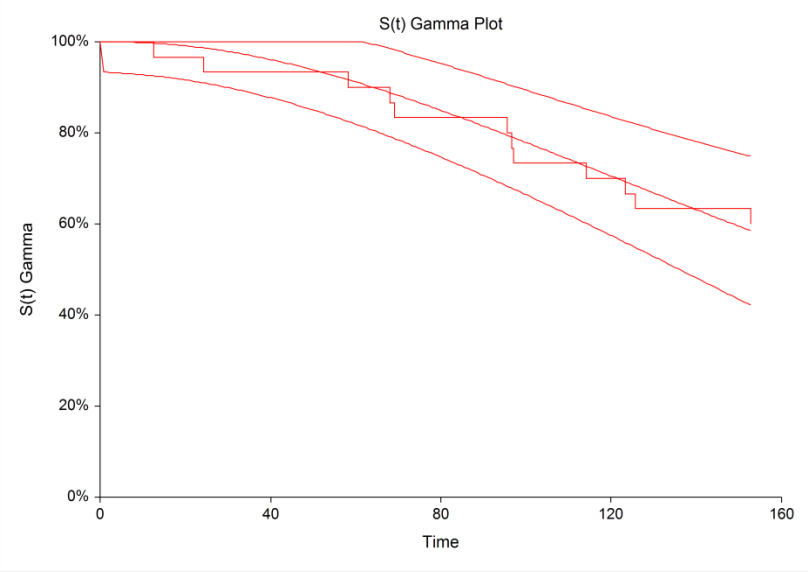
This plot shows the product-limit survivorship function for the data analyzed. If you have several groups, a separate line is drawn for each group. The step nature of the plot reflects the nonparametric product-limit survival curve.

Hazard Function Plot



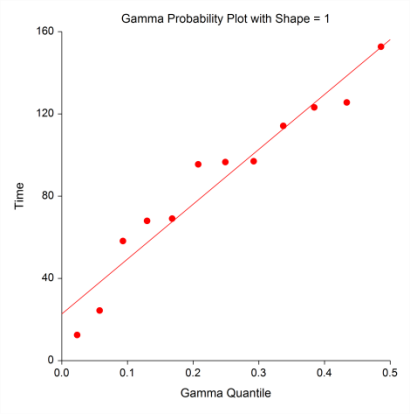
This plot shows the cumulative hazard function for the data analyzed. If you have several groups, then a separate line is drawn for each group. The shape of the hazard function is often used to determine an appropriate survival distribution.

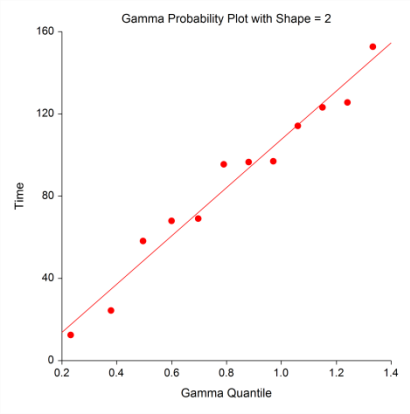
Gamma Reliability Plot

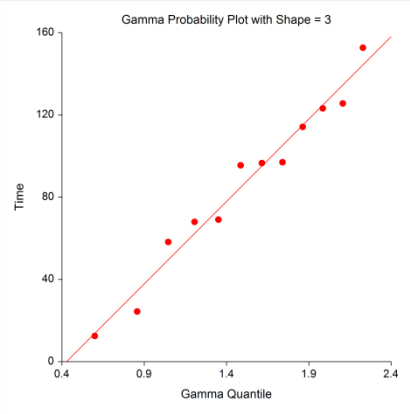


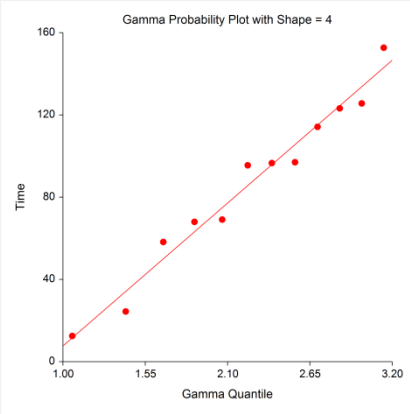
This plot shows the product-limit survival function (the step function) and the gamma distribution overlaid. The confidence limits are also displayed. If you have several groups, a separate line is drawn for each group.

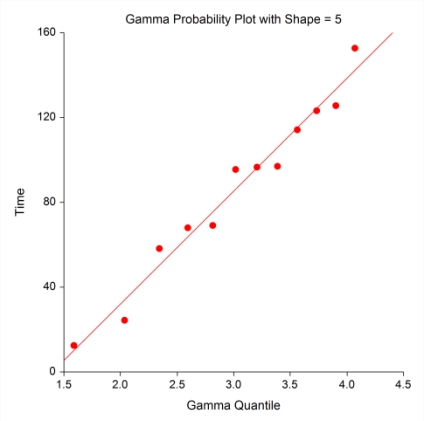
Gamma Probability Plots











There is a gamma probability plot for each specified value of the shape parameter (set in the Prob.Plot Shape Values option of the Search tab). The expected quantile of the theoretical distribution is plotted on the horizontal axis. The time value is plotted on the vertical axis. Note that censored points are not shown on this plot. Also note that for grouped data, only one point is shown for each group.

These plot let determine an appropriate value of *A*. They also let you investigate the goodness of fit of the gamma distribution to your data. You have to decide whether the gamma distribution is a good fit to your data by looking at these plots and by comparing the value of the log likelihood to that of other distributions.

For this particular set of data, it appears that *A* equal two or three would work just fine. Note that the maximum likelihood estimate of *A* is 2.4—right in between!

Multiple-Censored and Grouped Data

The case of grouped, or multiple-censored, data cause special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the gamma 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 or multiple-censored 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, both censored and uncensored; 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).