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 $A = v/2$, where v is an integer, the gamma becomes the chi-square distribution with v 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.

Gamma Probability Plot – F(t) Calculation Method

The user may specify 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 quantile on the horizontal axis).

The five calculation options are

- **Median (Approximate) ($F(t_j) = [j - 0.3]/[n + 0.4]$)**

  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) ($F(t_j) = 1/[1 + F(0.5, 2[n-j+1], 2j)\times[n-j+1]/j]$)**

  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 ($F(t_j) = j/[n + 1]$)**

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

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

- **White's Formula ($F(t_j) = [j - 3/8]/[n + 1/4]$)**

  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

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

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

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.

Setup

To run this example, complete the following steps:

1 Open the Weibull example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Weibull** and click **OK**.

2 Specify the Gamma Distribution Fitting procedure options

- Find and open the **Gamma Distribution Fitting** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Time Variable.....**Time**
Frequency Variable.....**Count**
Censor Variable.....**Censor**

Plots Tab

Gamma Reliability Plot Format (*Click the Button*)

Gamma Tab
Confidence Limits (Gamma Fit Line) .....**Checked**

At-Risk Table Tab
Show At-Risk Table .....**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Data Summary

Data Summary

Type of Censoring: Singly

Type of Observation	Rows	Count	Percent (%)	Minimum	Maximum	Average	Sigma
Failed	12	12	40%	12.5	152.7	86.41666	41.66633
Censored	1	18	60%	152.7	152.7		
Total	13	30	100%	12.5	152.7		

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.

Gamma Parameter Estimation

Gamma Parameter Estimation

		Maximum Likelihood (MLE)		
			95% Confidence Interval Limits	
Parameter	Probability Plot Estimate*	Estimate	Lower	Upper
A (Shape)	2	2.407362	0.598627	4.216096
C (Scale)	107.21	85.21823	36.96837	196.4422
D (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		

* Probability plot estimates were generated with F(t) calculated using the approximate median and using the model Time = A + B(F).

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

$$\tilde{C} = slope$$

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

$$\text{Mean} = D + AC$$

Median

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

$$\text{Median} = D + I(0.5, A, C)$$

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

Mode

The mode of the gamma distribution is given by

$$\text{Mode} = D + C(A - 1)$$

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	
Parameter	**A (Shape)**	**C (Scale)**
A (Shape)	0.8516349	-30.14704
C (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

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

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

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

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

Confidence Limits Method: Linear (Greenwood)

	Product-Limit Survival			Hazard Function			
		95% Confidence Interval Limits			95% Confidence Interval Limits		
Failure Time	Estimate	Lower	Upper	Estimate	Lower	Upper	Sample Size
12.5	0.9667	0.9024	1.0000	0.0339	0.0000	0.1027	30
24.4	0.9333	0.8441	1.0000	0.0690	0.0000	0.1695	29
58.2	0.9000	0.7926	1.0000	0.1054	0.0000	0.2324	28
68.0	0.8667	0.7450	0.9883	0.1431	0.0118	0.2943	27
69.1	0.8333	0.7000	0.9667	0.1823	0.0339	0.3567	26
95.5	0.8000	0.6569	0.9431	0.2231	0.0585	0.4203	25
96.6	0.7667	0.6153	0.9180	0.2657	0.0855	0.4856	24
97.0	0.7333	0.5751	0.8916	0.3102	0.1148	0.5532	23
114.2	0.7000	0.5360	0.8640	0.3567	0.1462	0.6236	22
123.2	0.6667	0.4980	0.8354	0.4055	0.1799	0.6972	21
125.6	0.6333	0.4609	0.8058	0.4568	0.2160	0.7746	20
152.7	0.6000	0.4247	0.7753	0.5108	0.2545	0.8564	19
152.7+							18

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.

Gamma Reliability

Gamma Reliability

Failure Time	Probability Plot Estimate	Maximum Likelihood (MLE) Estimate	95% Confidence Interval Limits Lower	Upper
8	0.9974	0.9990	0.9292	1.0000
16	0.9899	0.9948	0.9215	1.0000
24	0.9784	0.9871	0.9102	1.0000
32	0.9634	0.9758	0.8954	1.0000
40	0.9455	0.9611	0.8773	1.0000
48	0.9252	0.9434	0.8561	1.0000
56	0.9029	0.9231	0.8322	1.0000
64	0.8791	0.9004	0.8058	0.9950
72	0.8540	0.8757	0.7773	0.9742
80	0.8280	0.8495	0.7468	0.9522
88	0.8013	0.8220	0.7148	0.9293
96	0.7742	0.7936	0.6813	0.9058
104	0.7468	0.7645	0.6467	0.8822
112	0.7193	0.7349	0.6110	0.8588
120	0.6920	0.7052	0.5746	0.8357
128	0.6648	0.6754	0.5376	0.8132
136	0.6380	0.6458	0.5001	0.7915
144	0.6116	0.6166	0.4625	0.7707
152	0.5857	0.5878	0.4248	0.7508
160	0.5604	0.5595	0.3873	0.7318

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 Maximum Likelihood (MLE) Reliability

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

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

Confidence Interval Limits for Maximum Likelihood (MLE) 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.

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

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

where

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

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

and $\phi(z)$ is the standard normal density.

Gamma Percentiles

Gamma Percentiles

Percentile	MLE Failure Time
5	45.2
10	64.1
15	79.9
20	94.2
25	107.9
30	121.4
35	134.9
40	148.6
45	162.7
50	177.6
55	193.2
60	210.1
65	228.5
70	249.1
75	272.6
80	300.4
85	335.1
90	382.2
95	459.4

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

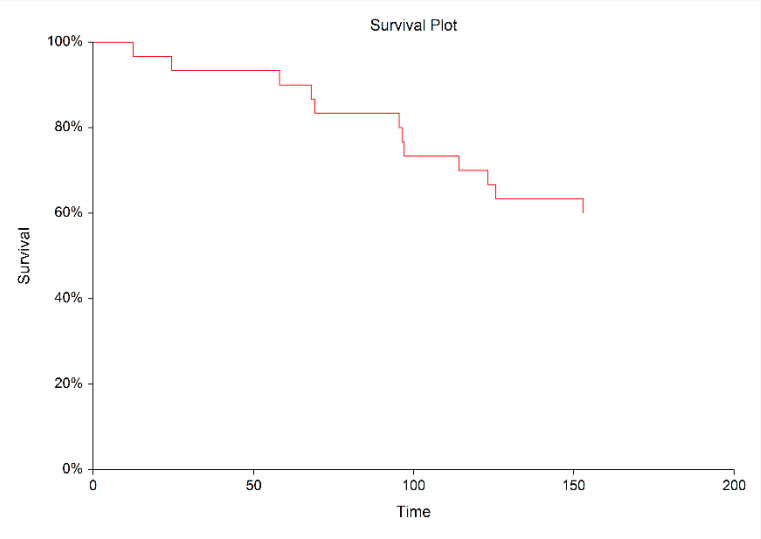
Estimated Percentile

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

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

Product-Limit Survivorship Plot

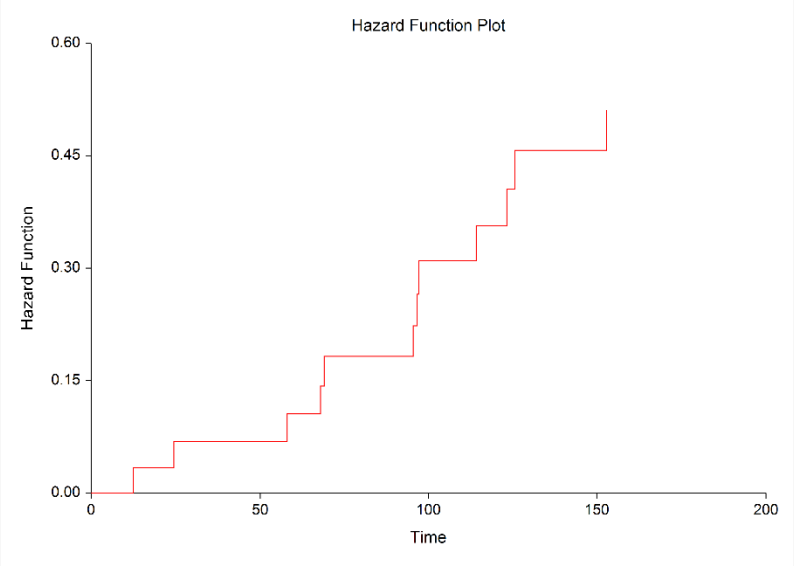
Product-Limit Survivorship Plot



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

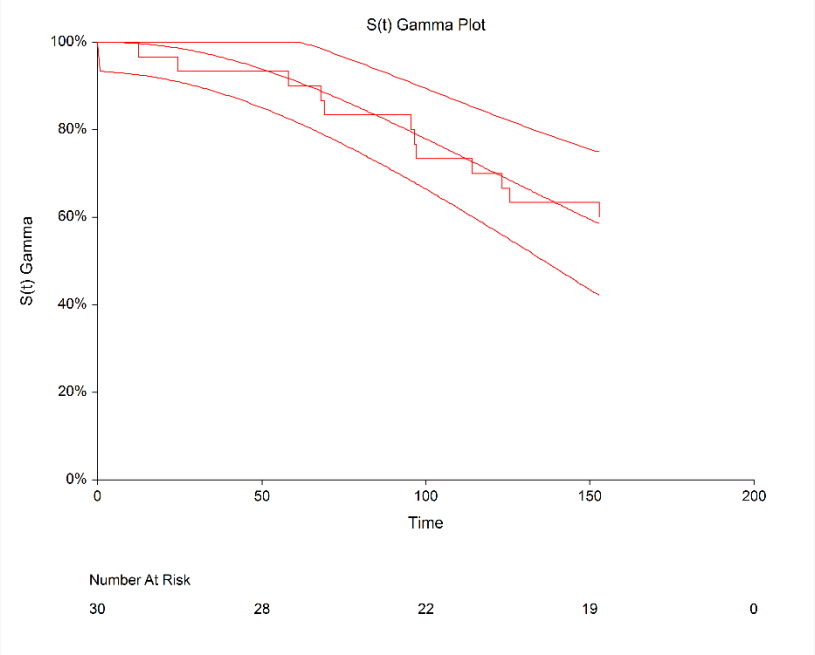
Hazard Function Plot



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

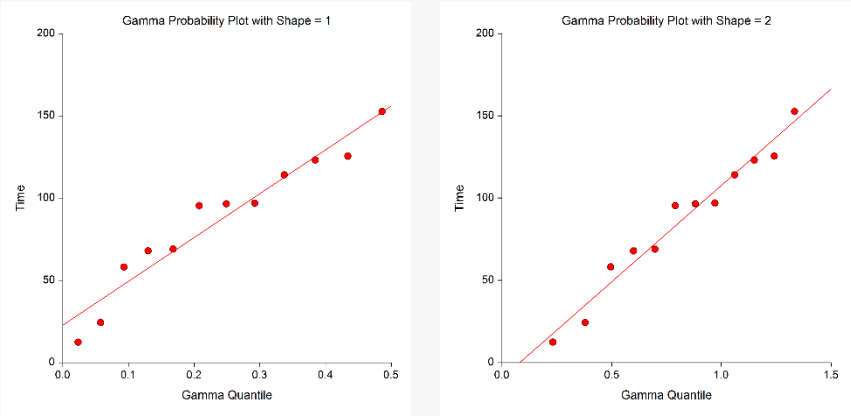
Gamma Reliability Plot

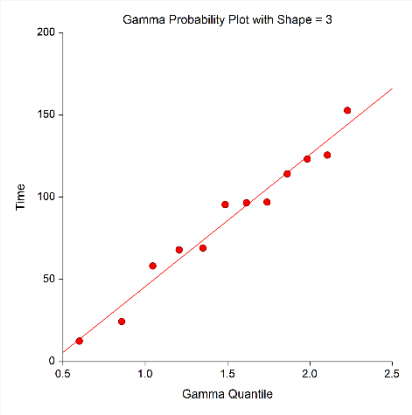


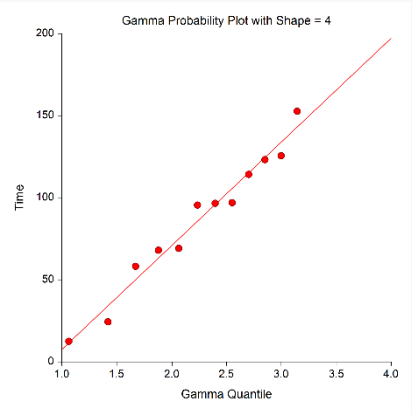
This plot shows the product-limit survival function (the step function) and the gamma distribution overlaid. The confidence limits are also displayed. If you have several groups, a separate line is drawn for each group. The plot includes the number at risk at several time points.

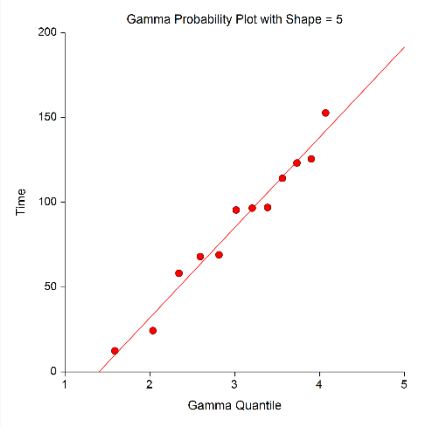
Gamma Probability Plots

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 plots help you determine an appropriate value of *A*. They also let you investigate the goodness of fit of the gamma distribution to your data. You must 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 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

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

where 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).