

www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

TextbookSurvival Failure Time Analysis

What can we help you find?

Search









Looking for info about statistics?

We wrote the book on it.  
And you can read it for free!

- Elementary Concepts
- Statistics Glossary
- Basic Statistics
- ANOVA / MANOVA
- Association Rules
- Boosting Trees
- Canonical Analysis
- CHAID Analysis
- C & R Trees
- Classification Trees
- Cluster Analysis
- Correspondence Analysis
- Data Mining Techniques
- Discriminant Analysis
- Distribution Fitting
- Experimental Design
- Factor Analysis
- General Discrim. Analysis
- General Linear Models
- Generalized Additive Mod.
- Generalized Linear Mod.
- General Regression Mod.
- Graphical Techniques
- Ind.Components Analysis
- Linear Regression
- Log-Linear Analysis
- MARSplines
- Machine Learning
- Multidimensional Scaling
- Neural Networks
- Nonlinear Estimation
- Nonparametric Statistics
- Partial Least Squares
- Power Analysis
- Process Analysis
- Quality Control Charts
- Reliability / Item Analysis
- SEPATH (Structural eq.)
- Survival Analysis
- Text Mining
- Time Series / Forecasting

## Survival/Failure Time Analysis

- General Information
- Censored Observations
- Analytic Techniques
- Life Table Analysis
  - Number of Cases at Risk
  - Proportion Failing
  - Proportion surviving
  - Cumulative Proportion Surviving (Survival Function)
  - Probability Density
  - Hazard rate
  - Median survival time
  - Required sample sizes
- Distribution Fitting
  - General Introduction
  - Estimation
  - Goodness-of-fit
  - Plots
- Kaplan-Meier Product-Limit Estimator
- Comparing Samples
  - General Introduction
  - Available tests
  - Choosing a two-sample test
  - Multiple sample test
  - Unequal proportions of censored data
- Regression Models
  - General Introduction
  - Cox's Proportional Hazard Model
  - Cox's Proportional Hazard Model with Time-Dependent Covariates
  - Exponential Regression

-  Variance Components
-  Statistical Advisor
-  Distribution Tables
-  References Cited
-  Send Comments
-  Business Solutions
-  Free Resources
-  About Textbook

- [Normal and Log-Normal Regression](#)
- [Stratified Analyses](#)

## General Information

These techniques were primarily developed in the medical and biological sciences, but they are also widely used in the social and economic sciences, as well as in engineering (reliability and failure time analysis).

Imagine that you are a researcher in a hospital who is studying the effectiveness of a new treatment for a generally terminal disease. The major variable of interest is the number of days that the respective patients survive. In principle, one could use the standard parametric and nonparametric statistics for describing the average survival, and for comparing the new treatment with traditional methods (see [Basic Statistics](#) and [Nonparametrics and Distribution Fitting](#)). However, at the end of the study there will be patients who survived over the entire study period, in particular among those patients who entered the hospital (and the research project) late in the study; there will be other patients with whom we will have lost contact. Surely, one would not want to exclude all of those patients from the study by declaring them to be missing data (since most of them are "survivors" and, therefore, they reflect on the success of the new treatment method). Those observations, which contain only partial information are called [censored observations](#) (e.g., "patient A survived at least 4 months before he moved away and we lost contact;" the term censoring was first used by Hald, 1949).

[To index](#)

## Censored Observations

In general, censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time. Censored observations may occur in a number of different areas of research. For example, in the social sciences we may study the "survival" of marriages, high school drop-out rates (time to drop-out), turnover in organizations, etc. In each case, by the end of the study period, some subjects will still be married, will not have dropped out, or are still working at the same company; thus, those subjects represent censored observations.

In economics we may study the "survival" of new businesses or the "survival" times of products such as automobiles. In quality control research, it is common practice to study the "survival" of parts under stress (failure time analysis).

[To index](#)

## Analytic Techniques

Essentially, the methods offered in *Survival Analysis* address the same research questions as many of the other procedures; however, all methods in *Survival Analysis* will handle [censored data](#). The *life table*, *survival distribution*, and *Kaplan-Meier* survival function estimation are all descriptive methods for estimating the distribution of survival times from a sample. Several techniques are available for comparing the survival in two or more groups. Finally, *Survival Analysis* offers several regression models for estimating the relationship of (multiple) continuous variables to survival times.

[To index](#)

## Life Table Analysis

The most straightforward way to describe the survival in a sample is to compute the *Life Table*. The life table technique is one of the oldest methods for analyzing survival (failure time) data (e.g., see Berkson & Gage, 1950; Cutler & Ederer, 1958; Gehan, 1969). This table can be thought of as an "enhanced" frequency distribution table. The distribution of survival times is divided into a certain number of intervals. For each interval we can then compute the number and proportion of cases or objects that entered the respective interval "alive," the number and proportion of cases that failed in the respective interval (i.e., number of terminal events, or number of cases that "died"), and the number of cases that were lost or **censored** in the respective interval.

Based on those numbers and proportions, several additional statistics can be computed:

- **Number of Cases at Risk**
- **Proportion Failing**
- **Proportion surviving**
- **Cumulative Proportion Surviving (Survival Function)**
- **Probability Density**
- **Hazard rate**
- **Median survival time**
- **Required sample sizes**

**Number of Cases at Risk.** This is the number of cases that entered the respective interval alive, minus half of the number of cases lost or censored in the respective interval.

**Proportion Failing.** This proportion is computed as the ratio of the number of cases failing in the respective interval, divided by the number of cases at risk in the interval.

**Proportion Surviving.** This proportion is computed as 1 minus the proportion failing.

**Cumulative Proportion Surviving (Survival Function).** This is the cumulative proportion of cases surviving up to the respective interval. Since the probabilities of survival are assumed to be independent across the intervals, this probability is computed by multiplying out the probabilities of survival across all previous intervals. The resulting function is also called the *survivorship* or *survival function*.

**Probability Density.** This is the estimated probability of failure in the respective interval, computed per unit of time, that is:

$$F_i = (P_i - P_{i+1}) / h_i$$

In this formula,  $F_i$  is the respective probability density in the  $i$ 'th interval,  $P_i$  is the estimated cumulative proportion surviving at the beginning of the  $i$ 'th interval (at the end of interval  $i-1$ ),  $P_{i+1}$  is the cumulative proportion surviving at the end of the  $i$ 'th interval, and  $h_i$  is the width of the respective interval.

**Hazard Rate.** The hazard rate (the term was first used by Barlow, 1963) is defined as the probability per time unit that a case that has survived to the beginning of the respective interval will fail in that interval. Specifically, it is computed as the number of failures per time units in the respective interval, divided by the average number of surviving cases at the mid-point of the interval.

**Median Survival Time.** This is the survival time at which the cumulative survival function is equal to 0.5. Other percentiles (25th and 75th percentile) of the cumulative survival function can be computed accordingly. Note that the 50th percentile (median) for the cumulative survival function is usually not the same as the point in time up to which 50% of the sample survived. (This would only be the case if there were no censored observations prior to this time).

**Required Sample Sizes.** In order to arrive at reliable estimates of the three major functions

(survival, probability density, and hazard) and their standard errors at each time interval the minimum recommended sample size is 30.

[To index](#)

## Distribution Fitting

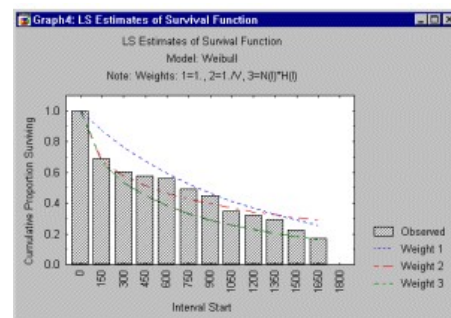
- [General Introduction](#)
- [Estimation](#)
- [Goodness-of-fit](#)
- [Plots](#)

**General Introduction.** In summary, the life table gives us a good indication of the distribution of failures over time. However, for predictive purposes it is often desirable to understand the shape of the underlying survival function in the population. The major distributions that have been proposed for modeling survival or failure times are the [exponential](#) (and linear exponential) distribution, the [Weibull](#) distribution of extreme events, and the Gompertz distribution.

**Estimation.** The parameter estimation procedure (for estimating the parameters of the theoretical survival functions) is essentially a least squares linear regression [algorithm](#) (see Gehan & Siddiqui, 1973). A linear regression algorithm can be used because all four theoretical distributions can be "made linear" by appropriate transformations. Such transformations sometimes produce different variances for the residuals at different times, leading to biased estimates.

**Goodness-of-Fit.** Given the parameters for the different distribution functions and the respective model, we can compute the likelihood of the data. One can also compute the likelihood of the data under the null model, that is, a model that allows for different hazard rates in each interval. Without going into details, these two likelihoods can be compared via an incremental [Chi-square](#) test statistic. If this *Chi-square* is statistically significant, then we conclude that the respective theoretical distribution fits the data significantly worse than the null model; that is, we reject the respective distribution as a model for our data.

**Plots.** You can produce plots of the survival function, hazard, and probability density for the observed data and the respective theoretical distributions. These plots provide a quick visual check of the goodness-of-fit of the theoretical distribution. The example plot below shows an observed survivorship function and the fitted Weibull distribution.



Specifically, the three lines in this plot denote the theoretical distributions that resulted from three different estimation procedures (least squares and two methods of weighted least squares).

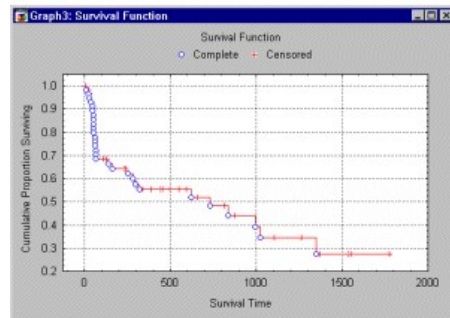
## Kaplan-Meier Product-Limit Estimator

Rather than classifying the observed survival times into a life table, we can estimate the survival function directly from the continuous survival or failure times. Intuitively, imagine that we create a life table so that each time interval contains exactly one case. Multiplying out the survival

probabilities across the "intervals" (i.e., for each single observation) we would get for the survival function:

$$S(t) = \prod_{j=1}^t [(n-j)/(n-j+1)]^{\delta(j)}$$

In this equation,  $S(t)$  is the estimated survival function,  $n$  is the total number of cases, and  $\prod$  denotes the multiplication (geometric sum) across all cases less than or equal to  $t$ ;  $\delta(j)$  is a constant that is either 1 if the  $j$ 'th case is uncensored (complete), and 0 if it is censored. This estimate of the survival function is also called the *product-limit estimator*, and was first proposed by Kaplan and Meier (1958). An example plot of this function is shown below.



The advantage of the Kaplan-Meier Product-Limit method over the life table method for analyzing survival and failure time data is that the resulting estimates do not depend on the grouping of the data (into a certain number of time intervals). Actually, the Product-Limit method and the life table method are identical if the intervals of the life table contain at most one observation.

[To index](#)

## Comparing Samples

- [General Introduction](#)
- [Available tests](#)
- [Choosing a two-sample test](#)
- [Multiple sample test](#)
- [Unequal proportions of censored data](#)

**General Introduction.** One can compare the survival or failure times in two or more samples. In principle, because survival times are not normally distributed, nonparametric tests that are based on the *rank ordering* of survival times should be applied. A wide range of nonparametric tests can be used in order to compare survival times; however, the tests cannot "handle" [censored observations](#).

**Available Tests.** The following five different (mostly nonparametric) tests for censored data are available: Gehan's generalized Wilcoxon test, the Cox-Mantel test, the Cox's  $F$  test, the log-rank test, and Peto and Peto's generalized Wilcoxon test. A nonparametric test for the comparison of multiple groups is also available. Most of these tests are accompanied by appropriate  $z$ -values (values of the standard normal distribution); these  $z$ -values can be used to test for the statistical significance of any differences between groups. However, note that most of these tests will only yield reliable results with fairly large samples sizes; the small sample "behavior" is less well understood.

**Choosing a Two Sample Test.** There are no widely accepted guidelines concerning which test to use in a particular situation. Cox's  $F$  test tends to be more powerful than Gehan's generalized Wilcoxon test when:

1. Sample sizes are small (i.e.,  $n$  per group less than 50);
2. If samples are from an [exponential](#) or [Weibull](#);
3. If there are no censored observations (see Gehan & Thomas, 1969).

Lee, Desu, and Gehan (1975) compared Gehan's test to several alternatives and showed that the Cox-Mantel test and the log-rank test are more powerful (regardless of censoring) when the samples are drawn from a population that follows an exponential or Weibull distribution; under those conditions there is little difference between the Cox-Mantel test and the log-rank test. Lee (1980) discusses the power of different tests in greater detail.

**Multiple Sample Test.** There is a multiple-sample test that is an extension (or generalization) of Gehan's generalized Wilcoxon test, Peto and Peto's generalized Wilcoxon test, and the log-rank test. First, a score is assigned to each survival time using Mantel's procedure (Mantel, 1967); next a *Chi-square* value is computed based on the sums (for each group) of this score. If only two groups are specified, then this test is equivalent to Gehan's generalized Wilcoxon test, and the computations will default to that test in this case.

**Unequal Proportions of Censored Data.** When comparing two or more groups it is very important to examine the number of [censored observations](#) in each group. Particularly in medical research, censoring can be the result of, for example, the application of different treatments: patients who get better faster or get worse as the result of a treatment may be more likely to drop out of the study, resulting in different numbers of censored observations in each group. Such systematic censoring may greatly bias the results of comparisons.

[To index](#)

## Regression Models

- [General Introduction](#)
- [Cox's Proportional Hazard Model](#)
- [Cox's Proportional Hazard Model with Time-Dependent Covariates](#)
- [Exponential Regression](#)
- [Normal and Log-Normal Regression](#)
- [Stratified Analyses](#)

### GENERAL INTRODUCTION

A common research question in medical, biological, or engineering (failure time) research is to determine whether or not certain continuous (independent) variables are correlated with the survival or failure times. There are two major reasons why this research issue cannot be addressed via straightforward multiple regression techniques (as available in [Multiple Regression](#)): First, the dependent variable of interest (survival/failure time) is most likely not normally distributed -- a serious violation of an assumption for ordinary least squares multiple regression. Survival times usually follow an [exponential](#) or [Weibull](#) distribution. Second, there is the problem of [censoring](#), that is, some observations will be incomplete.

### COX'S PROPORTIONAL HAZARD MODEL

The proportional hazard model is the most general of the regression models because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. The model assumes that the underlying hazard *rate* (rather than survival time) is a function of the independent variables (covariates); no assumptions are made about the nature or shape of the hazard function. Thus, in a sense, Cox's regression model may be considered to be a nonparametric method. The model may be written as:

$$h\{t, (z_1, z_2, \dots, z_m)\} = h_0(t) \cdot \exp(b_1 \cdot z_1 + \dots + b_m \cdot z_m)$$



where  $h(t, \dots)$  denotes the resultant hazard, given the values of the  $m$  covariates for the respective case ( $z_1, z_2, \dots, z_m$ ) and the respective survival time ( $t$ ). The term  $h_0(t)$  is called the *baseline hazard*; it is the hazard for the respective individual when all independent variable values are equal to zero. We can linearize this model by dividing both sides of the equation by  $h_0(t)$  and then taking the natural logarithm of both sides:

$$\log[h(t, (z, \dots))/h_0(t)] = b_1 z_1 + \dots + b_m z_m$$

We now have a fairly "simple" linear model that can be readily estimated.

**Assumptions.** While no assumptions are made about the shape of the underlying hazard function, the model equations shown above do imply two assumptions. First, they specify a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates. This assumption is also called the *proportionality assumption*. In practical terms, it is assumed that, given two observations with different values for the independent variables, the ratio of the hazard functions for those two observations does not depend on time. The second assumption of course, is that there is a log-linear relationship between the independent variables and the underlying hazard function.

## COX'S PROPORTIONAL HAZARD MODEL WITH TIME-DEPENDENT COVARIATES

An assumption of the proportional hazard model is that the hazard function for an individual (i.e., observation in the analysis) depends on the values of the covariates and the value of the baseline hazard. Given two individuals with particular values for the covariates, the ratio of the estimated hazards over time will be constant -- hence the name of the method: the *proportional hazard* model. The validity of this assumption may often be questionable. For example, age is often included in studies of physical health. Suppose you studied survival after surgery. It is likely, that age is a more important predictor of risk immediately after surgery, than some time after the surgery (after initial recovery). In accelerated life testing one sometimes uses a stress covariate (e.g., amount of voltage) that is slowly increased over time until failure occurs (e.g., until the electrical insulation fails; see Lawless, 1982, page 393). In this case, the impact of the covariate is clearly dependent on time. The user can specify arithmetic expressions to define covariates as functions of several variables and survival time.

**Testing the Proportionality Assumption.** As indicated by the previous examples, there are many applications where it is likely that the proportionality assumption does not hold. In that case, one can explicitly define covariates as functions of time. For example, the analysis of a data set presented by Pike (1966) consists of survival times for two groups of rats that had been exposed to a carcinogen (see also Lawless, 1982, page 393, for a similar example). Suppose that  $z$  is a grouping variable with codes 1 and 0 to denote whether or not the respective rat was exposed. One could then fit the proportional hazard model:

$$h(t, z) = h_0(t) \cdot \exp\{b_1 z + b_2 [z \cdot \log(t) - 5.4]\}$$

Thus, in this model the conditional hazard at time  $t$  is a function of (1) the baseline hazard  $h_0$ , (2) the covariate  $z$ , and (3) of  $z$  times the logarithm of time. Note that the constant 5.4 is used here for scaling purposes only: the mean of the logarithm of the survival times in this data set is equal to 5.4. In other words, the conditional hazard at each point in time is a function of the covariate and time; thus, the effect of the covariate on survival is dependent on time; hence the name *time-dependent covariate*. This model allows one to specifically test the proportionality assumption. If parameter  $b_2$  is statistically significant (e.g., if it is at least twice as large as its standard error), then one can conclude that, indeed, the effect of the covariate  $z$  on survival is dependent on time, and, therefore, that the proportionality assumption does not hold.

## EXPONENTIAL REGRESSION

Basically, this model assumes that the survival time distribution is exponential, and contingent on the values of a set of independent variables ( $z_i$ ). The rate parameter of the [exponential distribution](#) can then be expressed as:

$$S(z) = \exp(a + b_1 \cdot z_1 + b_2 \cdot z_2 + \dots + b_m \cdot z_m)$$

$S(z)$  denotes the survival times,  $a$  is a constant, and the  $b_i$ 's are the regression parameters.

Goodness-of-fit. The [Chi-square](#) goodness-of-fit value is computed as a function of the log-likelihood for the model with all parameter estimates ( $L1$ ), and the log-likelihood of the model in which all covariates are forced to 0 (zero;  $L0$ ). If this [Chi-square](#) value is significant, we reject the null hypothesis and assume that the independent variables are significantly related to survival times.

Standard exponential order statistic. One way to check the exponentiality assumption of this model is to plot the residual survival times against the standard exponential order statistic *theta*. If the exponentiality assumption is met, then all points in this plot will be arranged roughly in a straight line.

## NORMAL AND LOG-NORMAL REGRESSION

In this model, it is assumed that the survival times (or log survival times) come from a [normal distribution](#); the resulting model is basically identical to the ordinary multiple regression model, and may be stated as:

$$t = a + b_1 \cdot z_1 + b_2 \cdot z_2 + \dots + b_m \cdot z_m$$

where  $t$  denotes the survival times. For log-normal regression,  $t$  is replaced by its natural logarithm. The normal regression model is particularly useful because many data sets can be transformed to yield approximations of the normal distribution. Thus, in a sense this is the most general fully parametric model (as opposed to Cox's proportional hazard model which is non-parametric), and estimates can be obtained for a variety of different underlying survival distributions.

Goodness-of-fit. The [Chi-square](#) value is computed as a function of the log-likelihood for the model with all independent variables ( $L1$ ), and the log-likelihood of the model in which all independent variables are forced to 0 (zero,  $L0$ ).

## STRATIFIED ANALYSES

The purpose of a stratified analysis is to test the hypothesis whether identical regression models are appropriate for different groups, that is, whether the relationships between the independent variables and survival are identical in different groups. To perform a stratified analysis, one must first fit the respective regression model separately within each group. The sum of the log-likelihoods from these analyses represents the log-likelihood of the model with different regression coefficients (and intercepts where appropriate) in different groups. The next step is to fit the requested regression model to all data in the usual manner (i.e., ignoring group membership), and compute the log-likelihood for the overall fit. The difference between the log-likelihoods can then be tested for statistical significance (via the [Chi-square](#) statistic).