# Chapter 1 What Is Survival Analysis?



## 1. The Nature of Survival Data

Survival data are special and, thus, they require special methods for their analyses. Before going into what makes these data special and how they are analyzed, let's establish some terminology and explain what is meant by survival data.

Although you might naturally think of survival data as dealing with the time until death, actually the methods that are discussed in this book can be used for data that deal with the time until the occurrence of any well-defined event. In addition to death, that event can be, for example,

1. Relapse of a patient in whom disease had been in remission.
2. Death from a specific cause.
3. Development of a disease in someone at high risk.
4. Recovery of platelet count after bone marrow transplantation.
5. Relief from headache, rash, nausea, etc.

Note that for examples 1, 2, and 3, longer times until the event occurs are better. For examples 4 and 5, shorter times are better. Nevertheless, the methods that are described in this book can be applied to any of these examples. For the purpose of this book,

words like *survival* and *death* are used to describe these methods, but you should be aware of the broader areas of applicability.

This might be a good place for a few words about example 2 on the above list, cause-specific death. When you analyze the survival of patients with, for example, some form of cancer, you might want to focus on death caused by cancer, particularly in an older population in which we expect deaths from other causes as well. You would then count as "events" only those deaths caused by cancer. A death from any other cause would be treated the same as if the patient had suddenly moved out of state and could no longer be followed. Of course, this methodology requires that you establish rules and a mechanism for distinquishing between cause-specific and noncause-specific deaths. In the New York Health Insurance Plan study that was designed to assess the efficacy of mammography (Venet et al. 1988), women were randomized either to a group that received annual mammography or to a group that did not. Since the study's planners realized there would be considerable mortality not related to breast cancer, they took as their endpoint death caused by breast cancer. A committee was created to determine whether or not the death of a woman was due to breast cancer. This committee, which was blinded with respect to the woman's group assignment, followed a detailed algorithm that was described in the study protocol.

What makes analyses of these types of data distinctive is that often there are many subjects in whom the event did not occur during the time that the patient was followed. This can happen for several reasons. Here are some examples:

1. The event of interest is death, but at the time of analysis the patient is still alive.

2. A patient is lost to followup without having experienced the event of interest (death).

3. A competing event occurs that precludes the event of interest. For example, in a study designed to compare two treatments for prostate cancer, the event of interest might be death caused by the cancer. However, a patient might die of an unrelated cause instead, such as an automobile accident.

4. A patient is dropped from the study, without having experienced the event of interest, because of a major protocol violation or for reasons specified by the protocol.

In all of these situations, you don't know the time until the event occurs. Without knowledge of the methods that are described in this book, a researcher might simply exclude such cases. But clearly this throws out a great deal of useful information. In all of these cases, we know that the time to the event was at least some number. For example, a subject who was known to be alive three years into a study and then moved to another state and could no longer be followed is known to have a survival time of at least three years. This subject's time is said to be *right censored*. A subject's observed time, $t$, is right censored if, after time $t$, he or she is known to still be alive. Thus you know that this subject's survival time is at least $t$. A survival time might also be *left censored*. This happens if all that is known about the time to death is that it is less than or equal to some value. A death is *interval censored* if it is known only that it occurred during some time interval. Although much current research focuses on ways to deal with left- and interval-

censored data, most survival analytic methods deal only with right-censored data. Of the three SAS procedures that deal explicitly with survival data, two deal only with right censoring. This is the type of censoring most commonly seen in medical research. The third, the LIFEREG procedure, which is discussed in Chapter 5, "Parametric Methods," deals with left and interval censoring as well. Except for that chapter and a section in Chapter 3, this book does not consider left- or interval-censored times, and the term *censored* will always mean *right censored.*

Survival data, therefore, are described by a pair of variables, say ($t$, $d$). They can be interpreted as follows:

- $t$ represents the time that the subject was observed on study.
- $d$ is an indicator variable that specifies whether the event in question either ocurred or did not occur at the end of time $t$. The value of $d$ might be 0 to indicate that the event did not occur or 1 to indicate that it did. Then $d$=0 means that the corresponding $t$ is a censored time. Of course, you can substitute your choice of values.

The SAS survival analysis procedures, as well as the macros that are presented in this book, do allow the user to specify any set of values to indicate that a time is censored. This is convenient when you have data in which a variable indicating a subject's final status can have several values that indicate censoring. Subscripts will be used to distinguish the subjects. Thus, if there are $n$ subjects in a study, their survival data might be represented by the $n$ pairs $(t_1, d_1)$, $(t_2, d_2)$, . . . $(t_n, d_n)$.

Sometimes in textbooks or in journal articles, survival data are reported by using only the time variable. Censoring is indicated by adding a plus sign to the time. For example, reporting survival data as 2.6, 3.7+, 4.5, 7.2, 9.8+ would mean that the second and fifth observations are censored and the others are not. In a data set, you can store information about both the survival time and the censoring value using only one variable. Censoring is indicated by making the time negative. Using this convention, the above data would be 2.6, -3.7, 4.5, 7.2, - 9.8. A SAS DATA step can easily be written to convert such a data set to the desired forms, which contains separate variables for $t$ and $d$. This is illustrated by the example code and output below:

```
proc print data=original;
title 'Original Data Set';
data; set original;
d=1;
if time<0 then do;
        d=0;
        time=-time;
        end;
proc print;
title 'Modified Data Set';
run;
```

Original Data Set

| OBS | TIME |
|---|---|
| 1 | 2.6 |
| 2 | -3.7 |
| 3 | 4.5 |
| 4 | 7.2 |
| 5 | -9.8 |

Modified Data Set

| OBS | TIME | D |
|---|---|---|
| 1 | 2.6 | 1 |
| 2 | 3.7 | 0 |
| 3 | 4.5 | 1 |
| 4 | 7.2 | 1 |
| 5 | 9.8 | 0 |

There is another way of thinking about the variables *t* and *d*. Each patient on study is really subject to two random variables: the time until death (or the event of interest) and the time until censoring. Once one of these events happens, you can observe that time but not the other. The variable *t* can be thought of as the minimum of the time until death and the time until censoring. The variable *d* indicates whether that minimum is the time until death (*d*=1) or the time until censoring (*d*=0). An important assumption in any analysis that follows is that the time until death and the time until censoring are independent. This would generally be true, for example, if censoring occurred simply because the follow-up period ended. On the other hand, suppose you are analyzing the data from a study in which patients with some sort of cardiac disease are randomized to drug treatment or surgery. In some cases it might later be decided that a patient randomized to drug treatment now needs to get surgery. You might be tempted to take the patient off study with a censored survival time that is equal to the time until surgery. However, if the decision for surgery was based on the patient's deteriorating condition, to do so would create bias in favor of the drug treatment. That is because such a patient's death would not be counted as a death once he had been censored. A better approach might be to anticipate this possibility when planning the study. You might plan the study as a comparison of two treatment strategies: immediate surgery versus initial drug treatment with surgery under certain conditions that are established in advance.

## 2. Calendar Time and Study Time

Another concept we need to discuss is how time is defined in survival studies. In most survival studies, patients do not all begin their participation at the same time. Instead they are accrued over a period of time. Often they are also followed for a period of time after accrual has ended. Consider a study that starts accrual on February 1, 1996 and accrues for twenty-four months until January 31, 1998 with an additional twelve months of follow-up until January 31, 1999. In other words, no more patients are entered on study

after January 31, 1998, and those accrued are followed until January 31, 1999. Now consider the following patients:

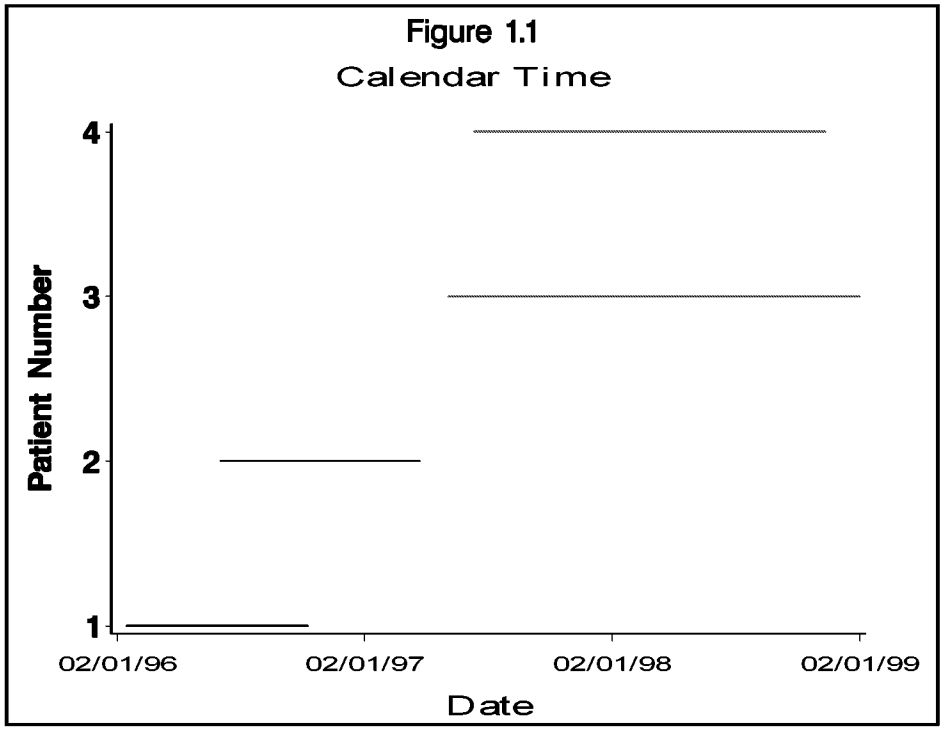
Patient #1: Enters on February 15, 1996 and dies on November 8, 1996.

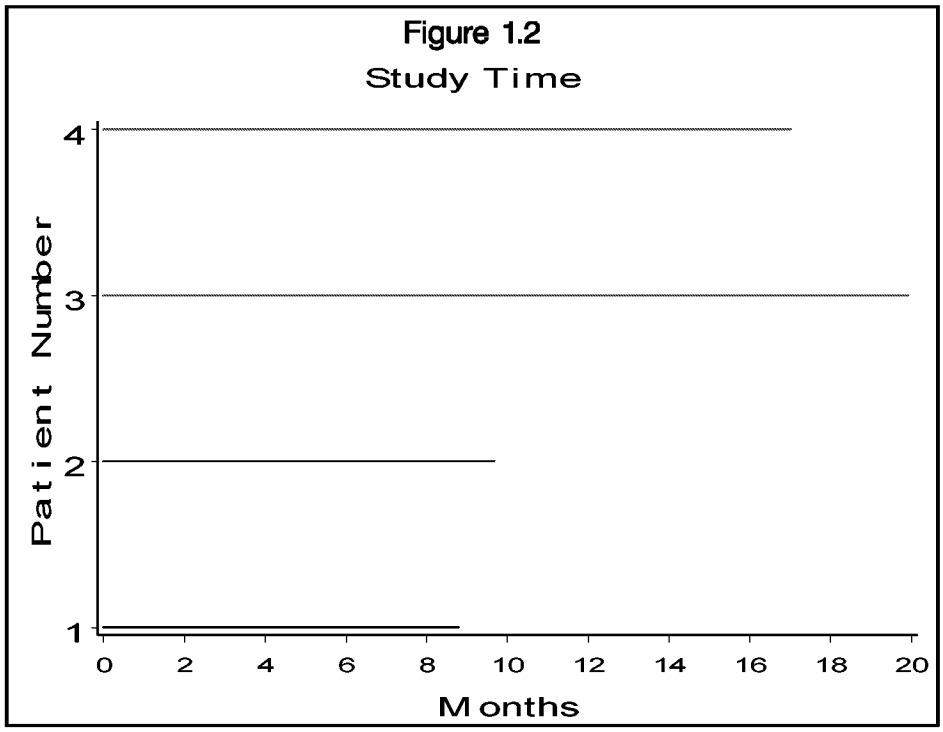
Patient #2: Enters on July 2, 1996 and is censored (lost to follow-up) on April 23, 1997.

Patient #3: Enters on June 5, 1997 and is still alive and censored at the end of the followup period.

Patient #4: Enters on July 13, 1997 and dies on December 12, 1998.

Their experience is shown graphically in Figure 1.1. In survival analyses, all patients are thought of as starting at time 0. Thus, their survival experience can be represented as in Figure 1.2. When reference is made to the number surviving or the number at risk at some time, the time referred to is the time from each patient's study entry–not the time since the study started. For example, among the above four patients, two of them (#3 and #4) are still at risk at 12 months. None are still at risk at 24 months. Both of these statements are made based on Figure 1.2. If, at some later date, you speak of those who are at risk at $t = 6$ months, that has nothing to do with the situation on July 31, 1996, 6 months after the start of the study. Rather you mean those who, as of the last date that the data were updated, had been on study for at least 6 months without dying or being censored.






## 3. An Example

Sometimes in studies that involve follow-up of patients, there is interest in more than one time variable. For example, an oncology study might examine the time until death, which will be called *survival time*, and the time until death or relapse, which will be called *disease-free survival time*. The database must then contain information on both endpoints. Because SAS handles dates internally as numeric constants (the number of days before or after January 1, 1960), it is often convenient for the data sets to contain the dates of interest and to include in a SAS DATA step the statements to calculate the time values that are needed. As an example, consider a sample of patients who are treated for malignant melanoma. Presumably they are rendered disease free surgically. Suppose that, in addition, they are treated with either treatment A or B, both of which are thought to inhibit relapse and improve survival. We might want to consider both survival and disease-free survival of these patients and how they are affected by treatment, tumor thickness, stage of disease, and tumor site. The database might look like this:

| PTID | DATESURG | DATEREL | DATEDTH | DATELAST | TRTMENT | SITE | STAGE | THICK |
|---|---|---|---|---|---|---|---|---|
| 13725 | 10/5/93 | 11/6/94 | 1/5/95 | 1/5/95 | A | 1 | III | 1.23 |
| 25422 | 3/7/93 | . | 2/6/94 | 2/6/94 | B | 3 | II | 1.13 |
| 34721 | 9/6/94 | . | . | 3/18/95 | B | 2 | III | 2.15 |
| etc. | | | | | | | | |

Note the inclusion of a unique patient identifying number, PTID. While this number will play no role in the analyses of this data set, it is a good idea to associate such a number with each patient on a trial. This will facilitate merging with other data sets if you decide to add other variables of interest later. Names are usually not good for this purpose because of the risk of spelling variations and errors. Also, note that treatment (TRTMENT), site (SITE), and stage (STAGE) are represented by codes or brief symbolic names. For obvious reasons, we should avoid having long words or phrases for things like disease site or tumor histology. TRTMENT is a dichotomous variable. Although numbers are used for the possible sites, SITE is categorical. The numbers used do not imply any ordering. STAGE is ordinal; stages I, II, III, and IV represent successively more extensive disease. Finally, tumor thickness in millimeters (THICK) is a continuous variable. Later chapters discuss SAS procedures that deal with all of these types of variables. In this case, missing values for date variables are used to indicate that the event did not occur. In order to analyze survival time and disease-free survival time, the following variables are needed:

DFSEVENT has the value 1 if the patient died or relapsed, 0 otherwise.

DFSTIME is the time, in months, from surgery to death or relapse, if either occurred. Otherwise, it is the time that the patient was observed after surgery.

SUREVENT has the value 1 if the patient died, 0 otherwise.

SURVTIME is the time, in months, from surgery to death, if the patient died. Otherwise, it is the time that the patient was observed after surgery.

To add the variables that are needed to analyze survival and disease-free survival to the data set, the code might look like this:

```
data melanoma; set melanoma;

  /*  Defining dfs time and event variables  */
dfsevent = 1 - (daterel = .)*(datedth = .);
  /*  Divide by 30.4 to convert from days to months */
if dfsevent = 0 then dfstime = (datelast - datesurg)/30.4;
else dfstime = (min(daterel, datedth) - datesurg)/30.4;
```

```
 /* Defining survival time and event variables */
surevent = (datedth ne .);
if surevent = 0 then survtime = (datelast - datesurg)/30.4;
else survtime = (datedth-datesurg)/30.4;
```

The divisions by 30.4 are simply to convert time from days to months, a more convenient time unit. Note that 30.4 is approximately 365/12. Also, when terms such as (daterel=.) or (datedth=.) are used in an arithmetic expression, they have the value 0 if false and 1 if true. The above statements create the variables DFSTIME and DFSEVENT to be used in analyses of disease-free survival, and the variables SURVTIME and SUREVENT to be used in analyses of survival. The first three observations of the resultant data set would look like this:

| PTID | DATESURG | DATEREL | DATEDTH | DATELAST | TRTMENT | SITE | STAGE |
|---|---|---|---|---|---|---|---|
| 13725 | 10/05/93 | 11/06/94 | 01/05/95 | 01/05/95 | A | 1 | III |
| 25422 | 03/07/93 | . | 02/06/94 | 02/06/94 | B | 3 | II |
| 34721 | 09/06/94 | . | . | 03/18/95 | B | 2 | III |

| PTID | THICK | DFSEVENT | DFSTIME | SUREVENT | SURVTIME |
|---|---|---|---|---|---|
| 13725 | 1.23 | 1 | 13.0592 | 1 | 15.0329 |
| 25422 | 1.13 | 1 | 11.0526 | 1 | 11.0526 |
| 34721 | 2.15 | 0 | 6.3487 | 0 | 6.3487 |

Now that these variables have been defined, there are several questions you might want to address. For example, using the methods described in Chapter 2, "Nonparametric Survival Function Estimation," you might want to estimate the survival and disease-free survival probabilities over time for the overall cohort and for subgroups defined by treatment, stage, site, etc. Standard errors and confidence intervals for those estimates can also be calculated. You might also want to perform statistical tests to assess the evidence for the superiority of one treatment over the other. This is discussed in Chapter 3, "Nonparametric Comparison of Survival Distributions." Now, it might happen that the patients who were treated with treatment A had a worse prognosis (as seen by their stages, perhaps) than did those treated with treatment B. If the treatment assignment was not randomized, this might happen if the treating physicians preferred treatment A for more advanced tumors. Even if the treatment assignment were randomized, it could happen by chance that one of the treatment groups had a higher proportion of patients with more advanced disease. Using methods that are discussed in Chapter 3 and in Chapter 4, "Proportional Hazards Regression," you will learn how to compare the two treatments after adjusting for the stage of the disease. In addition, you will be able, if you make certain assumptions, to create a model that produces estimated survival and disease-free survival probabilities for patients with specified values of the above variables. Techniques for doing this are presented in Chapters 4 and 5. For example, you will learn how to estimate the probability that a patient will survive for at least three years if that patient is treated with treatment A for a stage II tumor of thickness 1.5 mm at site 1.

## 4. Functions that Describe Survival

### 4.1 The Distribution Function and the Survival Function

The survival time of a subject being followed on a clinical study will be thought of as a random variable, $T$. As with random variables in other areas of statistics, this random variable can be characterized by its cumulative distribution function, often simply called *distribution function*, denoted $F(t)$ and defined by

$$F(t) = Pr[T < t], \; t \geq 0 \tag{1}$$

That is, for any nonnegative value of $t$, $F(t)$ is the probability that survival time will be less than $t$. Of course, you could just as well describe the random variable, $T$, in terms of the probability that survival time will be at least $t$. This function is called the *survival function* and will be denoted $S(t)$. We then have

$$S(t) = 1 - F(t) = Pr[T \geq t], \; t \geq 0. \tag{2}$$

By convention, $S(t)$ is usually used in survival analysis, although $F(t)$ is more commonly used in other areas of statistics.

### 4.2 The Density Function

Another function that is useful in describing a random variable is the *density function*. To understand how this function is defined, think of the change in the value of a cumulative distribution function (as defined in the previous section) as $t$ increases by a small amount, say from $t$ to $t + \Delta t$. Symbolically, this change can be written as $F(t + \Delta t) - F(t)$. The average change over the interval is simply this value divided by the interval length, that is

$$\frac{F(t + \Delta t) - F(t)}{\Delta t} \tag{3}$$

Now consider the limit of this ratio as $\Delta t$ approaches 0, which is written as

$$f(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \tag{4}$$

Those who are familiar with calculus will recognize this limit, $f(t)$, as the derivative of $F(t)$ with respect to $t$, generally written $F'(t)$. This function is known as the *probability density function*, or simply the density function, of the random variable $T$. You might think of it as the instantaneous rate of change of the death probability with respect to time. Since $S(t) = 1 - F(t)$, it is not surprising that its instantaneous rate of change, $S'(t)$, is $-f(t) = -F'(t)$.

### 4.3 The Hazard Function

A third very useful way to characterize survival is by using a function called the *hazard function*, which we will usually denote by $h(t)$. It is the instantaneous rate of change of the death probability (as described in the previous section) conditioned on the patient's having survived to time $t$. The formula for the hazard is

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}, \quad t \geq 0 \tag{5}$$

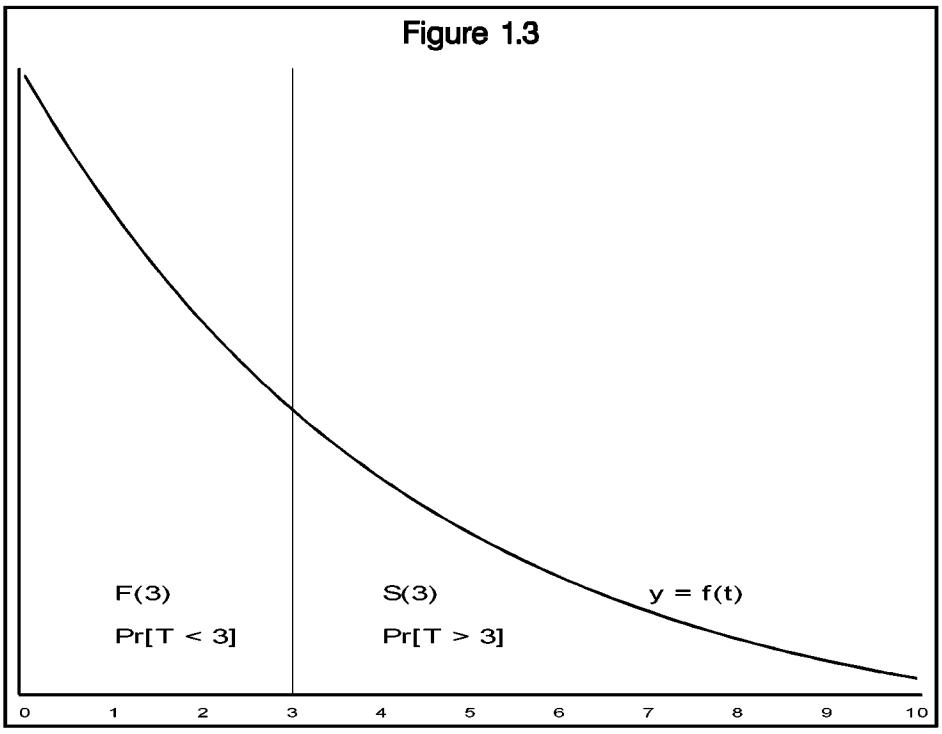
To understand why $f(t)$ is divided by $S(t)$, consider the probability of rolling a 3 on a toss of a six-sided die. Of course, that probability is 1/6. But suppose somebody tossed a die and told you that the result was an odd number. Then, what is the probability that the result is a 3, conditioned on the fact that the result is odd? Since there are now only three possible outcomes (1, 3, and 5) and all are equally likely, the answer is 1/3. That answer can be obtained by dividing 1/6 by 1/2, which is the probability of rolling an odd number. In the same manner, you can calculate the instantaneous change in the death probability at time $t$, conditioned on survival to time $t$, by dividing $f(t)$ by the probability of surviving to time $t$, $S(t)$.

Although the hazard at time $t$ conveys information about the risk of death at that time for a patient who has survived for that long, you should not think of the hazard as a probability. In fact, it may exceed 1.0. A way to associate the hazard, $h(t)$, at time $t$, with a probability is to note that, based on equation 5 and the definition of $f(t)$, you can calculate the approximation for $\Delta t$ near 0, of

$$h(t)\Delta t \doteq \frac{F(t + \Delta t) - F(t)}{S(t)} \tag{6}$$

The numerator in equation 6 is the probability that the patient dies by time $t + \Delta t$ minus the probability that he or she dies by time $t$; that is, the numerator is the probability that the patient dies at the time between $t$ and $t + \Delta t$. As noted above, dividing by $S(t)$ conditions on surviving to time $t$. Thus the hazard at time $t$ multiplied by a small increment of time approximates the probability of dying within that increment of time after $t$ for a patient who survived to time $t$.

Using a fundamental theorem of calculus, if we plot the graph of the function $y = f(t)$, then for any value, $t_0$, of $t$, $F(t_0)$ is the area above the horizontal axis, under the curve, and to the left of a vertical line at $t_0$. $S(t_0)$ is the area to the right of $t_0$. Figure 1.3 illustrates this property for $t_0 = 3$ and an arbitrary density function $f(t)$.




Another important relationship between the functions that describe survival is given by

$$S(t) = \exp[-\int_0^t h(u)du] \tag{7}$$

The integral in equation 7 is called the cumulative hazard at time *t,* and it plays a critical role in long-term survival. If this integral increases without bound as $t \to \infty$, then $S(t)$ approaches 0 as $t \to \infty$. In other words, there are no long-term survivors or "cures." If, however, the integral approaches a limit, $c < \infty$, as $t \to \infty$, then $S(t)$ approaches exp(-c) as $t \to \infty$. In this case, we can think of exp(-*c*) as the *cure rate*. Estimation of a cure rate is one of the most important and challenging problems of survival analysis. An approach to this problem will be presented in Chapter 5.

## 5. Some Commonly Used Survival Functions

### 5.1 The Exponential Function

The simplest function that you can use to describe survival is the exponential function given by

$$S(t) = \exp(-\lambda t), \; t \geq 0 \tag{8}$$

This survival function has only one parameter, the constant hazard, $\lambda$. The median survival time, defined as the solution of $S(t) = 0.5$, is $t = -\log_e(0.5)/\lambda$. Also, if we assume a probability of $p$ of surviving for time $t$, then $\lambda$ is determined by $\lambda = -\log_e(p)/t$.

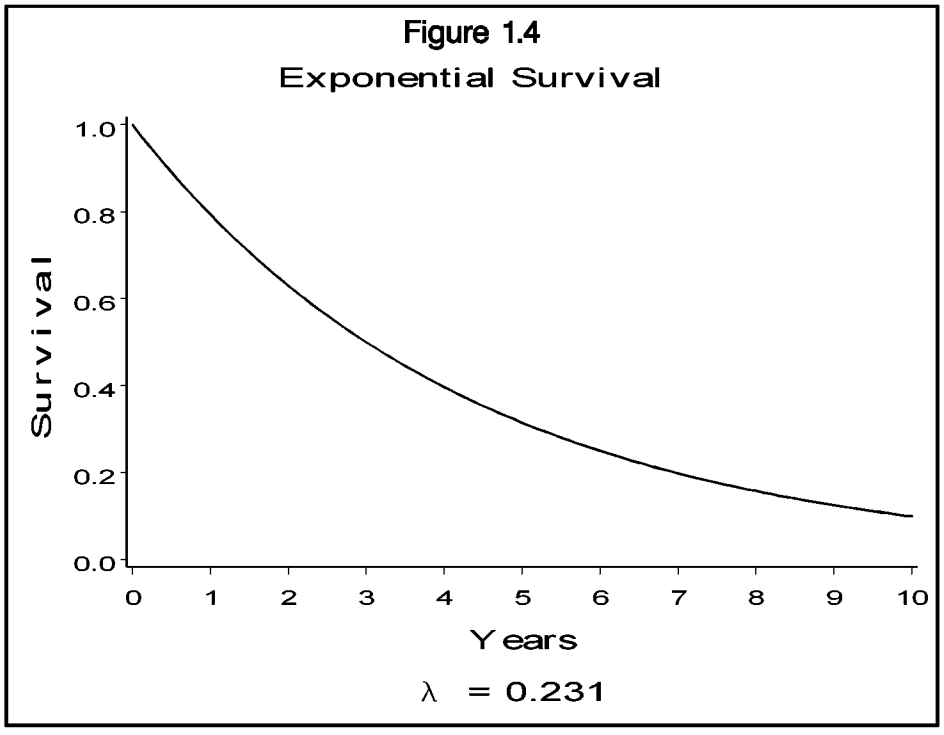
### 5.2 The Weibull Function

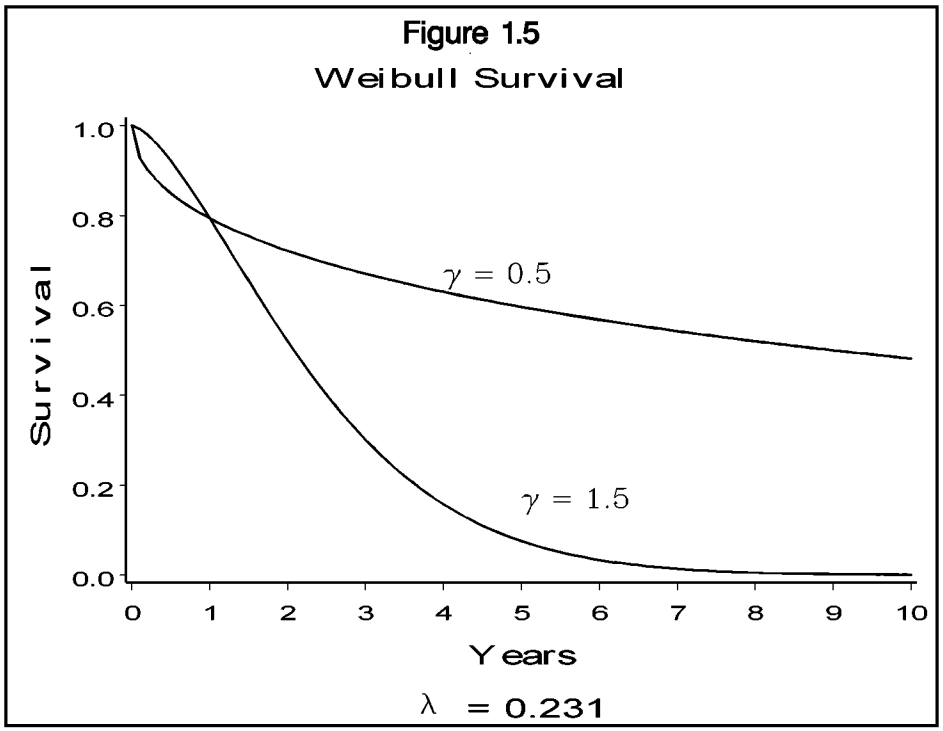
A more complex, but often more realistic, model for survival is given by the Weibull function

$$S(t) = \exp(-\lambda t^{\gamma}), \; t \geq 0 \tag{9}$$

Note that the exponential survival function is a special case of the Weibull function where $\gamma=1$. The hazard function is given by $h(t)=\lambda\gamma t^{\gamma-1}$. It increases as $t$ increases if $\gamma > 1$, and decreases as $t$ increases if $0 < \gamma < 1$

Graphs of survival functions of each type are shown in Figures 1.4 and 1.5.






Other functions, such as the lognormal, gamma, and Rayleigh, are also sometimes used to describe survival, but will not be discussed in this chapter.

## 6. Functions that Allow for Cure

### 6.1 The Idea of *Cure Models*

The survival functions described in the previous section are all based on proper distribution functions, that is $F(t) \to 1$ as $t \to \infty$. Of course, this means that $S(t) \to 0$ as $t \to \infty$. Often, however, a model, to be realistic, must allow for a nonzero probability of indefinite survival - that is, a nonzero probability of cure. Suppose you were analyzing survival data for a cohort of children who had Hodgkin's Disease. You might find that a considerable number of patients were alive, apparently free of disease and still being followed after ten years, and that no deaths had occurred after four years. You could assume that a nonzero proportion had been cured in this case. A survival function that goes to zero with increasing time is not a good model for such data.

Figures 1.6, 1.7, and 1.8 graphically illustrate three types of survival functions that allow for cure. For purposes of comparison, the parameters in each case are chosen so that the cure rate is 30% and the noncures have a median survival time of one year.

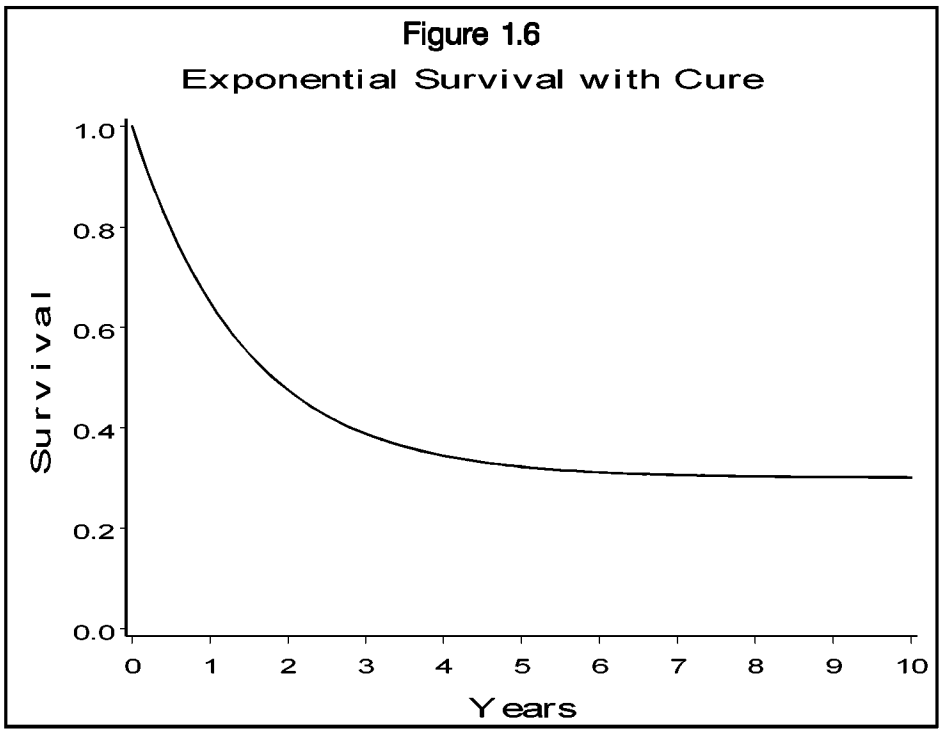
### 6.2 Mixed Models

One way to model such data is to assume that the population being studied is a mixture of two subpopulations. A proportion, $\pi$, is cured, and the remaining proportion, $1-\pi$, has a

survival function as described in section 6.1. If, for example, the survival function of the non-cured patients is exponential, the survival of the entire population can be given by

$$S(t) = \pi + (1 - \pi)\exp(-\lambda t), \; t \geq 0 \tag{10}$$

The graph of such a survival function approaches a plateau at $S(t) = \pi$ as $t \to \infty$. This model has been studied by Goldman (1984) and Sposto and Sather (1985). Figure 1.6 illustrates this example. Of course, the exponential function in equation 10 can be replaced by any survival function. For example, Gamel et al. (1994) have considered such a model based on a lognormal survival function for the non-cured patients.

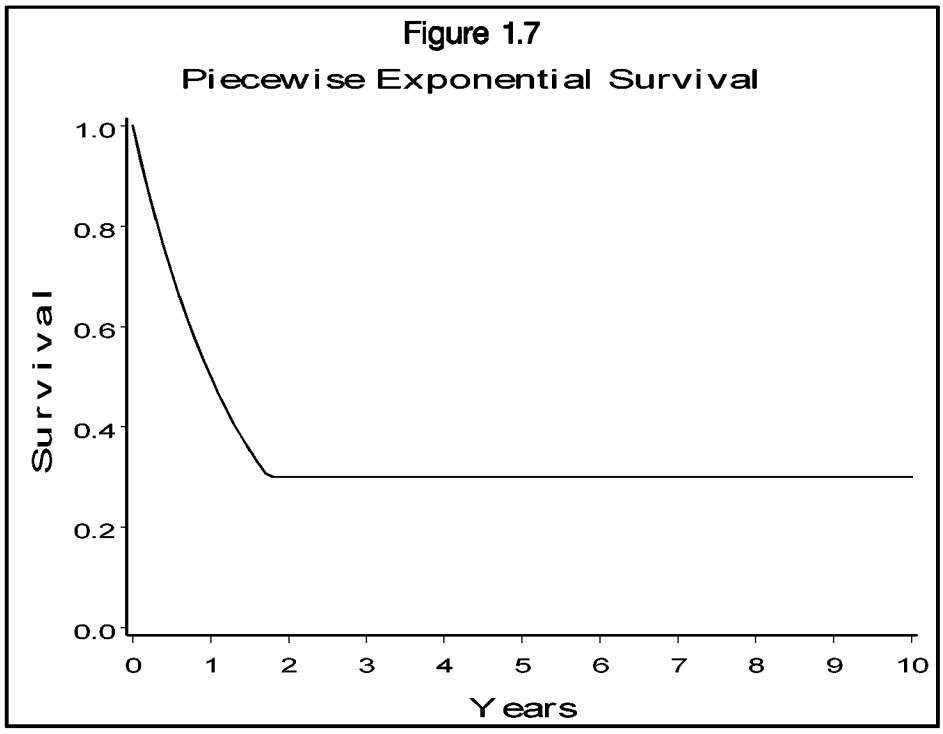


### 6.3 The Piecewise Exponential Model

Another model that can allow for cure is the piecewise exponential model as described by Shuster (1990). This model assumes that the hazard is constant over intervals, but can be different for different intervals. For example, we suppose that $h(t)=\lambda$ for $0 \le t < t_0$ and $h(t)=0$ for $t \ge t_0$. For this model, the survival function is given by

$$\begin{aligned} S(t) &= \exp(-\lambda t) \text{ for } 0 \le t < t_0 \\ S(t) &= \exp(-\lambda t_0) \text{ for } t \ge t_0 \end{aligned} \tag{11}$$

Figure 1.7 illustrates this example.



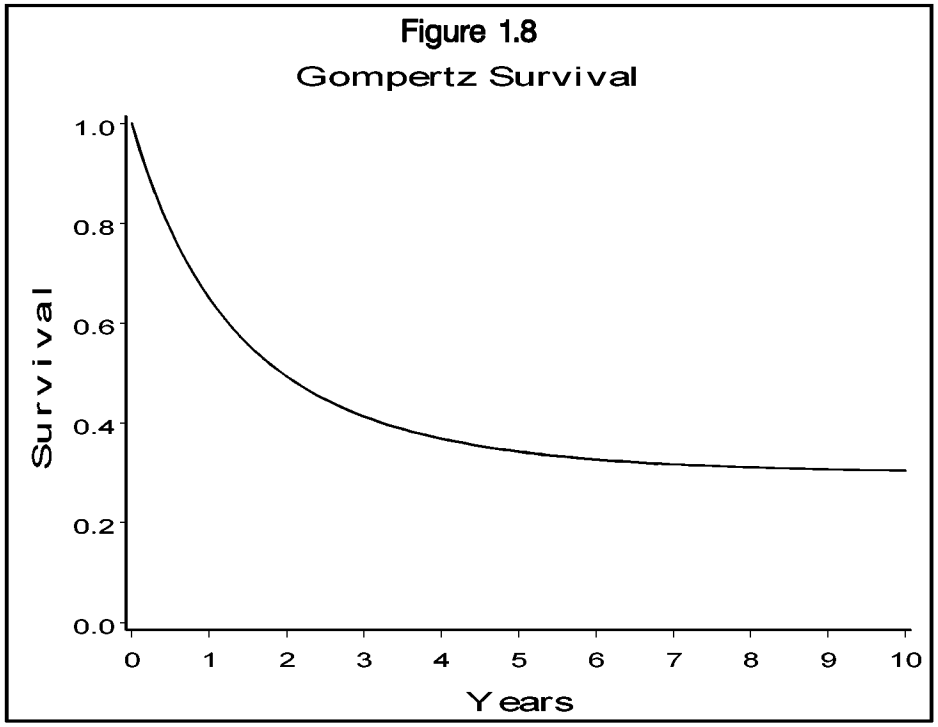

### 6.4 The Gompertz Model

Still another model for survival that allows for cure is given by the Gompertz function, which is defined by

$$S(t) = \exp\{-\frac{\gamma}{\theta}[\exp(\theta t) - 1]\} \quad \gamma > 0,\ t \ge 0 \tag{12}$$

Although this function appears to be rather complicated, it follows by equation 7 from the assumption that $h(t)$ is increasing or decreasing exponentially with rate $\theta$ as $t$ increases. In fact, this function was first used by Gompertz (1825) to describe mortality

in an aging male population in which he observed an exponentially increasing hazard. With $\theta < 0$, it's not hard to see that $S(t) \rightarrow \exp(\gamma/\theta)$ as $t \rightarrow \infty$. This function was first used to describe survival of patients by Haybittle (1959). It has also been studied in this context by others (Gehan and Siddiqui, 1973; Cantor and Shuster, 1990; and Garg, Rao, and Redmond, 1970 ). $S(t) \rightarrow \exp(-\gamma t)$ as $\theta \rightarrow 0$, so that the exponential function can be thought of as a special case of the Gompertz function with $\theta = 0$. Figure 1.8 illustrates this example.




## 7. Parametric and Nonparametric Methods

If you are willing to assume that survival can be described by a distribution of one of the types described in this chapter, then the way to use a set of sample data to make estimates or inferences about the underlying population is clear. You need to somehow obtain estimates of the parameter(s) that determine the distribution. Then, the desired estimates follow almost immediately. For example, assume that survival in a particular cohort can be described by the exponential model in equation 8. Then, if $\lambda$ is estimated (by using methods to be described later in this book) to be 0.043 where the unit of time is years, then $\exp(-3 \times 0.043) = \exp(-0.129) = 0.879$ is the estimated probability of survival for at least three years. Furthermore, if the standard error of the estimate of $\lambda$ is known, then standard methods permit us to calculate the standard error of the estimate of $S(t)$ and confidence intervals for $S(3)$. Similarly, if exponentiality is assumed for survival of patients in each of two treatment arms, then the superiority of one of the treatments is equivalent to its having a smaller $\lambda$. These matters will be discussed in more detail in Chapter 5.

Often, however, statisticians are reluctant to base analyses on assumed types of distributions. Many statistical methods, such as *t* tests and ANOVA, are rather robust to the assumption of normality for reasonably large sample sizes. Thus, inferences can often be made with some confidence even if you are not confident of normality. This is not true for methods of survival analysis. An estimate or inference based on the assumption of exponentiality might be grossly erroneous if that assumption does not hold. Thus, you would rather make statements that hold regardless of the underlying distribution. To enable us to make such statements, a class of methods has been developed that are valid without any distributional assumptions, or sometimes with only very modest distributional assumptions. Because of their power and, in many cases, their simplicity and intuitive appeal, these methods have come to dominate. Most SAS procedures that are used for survival analysis, and most of this book, is based upon such methods.

## 8. Parameters, Estimates, and the "hat" Notation

The parametric models described in sections 5 and 6 are each characterized by one or more parameters that are generally denoted by Greek letters. Their values, of course, are not known to us, but later in this book you learn about ways to estimate them. It is important that you keep in mind the distinction between parameters and their estimates. A *parameter* is an unknown and unknowable characteristic of a population. We study a sample drawn from a population in order to derive estimates of parameters. Sometimes these estimates also lead to hypothesis tests about the parameters. It is helpful to have a notation for estimates that reminds us of the parameter we are estimating. This book uses the name of a parameter with a "hat" (^) over it to represent an estimate for the parameter. For example, $\hat{\lambda}$ is the notation used for an estimate of $\lambda$. How to calculate $\hat{\lambda}$ from a sample is discussed in Chapters 4 and 5. Similarly, if survival time in a given population is described by a survival function $S(t)$, the notation $\hat{S}(t)$ will be used for estimates of $S(t)$. Methods of calculating $\hat{S}(t)$ from a sample are discussed in Chapters 2, 4, and 5.

## 9. Some Common Assumptions

In the analysis of survival data, we are frequently concerned about the effect of a variable on survival. The treatment to which a patient is assigned might be such a variable. Other variables might be demographic - age, race, or sex, for example. Still others might be associated with the patient's disease - cancer stage, number of blocked arteries, Karnofsky status, and so on. There are many ways in which a variable can impact on survival. Suppose a variable that is thought to impact on survival is observed, and let $h(t, x)$ be the hazard at time $t$ for a patient with a value of $x$ for that variable. The survival function has the proportional hazards property if, for some positive number, $c$, we have for all values of $t$ and $x$

$$\frac{h(t,\ x\ +\ 1)}{h(t,\ x)}\ =\ c \tag{13}$$

According to this assumption, the effect of the variable is multiplicative on the hazard function. Then, whether the hazard is increasing, decreasing, or not affected by increasing values of the variable depends upon whether that constant multiple is less than, greater than, or equal to 1. Another possible assumption is that the effect of a variable might be to accelerate (or decelerate) mortality. Let $S_1(t)$ and $S_2(t)$ be the survival functions for two values of a variable. Often, this variable is the treatment a patient received, so that $S_1(t)$ and $S_2(t)$ are the survival functions for the two treatments. Then, for some positive number, $b$, you might have $S_1(t) = S_2(bt)$ for all nonnegative values of $t$. In other words, the probability of surviving to time $t$ for one value of the variable is the same as the probability of surviving to time $bt$ for the other. This is called the *accelerated failure time assumption*. Whether the value associated with $S_1(t)$ is better than, worse than, or equivalent to the value associated with $S_2(t)$ depends upon whether $b$ is less than, greater than, or equal to 1. While there are statistical methods that do not require such assumptions, as you shall see later in this book, often these assumptions are reasonable and can lead to more powerful and informative analyses. Assumptions such as these, which attribute certain properties to the underlying survival distribution without specifying its form, are said to be semiparametric. Methods based on such assumptions occupy a place between the nonparametric and parametric methods.