Applied Survival Analysis

Regression Modeling of Time to Event Data

DAVID W. HOSMER, Jr.

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

CHAPTER 2

Descriptive Methods for Survival Data

2.1 INTRODUCTION

In any applied setting, a statistical analysis should begin with a thoughtful and thorough univariate description of the data. The fundamental building block of this analysis is an estimate of the cumulative distribution function. Typically, not much attention is paid to this fact in an introductory course on statistical methods, where directly computed estimators of measures of central tendency and variability are more easily explained and understood. However, routine application of standard formulas for estimators of the sample mean, variance, median, etc., will not yield estimates of the desired parameters when the data include censored observations. In this situation, we must obtain an estimate of the cumulative distribution function in order to obtain values of statistics which do estimate the parameters of interest.

In the HMO-HIV+ study described in Chapter 1, we assume that the recorded data are continuous and are subject to right censoring only. Remember that time itself is always continuous, but our inability to measure it precisely is an issue that we must deal with. We introduced the cumulative distribution function in Chapter 1 along with its complement, the survivorship function. Simply stated, the cumulative distribution function is the probability that a subject selected at random will have a survival time less than some stated value, t. This is denoted as $F(t) = \Pr(T < t)$. The survivorship function is the probability of observing a survival time greater than or equal to some stated value t, denoted $S(t) = \Pr(T \geq t)$. In most applied settings we are more interested in describing how long the study subjects live, than how quickly they die. Thus estimation (and inference) focuses on the survivorship function.

2.2 ESTIMATION OF THE SURVIVORSHIP FUNCTION

The Kaplan-Meier estimator of the survivorship function [Kaplan and Meier (1958)], also called *the product limit* estimator, is the estimator used by most software packages. This estimator incorporates information from all of the observations available, both uncensored and censored, by considering survival to any point in time as a series of steps defined by the observed survival and censored times. It is analogous to considering a toddler who must take five steps to walk from a chair to a table. This journey of five steps must begin with one successful step. The second step can only be taken if the first was successful. The third step can be taken only if the second (and also the first) was successful. Finally the fifth step is possible only if the previous four were completed successfully. In an analysis of survival time, we estimate the conditional probabilities of "successful steps" and then multiply them together to obtain an estimate of the overall survivorship function.

To illustrate these ideas in the context of survival analysis, we describe estimation of the survivorship function in detail using data for the first five subjects in the HMO-HIV+ study in Table 1.1, as shown in Table 2.1.

The "steps" are intervals defined by a rank ordering of the survival times. Each interval begins at an observed time and ends just before the next ordered time and is indexed by the rank order of the time point defining its beginning. Subject 4's survival time of 3 months is the shortest and is used to define the interval $I_0 = \{t : 0 \le t < 3\} = [0,3)$. The expression in curly brackets, { }, defines a collection or set of values that includes all times beginning with and including 0 and up to, but not including, 3. This is more concisely denoted using the mathematical notation of a square bracket to mean the value is included, a parenthesis to mean the value is not included, and the comma to mean all values in between. We use both notations in this text. The second rank-ordered

Table 2.1 Survival Times and Vital Status (Censor) for Five Subjects from the HMO-HIV+ Study

Subject	Time	Censor
1	5	1
2	6	0
3	8	1
4	3	1
5	22	1

time is subject 1's survival time of 5 months. This survival time, in conjunction with the ordered survival time of subject 4, defines interval $I_1 = \{t : 3 \le t < 5\} = [3,5)$. The next ordered time is subject 2's censored time of 6 months and, in conjunction with subject 1's value of 5 months, defines interval $I_2 = \{t : 5 \le t < 6\} = [5,6)$. The next interval uses subject 3's value of 8 months and the previous value of 6 months and defines $I_3 = \{t : 6 \le t < 8\} = [6,8)$. Subject 5's value of 22 months and subject 3's value of 8 months are used to define the next to last interval $I_4 = \{t : 8 \le t < 22\} = [8,22)$. The last interval is defined as $I_5 = \{t : t \ge 22\} = [22,\infty)$.

All subjects were alive at time $t = 0$ and remained so until subject 4 died at 3 months. Thus, the estimate of the probability of surviving through interval I_0 is 1.0; thus, the estimate of the survivorship function is

$$\hat{S}(t) = 1.0$$

at each t in I_0. Just before time 3 months, five subjects were alive, and at 3 months one subject died. In order to describe the value of the estimator at 3 months, consider a small interval beginning just before 3 months and ending at 3 months. We designate such an interval as $(3-\delta, 3]$. The estimated conditional probability of dying in this small interval is 1/5 and the probability of surviving through it is $1-1/5 = 4/5$. At any specified time point, the number of subjects alive is called the number at risk of dying or simply the number at risk. At time 3 months this number is denoted as n_1, the 1 referring to the fact that 3 months is the first observed time. The number of deaths observed at 3 months was 1 but, with a larger sample, more than one could have been observed. To allow for this, we denote the number of deaths observed as d_1. In this more general notation, the estimated probability of dying in the small interval around 3 is d_1/n_1 and the estimated probability of surviving is $(n_1 - d_1)/n_1$. The probability that a subject survives to 3 months is estimated as the probability of surviving through interval I_0 times the conditional probability of surviving through the small interval around 3. Throughout the discussion of the Kaplan-Meier estimator, the word "conditional" refers to the fact that the probability applies to those who survived to the point or interval under consideration. Since we observed the death at exactly 3 months, this estimated probability would be the same no matter how small a value of δ we use to define the interval around 3 months. Thus, we consider the estimate of the survival probability to be at exactly 3 months. The value of this estimate is

$$\hat{S}(3) = 1.0 \times (4/5) = 0.8.$$

We now consider estimation of the survivorship function at each time point in the remainder of interval I_1. No other failure times (deaths) were observed, hence the estimated conditional probability of survival through small intervals about every time point in the interval is 1.0. Cumulative multiplication of these times the estimated survivorship function leaves it unchanged from its value at 3 months.

The next observed failure time is 5 months. The number at risk is $n_2 = 4$ and the number of deaths is $d_2 = 1$. The estimated conditional probability of surviving through a similarly defined small interval at 5 months, $(5-\delta, 5]$, is $(4-1)/4 = 0.75$. By the same argument used at 3 months, the estimate of the survivorship function at 5 months is the product of the respective estimated conditional probabilities,

$$\hat{S}(5) = 1.0 \times (4/5) \times (3/4) = 0.6.$$

No other failure times were observed in I_2, thus the estimate remains at 0.6 through the interval.

The number at risk at the next observed time, 6 months, is $n_3 = 3$ and the number of deaths is zero since subject 2 was lost to follow-up at 6 months. The estimated conditional probability of survival through a small interval at 6 months is $(3-0)/3 = 1.0$. Again, the estimated survivorship function is obtained by successive multiplication of the estimated conditional probabilities and is

$$\hat{S}(6) = 1.0 \times (4/5) \times (3/4) \times (3/3) = 0.6.$$

No failure times were observed in I_3 and the estimate remains the same until the next observed failure time.

The number at risk 8 months after the beginning of the study is $n_4 = 2$ and the number of deaths is $d_4 = 1$. The estimated conditional probability of survival through a small interval at 8 months is $(2-1)/2 = 0.5$. Hence, by the same argument used at 3, 5 and 6 months, the estimated survivorship function at 8 months after the beginning of the study is

$$\hat{S}(8) = 1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) = 0.3.$$

No other failure times were observed in I_4, thus the estimated survivorship function remains constant and equal to 0.3 throughout the interval.

The last observed failure time was 22 months. There was a single subject at risk and this subject died, hence $n_5 = 1$ and $d_5 = 1$. The estimated conditional probability of surviving through a small interval at 22 months is $(1-1)/1 = 0.0$. The estimated survivorship function at 22 months is

$$\hat{S}(22) = 1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) \times (0/1) = 0.0.$$

No subjects were alive after 22 months; thus the estimated survivorship function is equal to zero after that point.

Through this example, we have demonstrated the essential features of the Kaplan-Meier estimator of the survivorship function. The estimator at any point in time is obtained by multiplying a sequence of conditional survival probability estimators. Each conditional probability estimator is obtained from the observed number at risk of dying and the observed number of deaths and is equal to " $(n-d)/n$." This estimator allows each subject to contribute information to the calculations as long as they are known to be alive. Subjects who die contribute to the number at risk up until their time of death, at which point they also contribute to the number of deaths. Subjects who are censored contribute to the number at risk until they are lost to follow-up.

The estimate obtained from the data in Table 2.1 is presented in tabular form in Table 2.2. Computer software packages often present an abbreviated version of this table containing only the observed failure times and estimates of the survivorship function at these times with the implicit understanding that it is constant between failure times.

A graph is an effective way to display an estimate of a survivorship function. The graph shown in Figure 2.1 is obtained from the survivorship function in Table 2.2. The graph shows the decreasing step function defined by the estimated survivorship function. It drops at the values of the observed failure times and is constant between observed failure times. An embellishment provided by some software packages, but rarely presented in published articles, is an indicator on the graph where censored observations occurred. The censored time of 6 months appears as a small × in the figure.

Table 2.2 Estimated Survivorship Function Computed from the Survival Times for the Five Subjects from the HMO-HIV+ Study Shown in Table 2.1

Interval	$\hat{S}(t)$
$0 \le t < 3$	1.0
$3 \le t < 5$	0.8
$5 \le t < 6$	0.6
$6 \le t < 8$	0.6
$8 \le t < 22$	0.3
$t \ge 22$	0.0

In our example, no two subjects shared an observation time, and the longest observed time was a failure. Simple modifications to the method described above are required when either of these conditions is not met. Consider a case where a failure and a censored observation have the same recorded value. We assume that, since the censored observation was known to be alive when last seen, its survival time is longer than the recorded time. Thus a censored subject contributes to the number at risk at the recorded time but is not among those at risk immediately after that time. Along the same lines, suppose we have multi-

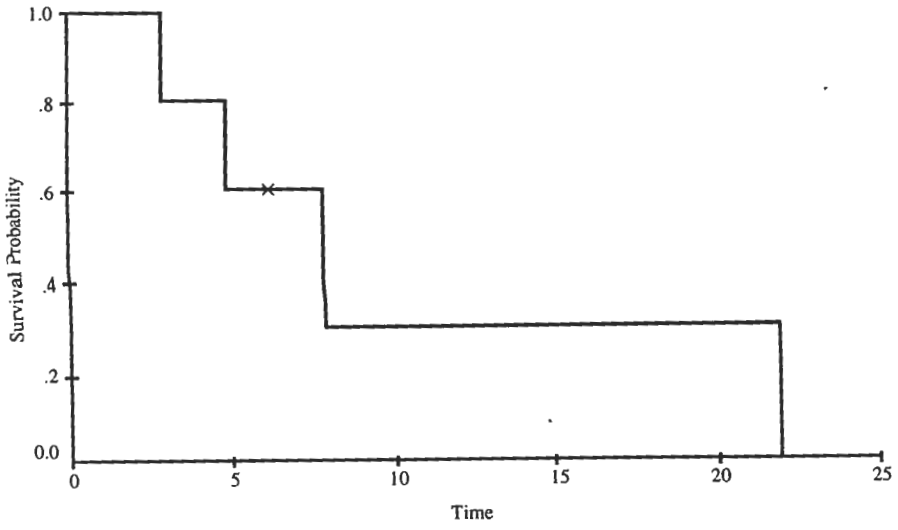

Figure 2.1 Graph of the estimated survivorship function from Table 2.2.

ple failures, $d > 1$, at some time t. It is unlikely that each subject died at the exact same time t; however, we were unable to record the data with any more accuracy. One way to break these ties artificially would be to order the d tied failure times randomly by subtracting a tiny random value from each. For example, if we had observed three values at 8 months we could subtract from each failure time the value of a uniformly distributed random variable on the interval (0, 0.01). This would artificially order the times, yet not change their respective positions relative to the rest of the observed failure times. We would estimate the survivorship function with $d = 1$ at each of the randomly ordered times. The resulting estimate of the survivorship function at the last of the d times turns out to be identical to that obtained using $(n-d)/n$ as the estimate of the conditional probability of survival for all d considered simultaneously. Thus, it is unnecessary to make adjustments for ties when estimating the survivorship function. However, if there are extensive numbers of tied failure times, then a discrete time model may be a more appropriate model choice (see Chapter 7).

If the last observed time corresponds to a censored observation, then the estimate of the survivorship function does not go to zero. Its smallest value is that estimated at the last observed survival time. In this case the estimate is considered to be undefined beyond the last observed time. If both censored and non-censored values occur at the longest observed time, then the protocol of assuming that censoring takes place after failures dictates that $(n-d)/n$ is used to estimate the conditional survival probability at this time. The estimated survivorship function does not go to zero and is undefined after this point. When these types of ties occur, software packages, which provide a tabular listing of the observed survival times and estimated survivorship function, list the censored observations after the survival time, with the value of the estimated survivorship function at the survival time. Simple examples demonstrating each of these situations are obtained by adding additional subjects to the five shown in Table 2.1.

In order to use the Kaplan–Meier estimator in other contexts, we need a more general formulation. Assume we have a sample of n independent observations denoted (t_i, c_i), $i = 1, 2, \ldots, n$ of the underlying survival time variable T and the censoring indicator variable C.[1] Assume that among the n observations there are $m \leq n$ recorded times of failure.

---

[1] Unless stated otherwise we assume recorded values of time are continuous and subject only to right censoring.

We denote the rank-ordered survival times as $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. In this text, when quantities are placed in rank order we use the same variable notation but place subscripts in parentheses. Let the number at risk of dying at $t_{(i)}$ be denoted n_i and the observed number of deaths be denoted d_i. The Kaplan–Meier estimator of the survivorship function at time t is obtained from the equation

$$\hat{S}(t) = \prod_{t_{(i)} \le t} \frac{n_i - d_i}{n_i} \tag{2.1}$$

with the convention that

$$\hat{S}(t) = 1 \text{ if } t < t_{(1)}.$$

This formulation differs slightly from that described using the data in Table 2.1 in that intervals defined by censored observations are not considered. We saw in the example that conditional survival probabilities are equal to one at censored observations and that the estimate of the survivorship function is unchanged from the value at the previous survival time. Thus the general formula in (2.1) uses only the points at which the value of the estimator changes.

Figure 2.2 presents the graph of the Kaplan–Meier estimate of the

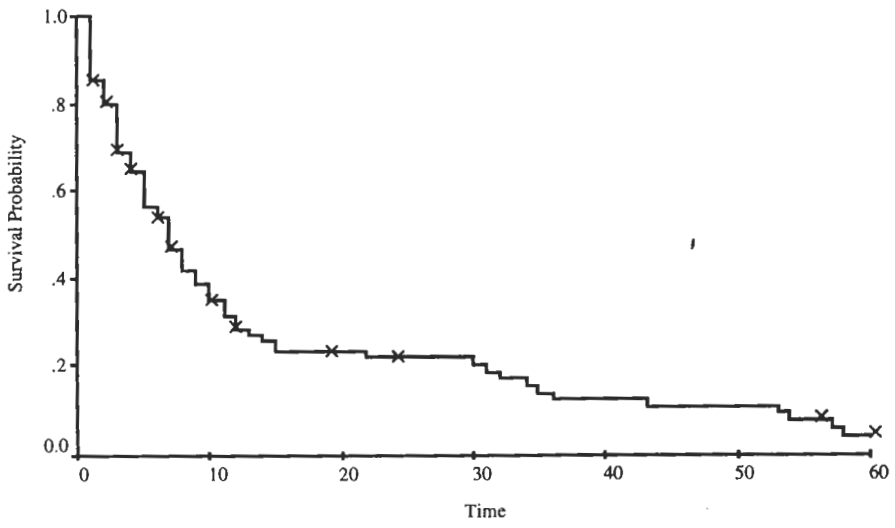

Figure 2.2 Kaplan–Meier estimate of the survivorship function for the HMO-HIV+ study.

survivorship function in (2.1), using all subjects in the HMO-HIV+ study. The construction of the estimate in this case demonstrates conventions for handling tied survival times as well as tied survival and censored times. The data, along with calculations for the beginning and end of the survivorship function, are presented in Table 2.3. The columns in Table 2.3 present the time interval, the number at risk of dying (n), the number of deaths (d), the number of subjects lost to follow-up (c), the estimate of the conditional survival probability $[(n-d)/n]$ and the estimate of the survivorship function $[\hat{S}(t)]$. All quantities are evaluated at the time point defined by the end of the previous interval and the beginning of the current interval.

The first observed survival time is 1 month; thus the value of the estimated survivorship function at each point in the interval [0,1) is 1.0. At 1 month there were 100 subjects at risk. Of these, 15 died and 2 were lost to follow-up (censored), yielding an estimate of the conditional survival probability of $0.85 = (100-15)/100$. The estimate of the survivorship function at 1 month is $0.85 = 1.0 \times 0.85$. The estimate remains at this value at each point in the interval [1,2). At the next observed survival time, 2 months, there were only 83 subjects at risk since 15 died and 2 were lost to follow-up one month before. At 2 months, 5 subjects died and 5 more were lost to follow-up; thus the estimate of the conditional survival probability is $(83-5)/83 = 0.9398$. The estimate of the survivorship function is obtained as the product of the value of the survivorship function just prior to 2 months and the conditional survival probability at 2 months and is $0.85 \times 0.9398 = 0.7988$. The estimate remains at this value throughout the interval [2,4). At the next observed survival time, 4 months, there were 73 subjects at risk, since 5 died and 5 were censored at 2 months. At 4 months, 10 subjects died and 2 were censored. The estimate of the conditional survival probability is

Table 2.3 Partial Calculations of the Kaplan–Meier Estimate Shown in Figure 1.2

Interval	n	d	c	$(n-d)/n$	$\hat{S}$
[0,1)	100	0	0	1.0	1.0
[1,2)	100	15	2	0.85	0.85
[2,4)	83	5	5	0.9398	0.7988
[4,5)	73	10	2	0.8630	0.6894
⋮	⋮	⋮	⋮	⋮	⋮
[58,60)	3	1	0	0.6667	0.0389
[60,60]	2	0	2	1.0	0.0389

$(73-10)/73 = 0.8630$ and the estimate of the survivorship function is $0.7988 \times 0.8630 = 0.6894$. The estimate remains at this value until the next observed survival time, 5 months, at which time 61 subjects are at risk. This process continues, sequentially, considering each observed survival time, until the last observed survival time, which was 58 months. At that time 3 subjects were at risk, 1 died and none were censored. The estimate of the conditional survival probability is $(3-1)/3 = 0.6667$. The estimate of the survivorship function is $0.0584 \times 0.6667 = 0.0389$, where 0.0584 is the value just prior to 58 months. The largest observed time is 60 months, when 2 subjects remained at risk and both were censored. Thus, the estimate of the conditional survival probability is $(2-0)/2 = 1.0$ and the estimate of the survivorship function remains at the value 0.0389. The function is undefined beyond 60 months, which is denoted in Table 2.3 by recording the last interval as [60,60].

When we have a large study whose mortality experience is presented in calendar time units (such as quarterly, semi-annually, etc.), the life-table estimator of the survivorship function may be used as an alternative to the Kaplan–Meier estimator. The life-table estimator has been used for more than 100 years to describe human mortality experience and is among the earliest examples of the application of statistical methods. It will not play a large role in the analysis of survival data in this text, but we present it because of its historical importance and the fact that it is a grouped-data analog of the Kaplan–Meier estimator. More detail on the various types of life-table estimators may be found in Lee (1992).

In some applied settings the data may be quite extensive with sample sizes in the many hundreds of subjects. In these situations it can be quite cumbersome to tabulate or graph the Kaplan–Meier estimator of the survivorship function. In a sense, the problem faced is similar to one addressed in a first course on statistical methods: how best to reduce the volume of data but not the statistical information that can be gleaned from it. To this end the histogram is usually introduced as an estimator of the density function and the resulting cumulative percent distribution polygon as an estimator of the cumulative distribution function. This process could be reversed. That is, we might first derive the estimator of the cumulative distribution and, afterwards, compute the histogram as a function of the cumulative distribution. When the data contain censored observations, using the second approach and deriving an estimator of the survivorship function (instead of the cumulative distribution function) is the more feasible tactic. The first step is to define the intervals that will be used to group the data. The goal in the choice of intervals is

the same as for the construction of a histogram—the intervals should be biologically meaningful, yield an adequate description of the data and, if convenient, be of equal width. There are no mechanized rules for construction of the histogram, to guide in the choice of number of intervals. However, the meaningful unit will likely be some multiple of a year.

Once a set of intervals has been chosen, the construction of the estimator follows the basic idea used for the Kaplan–Meier estimator. Suppose we decide to use 6-month intervals. A typical interval will be of the form $[t, t+6)$. As before, let n denote the number of subjects at risk of dying at time t. These subjects are often described as the number who enter the interval alive. As we follow these subjects across the interval, d subjects have survival times and c subjects have censored times in this interval. Thus, not all subjects were at risk of dying for the entire interval. A modification typically employed is to reduce the size of the risk set by one-half of those censored in the interval. The rationale behind this adjustment is that if we assume the censored observations were uniformly distributed over the interval, then the average size of the risk set in the interval is $n-(c/2)$. This average risk set size is used to calculate the estimate of the conditional probability of survival through the interval as $\big(n-(c/2)-d\big)/\big(n-(c/2)\big)$. These estimates of the conditional probabilities are multiplied to obtain the life-table estimator of the survivorship function.

The life-table estimator of the survivorship function for the HMO-HIV+ data using 6-month intervals is shown in Table 2.4. The estimated value of the survivorship function in the first interval is

$$0.5684 = (100-(10/2)-41)/(100-(10/2)).$$

The value in the second interval is computed as

$$0.3171 = 0.5684 \times (49-(3/2)-21)/(49-(3/2)).$$

The remaining values are calculated in a similar fashion.

When we graph the estimate, we have to decide how to represent the actual values. Consider the first interval [0, 6), where the value of the estimated survivorship function is reported in Table 2.4 as 0.5684. If, as in Figure 2.3, we were to represent the graph as a step function, then this interval would be represented by a horizontal straight line of height

Table 2.4 Life-Table Estimator of the Survivorship Function for the HMO-HIV+ Study

Interval	Enter	Die	Censored	$\hat{S}$
[0 , 6)	100	41	10	0.5684
[6 , 12)	49	21	3	0.3171
[12 , 18)	25	6	2	0.2378
[18 , 24)	17	1	1	0.2234
[24 , 30)	15	0	1	0.2234
[30 , 36)	14	5	0	0.1436
[36 , 42)	9	1	0	0.1277
[42 , 48)	8	1	0	0.1117
[48 , 54)	7	1	0	0.0958
[54 , 60)	6	3	1	0.0435
[60 , 66)	2	0	2	0.0435

1 until 6 months when it would drop to 0.5684. Other intervals would be represented in a similar manner. An alternative representation, used by some software packages, is a polygon connecting the value of the estimator drawn at the end of the interval. The first interval would be

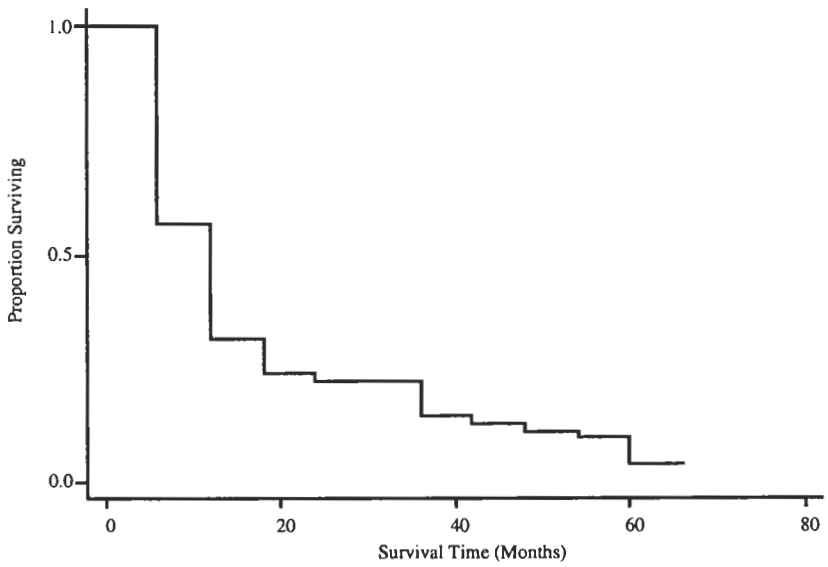

Figure 2.3 Step function representation of life-table estimate of the survivorship function for the HMO-HIV+ study in Table 2.4.

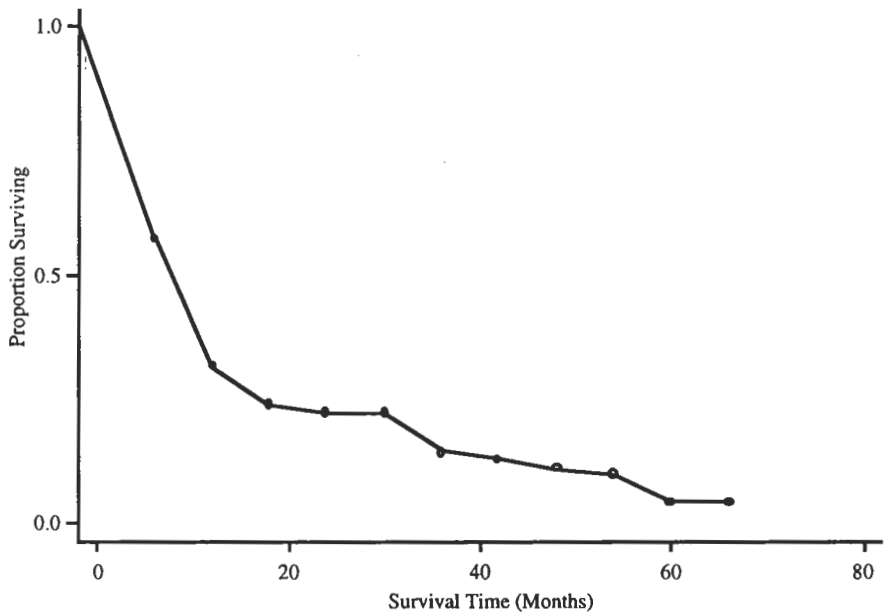

Figure 2.4 Polygon representation of life-table estimate of the survivorship function for the HMO-HIV+ study in Table 2.4.

represented by a point of height 0.5684 plotted at 6 months, the second by a point of height 0.3171 at 12 months, the third by a point of height 0.2378 at 18 months, and so on. These points are then connected by straight lines. The rationale for using the polygon is to better represent the assumed underlying continuous distribution of survival time. Some, but not all, programs will plot a point equal to 1.0 at time zero since, by definition, that is the value of the survivorship function at zero. This point is then connected to the point representing the first interval. The polygonal representation of the life-table estimator from Table 2.4 is shown in Figure 2.4.

Because the graph in Figure 2.4 has been drawn as a polygon, it looks smoother than the step function of the Kaplan–Meier estimator. The life-table estimate in Figure 2.3 in this example does a reasonable job of estimating the survivorship function. Since it is a grouped-data statistic, it is not as precise an estimate as the Kaplan–Meier estimator, which uses the individual values. Later in this chapter we discuss estimation of percentiles of the survival time distribution and these use the Kaplan–Meier estimator.

2.3 USING THE ESTIMATED SURVIVORSHIP FUNCTION

In Section 2.2 we described in detail how to calculate the Kaplan–Meier and life-table estimators of the survivorship function with little if any discussion of how to interpret the resulting estimate or how it may be used to derive point estimates of quantiles of the distribution. One of the biggest challenges in survival analysis is becoming accustomed to using the survivorship function as a descriptive statistic. This function describes the complement of what we typically describe in a set of data. The change from thinking about the percentage of observations less than a value to thinking about the percentage greater than that value, like many things, becomes easier with practice.

The survivorship function estimate shown in Figure 2.2 descends sharply at first and then tails off gradually, reaching its minimum value of 0.04 at 60 months. The initial steep descent shows that there were many subjects who died shortly after enrollment in the study. The relatively long right tail is a result of the few subjects who had long survival times. The minimum value of the survivorship function is not zero since the largest observed time was a censored observation. The shape of the curve depends on the observed survival times and the proportion of observations that are censored. If many subjects in the HMO-HIV+ study had long survival times with the same pattern of censored observations, then the curve would descend slowly at first and then more rapidly until the minimum is reached. If the survival times were more evenly distributed over the 60 months, then the curve would descend gradually to its minimum value. The pattern of enrollment in a follow-up study can influence the shape of the curve. A study with a 2-year enrollment period and 5 years overall length with many late entries is likely to have more censored observations and thus a different looking estimated survivorship function than the same study with many early entries. Many factors influence the shape of the survivorship function, and thus it is difficult to make accurate statements about what a "typical" survivorship function will look like.

In most, if not all, applied settings we will need a confidence interval estimate for the survivorship function as well as point and confidence interval estimates of various quantiles of the survival time distribution. We begin by discussing confidence interval estimation of the survivorship function.

Several different approaches may be taken when deriving an estimator for the variance of the Kaplan–Meier estimator. We derive it

from a technique which is referred to as the *delta method* and is based on a first-order Taylor series expansion. This method is presented in general terms in Appendix 1. The Kaplan–Meier estimator at any time t may be viewed as a product of proportions. Rather than derive a variance estimator of this product, we derive one for its log since the variance of a sum is simpler to calculate than variance of a product. The log of the Kaplan–Meier estimator is

$$\ln\left(\hat{S}(t)\right) = \sum_{t_{(i)} \le t} \ln\left(\frac{n_i - d_i}{n_i}\right)$$
$$= \sum_{t_{(i)} \le t} \ln\left(\hat{p}_i\right),$$

where

$$\hat{p}_i = \left(n_i - d_i\right)/n_i .$$

If we consider the observations in the risk set at time $t_{(i)}$ to be independent Bernoulli observations with constant probability, then $\hat{p}_i$ is an estimator of this probability and an estimator of its variance is $\left(\hat{p}_i(1-\hat{p}_i)\right)/n_i$. As shown in Appendix 1, the variance of the log of variable X is approximately:

$$\mathrm{Var}[\ln(X)] \cong \frac{1}{\mu_X^2}\sigma_X^2, \tag{2.2}$$

where the mean and variance of X are denoted μ_X and σ_X^2, respectively. An estimator for the variance is obtained by replacing μ_X and σ_X^2 in (2.2) with estimators of their respective values. Applying this result to $\ln\left(\hat{p}_i\right)$ yields the estimator

$$\hat{\mathrm{Var}}\left[\ln(\hat{p}_i)\right] \cong \frac{1}{\hat{p}_i^2}\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$$
$$\cong \frac{d_i}{n_i(n_i - d_i)} .$$

If we assume that observations at each time are independent, then the estimator of the variance of the log of the survivorship function is

$$\widehat{\mathrm{Var}}\left[\ln\left(\hat{S}(t)\right)\right] = \sum_{t_{(i)} \le t} \widehat{\mathrm{Var}}\left[\ln(\hat{p}_i)\right]$$
$$= \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} . \qquad (2.3)$$

An estimator of the variance of the survivorship function is obtained by another application of the delta method shown in Appendix 1. This time an approximation is applied to find the variance of an exponentiated variable and is

$$\mathrm{Var}(e^X) \cong \left(e^{\mu_X}\right)^2 \sigma_X^2 . \qquad (2.4)$$

Using the fact that $\hat{S}(t) = e^{\ln\left(\hat{S}(t)\right)}$, we let X stand for $\ln\left(\hat{S}(t)\right)$, σ_X^2 stand for the variance estimator in (2.3) and approximate μ_X by $\ln\left(\hat{S}(t)\right)$ in expression (2.4). Then we obtain Greenwood's formula [Greenwood (1926)] for the variance of the survivorship function:

$$\widehat{\mathrm{Var}}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} . \qquad (2.5)$$

The method shown to derive the estimator in (2.5) is, in some sense, the "traditional" approach in that it may be found in most texts on survival analysis published prior to 1990. In contrast, the texts by Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993) consolidate a large number of results derived from applications of theory based on counting processes and martingales. This theory is well beyond the scope of this text, but we mention it here as it has allowed development of many useful tools and techniques for the study of survival time. The current thrust in the development of software is based on the counting process paradigm as its methods and tools may be used to analyze, in a relatively uncomplicated manner, some rather complex problems. The estimator in (2.5) may also be obtained from the counting process approach.

The counting process approach to the analysis of survival time plays a central role in many of the methods discussed in this text. A brief presentation of the central ideas behind the counting process formulation of survival analysis is given in Appendix 2. We will use results

from this theory to provide justification for estimators, confidence interval estimators and hypothesis testing methods.

After obtaining the estimated survivorship function, we may wish to obtain pointwise confidence interval estimates. The counting process theory has been used to prove that the Kaplan–Meier estimator and functions of it are asymptotically normally distributed [Andersen, Borgan, Gill and Keiding (1993, Chapter IV) or Fleming and Harrington (1991, Chapter 6)]. Thus, we may obtain pointwise confidence interval estimates for functions of the survivorship function by adding and subtracting the product of the estimated standard error times a quantile of the standard normal distribution. We could apply this theory directly to the Kaplan–Meier estimator using the variance estimator in (2.5). However, this approach could easily lead to confidence interval endpoints that are less than zero or greater than one. In addition, the assumption of normality implicit in the use of the procedure may not hold for the small to moderate sample sizes often seen in typical problems. To address these problems, Kalbfleisch and Prentice (1980, page 15) suggest that confidence interval estimation should be based on the function

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right],$$

called the *log-log survivorship function*. One advantage of this function over the survivorship function is that its possible range is from minus to plus infinity. The expression for the variance of the log-log survivorship function is obtained from a second application of the delta method for a log transformed variable shown in (2.2). The estimator of the variance of the log-log survivorship function is

$$\hat{\mathrm{Var}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\} = \frac{1}{\left[\ln\left(\hat{S}(t)\right)\right]^2}\sum_{t_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}. \tag{2.6}$$

The endpoints of a $100(1-\alpha)$ percent confidence interval for the log-log survivorship function are given by the expression

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm z_{1-\alpha/2}\hat{\mathrm{SE}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\}, \tag{2.7}$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution and $\hat{\mathrm{SE}}(\cdot)$ represents the estimated standard error of the argu-

ment, which in this case is the positive square root of (2.6). If we denote the lower and upper endpoints of this confidence interval as $\hat{c}_l$ and $\hat{c}_u$, it follows that the lower and upper endpoints of the confidence interval for the survivorship function are

$$\exp\left[-\exp\left(\hat{c}_u\right)\right] \text{ and } \exp\left[-\exp\left(\hat{c}_l\right)\right], \tag{2.8}$$

respectively. That is, the lower endpoint from (2.7) yields the upper endpoint in (2.8). These are the endpoints reported by most, if not all, software packages for each observed value of survival time. The confidence interval is valid only for values of time over which the Kaplan–Meier estimator is defined, which is basically the observed range of survival times. Borgan and Leistøl (1990) studied this confidence interval and found that it performed well for sample sizes as small as 25 with up to 50 percent right-censored observations.

Figure 2.5 presents the Kaplan–Meier estimator of the survivorship function for the HMO-HIV+ study and the upper and lower pointwise 95 percent confidence bands computed using (2.8). The endpoints of the pointwise confidence intervals are connected to form a "confidence band." (Recall that any time one has a collection of individual 95 percent confidence interval estimates, the probability that they all contain their respective parameters is much less than 95 percent.) An alternative presentation used by some software packages connects the endpoints of the confidence intervals with vertical lines. This is useful for small data sets, but for large data sets the resulting graph becomes cluttered with too many lines, and we lose the visual conciseness seen in Figure 2.5. This figure demonstrates some of the properties of the log-log-based confidence interval estimator. The intervals are skewed for large and, though harder to see in Figure 2.5, small values of the estimated survivorship function and are fairly symmetric around 0.5. The direction of skewness is opposite for the two tails, toward zero for values of the estimated survivorship function near one and toward one for values near zero. In all cases, the endpoints lie between zero and one. In Figure 2.5, the confidence intervals further support the observation of a survivorship function describing many early deaths with a few deaths near the maximum of 5 years of follow-up.

Simultaneous confidence bands for the entire survivorship function are not as readily available as the pointwise estimates, since they require percentiles for statistical distributions not typically computed by software packages. The band proposed by Hall and Wellner (1980) is discussed in some detail in Andersen, Borgan, Gill and Keiding (1993) and

Fleming and Harrington (1991). It is also discussed in Marubini and Valsecchi (1995). A table of percentiles obtained from Hall and Wellner (1980) is provided in Appendix 3. Given the tabled percentiles, confidence bands based on the estimated survivorship function itself, or its log-log transformation, are not difficult to calculate. Borgan and Leistøl (1990) show that the performance of the Hall and Wellner confidence bands is comparable for both functions and is adequate for samples as small as 25 with up to 50 percent censoring. To maintain consistency with the pointwise intervals calculated in (2.8), which are based on the log-log transformation, we present the Hall and Wellner bands for the transformed function. Hall and Wellner, as well as Borgan and Leistøl, recommend that these confidence bands be restricted to values of time smaller than or equal to the largest observed survival time, e.g., the largest non-censored value of time denoted $t_{(m)}$. The endpoints of the $100(1-\alpha)$ percent confidence bands in the interval $[0, t_{(m)}]$ for the log-log transformation are

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm H_{\hat{a},\alpha} \frac{\left(1+n\hat{\sigma}^2(t)\right)}{\sqrt{n}\left|\ln\left(\hat{S}(t)\right)\right|}, \tag{2.9}$$

where

$$\hat{\sigma}^2(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)},$$

the estimator of the variance of the log of the Kaplan–Meier estimator from (2.3), and $H_{\hat{a},\alpha}$ is a percentile from Appendix 3, where

$$\hat{a} = n\hat{\sigma}^2(t_{(m)}) \Big/ \left[1+n\hat{\sigma}^2(t_{(m)})\right].$$

If we denote the lower and upper endpoints of this confidence band as $\hat{b}_l$ and $\hat{b}_u$, then the lower and upper endpoints of the confidence band for the survivorship function are

$$\exp\left[-\exp\left(\hat{b}_u\right)\right] \text{ and } \exp\left[-\exp\left(\hat{b}_l\right)\right]. \tag{2.10}$$

To obtain the bands for the survivorship function from the HMO-HIV+ study, we note that the largest observed survival time is 58 months and $\hat{\sigma}^2(58) = 0.423$. Most software packages will provide either the values of the estimated variance of the log of the Kaplan–Meier estimator

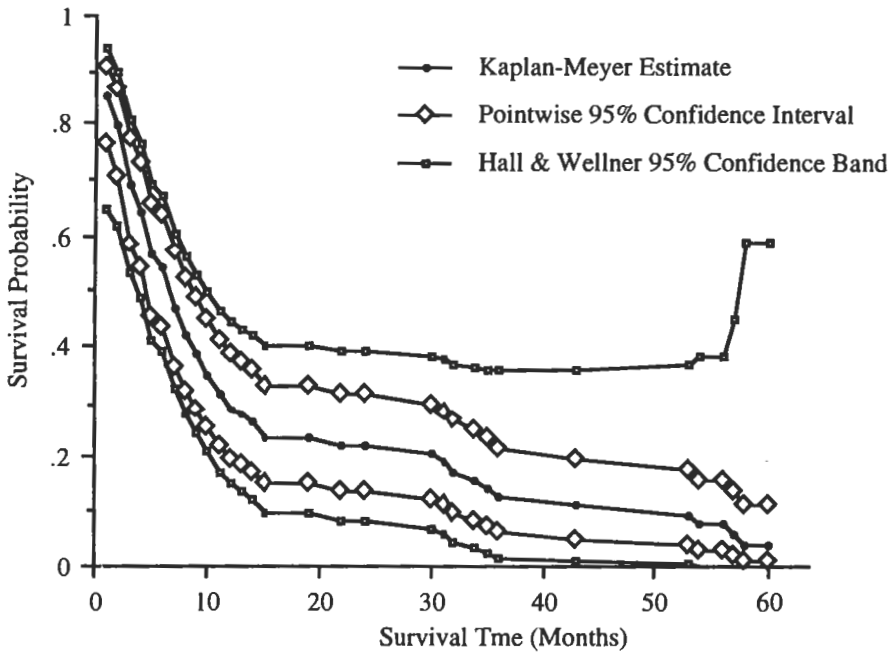

Figure 2.5 Kaplan–Meier estimate, pointwise 95% confidence intervals, and Hall and Wellner 95% confidence bands for the survivorship function for the HMO–HIV+ study.

or those of the Greenwood estimator of the variance of the survivorship function. The values of $\hat{\sigma}^2(t)$ are easily obtained by dividing the Greenwood estimator by the square of the Kaplan–Meier estimator. To obtain the percentile from Appendix 3 we compute

$$\hat{a} = (100 \times 0.423)/(1 + 100 \times 0.423) = 0.98$$

and note that, since both $H_{0.9,0.95}$ and $H_{1.0,0.95}$ equal 1.358, linear interpolation of tabled values is not necessary and we use 1.358. In cases when $\hat{a} < 0.9$, linear interpolation between two tabled values may be required to obtain the most accurate value. To obtain the confidence bands, we compute the endpoints in (2.9) and (2.10) for each observed value of time. We can ignore the censoring since the estimated survivorship function and its variance are constant between observed failure times. These endpoints may be plotted, along with the estimated survivorship function, restricting the plot to the interval [0,58]. This plot is

also shown in Figure 2.5. The increased width of the confidence bands relative to the pointwise confidence intervals is seen in this figure. The increased width is needed to assure that the probability is 95 percent that each of the individual 95 percent confidence interval estimates simultaneously covers its respective parameter. In particular, we note the lack of precision in the band for times near the maximum of 58 months. The bands do support the observation of many early deaths and a few at or near the maximum follow-up time of 60 months.

The estimated survivorship function and its confidence intervals and/or bands provide a useful descriptive measure of the overall pattern of survival times. However, it is often useful to supplement the presentation with point and interval estimates of key quantiles. The estimated survivorship function may be used to estimate quantiles of the survival time distribution in the same way that the estimated cumulative distribution of, say, height or weight may be used to estimate quantiles of its distribution. This may be done graphically and the graphical procedure can be codified into a formula for analytic calculations based on the tabular form of the estimate.

The quantiles most frequently reported by software packages are the three quartile boundaries of the survival time distribution. To obtain graphical estimates, begin on the percent survival (y) axis at the quartile of interest and draw a horizontal line until it first touches the estimated survivorship function. A vertical line is drawn down to the time axis to obtain the estimated quartile. In order for the estimate to be finite, the horizontal line must hit the survivorship function. Thus, the minimum possible estimated quantile which has a finite value is the observed minimum of the survivorship function, and only quantiles within the observed range of the estimated survivorship function may be estimated. For example, if the range was from 1.0 to 0.38 then we could estimate the 75th and 50th percentiles but not the 25th percentile. Graphically determined estimates of the three quartile boundaries, denoted $\hat{t}_{75}$, $\hat{t}_{50}$ and $\hat{t}_{25}$, based on the Kaplan–Meier estimate of the survivorship function for the data in Table 2.1 are shown in Figure 2.6.

The graphical method is easy to use, but it is not especially precise. The method may be described in a formula, from which a more accurate numerical value may be determined from a tabular presentation of the estimated survivorship function. We illustrate the method by estimating the median or second quartile, $\hat{t}_{50}$, and we then generalize it into a formula that may be used for any quantile. By referring to Table 2.2, and Figure 2.6 we see that the horizontal line hits the survivorship funct-