Chapter 550

Distribution (Weibull) Fitting

Introduction

This procedure estimates the parameters of the exponential, extreme value, logistic, log-logistic, lognormal, normal, and Weibull probability distributions by maximum likelihood. It can fit complete, right censored, left censored, interval censored (readout), and grouped data values. It also computes the nonparametric Kaplan-Meier and Nelson-Aalen estimates of survival and associated hazard rates. It outputs various statistics and graphs that are useful in reliability and survival analysis. When the choice of the probability distribution is in doubt, the procedure helps select an appropriate probability distribution from those available.

Features of this procedure include:

1. Probability plotting, hazard plotting, and reliability plotting for the common life distributions. The data may be any combination of complete, right censored, left censored, and interval censored data.
2. Maximum likelihood and probability plot estimates of distribution parameters, percentiles, reliability (survival) functions, hazard rates, and hazard functions.
3. Confidence intervals for distribution parameters and percentiles.
4. Nonparametric estimates of survival using the Kaplan-Meier procedure.

Overview of Survival and Reliability Analysis

This procedure may be used to conduct either survival analysis or reliability analysis. When a study concerns a biological event associated with the study of animals (including humans), it is usually called *survival analysis*. When a study concerns machines in an industrial setting, it is usually called *reliability analysis*. Survival analysis emphasizes a nonparametric estimation approach (Kaplan-Meier estimation), while reliability analysis emphasizes a parametric approach (Weibull or lognormal estimation). In the rest of this chapter, when we refer to survival analysis, you can freely substitute 'reliability' for 'survival.' The two terms refer to the same type of analysis.

We will give a brief introduction to the subject in this section. For a complete account of survival analysis, we suggest the book by Klein and Moeschberger (1997).

Survival analysis is the study of the distribution of lifetimes. That is, it is the study of the elapsed time between an initiating event (birth, start of treatment, diagnosis, or start of operation) and a terminal event (death, relapse, cure, or machine failure). The data values are a mixture of complete (terminal event occurred) and censored (terminal event has not occurred) observations. From the data values, the survival analyst makes statements about the survival distribution of the failure times. This distribution allows questions about such quantities as survivability, expected lifetime, and mean time to failure to be answered.

Let *T* be the elapsed time until the occurrence of a specified event. The event may be death, occurrence of a disease, disappearance of a disease, appearance of a tumor, etc. The probability distribution of *T* may be specified using one of the following basic functions. Once one of these functions has been specified, the others may be derived using the mathematical relationships presented.

1. Probability density function, *f(t)*. This is the probability that an event occurs at time *t*.
2. Cumulative distribution function, *F(t)*. This is the probability that an individual survives until time *t.*

$$F(t) = \int_0^t f(x)dx$$

3. Survival function, *S(t)* or Reliability function, *R(t).* This is the probability that an individual survives beyond time *t*. This is usually the first quantity that is studied. It may be estimated using the nonparametric Kaplan-Meier curve or one of the parametric distribution functions.

$$R(t) = S(t) = \int_t^\infty f(x)dx = 1 - F(t)$$

$$S(t) = \exp\left[-\int_0^t h(x)dx\right] = \exp[-H(t)]$$

4. Hazard rate, *h(t)*. This is the probability that an individual at time *t* experiences the event in the next instant. It is a fundamental quantity in survival analysis. It is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, and the inverse of Mill's ratio in economics. The empirical hazard rate may be used to identify the appropriate probability distribution of a particular mechanism, since each distribution has a different hazard rate function. Some distributions have a hazard rate that decreases with time, others have a hazard rate that increases with time, some are constant, and some exhibit all three behaviors at different points in time.

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

5. Cumulative hazard function, *H(t)*. This is integral of *h(t)* from 0 to *t*.

$$H(t) = \int_0^t h(x)dx = -\ln[S(t)]$$

Nonparametric Estimators of Survival

There are two competing nonparametric estimators of the survival distribution, *S(t),* available in this procedure. The first is the common Kaplan-Meier Product limit estimator. The second is the Nelson-Aalen estimator of the cumulative hazard function, *H(t)*.

Kaplan-Meier Product-Limit Estimator

The most common nonparametric estimator of the survival function is called the Kaplan-Meier product limit estimator. This estimator is defined as follows in the range of time values for which there are data.

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_1 \le t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t_1 \le t \end{cases}$$

In the above equation, d_i represents the number of deaths at time t_i and Y_i represents the number of individuals who are at risk at time t_i.

The variance of *S(t)* is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)}$$

The product limit estimator may be used to estimate the cumulative hazard function *H(t)* using the relationship

$$\hat{H}(t) = -\log\left[\hat{S}(t)\right]$$

Linear (Greenwood) Confidence Limits

This estimator may be used to create confidence limits for *S(t)* using the formula

$$\hat{S}(t) \pm z_{1-\alpha/2}\sigma_S(t)\hat{S}(t)$$

where

$$\sigma_S^2(t) = \frac{\hat{V}\left[\hat{S}(t)\right]}{\hat{S}^2(t)}$$

and *z* is the appropriate value from the standard normal distribution. We call this the *Linear (Greenwood) confidence interval*.

Log Hazard Confidence Limits

Better confidence limits may be calculated using the logarithmic transformation of the hazard functions. These limits are

$$\hat{S}(t)^{1/\theta}, \hat{S}(t)^{\theta}$$

where

$$\theta = \exp\left\{\frac{z_{1-\alpha/2}\sigma_S(t)}{\log[\hat{S}(t)]}\right\}$$

ArcSine-Square Root Hazard Confidence Limits

Another set of confidence limits using an improving transformation is given by the (intimidating) formula

$$\sin^2\left\{\max\left[0, \arcsin\left(\hat{S}(t)^{1/2} - 0.5 z_{1-\alpha/2}\sigma_S(t)\left(\frac{\hat{S}(t)}{1-\hat{S}(t)}\right)^{1/2}\right]\right\} \le S(t)$$

$$\le \sin^2\left\{\min\left[\frac{\pi}{2}, \arcsin(\hat{S}(t)^{1/2} + 0.5 z_{1-\alpha/2}\sigma_S(t)\left(\frac{\hat{S}(t)}{1-\hat{S}(t)}\right)^{1/2}\right]\right\}$$

Nelson-Aalen Hazard Confidence Limits

An alternative estimator of *H(t)*, which has better small sample size properties is the Nelson-Aalen estimator given by

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t < t_1 \\ \sum_{t_1 \le t} \frac{d_i}{Y_i} & \text{if } t \le t_1 \end{cases}$$

The variance of this estimate is given by the formula

$$\sigma_{\tilde{H}}^2(t) = \sum_{t_i \le t} \frac{(Y_i - d_i)d_i}{(Y_i - 1)Y_i^2}$$

The 100(1-alpha)% confidence limits for *H(t)* are calculated using

$$\tilde{H}(t)\exp\left(\pm z_{1-\alpha/2}\sigma_H(t) \,/\, \tilde{H}(t)\right)$$

This hazard function may be used to generate the Nelson-Aalen estimator of *S(t)* using the formula

$$\tilde{S}(t) = e^{\tilde{H}(t)}$$

Using these formulas, a fourth set of confidence limits for *S(t)* may be calculated as

$$\exp\left\{\tilde{H}(t) \pm z_{1-\alpha/2}\sigma_H(t)\right\}$$

Parametric Survival Distributions

This section presents the parametric probability distributions that may be analyzed with this procedure.

Normal Distribution

The normal distribution is one of the most commonly used in statistics. However, it is used infrequently as a lifetime distribution because it allows negative values while lifetimes are always positive. It has been found that the logarithms of failure times may be fit by the normal distribution. Hence the lognormal has become a popular distribution in reliability work, while the normal has been put on the sideline.

The normal distribution is indexed by a location (M) and a scale (S) parameter. A threshold parameter is meaningless in this case, since it is an adjustment to the location parameter. Using these symbols, the normal density function may be written as

$$f(t|M,S) = \frac{1}{S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-M}{S}\right)^2}, \quad -\infty < M < \infty, S > 0, -\infty < t < \infty$$

Location Parameter - M

The location parameter of the normal distribution is often called the mean.

Scale Parameter - S

The scale parameter of the normal distribution is usually called the standard deviation.

Lognormal Distribution

The normal distribution is one of the most commonly used in statistics. Although the normal distribution itself does not often work well with time-to-failure data, it has been found that the logarithm of failure time often does. Hence the lognormal has become a popular distribution in reliability work.

The lognormal distribution is indexed by a shape (S) , a scale (M), and a threshold (D) parameter. Using these symbols, the three parameter *lognormal* density function may be written as

$$f(t|M,S,D) = \frac{1}{(t-D)S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(t-D)-M}{S}\right)^2}, -\infty < M < \infty, S > 0, -\infty < D < \infty, t > D$$

It is often more convenient to work with logarithms to the base 10 (denoted by *log*) rather than logarithms to the base e (denoted by *ln*). The *lognormal10* density function is written as

$$f(t|M,S,D) = \frac{1}{\ln(10)(t-D)S\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(t-D)-M}{S}\right)^2}, -\infty < M < \infty, S > 0, -\infty < D < \infty, t > D$$

Shape Parameter - S

The shape parameter of the lognormal distribution of *t* is the standard deviation in the normal distribution of *ln(t-D)* or *log(t-D).* That is, the scale parameter of the normal distribution is the shape parameter of the lognormal distribution.

Scale Parameter - M

The scale parameter of the lognormal distribution *t* is the mean in the normal distribution of *ln(t-D)* or *log(t-D).* That is, the location parameter of the normal distribution is the scale parameter of the lognormal 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 lognormal distribution. When a positive value is given to *D,* we are inferring that no failures can occur between zero and *D.*

Reliability Function

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

$$R(t) = 1 - \Phi\left(\frac{\ln(t - D) - M}{S}\right)$$

where $\Phi(z)$ is the standard normal distribution function.

The conditional reliability function, *R(t,T),* may also be of interest. This 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

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

Weibull Distribution

The Weibull distribution is named for Professor Waloddi Weibull whose papers led to the wide use of the distribution. He demonstrated that the Weibull distribution fit many different datasets and gave good results, even for small samples. The Weibull distribution has found wide use in industrial fields where it is used to model time to failure data.

The three parameter Weibull distribution is indexed by a shape (B), a scale (C), and a threshold (D) parameter. Using these symbols, the three parameter Weibull density function may be written as

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

The symbol *t* represents the random variable (usually elapsed time). The threshold parameter *D* represents the minimum value of *t* that can occur. Setting the threshold to zero results in the common, two parameter Weibull distribution.

Shape Parameter - B

The shape (or power) parameter controls the overall shape of the density function. Typically, this value ranges between 0.5 and 8.0. The estimated standard errors and confidence limits displayed by the program are only valid when *B* > 2.0.

One of the reasons for the popularity of the Weibull distribution is that it includes other useful distributions as special cases or close approximations. For example, if

B = 1 The Weibull distribution is identical to the exponential distribution.

B = 2 The Weibull distribution is identical to the Rayleigh distribution.

B = 2.5 The Weibull distribution approximates the lognormal distribution.

B = 3.6 The Weibull distribution approximates the normal distribution.

Scale Parameter - C

The scale parameter only changes the scale of the density function along the time axis. Hence, a change in this parameter has the same effect on the distribution as a change in the scale of time—for example, from days to months or from hours to days. However, it does not change the actual shape of the distribution.

C is known as the *characteristic life.* No matter what the shape, 63.2% of the population fails by *t* = *C*+*D*.

Some authors use 1/*C* instead of *C* as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how you usually scale a set of numbers. For example, remember how you create a z-score when dealing with the normal data or create a percentage by dividing by the maximum.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. Often, this parameter is referred to as the *location* parameter. We use 'threshold' rather than 'location' to stress that this parameter sets the minimum time. We reserve 'location' to represent the center of the distribution. This is a fine point and we are not upset when people refer to this as the location parameter.

When *D* is set to zero, we obtain the two parameter Weibull distribution. It is possible, but unusual, for *D* to have a negative value. When using a search algorithm to find the estimated value of *D*, a nonzero value will almost certainly be found. However, you should decide physically if a zero probability of failure in the interval between 0 and *D* is truly warranted.

A downward or upward sloping tail on the Weibull probability plot or values of $B > 6.0$ are indications that a nonzero threshold parameter will produce a better fit to your data.

Negative values of *D* represent an amount of time that has been subtracted from the actual times. On the other hand, positive values of *D* represent a period of time between the starting point and when any failures can occur. For example, positive values of *D* may represent the amount of time between the production of an item and when it is placed in service.

Relationship to the Extreme Value Distribution

The extreme value distribution is directly related to the Weibull distribution. If $x = ln(t)$ and *t* follows the Weibull distribution, *x* follows the extreme value distribution.

Reliability Function

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

$$R(t) = e^{-\left(\frac{t-D}{C}\right)^B}$$

The reliability function is one minus the cumulative distribution function. That is,

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

The conditional reliability function, *R(t,T),* may also be of interest. This 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

$$h(t) = \frac{f(t)}{R(t)} = \frac{B}{C}\left(\frac{t-D}{C}\right)^{B-1}$$

Depending on the values of the distribution's parameters, the Weibull's hazard function can be decreasing (when $B<1$), constant (when $B=1$ at $1/C$), or increasing (when $B>1$) over time.

Extreme Value Distribution

The extreme value distribution is occasionally used to model lifetime data. It is included here because of its close relationship to the Weibull distribution. It turns out that if *t* is distributed as a Weibull random variable, then *ln(t)* is distributed as the extreme value distribution.

The density of the extreme value distribution may be written as

$$f(t|M,S) = \frac{1}{S}\exp\left(\frac{t-M}{S}\right)\exp\left(-\exp\left(\frac{t-M}{S}\right)\right), \quad S > 0$$

Exponential Distribution

The exponential distribution was one of the first distributions used to model lifetime data. It has now been superceded by the Weibull distribution, but is still used occasionally. The exponential distribution may be found from the Weibull distribution by setting *B* = 1.

The exponential distribution is a model for the life of products with a constant failure rate. The two parameter exponential distribution is indexed by both a scale and a threshold parameter. The density of the exponential distribution may be written as

$$f(t|S,D) = \frac{1}{S}\exp\left(-\frac{t-D}{S}\right), \quad S > 0$$

Scale Parameter - S

The scale parameter changes the scale of the density function along the time axis. Hence, a change in this parameter has the same effect on the distribution as a change in the scale of time—for example, from days to months or from hours to days. However, it does not change the actual shape of the distribution.

Some authors use 1/*S* instead of *S* as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how a set of numbers is usually scaled. For example, remember how z-scores are created when dealing with the normal distribution.

Threshold Parameter - D

The threshold parameter is the minimum value of the random variable *t*. Often, this parameter is referred to as the *location* parameter. We use 'threshold' rather than 'location' to stress that this parameter sets the minimum time. We reserve 'location' to represent the center of the distribution. This is a fine point and we are not upset when people refer to this as the location parameter.

When *D* is set to zero, we obtain the two parameter exponential distribution. It is possible, but unusual, for *D* to have a negative value. When using a search algorithm to find the estimated value of *D*, a nonzero value will almost certainly be found. However, you should decide physically if a zero probability of failure in the interval between 0 and *D* is truly warranted.

A downward or upward sloping tail on the exponential probability plot is an indication that a nonzero threshold parameter will produce a better fit to your data.

Negative values of *D* represent an amount of time that has been subtracted from the actual times. On the other hand, positive values of *D* represent a period of time between the starting point and when any failures can occur. For example, positive values of *D* may represent the amount of time between the production of an item and when it is placed in service.

Reliability Function

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

$$R(t) = e^{-\left(\frac{t-D}{S}\right)}$$

Hazard Function

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

$$h(t) = \frac{f(t)}{R(t)} = \frac{1}{S}$$

Logistic Distribution

The density of the logistic distribution may be written as

$$f(t|M,S) = \frac{\exp\left(\frac{t-M}{S}\right)}{S\left[1+\exp\left(\frac{t-M}{S}\right)\right]^2}, \quad S > 0$$

Log-logistic Distribution

The density of the log-logistic distribution may be written as

$$f(t|M,S,D) = \frac{\exp\left(\frac{\ln(t-D)-M}{S}\right)}{(t-D)S\left[1+\exp\left(\frac{\ln(t-D)-M}{S}\right)\right]^2}, \quad S > 0$$

The log-logistic distribution is used occasionally to model lifetime data, but it is so similar to the lognormal and the Weibull distributions that it adds little and is thus often dropped from consideration.

Parameter Estimation

The parameters of the reliability distributions may be estimated by maximum likelihood or by applying least squares regression to the probability plot. The probability plot method is popular because it uses a nice graphic (the probability plot) which allows a visual analysis of the goodness of fit of the distribution to the data. Maximum likelihood estimation is usually favored by statisticians because it has been shown to be optimum in most situations and because it provides estimates of standard errors and confidence limits. However, there are situations in which maximum likelihood does not do as well as the regression approach. For example, maximum likelihood does not do a good job of estimating the threshold parameter. When you want to include the threshold parameter in your model, we suggest that you use the regression approach to estimate it and then treat the threshold value as a known quantity in the maximum likelihood estimation.

Maximum Likelihood

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters which maximize the probability that the current set of data values occur.

The general form of the log-likelihood function is given by

$$L(\underline{P}) = \sum_{F} \ln\left(f(\underline{P}, t_k)\right) + \sum_{R} \ln\left(S(\underline{P}, t_k)\right) + \sum_{L} \ln\left(F(\underline{P}, t_k)\right) + \sum_{I} \ln\left(f(\underline{P}, t_{uk}) - f(\underline{P}, t_{lk})\right)$$

where *F* stands for the set of failed items, *R* stands for the set of right censored items, *L* stands for the set of left censored items, and *I* stands for the set of interval censored items. In the case of interval censored observations, t_{lk} represents the first time of the interval and t_{uk} represents the last time of the interval. Also, $\underline{P}$ represents one or two parameters as the case may be.

L($\underline{P}$*)* is maximized using two numerical procedures. First, a recently developed method called *differential evolution* is used. This is a robust maximization procedure that only requires evaluations of the function, but not its derivatives. The solution of the differential evolution phase is then used as starting values for the Newton-Raphson algorithm. Newton-Raphson is used because it provides an estimate of the variance-covariance matrix of the parameters, which is needed in computing confidence limits. The amount of emphasis on the differential evolution phase as opposed to the Newton-Raphson phase may be controlled using the maximum number of iterations allowed for each. Numerical differentiation is used to compute the first and second order derivatives that are needed by the Newton-Raphson procedure.

Probability Plot – F(t) Calculation Method

This option specifies the method used to determine F(t), which influences the probability plot estimates of the parameters and 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) ($F(t_j) = [j - 0.3]/[n + 0.4]$)**

  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

Survival data is somewhat more difficult to enter because of the presence of various types of censoring.

Failed or Complete

A failed observation is one in which the time until the terminal event was measured exactly; for example, the machine stopped working or the mouse died of the disease being studied.

Right Censored

A right censored observation provides a lower bound for the actual failure time. All that is known is that the failure occurred (or will occur) at some point after the given time value. Right censored observations occur when a study is terminated before all items have failed. They also occur when an item fails due to an event other than the one of interest.

Left Censored

A left censored observation provides an upper bound for the actual failure time. All we know is that the failure occurred at some point before the time value. Left censoring occurs when the items are not checked for failure until sometime after the study has begun. When a failed item is found, we do not know exactly when it failed, only that it was at some point before the left censor time.

Interval Censored or Readout

An interval censored observation is one in which we know that the failure occurred between two time values, but we do not know exactly when. This type of data is often called *readout* data. It occurs in situations where items are checked periodically for failures.

Sample Dataset

Most data sets require two (and often three) variables: the failure time variable, an optional censor variable indicating the type of censoring, and an optional count variable which gives the number of items occurring at that time. If the censor variable is omitted, all time values represent 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 the failure rate of a particular machine. This particular experiment began with 30 items being tested. After the twelfth item failed at 152.7 hours, the experiment was stopped. The remaining eighteen observations were right censored. That is, we know that they will fail at some time in the future. These data are contained in the WEIBULL database.

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 Weibull Distribution

This section presents an example of how to fit the Weibull 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 Distribution (Weibull) Fitting procedure options

- Find and open the **Distribution (Weibull) 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**

Options Tab

Random Seed.....**4002922** (for reproducibility)
Derivatives.....**0.00006**

Plots Tab

Hazard Function Plot.....**Checked**
Hazard Rate Plot.....**Checked**

Survival/Reliability Plot Format (*Click the Button*)

Confidence Limits (Distribution Fit Line).....**Checked**

Hazard Function Plot Format (*Click the Button*)

Confidence Limits (Distribution Fit Line).....**Checked**

Hazard Rate Plot Format (*Click the Button*)

Confidence Limits (Distribution Fit Line).....**Checked**

3 Run the procedure

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

Data Summary

Data Summary

Type of Observation	Rows	Count	Percent (%)	Minimum	Maximum
Failed	12	12	40.00%	12.5	152.7
Right Censored	1	18	60.00%	152.7	152.7
Left Censored	0	0			
Interval Censored	0	0			
Total	13	30	100.00%	12.5	152.7

This report displays a summary of the amount of 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

Weibull Parameter Estimation

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
B (Shape)	1.26829	1.511543	0.4130418	0.8847539	2.58237
C (Scale)	279.7478	238.3481	57.24827	148.8551	381.6452
D (Threshold)	0	0			
Log-Likelihood		-80.05649			
Mean	259.7101	214.9709			
Median	209.5383	187.0276			
Mode	82.19741	116.3898			
Sigma	206.2112	144.9315			

Estimation Details

Differential Evolution Iterations	39
Newton-Raphson Restart	1
Newton-Raphson Iterations	10
User-Entered Random Seed	4002922

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 for the maximum likelihood estimates. In this example, we have set the threshold parameter to zero so we are fitting the two-parameter Weibull distribution.

Probability Plot Estimate

This estimation procedure uses the data from the probability plot to estimate the parameters. The estimation formula depends on which option was selected for the Prob Plot Model (in the Estimation tab window).

Probability Plot Model: F = A + B(Time)

The cumulative distribution function *F(t)*

$$F(t) = 1 - e^{-\left(\frac{t-D}{C}\right)^B}$$

may be rearranged as (assuming *D* is zero)

$$\ln\left(-\ln\left(1 - F(t)\right)\right) = -B[\ln(C)] + B[\ln(t)]$$

This is now in a linear form. If we let *y = ln(-ln(1-F(t)))* and *x = ln(t)*, the above equation becomes

$$y = -B[\ln(C)] + Bx$$

Using simple linear regression, we can estimate the intercept and slope. Using these estimates, we obtain estimates of the Weibull parameters *B* and *C* as

$$B = \text{slope}$$

and

$$C = \exp\left(\frac{-\text{intercept}}{B}\right)$$

We assumed that *D* was zero. If *D* is not zero, it is treated as a known value and subtracted from the time values before the above regression is calculated.

Probability Plot Model: Time = A + B(F)

The cumulative distribution function *F(t)*

$$F(t) = 1 - e^{-\left(\frac{t-D}{C}\right)^B}$$

may be rearranged as (assuming *D* is zero)

$$\left(\frac{1}{B}\right)\ln\left(-\ln\left(1 - F(t)\right)\right) + \ln(C) = \ln(t)$$

This is now in a linear form. If we let *x = ln(-ln(1-F(t)))* and *y = ln(t)*, the above equation becomes

$$y = \frac{1}{B}x + \ln(C)$$

Using simple linear regression, we can estimate the intercept and slope. Using these estimates, we obtain estimates of the Weibull parameters *B* and *C* as

$$B = \frac{1}{\text{slope}}$$

and

$$C = \exp(\text{intercept})$$

Parameter estimates for the other distributions are found in a similar manner.

Maximum Likelihood Estimates of B, C, M, and S

These are the usual maximum likelihood estimates (MLE) of the parameters. The formulas for the standard errors and confidence limits use the estimated variance covariance matrix, which is the inverse of the Fisher information matrix, $\{vc_{i,j}\}$. The standard errors are given as the square roots of the diagonal elements $vc_{1,1}$ and $vc_{2,2}$.

In the case of the Weibull distribution, the confidence limits for *B* are

$$\hat{B}_{lower,1-\alpha/2} = \frac{\hat{B}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{1,1}}}{\hat{B}}\right\}}$$

$$\hat{B}_{upper,1-\alpha/2} = \hat{B}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{1,1}}}{\hat{B}}\right\}$$

In the case of the Weibull distribution, the confidence limits for *C* are

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

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

In the case of all other distributions, the confidence limits for *M* are

$$\hat{M}_{lower,1-\alpha/2} = \hat{M} - z_{1-\alpha/2}\sqrt{vc_{1,1}}$$

$$\hat{M}_{upper,1-\alpha/2} = \hat{M} + z_{1-\alpha/2}\sqrt{vc_{1,1}}$$

In the case of all other distributions, the confidence limits for *S* are

$$\hat{S}_{lower,1-\alpha/2} = \frac{\hat{S}}{\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{S}}\right\}}$$

$$\hat{S}_{upper,1-\alpha/2} = \hat{S}\exp\left\{\frac{z_{1-\alpha/2}\sqrt{vc_{2,2}}}{\hat{S}}\right\}$$

Log-Likelihood

This is the value of the log-likelihood function calculated using the maximum likelihood parameter estimates. This is the value of the function being maximized. It is often used as a goodness-of-fit statistic. You can compare the log-likelihood values achieved by each distribution and select as the best fitting the one with the maximum value.

Note that we have found that several popular statistical programs calculate this value without including all of the terms. Hence, they present erroneous values. The error occurs because they omit the 1/*t* term in the denominator of the lognormal and the Weibull log-likelihood functions. Also, they may fail to include a correction for using the logarithm to the base 10 in the lognormal10. Hopefully, future editions of these programs will calculate the likelihood correctly.

Mean

This is the mean time to failure (MTTF). It is the mean of the random variable (failure time) being studied given that the fitted distribution provides a reasonable approximation to your data's actual distribution. In the case of the Weibull distribution, the formula for the mean is

$$Mean = D + C\Gamma\left(1 + \frac{1}{B}\right)$$

where $\Gamma(x)$ is the gamma function.

Median

The median is the value of *t* where *F(t)=0.5*. In the case of the Weibull distribution, the formula for the median is

$$Median = D + C(\log 2)^{1/B}$$

Mode

The mode of the Weibull distribution is given by

$$Mode = D + C\left(1 - \frac{1}{B}\right)^{1/B}$$

Sigma

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

$$\sigma = C\sqrt{\Gamma\left(1+\frac{2}{B}\right) - \Gamma^2\left(1+\frac{1}{B}\right)}$$

where $\Gamma(x)$ is the gamma function.

Differential Evolution Iterations

This is the number of iterations used in the differential evolution phase of the maximum likelihood algorithm. If this value is equal to the maximum number of generations allowed, the algorithm did not converge, so you should increase the maximum number of generations and re-run the procedure.

Newton Raphson Restarts

This is the number of times the Newton Raphson phase of the maximum likelihood algorithm was restarted because the algorithm diverged. Make sure that the maximum number of restarts was not reached.

Newton Raphson Iterations

This is the number of iterations used in the Newton Raphson phase of the maximum likelihood algorithm. If this value is equal to the maximum number of iterations allowed, the algorithm did not converge, so you should increase the maximum number of iterations and re-run the procedure.

Variance-Covariance Matrix

Weibull Variance-Covariance Matrix

	Parameter	
Parameter	**B (Shape)**	**C (Scale)**
B (Shape)	0.1706035	-14.33368
C (Scale)	-14.33368	3277.364

This table gives the inverse of the Fisher information matrix evaluated at the maximum likelihood estimates which is an asymptotic estimate of the variance-covariance matrix of the two parameters. These values are calculated using numerical second-order derivatives.

Note that because these are numerical derivatives based on a random start provided by differential evolution, the values of the last two decimal places may vary from run to run. You can stabilize the values by changing the value of Derivatives constant, but this will have little effect on the overall accuracy of your results.

Kaplan-Meier Product-Limit Survival Distribution

Kaplan-Meier Product-Limit Survival Distribution

Confidence Limits Method: Linear (Greenwood)

Failure Time	Survival	Lower 95% C.L.	Upper 95% C.L.	Hazard Function	Lower 95% C.L.	Upper 95% C.L.	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 earlier. Note that these estimates do not use the selected parametric distribution in any way. They are the nonparametric estimates and are completely independent of the distribution that is being fit.

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 left censored and interval censored observations are treated as failed observations for the calculations on this report.

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 precise. This shows up as a widening of the confidence limits.

Nonparametric Hazard Rate

Nonparameteric Hazard Rate

Failure Time	Nonparametric Hazard Rate	Std Error of Hazard Rate	95% Lower Conf. Limit of Hazard Rate	95% Upper Conf. Limit of Hazard Rate
8.0	0.0016	0.0014	0.0003	0.0090
16.0	0.0019	0.0014	0.0005	0.0078
24.0	0.0016	0.0012	0.0004	0.0066
32.0	0.0015	0.0010	0.0004	0.0052
40.0	0.0016	0.0009	0.0005	0.0047
48.0	0.0021	0.0011	0.0008	0.0059
56.0	0.0024	0.0014	0.0008	0.0075
64.0	0.0027	0.0015	0.0009	0.0082
72.0	0.0036	0.0016	0.0015	0.0086
80.0	0.0042	0.0017	0.0019	0.0095
88.0	0.0043	0.0018	0.0018	0.0099
96.0	0.0045	0.0019	0.0019	0.0105
104.0	0.0052	0.0022	0.0023	0.0117
112.0	0.0054	0.0022	0.0024	0.0121
120.0	0.0047	0.0021	0.0019	0.0113
128.0	0.0038	0.0020	0.0014	0.0105
136.0	0.0038	0.0021	0.0013	0.0111
144.0	0.0038	0.0036	0.0006	0.0246
152.0	0.0081	0.0081	0.0011	0.0575

This report displays the nonparametric estimate of the hazard rate, *h(t).* Note that this is not the cumulative hazard function *H(t)* shown in the last report. It is the derivative of *H(t)*. Since it is *h(t)* that needs to be studied in order to determine the characteristics of the failure process, this report and its associated plot (which is shown below) become very import.

The formula for the Nelson-Aalen estimator of the cumulative hazard is

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t < t_1 \\ \sum_{t_1 \le t} \frac{d_i}{Y_i} & \text{if } t_1 \le t \end{cases}$$

The variance of this estimate is

$$\sigma_H^2(t) = \sum_{t_i \le t} \frac{(Y_i - d_i)d_i}{(Y_i - 1)Y_i^2}$$

In the above equation, d_i represents the number of deaths at time t_i and Y_i represents the number of individuals who are at risk at time t_i.

The hazard rate is estimated using kernel smoothing of the Nelson-Aalen estimator as given in Klein and Moeschberger (1997). The formulas for the estimated hazard rate and its variance are given by

$$\hat{h}(t) = \frac{1}{b}\sum_{D} K\left(\frac{t - t_k}{b}\right)\Delta\tilde{H}(t_k)$$

$$\sigma^2\left[\hat{h}(t)\right] = \frac{1}{b^2}\sum_{D} K\left(\frac{t - t_k}{b}\right)^2 \Delta\hat{V}\left[\tilde{H}(t_k)\right]$$

where b is the bandwidth about t and

$$\Delta\tilde{H}(t_k) = \tilde{H}(t_k) - \tilde{H}(t_{k-1})$$

$$\Delta\hat{V}\left[\tilde{H}(t_k)\right] = \hat{V}\left[\tilde{H}(t_k)\right] - \hat{V}\left[\tilde{H}(t_{k-1})\right]$$

Three choices are available for the kernel function *K(x)* in the above formulation. These are defined differently for various values of t. Note that the t_k's are for failed items only and that t_D is the maximum failure time. For the *uniform kernel* the formulas for the various values of t are

$$K(x) = \frac{1}{2} \quad \text{for} \quad t - b \le t \le t + b$$

$$K_L(x) = \frac{4(1+q^3)}{(1+q)^4} + \frac{6(1-q)}{(1+q)^3}x \quad \text{for} \quad t < b$$

$$K_R(x) = \frac{4(1+r^3)}{(1+r)^4} - \frac{6(1-r)}{(1+r)^3}x \quad \text{for} \quad t_D - b < t < t_D$$

where

$$q = \frac{t}{b}$$

$$r = \frac{t_D - t}{b}$$

For the *Epanechnikov kernel* the formulas for the various values of t are

$$K(x) = \frac{3}{4}(1 - x^2) \quad \text{for} \quad t - b \le t \le t + b$$

$$K_L(x) = K(x)(A + Bx) \quad \text{for} \quad t < b$$

$$K_R(x) = \frac{4(1+r^3)}{(1+r)^4} - \frac{6(1-r)}{(1+r)^3}x \quad \text{for} \quad t_D - b < t < t_D$$

where

$$A = \frac{64(2 - 4q + 6q^2 - 3q^3)}{(1+q)^4(19 - 18q + 3q^2)}$$

$$B = \frac{240(1-q)^2}{(1+q)^4(19 - 18q + 3q^2)}$$

$$q = \frac{t}{b}$$

$$r = \frac{t_D - t}{b}$$

For the *biweight kernel* the formulas for the various values of *t* are

$$K(x) = \frac{15}{16}(1 - x^2)^2 \quad \text{for} \quad t - b \le t \le t + b$$

$$K_L(x) = K(x)(A + Bx) \quad \text{for} \quad t < b$$

$$K_R(x) = K(-x)(A - Bx) \quad \text{for} \quad t_D - b < t < t_D$$

where

$$A = \frac{64(8 - 24q + 48q^2 - 45q^3 + 15q^4)}{(1+q)^5(81 - 168q + 126q^2 - 40q^3 + 5q^4)}$$

$$B = \frac{1120(1-q)^3}{(1+q)^5(81 - 168q + 126q^2 - 40q^3 + 5q^4)}$$

$$q = \frac{t}{b}$$

$$r = \frac{t_D - t}{b}$$

Confidence intervals for *h(t)* are given by

$$\hat{h}(t)\exp\left[\pm\frac{z_{1-\alpha/2}\sigma\left[\hat{h}(t)\right]}{\hat{h}(t)}\right]$$

Care must be taken when using these kernel-smoothed estimators since they are actually estimating a smoothed version of the hazard rate, not the hazard rate itself. Thus, they may be biased and are greatly influenced by the choice of the bandwidth *b*. We have found that you must experiment with *b* to find an appropriate value for each dataset.

Parametric Hazard Rate

Weibull Hazard Rate

Failure Time	Weibull Hazard Rate	95% Lower Conf. Limit of Hazard Rate	95% Upper Conf. Limit of Hazard Rate
8.0	0.0011	0.0005	0.0023
16.0	0.0016	0.0008	0.0032
24.0	0.0020	0.0010	0.0040
32.0	0.0023	0.0011	0.0046
40.0	0.0025	0.0012	0.0052
48.0	0.0028	0.0014	0.0057
56.0	0.0030	0.0015	0.0062
64.0	0.0032	0.0016	0.0066
72.0	0.0034	0.0017	0.0070
80.0	0.0036	0.0018	0.0074
88.0	0.0038	0.0019	0.0078
96.0	0.0040	0.0020	0.0081
104.0	0.0041	0.0020	0.0085
112.0	0.0043	0.0021	0.0088
120.0	0.0045	0.0022	0.0091
128.0	0.0046	0.0023	0.0094
136.0	0.0048	0.0023	0.0097
144.0	0.0049	0.0024	0.0100
152.0	0.0050	0.0025	0.0103
160.0	0.0052	0.0025	0.0105

This report displays the maximum likelihood estimates of the hazard rate, *h(t)*, based on the selected probability distribution and the definition of the hazard rate

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

Asymptotic confidence limits are computed using the formula from Nelson (1991) page 294.

$$\hat{h}(t)\exp\left[\pm\frac{z_{1-\alpha/2}s\left[\hat{h}(t)\right]}{\hat{h}(t)}\right]$$

where

$$s^2\left[\hat{h}(t)\right] = \left(\frac{\partial\hat{h}}{\partial P_1}\right)^2 vc_{1,1} + \left(\frac{\partial\hat{h}}{\partial P_2}\right)^2 vc_{2,2} + 2\left(\frac{\partial\hat{h}}{\partial P_1}\right)\left(\frac{\partial\hat{h}}{\partial P_2}\right)vc_{1,2}$$

The partial derivatives are evaluated using numerical differentiation.

Note that we have found that the above approximation behaves poorly for some distributions. However, this is the only formula that we have been able to find, so this is what we provide. If you find that the confidence limits have a strange appearance (especially, in that the width goes to zero), please ignore them. They should appear as nice expanding lines about the estimated hazard rate.

Parametric Failure Distribution

Weibull Failure Distribution

Failure Time	Prob Plot Estimate of Failure	Max Like Estimate of Failure	95% Lower Conf. Limit of Failure	95% Upper Conf. Limit of Failure
8.0	0.0110	0.0059	0.0005	0.0622
16.0	0.0262	0.0167	0.0027	0.1011
24.0	0.0434	0.0306	0.0067	0.1345
32.0	0.0619	0.0469	0.0127	0.1649
40.0	0.0814	0.0651	0.0209	0.1935
48.0	0.1014	0.0849	0.0310	0.2209
56.0	0.1219	0.1060	0.0431	0.2477
64.0	0.1427	0.1281	0.0569	0.2741
72.0	0.1637	0.1510	0.0723	0.3005
80.0	0.1849	0.1747	0.0888	0.3272
88.0	0.2060	0.1989	0.1064	0.3543
96.0	0.2271	0.2235	0.1245	0.3819
104.0	0.2480	0.2483	0.1431	0.4102
112.0	0.2689	0.2734	0.1617	0.4391
120.0	0.2895	0.2984	0.1801	0.4688
128.0	0.3099	0.3234	0.1982	0.4991
136.0	0.3301	0.3483	0.2158	0.5298
144.0	0.3500	0.3730	0.2328	0.5607
152.0	0.3695	0.3975	0.2491	0.5917
160.0	0.3888	0.4216	0.2648	0.6225

This report displays the estimated values of the cumulative failure distribution, *F(t)*, at the time values that were specified in the Times option of the Reports Tab. These failure values are the estimated probability that failure occurs by the given time point. For example, the maximum likelihood estimate that a unit will fail within 88 hours is 0.1989. The 95% confidence estimate of this probability is 0.1064 to 0.3543.

The asymptotic confidence limits are computed using the following formula:

$$\hat{F}_L(t) = \hat{F}\left(\hat{u} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

$$\hat{F}_U(t) = \hat{F}\left(\hat{u} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

where

$$\hat{u} = \frac{t - \hat{\mu}}{\hat{\sigma}}$$

$$\hat{V}(\hat{u}) = \frac{V(\hat{\mu}) + \hat{u}^2 V(\hat{\sigma}) + 2\hat{u}Cov(\hat{\mu}, \hat{\sigma})}{\hat{\sigma}^2}$$

Note that limits for the Weibull, lognormal, and log-logistic are found using the corresponding extreme value, normal, and logistic probability functions using the substitution *y=ln(t)*.

Parametric Reliability

Weibull Reliability

Failure Time	Prob Plot Estimate of Survival	Max Like Estimate of Survival	95% Lower Conf. Limit of Survival	95% Upper Conf. Limit of Survival
8.0	0.9890	0.9941	0.9378	0.9995
16.0	0.9738	0.9833	0.8989	0.9973
24.0	0.9566	0.9694	0.8655	0.9933
32.0	0.9381	0.9531	0.8351	0.9873
40.0	0.9186	0.9349	0.8065	0.9791
48.0	0.8986	0.9151	0.7791	0.9690
56.0	0.8781	0.8940	0.7523	0.9569
64.0	0.8573	0.8719	0.7259	0.9431
72.0	0.8363	0.8490	0.6995	0.9277
80.0	0.8151	0.8253	0.6728	0.9112
88.0	0.7940	0.8011	0.6457	0.8936
96.0	0.7729	0.7765	0.6181	0.8755
104.0	0.7520	0.7517	0.5898	0.8569
112.0	0.7311	0.7266	0.5609	0.8383
120.0	0.7105	0.7016	0.5312	0.8199
128.0	0.6901	0.6766	0.5009	0.8018
136.0	0.6699	0.6517	0.4702	0.7842
144.0	0.6500	0.6270	0.4393	0.7672
152.0	0.6305	0.6025	0.4083	0.7509
160.0	0.6112	0.5784	0.3775	0.7352

This report displays the estimated reliability (survival) 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 0.9531 that failure will not occur until after 32 hours. The 95% confidence for this estimated probability is 0.8351 to 0.9873.

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. (They have not been derived for the probability plot estimates for all data situations). The formulas used are as follows.

$$\hat{R}_L(t) = \hat{R}\left(\hat{u} - z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

$$\hat{R}_U(t) = \hat{R}\left(\hat{u} + z_{1-\alpha/2}\sqrt{\hat{V}(\hat{u})}\right)$$

where

$$\hat{u} = \frac{t - \hat{M}}{\hat{S}}$$

$$\hat{V}(\hat{u}) = \frac{V(\hat{M}) + \hat{u}^2 V(\hat{S}) + 2\hat{u}Cov(\hat{M},\hat{S})}{\hat{S}^2}$$

Note that limits for the Weibull, lognormal, and log-logistic are found using the corresponding extreme value, normal, and logistic probability functions using the substitution *y=ln(t)*.

Parametric Percentiles

Weibull Percentiles

Failure Time Percentage	Prob Plot Estimate of Failure Time	Max Like Estimate of Failure Time	95% Lower Conf. Limit of Failure Time	95% Upper Conf. Limit of Failure Time
5.00	26.9	33.4	14.2	78.4
10.00	47.4	53.8	28.5	101.4
15.00	66.8	71.6	42.7	120.3
20.00	85.7	88.4	56.5	138.3
25.00	104.7	104.5	69.7	156.8
30.00	124.1	120.5	82.2	176.7
35.00	144.0	136.5	93.9	198.6
40.00	164.7	152.8	104.8	222.9
45.00	186.5	169.6	115.0	250.1
50.00	209.5	187.0	124.7	280.5
55.00	234.3	205.4	134.0	314.8
60.00	261.1	225.0	143.1	353.7
65.00	290.7	246.1	152.1	398.3
70.00	323.8	269.5	161.3	450.3
75.00	361.9	295.8	170.9	512.1
80.00	407.1	326.5	181.3	588.2
85.00	463.5	364.1	193.0	686.7
90.00	540.0	413.9	207.4	826.0
95.00	664.5	492.6	227.8	1065.0

This report displays failure time percentiles and, for the maximum likelihood estimates, confidence intervals for those percentiles. For example, the estimated median failure time is 187 hours. The 95% confidence limits for the median time are 124.7 to 280.5. Note that these limits are very wide for two reasons. First, the sample size is small. Second, the shape parameter is less than 2.0.

The estimated $100p^{th}$ percentile and associated confidence interval is computed using the following steps:

1. Compute $w_p = F^{-1}(p)$
2. Compute $y_p = \hat{M} + w_p\hat{S}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1 / \hat{B}$.
3. Compute $V(y_p) = VC_{1,1} + y_p^2 VC_{2,2} + 2y_p VC_{1,2}$.
4. For the normal, extreme value, and logistic distributions, the confidence interval for the percentile is given by

$$T_{Lower,p} = y_p - z_{1-\alpha/2}\sqrt{V(y_p)} + D$$

$$T_{Upper,p} = y_p + z_{1-\alpha/2}\sqrt{V(y_p)} + D$$

For the lognormal, exponential, Weibull, and log-logistic distributions, the confidence interval for the percentile is given by

$$T_{Lower,p} = \exp\left(y_p - z_{1-\alpha/2}\sqrt{V(y_p)}\right) + D$$

$$T_{Upper,p} = \exp\left(y_p + z_{1-\alpha/2}\sqrt{V(y_p)}\right) + D$$

For the lognormal base 10 distribution, the confidence interval for the percentile is given by

$$T_{Lower,p} = 10^{y_p - z_{1-\alpha/2}\sqrt{V(y_p)}} + D$$

$$T_{Upper,p} = 10^{y_p + z_{1-\alpha/2}\sqrt{V(y_p)}} + D$$

Parametric Residual Life

Weibull Residual Life

Failure Time	Proportion Failing	25.0th %tile Residual Life	50.0th %tile Residual Life	75.0th %tile Residual Life	90.0th %tile Residual Life
8.0	0.0059	97.9	180.1	288.7	406.6
16.0	0.0167	92.5	174.0	282.2	399.9
24.0	0.0306	87.9	168.5	276.2	393.5
32.0	0.0469	83.8	163.5	270.6	387.5
40.0	0.0651	80.1	158.9	265.3	381.8
48.0	0.0849	76.9	154.5	260.2	376.3
56.0	0.1060	73.9	150.5	255.4	371.1
64.0	0.1281	71.3	146.7	250.9	366.0
72.0	0.1510	68.8	143.2	246.5	361.1
80.0	0.1747	66.6	139.9	242.4	356.4
88.0	0.1989	64.6	136.7	238.4	351.8
96.0	0.2235	62.7	133.8	234.5	347.4
104.0	0.2483	60.9	131.0	230.9	343.1
112.0	0.2734	59.3	128.3	227.3	339.0
120.0	0.2984	57.8	125.8	223.9	335.0
128.0	0.3234	56.4	123.4	220.7	331.1
136.0	0.3483	55.1	121.1	217.5	327.3
144.0	0.3730	53.8	118.9	214.5	323.6
152.0	0.3975	52.7	116.9	211.5	320.0
160.0	0.4216	51.6	114.9	208.7	316.6

This report gives percentiles of the estimated life remaining after a certain time period. For example, the estimated median remaining life of items reaching 80.0 hours is 139.9 hours.

The percentile and associated confidence interval of residual (remaining) life is computed using the following steps:

1. Compute $z_p = \frac{y_p - \hat{M}}{\hat{S}}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1 / \hat{B}$. Also note that for the normal, extreme value, and logistic distributions, $y_p = t$. For the lognormal, Weibull, and log-logistic distributions, $y_p = e^t$. And for the lognormal base 10 distribution, $y_p = 10^t$.
2. Compute $p_0 = F(z_p)$
3. Compute $p_1 = p_0(1 + P)$, where *P* is the percentile of residual life to be estimated.
4. Compute $w_p = \hat{M} + F^{-1}(p_1)\hat{S}$. Note that in the case of the Weibull and exponential distributions, we let $\hat{M} = \ln(\hat{C})$ and $\hat{S} = 1 / \hat{B}$.

   For the normal, extreme value, and logistic distributions, the estimate is given by

   $$T_p = w_p$$

   For the lognormal, exponential, Weibull, and log-logistic distributions, the estimate is given by

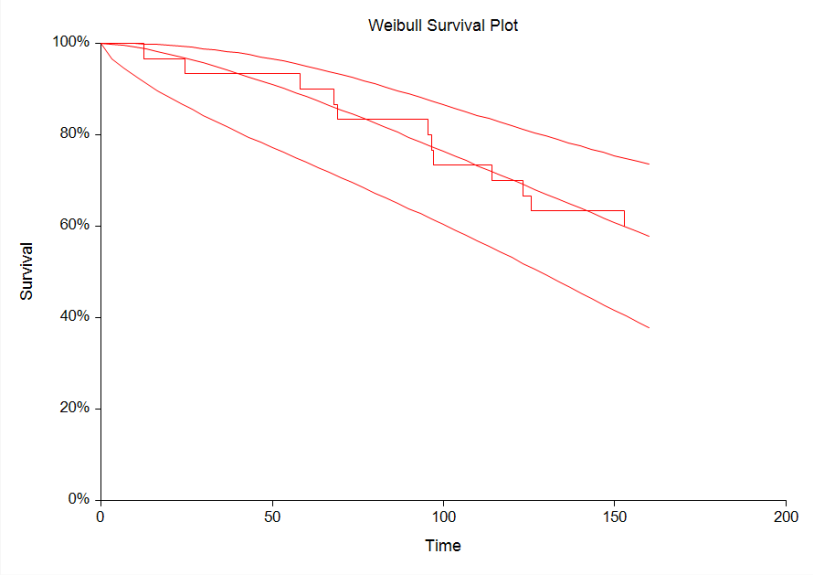
   $$T_p = e^{w_p}$$

   For the lognormal base 10 distribution, the estimate is given by

   $$T_p = 10^{w_p}$$

Survival Plot

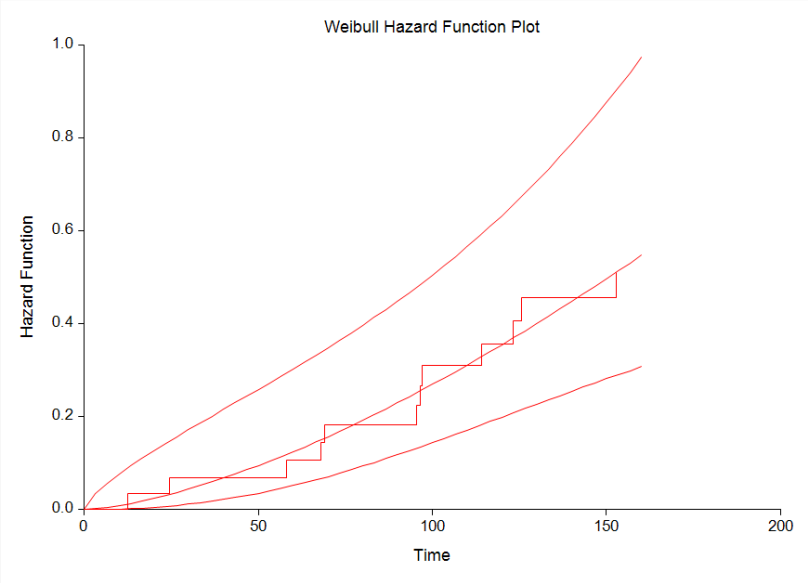
Survival Plot



This plot shows the product-limit survivorship function (the step function) as well as the parametric survival plot and associated confidence intervals. If there are several groups, a separate line is drawn for each group.

Hazard Function Plot

Hazard Function Plot

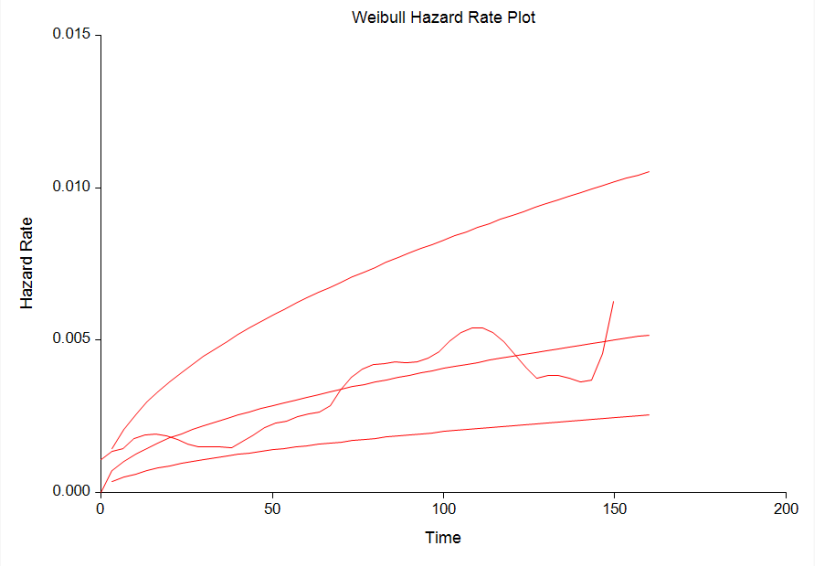


This plot shows the parametric and nonparametric cumulative hazard functions for the data analyzed. Confidence limits for the parametric cumulative hazard function are also given.

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.

Hazard Rate Plot

Hazard Rate Plot

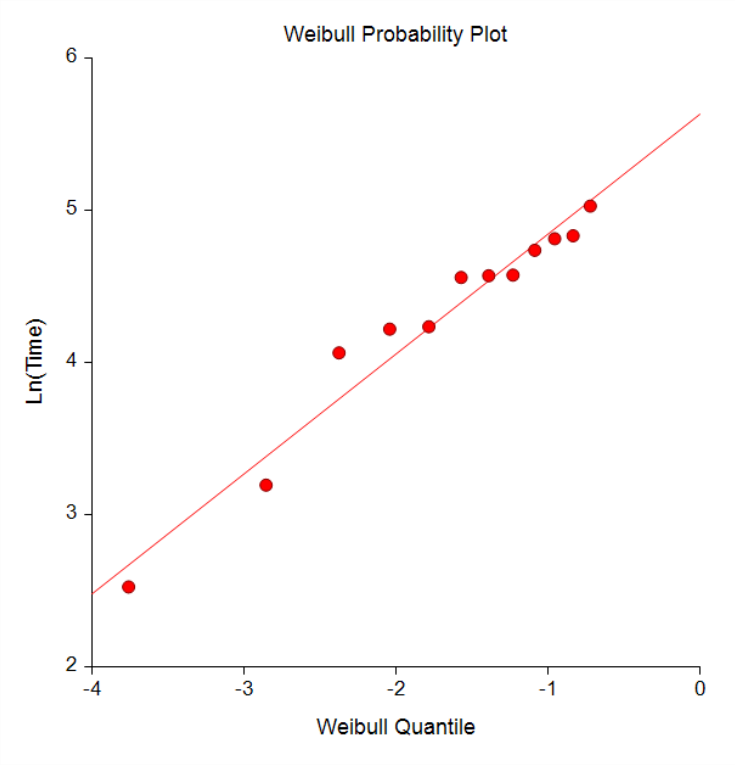


This plot shows the parametric and nonparametric hazard rate plots with confidence limits for the parametric hazard rate. This plot is especially useful for studying the shape of the nonparametric hazard rate and comparing that with the parametric hazard rate. When selecting a probability distribution to represent a set of data, it is important to determine if the parametric hazard rate plot has a general shape that is consistent both with the nonparametric hazard rate and with your prior knowledge of the hazard distribution. This plot allows you to make this comparison.

Note that the asymptotic confidence intervals are not well behaved for some distributions. If the confidence intervals seem to have zero width at some point along the plot, you should realize that they fall into this category and ignore them.

Probability Plot

Probability Plot(s)



This is the Weibull probability plot of these data. The expected quantile of the theoretical distribution is plotted on the horizontal axis. The natural logarithm of the time value is plotted on the vertical axis. Note that censored points are not shown on this plot. Also note that for tied data, only one point is shown for each set of ties.

This plot lets you investigate the goodness of fit of the selected probability distribution to your data. If the points seem to fall along a straight line, the selected probability model may be useful. If the plot shows a downward curve, the value of the threshold parameter, *D*, may need to be increased. If the plot shows an upward curve, the value of the threshold parameter may need to be decreased. Or you may need to select a different distribution.

You must decide whether the probability 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.

Multiple-Censored and Grouped Data

The case of grouped and multiple-censored data cause special problems when creating a probability plot. Remember that the horizontal axis represents the expected quantile from the selected distribution for each (sorted) failure time. In the regular case, we use the rank of the observation in the overall dataset. However, in the case of grouped or multiple-censored data, we 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).

Left censored and interval censored data are treated as failures for making the probability plots.

Example 2 – Distribution Selection

This section presents an example of how to let the program help you pick an appropriate parametric 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 Distribution (Weibull) Fitting procedure options

- Find and open the **Distribution (Weibull) Fitting** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** 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**
Censor Variable	**Censor**
Frequency Variable	**Count**
Distribution	**Find Best**
Options Tab	
Derivatives	**0.00006**
Plots Tab	
Two Plots Per Line	**Checked**

3 Run the procedure

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

Data Summary

Data Summary

Type of Observation	Rows	Count	Percent (%)	Minimum	Maximum
Failed	12	12	40.00%	12.5	152.7
Right Censored	1	18	60.00%	152.7	152.7
Left Censored	0	0			
Interval Censored	0	0			
Total	13	30	100.00%	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.

Distribution Fit Summary

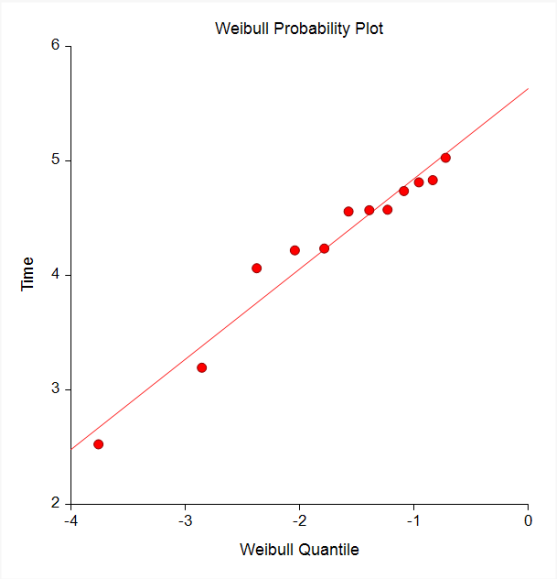
Distribution Fit Summary

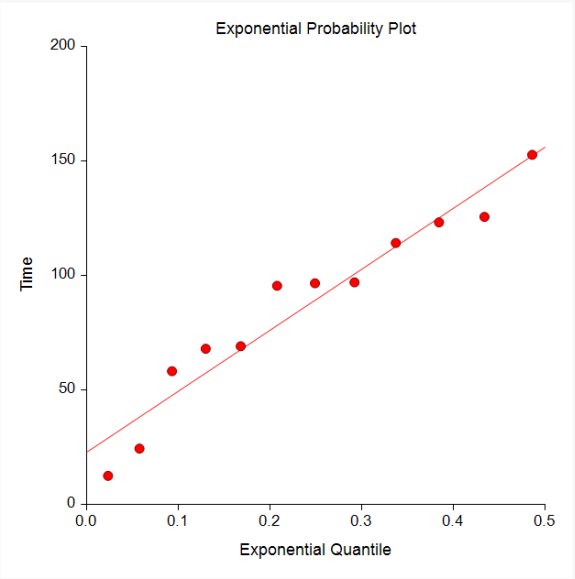
Distribution	Likelihood	Shape	Scale	Threshold
Weibull	-80.05649	1.511543	238.3481	0.0
Loglogistic	-80.11679	5.28008	0.5909371	0.0
Lognormal10	-80.38821	0.4941201	2.323475	0.0
Lognormal	-80.38821	1.137753	5.349999	0.0
Exponential	-81.04864	1	315.4667	0.0
Normal	-81.24539	171.1062	84.88175	0.0
Logistic	-81.74763	169.1118	49.77026	0.0
Extreme Value	-82.1103	189.3399	57.44398	0.0

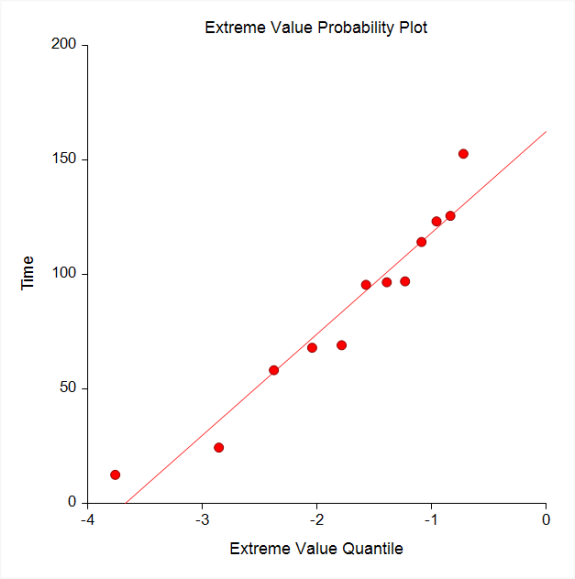
This report displays the values of the log-likelihood for each distribution along with the estimated values of its parameters. Since our desire is to maximize the likelihood, under normal circumstances, we would pick the distribution at the top of the report since it has the largest likelihood value. In this example, we would select the Weibull distribution.

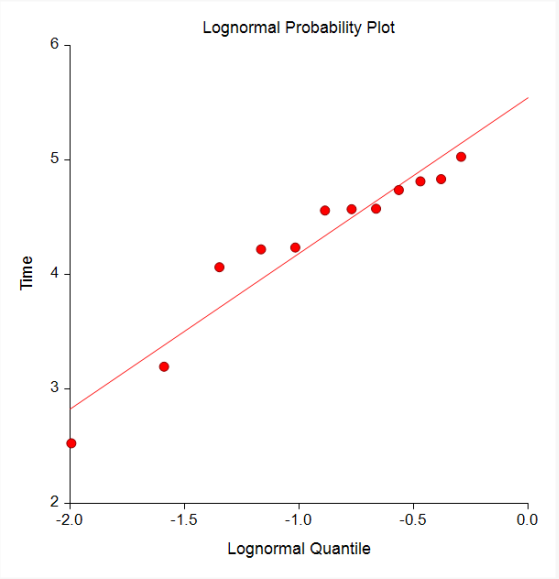
Probability Plots

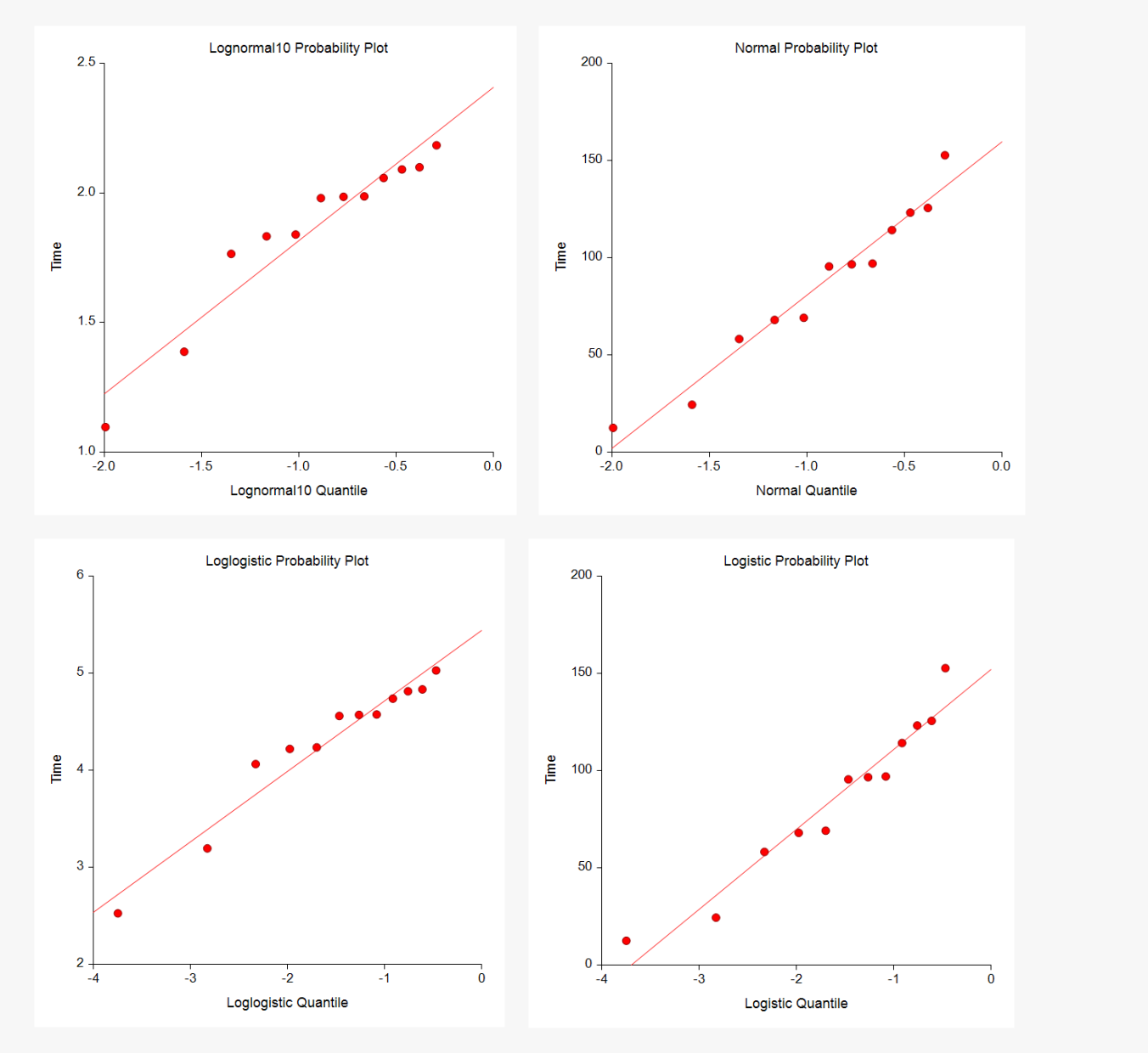
Probability Plot(s)











By studying these probability plots, we can determine which distributions fit the data the best. In this example, since there are only a few observations, it is difficult to select one distribution over another. We can see that our candidate from the last section, the Weibull distribution, certainly cannot be removed on the basis of its probability plot. Without further information, our decision would be to select the Weibull distribution to fit these data.

Example 3 – Readout Data

This section presents an example of how to analyze readout data. The data used are found in the Readout105 dataset. The table below shows the results of a study to test the failure rate of a particular machine. This study began with 40 test machines. After each time period (24, 72, 168, etc.) the number of machines that had failed since the period began was recorded. This number is entered into the Count variable. Hence, two machines failed during the first 24 hours, one machine failed between 24 and 72 hours, and so on.

After 1500 hours, the study was terminated. Sixteen machines still had not failed. The data are entered in the spreadsheet as shown below.

We have used obvious indicators for censoring. Since the first period begins with a zero time, this entry represents left censored data. We indicate left censoring with an 'L.' The next eight rows represent interval censored data. Both beginning and ending times are needed for these entries. We indicate interval censoring with an 'I.' The last row corresponds to the sixteen machines that did not fail. These are entered as right censored data, which is indicated with an 'R.'

Readout105 Dataset

Time1	Time2	Censor	Count
24	0	L	2
72	24	I	1
168	72	I	3
300	168	I	2
500	300	I	2
750	500	I	4
1000	750	I	5
1250	1000	I	1
1500	1250	I	4
1500		R	16

Setup

To run this example, complete the following steps:

1 Open the Readout105 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Readout105** and click **OK**.

2 Specify the Distribution (Weibull) Fitting procedure options

- Find and open the **Distribution (Weibull) Fitting** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** 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.....**Time1**
Start Time Variable.....**Time2**
Frequency Variable.....**Count**
Censor Variable.....**Censor**
Failed.....**F**
Right.....**R**
Left.....**L**
Interval.....**I**
Distribution.....**Find Best**

Options Tab

Derivatives.....**0.00006**

Plots Tab

Two Plots Per Line.....**Checked**

3 Run the procedure

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

Data Summary

Data Summary

Type of Observation	Rows	Count	Percent (%)	Minimum	Maximum
Failed	0	0			
Right Censored	1	16	40.00%	1500	1500
Left Censored	1	2	5.00%	24	24
Interval Censored	8	22	55.00%	24	1500
Total	10	40	100.00%	24	1500

This report displays a summary of the data that were analyzed. We note that the number of rows and the total count appear to be correct.

Distribution Fit Summary

Distribution Fit Summary

Distribution	Likelihood	Shape	Scale	Threshold
Weibull	-79.42889	0.8222772	1746.067	0.0
Exponential	-79.96207	1	1631.161	0.0
Loglogistic	-80.27086	7.044066	1.030881	0.0
Lognormal10	-81.19075	0.8194178	3.046982	0.0
Lognormal	-81.19075	1.886779	7.015936	0.0
Normal	-81.44245	1213.697	913.7082	0.0
Logistic	-82.05516	1199.686	563.52	0.0
Extreme Value	-83.09204	1525.271	726.1455	0.0

It appears that the Weibull distribution is a reasonable choice for the parametric distribution, although the shape parameter is less than one. This may point to the need for a nonzero threshold value.

To finish this example, you would view the probability plots. Finally, you would try fitting the Weibull distribution to these data. We will leave that to you to do. Simply change the Distribution box to Weibull and rerun the procedure.

Example 4 – Engine Fan Data

Nelson (1982) gives data on the failure times of seventy diesel engine fans. Twelve of the fans failed during the duration of the test. Fifty-eight of the fans completed the test without failure, so only their running times were recorded. These data are contained in the FanFailure dataset. You can observe the data by opening this dataset. Note that 'F' designates a failure and 'C' designates a censored (non-failed) fan.

Two questions were to be answered from these data. First of all, the warranty period for the fan is 8000 hours. Management wanted to know what percentage would fail on or before the warranty period ended. Second, management wanted to know what happens to the failure rate as the fans age.

The following steps will set up the procedure to analyze these data and answer the two questions given above.

Setup

To run this example, complete the following steps:

1 Open the FanFailure example dataset

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

2 Specify the Distribution (Weibull) Fitting procedure options

- Find and open the **Distribution (Weibull) Fitting** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** 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	**Hours**
Frequency Variable	**Count**
Censor Variable	**Censor**
Failed	**F**
Right	**C**
Distribution	**Weibull**
Reports Tab	
Times	**1000:15000(1000)**

3 Run the procedure

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

We will show only those portions of the printout that are necessary to answer the two questions that were posed at the beginning of this section.

Parameter Estimation

Weibull Parameter Estimation

Parameter	Probability Plot Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
B (Shape)	1.202295	1.058446	0.2682645	0.6440661	1.739429
C (Scale)	17283.81	26296.85	12252.48	10551.25	65539.57
D (Threshold)	0	0			
Log-Likelihood		-135.1527			
Mean	16250.13	25715.61			
Median	12742.28	18600.24			
Mode	3925.094	1703.919			
Sigma	13574.96	24306.58			

Estimation Details

Differential Evolution Iterations	32
Newton-Raphson Restart	1
Newton-Raphson Iterations	7
User-Entered Random Seed	2928475

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

This report shows the estimated parameters. We are particularly interested to see that the shape parameter is almost exactly one. The confidence limits for the estimated shape parameter include one between them. Remember that when the shape parameter is one, the Weibull distribution reduces to the exponential distribution, a distribution which 'has no memory.' From this, we get an indication that the failure pattern of the fans does not change over time. That is, the failure rate does not change as the fans get older.

Parametric Failure Distribution

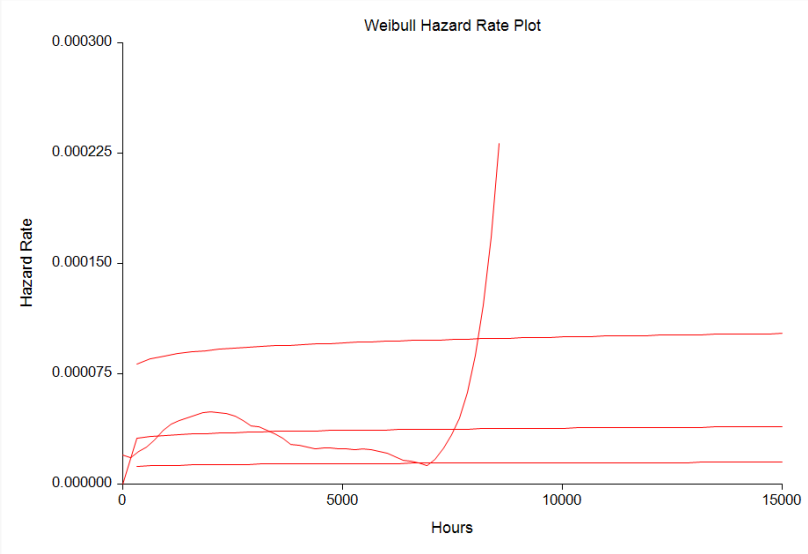
Weibull Failure Distribution

Failure Time	Prob Plot Estimate of Failure	Max Like Estimate of Failure	95% Lower Conf. Limit of Failure	95% Upper Conf. Limit of Failure
1000.0	0.0320	0.0309	0.0105	0.0895
2000.0	0.0721	0.0633	0.0289	0.1360
3000.0	0.1147	0.0956	0.0501	0.1782
4000.0	0.1581	0.1274	0.0719	0.2203
5000.0	0.2016	0.1585	0.0927	0.2636
6000.0	0.2444	0.1888	0.1121	0.3083
7000.0	0.2863	0.2184	0.1297	0.3539
8000.0	**0.3270**	**0.2471**	**0.1459**	**0.3999**
9000.0	0.3664	0.2749	0.1607	0.4456
10000.0	0.4043	0.3019	0.1743	0.4905
11000.0	0.4406	0.3280	0.1870	0.5340
12000.0	0.4753	0.3533	0.1987	0.5758
13000.0	0.5084	0.3778	0.2098	0.6156
14000.0	0.5399	0.4014	0.2202	0.6531
15000.0	0.5697	0.4242	0.2300	0.6883

This report presents the estimated failure proportions at various time periods. We note that at 8000 hours, the maximum likelihood estimate for the proportion failing is 0.247. The 95% confidence limits are 0.146 to 0.400. That is, almost 25% of the fans can be expected to fail by 8000 hours—a very high failure rate. Management will have to change the fans to decrease the proportion failing!

Hazard Rate Plot

Hazard Rate Plot



This plot shows both the parametric and nonparametric estimates of the hazard rates. First, we analyze the nonparametric estimate. Notice that the line wanders up and then down, but it does not extend outside the confidence limits of the parametric hazard rate. The sharp rise at the end of the plot is due to a lack of data in this region and should be ignored. We see that the parametric estimate of the hazard rate, the middle horizontal line, is a reasonable approximation for the nonparametric line. The above considerations again lead us to conclude that the failure rates do not change with age.

This ends this example. Notice how quickly we have been able to answer the two questions posed by management. The only task that we did not complete was to make sure that the Weibull distribution was appropriate for these data. A quick look at the probability plot will show you that it is.

Example 5 – Adding an At-Risk Table to a Survival Plot and a Hazard Function Plot

This section demonstrates how to add a table containing the number of subjects at risk, the cumulative number of censored observations, and the cumulative number of events to the bottom of a survival plot and a hazard function plot. The hazard function plot will be created with default settings, but we'll modify the survival plot to highlight some of the available options. The data used in this example are contained 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 Distribution (Weibull) Fitting procedure options

- Find and open the **Distribution (Weibull) Fitting** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5** 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**
Options Tab	
Derivatives	**0.00006**
Reports Tab	
All Reports	**Unchecked**

Plots Tab

Survival/Reliability Plot .....**Checked**
Hazard Function Plot .....**Checked**

Survival/Reliability Plot Format (*Click the Button*)

Survival Tab
Confidence Limits (Distribution Fit Line) .....**Checked**

At-Risk Table Tab

General Sub-Tab
Show At-Risk Table.....**Checked**
Display .....**Number At Risk (Number Censored) (Number of Events)**

Layout Sub-Tab
Top Outside Margin.....**0**

Groups Sub-Tab
Value Colors From .....**Line Fill**

Hazard Function Plot Format (*Click the Button*)

Haz. Function Tab
Confidence Limits (Distribution Fit Line) .....**Checked**

At-Risk Table Tab

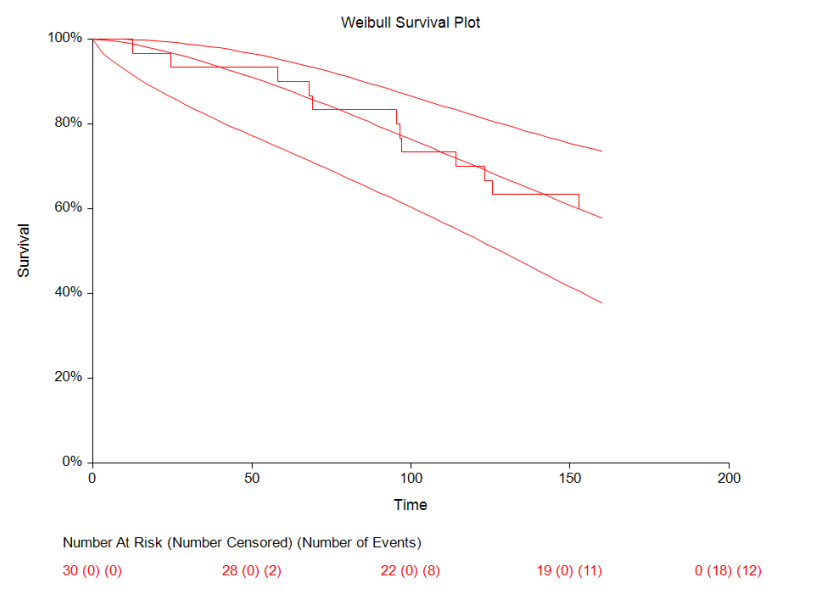
General Sub-Tab
Show At-Risk Table.....**Checked**

3 Run the procedure

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

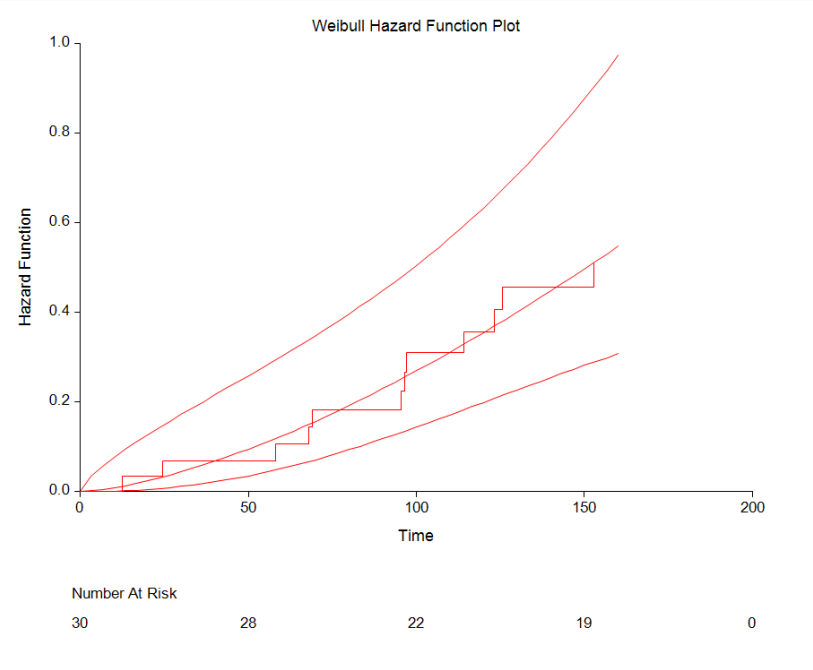
Output

Survival Plot



The modified survival plot is displayed with numbers at risk, cumulative censoring, and cumulative events at each reference time point. You can further modify the plot as required to suit your needs.

Hazard Function Plot



This plot is displayed with default at-risk table settings to give you an idea of what you get if you make no modifications to the table.