used, as in Example 2.2, but the grouping of the survival times does result in some loss of information. This is particularly so when the number of patients is small, less than about 30, say.

2.1.2 *Kaplan-Meier estimate of the survivor function*

The first step in the analysis of ungrouped censored survival data is normally to obtain the *Kaplan-Meier estimate* of the survivor function. This estimate is therefore considered in some detail. To obtain the Kaplan-Meier estimate, a series of time intervals is constructed, as for the life-table estimate. However, each of these intervals is designed to be such that one death time is contained in the interval, and this death time is taken to occur at the start of the interval.

As an illustration, suppose that $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$ are three observed survival times arranged in rank order, so that $t_{(1)} < t_{(2)} < t_{(3)}$, and that c is a censored survival time that falls between $t_{(2)}$ and $t_{(3)}$. The constructed intervals then begin at times $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$, and each interval includes the one death time, although there could be more than one individual who dies at any particular death time. Notice that no interval begins at the censored time of c. The situation is illustrated diagrammatically in Figure 2.3, in which D represents a death and C a censored survival time. Notice that two individuals die at $t_{(1)}$, one dies at $t_{(2)}$, and three die at $t_{(3)}$.

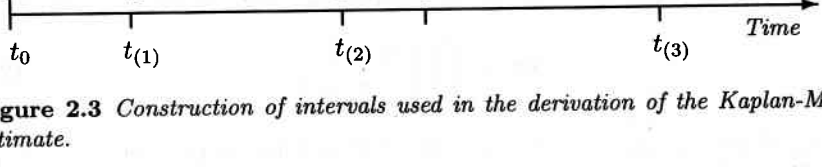

Figure 2.3 *Construction of intervals used in the derivation of the Kaplan-Meier estimate.*

The time origin is denoted by t_0, and so there is an initial period commencing at t_0, which ends just before $t_{(1)}$, the time of the first death. This means that the interval from t_0 to $t_{(1)}$ will not include a death time. The first constructed interval extends from $t_{(1)}$ to just before $t_{(2)}$, and since the second death time is at $t_{(2)}$, this interval includes the single death time at $t_{(1)}$. The second interval begins at time $t_{(2)}$ and ends just before $t_{(3)}$, and includes the death time at $t_{(2)}$ and the censored time c. There is also a third interval beginning at $t_{(3)}$, which contains the longest survival time, $t_{(3)}$.

In general, suppose that there are n individuals with observed survival times $t_1, t_2, \ldots, t_n$. Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time. We therefore suppose that there are r death times amongst the individuals, where $r \leqslant n$. After arranging these death times in ascending order, the jth is denoted $t_{(j)}$, for $j = 1, 2, \ldots, r$, and so the r ordered death times are $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. The number of individuals who are alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted n_j, for $j = 1, 2, \ldots, r$ and d_j will denote the number who die at this time. The time interval from

$t_{(j)} - \delta$ to $t_{(j)}$, where δ is an infinitesimal time interval, then includes one death time. Since there are n_j individuals who are alive just before $t_{(j)}$ and d_j deaths at $t_{(j)}$, the probability that an individual dies during the interval from $t_{(j)} - \delta$ to $t_{(j)}$ is estimated by d_j/n_j. The corresponding estimated probability of survival through that interval is then $(n_j - d_j)/n_j$.

It sometimes happens that there are censored survival times that occur at the same time as one or more deaths, so that a death time and a censored survival time appear to occur simultaneously. In this event, the censored survival time is taken to occur immediately after the death time when computing the values of the n_j.

From the manner in which the time intervals are constructed, the interval from $t_{(j)}$ to $t_{(j+1)} - \delta$, the time immediately before the next death time, contains no deaths. The probability of surviving from $t_{(j)}$ to $t_{(j+1)} - \delta$ is therefore unity, and the joint probability of surviving from $t_{(j)} - \delta$ to $t_{(j)}$ and from $t_{(j)}$ to $t_{(j+1)} - \delta$ can be estimated by $(n_j - d_j)/n_j$. In the limit, as δ tends to zero, $(n_j - d_j)/n_j$ becomes an estimate of the probability of surviving the interval from $t_{(j)}$ to $t_{(j+1)}$.

We now make the assumption that the deaths of the individuals in the sample occur independently of one another. Then, the estimated survivor function at any time, t, in the kth constructed time interval from $t_{(k)}$ to $t_{(k+1)}$, $k = 1, 2, \ldots, r$, where $t_{(r+1)}$ is defined to be ∞, will be the estimated probability of surviving beyond $t_{(k)}$. This is actually the probability of surviving through the interval from $t_{(k)}$ to $t_{(k+1)}$, and all preceding intervals, and leads to the Kaplan-Meier estimate of the survivor function, which is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{2.4}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, with $\hat{S}(t) = 1$ for $t < t_{(1)}$, and where $t_{(r+1)}$ is taken to be ∞. Strictly speaking, if the largest observation is a censored survival time, t^*, say, $\hat{S}(t)$ is undefined for $t > t^*$. On the other hand, if the largest observed survival time, $t_{(r)}$, is an uncensored observation, $n_r = d_r$, and so $\hat{S}(t)$ is zero for $t \geqslant t_{(r)}$. A plot of the Kaplan-Meier estimate of the survivor function is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death time.

Equation (2.4) shows that, as for the life-table estimate of the survivor function in equation (2.3), the Kaplan-Meier estimate is formed as a product of a series of estimated probabilities. In fact, the Kaplan-Meier estimate is the limiting value of the life-table estimate in equation (2.3) as the number of intervals tends to infinity and their width tends to zero. For this reason, the Kaplan-Meier estimate is also known as the *product-limit estimate* of the survivor function.

Note that if there are no censored survival times in the data set, $n_j - d_j = n_{j+1}$, $j = 1, 2, \ldots, k$, in equation (2.4), and on expanding the product we get

$$\hat{S}(t) = \frac{n_2}{n_1} \times \frac{n_3}{n_2} \times \cdots \times \frac{n_{k+1}}{n_k}. \tag{2.5}$$

This reduces to n_{k+1}/n_1, for $k = 1, 2, \ldots, r-1$, with $\hat{S}(t) = 1$ for $t < t_{(1)}$ and $\hat{S}(t) = 0$ for $t \geqslant t_{(r)}$. Now, n_1 is the number of individuals at risk just before the first death time, which is the number of individuals in the sample, and n_{k+1} is the number of individuals with survival times greater than or equal to $t_{(k+1)}$. Consequently, in the absence of censoring, $\hat{S}(t)$ is simply the empirical survivor function defined in equation (2.1). The Kaplan-Meier estimate is therefore a generalisation of the empirical survivor function that accommodates censored observations.

Example 2.3 Time to discontinuation of the use of an IUD

Data from 18 women on the time to discontinuation of the use of an IUD were given in Table 1.1. For these data, the survivor function, $S(t)$, represents the probability that a woman discontinues the use of the contraceptive device after any time t. The Kaplan-Meier estimate of the survivor function is readily obtained using equation (2.4), and the required calculations are set out in Table 2.2. The estimated survivor function, $\hat{S}(t)$, is plotted in Figure 2.4.

Table 2.2 *Kaplan-Meier estimate of the survivor function for the data from Example 1.1.*

Time interval	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0–	18	0	1.0000	1.0000
10–	18	1	0.9444	0.9444
19–	15	1	0.9333	0.8815
30–	13	1	0.9231	0.8137
36–	12	1	0.9167	0.7459
59–	8	1	0.8750	0.6526
75–	7	1	0.8571	0.5594
93–	6	1	0.8333	0.4662
97–	5	1	0.8000	0.3729
107	3	1	0.6667	0.2486

Note that since the largest discontinuation time of 107 days is censored, $\hat{S}(t)$ is not defined beyond $t = 107$.

2.1.3 Nelson-Aalen estimate of the survivor function

An alternative estimate of the survivor function, which is based on the individual event times, is the *Nelson-Aalen estimate*, given by

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp(-d_j/n_j). \tag{2.6}$$

This estimate can be obtained from an estimate of the cumulative hazard function, as shown in Section 2.3.3. Moreover, the Kaplan-Meier estimate of the survivor function can be regarded as an approximation to the Nelson-

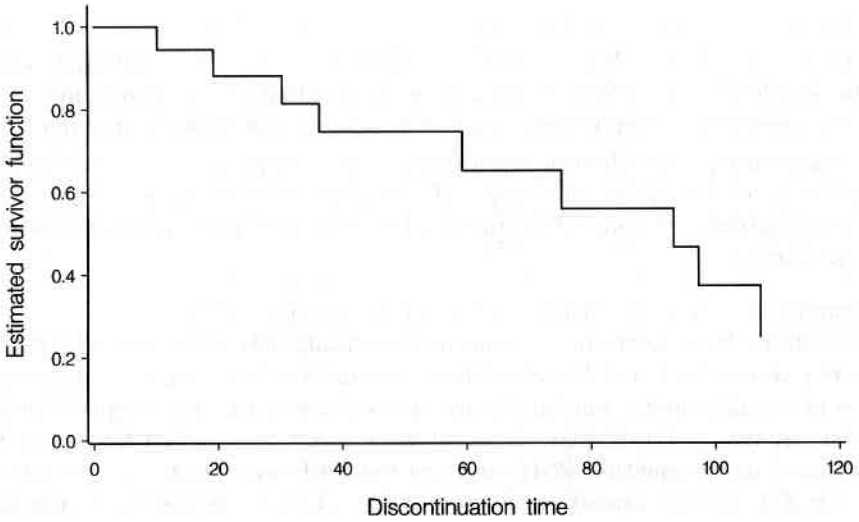

Figure 2.4 *Kaplan-Meier estimate of the survivor function for the data from Example 1.1.*

Aalen estimate. To show this, we use the result that

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots,$$

which is approximately equal to $1 - x$ when x is small. It then follows that $\exp(-d_j/n_j) \approx 1-(d_j/n_j) = (n_j-d_j)/n_j$, so long as d_j is small relative to n_j, which it will be except at the latest survival times. Consequently, the Kaplan-Meier estimate, $\hat{S}(t)$, in equation (2.4), approximates the Nelson-Aalen estimate, $\tilde{S}(t)$, in equation (2.6).

The Nelson-Aalen estimate of the survivor function, also known as *Altshuler's estimate*, will always be greater than the Kaplan-Meier estimate at any given time, since $e^{-x} \geqslant 1 - x$, for all values of x. Although the Nelson-Aalen estimate has been shown to perform better than the Kaplan-Meier estimate in small samples, in many circumstances, the estimates will be very similar, particularly at the earlier survival times. Since the Kaplan-Meier estimate is a generalisation of the empirical survivor function, the latter estimate has much to commend it.

Example 2.4 Time to discontinuation of the use of an IUD

The values shown in Table 2.2, which gives the Kaplan-Meier estimate of the survivor function for the data on the time to discontinuation of the use of an IUD, can be used to calculate the Nelson-Aalen estimate. This estimate is shown in Table 2.3.

From this table we see that the Kaplan-Meier and Nelson-Aalen estimates of the survivor function differ by less than 0.04. However, when we consider

Table 2.3 *Nelson-Aalen estimate of the survivor function for the data from Example 1.1.*

Time interval	$\exp(-d_j/n_j)$	$\tilde{S}(t)$
0–	1.0000	1.0000
10–	0.9460	0.9460
19–	0.9355	0.8850
30–	0.9260	0.8194
36–	0.9200	0.7539
59–	0.8825	0.6653
75–	0.8669	0.5768
93–	0.8465	0.4882
97–	0.8187	0.3997
107	0.7165	0.2864

the precision of these estimates, which we do in Section 2.2, we see that a difference of 0.04 is of no practical importance.

2.2 Standard error of the estimated survivor function

An essential aid to the interpretation of an estimate of any quantity is the precision of the estimate, which is reflected in the *standard error* of the estimate. This is defined to be the square root of the estimated variance of the estimate, and is used in the construction of an interval estimate for a quantity of interest. In this section, the standard error of estimates of the survivor function are given.

Because the Kaplan-Meier estimate is the most important and widely used estimate of the survivor function, the derivation of the standard error of $\hat{S}(t)$ will be presented in detail in this section. The details of this derivation can be omitted on a first reading.

2.2.1 Standard error of the Kaplan-Meier estimate*

The Kaplan-Meier estimate of the survivor function for any value of t in the interval from $t_{(k)}$ to $t_{(k+1)}$ can be written as

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j,$$

for $k = 1, 2, \ldots, r$, where $\hat{p}_j = (n_j - d_j)/n_j$ is the estimated probability that an individual survives through the time interval that begins at $t_{(j)}$, $j = 1, 2, \ldots, r$. Taking logarithms,

and so the variance of $\log \hat{S}(t)$ is given by

$$\text{var}\left\{\log \hat{S}(t)\right\} = \sum_{j=1}^{k} \text{var}\left\{\log \hat{p}_j\right\}. \qquad (2.7$$

Now, the number of individuals who survive through the interval beginning at $t_{(j)}$ can be assumed to have a *binomial distribution* with parameters n_j and p_j, where p_j is the true probability of survival through that interval. The observed number who survive is $n_j - d_j$, and using the result that the variance of a binomial random variable with parameters n, p is $np(1-p)$, the variance of $n_j - d_j$ is given by

$$\text{var}\,(n_j - d_j) = n_j p_j (1 - p_j).$$

Since $\hat{p}_j = (n_j - d_j)/n_j$, the variance of $\hat{p}_j$ is $\text{var}\,(n_j - d_j)/n_j^2$, that is, $p_j(1 - p_j)/n_j$. The variance of $\hat{p}_j$ may then be estimated by

$$\hat{p}_j(1-\hat{p}_j)/n_j. \tag{2.8}$$

In order to obtain the variance of $\log \hat{p}_j$, we make use of a general result for the approximate variance of a function of a random variable. According to this result, the variance of a function $g(X)$ of the random variable X is given by

$$\text{var}\,\{g(X)\} \approx \left\{\frac{\mathrm{d}g(X)}{\mathrm{d}X}\right\}^2 \text{var}\,(X). \tag{2.9}$$

This is known as the *Taylor series approximation* to the variance of a function of a random variable. Using equation (2.9), the approximate variance of $\log \hat{p}_j$ is $\text{var}\,(\hat{p}_j)/\hat{p}_j^2$, and using expression (2.8), the approximate estimated variance of $\log \hat{p}_j$ is $(1-\hat{p}_j)/(n_j\hat{p}_j)$, which on substitution for $\hat{p}_j$, reduces to

$$\frac{d_j}{n_j(n_j-d_j)}. \tag{2.10}$$

From equation (2.7),

$$\text{var}\left\{\log \hat{S}(t)\right\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}, \tag{2.11}$$

and a further application of the result in equation (2.9) gives

$$\text{var}\left\{\log \hat{S}(t)\right\} \approx \frac{1}{[\hat{S}(t)]^2}\,\text{var}\left\{\hat{S}(t)\right\},$$

$$\text{se}\{S^*(t)\} \approx S^*(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n'_j(n'_j-d_j)}\right\}^{\frac{1}{2}}, \tag{2.14}$$

in the notation of Section 2.1.1.

The standard error of the Nelson-Aalen estimator is

$$\text{se}\left\{\tilde{S}(t)\right\} \approx \tilde{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n_j^2}\right\}^{\frac{1}{2}}, \tag{2.15}$$

although other expressions have been proposed.

2.2.3 *Confidence intervals for values of the survivor function*

Once the standard error of an estimate of the survivor function has been calculated, a *confidence interval* for the corresponding value of the survivor function, at a given time t, can be found. A confidence interval is an interval estimate of the survivor function, and is the interval which is such that there is a prescribed probability that the value of the true survivor function is included

is given by

$$\text{se}\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}\right\}^{\frac{1}{2}}, \tag{2.13}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$. This result is known as *Greenwood's formula.*

If there are no censored survival times, $n_j - d_j = n_{j+1}$, and expression (2.10) becomes $(n_j - n_{j+1})/n_j n_{j+1}$. Now,

$$\sum_{j=1}^{k}\frac{n_j - n_{j+1}}{n_j n_{j+1}} = \sum_{j=1}^{k}\left(\frac{1}{n_{j+1}} - \frac{1}{n_j}\right) = \frac{n_1 - n_{k+1}}{n_1 n_{k+1}},$$

which can be written as

$$\frac{1-\hat{S}(t)}{n_1\hat{S}(t)},$$

since $\hat{S}(t) = n_{k+1}/n_1$ for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$, in the absence of censoring. Hence, from equation (2.12), the estimated variance of $\hat{S}(t)$ is $\hat{S}(t)[1-\hat{S}(t)]/n_1$. This is an estimate of the variance of the empirical survivor function, given in equation (2.1), on the assumption that the number of individuals at risk at time t has a binomial distribution with parameters $n_1, S(t)$.

2.2.2* *Standard error of the life-table and Nelson-Aalen estimates*

The life-table estimate of the survivor function is similar in form to the Kaplan-Meier estimate, and so the standard error of this estimator is obtained in a similar manner. The standard error of the life-table estimate is given by

within it. The intervals constructed in this manner are sometimes referred to as *pointwise confidence intervals*, since they apply to a specific survival time.

A confidence interval for the true value of the survivor function at a given time t is obtained by assuming that the estimated value of the survivor function at t is normally distributed with mean $S(t)$ and estimated variance given by equation (2.12). The interval is computed from *percentage points* of the standard normal distribution. Thus, if Z is a random variable that has a standard normal distribution, the upper (one-sided) $\alpha/2$-point, or the two-sided α-point, of this distribution is that value $z_{\alpha/2}$ which is such that $P(Z > z_{\alpha/2}) = \alpha/2$. This probability is the area under the standard normal curve to the right of $z_{\alpha/2}$, as illustrated in Figure 2.5. For example, the two-sided 5% and 1% points of the standard normal distribution, $z_{0.025}$ and $z_{0.005}$ are 1.96 and 2.58, respectively.

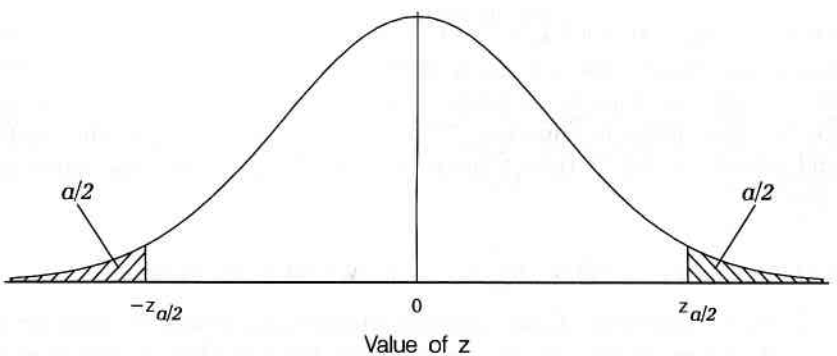

Figure 2.5 *Upper and lower $\alpha/2$-points of the standard normal distribution.*

A $100(1-\alpha)\%$ confidence interval for $S(t)$, for a given value of t, is the interval from $\hat{S}(t) - z_{\alpha/2}\,\mathrm{se}\{\hat{S}(t)\}$ to $\hat{S}(t) + z_{\alpha/2}\,\mathrm{se}\{\hat{S}(t)\}$, where $\mathrm{se}\{\hat{S}(t)\}$ is found from equation (2.13). These intervals for $S(t)$ can be superimposed on a graph of the estimated survivor function, as shown in Example 2.5.

One difficulty with this procedure arises from the fact that the confidence intervals are symmetric. When the estimated survivor function is close to zero or unity, symmetric intervals are inappropriate, since they can lead to confidence limits for the survivor function that lie outside the interval (0,1). A pragmatic solution to this problem is to replace any limit that is greater than unity by 1.0, and any limit that is less than zero by 0.0.

An alternative procedure is to transform $\hat{S}(t)$ to a value in the range $(-\infty, \infty)$, and obtain a confidence interval for the transformed value. The resulting confidence limits are then back-transformed to give a confidence interval for $S(t)$ itself. Possible transformations are the logistic transformation, $\log[S(t)/\{1 - S(t)\}]$, and the complementary log-log transformation $\log\{-\log S(t)\}$. Note that from equation (1.7), the latter quantity is the logarithm of the cumulative hazard function. In either case, the standard

error of the transformed value of $\hat{S}(t)$ can be found using the approximation in equation (2.9).

For example, the variance of $\log\{-\log \hat{S}(t)\}$ is obtained from the expression for var $\{\log \hat{S}(t)\}$ in equation (2.11). Using the general result in equation (2.9),

$$\text{var}\,\{\log(-X)\} \approx \frac{1}{X^2}\,\text{var}\,(X),$$

and setting $X = \log \hat{S}(t)$ gives

$$\text{var}\left[\log\{-\log \hat{S}(t)\}\right] \approx \frac{1}{\{\log \hat{S}(t)\}^2} \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$

The standard error of $\log\{-\log \hat{S}(t)\}$ is the square root of this quantity. This leads to $100(1-\alpha)\%$ limits of the form

$$\hat{S}(t)^{\exp[\pm z_{\alpha/2}\,\text{se}\{\log[-\log \hat{S}(t)]\}]},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution.

A further problem is that in the tails of the distribution of the survival times, that is, when $\hat{S}(t)$ is close to zero or unity, the variance of $\hat{S}(t)$ obtained using Greenwood's formula can underestimate the actual variance. In these circumstances, an alternative expression for the standard error of $\hat{S}(t)$ may be used. Peto *et al.* (1977) propose that the standard error of $\hat{S}(t)$ should be obtained from the equation

$$\text{se}\,\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{\{1 - \hat{S}(t)\}}}{\sqrt{(n_k)}},$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, where $\hat{S}(t)$ is the Kaplan-Meier estimate of $S(t)$ and n_k is the number of individuals at risk at $t_{(k)}$, the start of the kth constructed time interval.

This expression for the standard error of $\hat{S}(t)$ is conservative, in the sense that the standard errors obtained will tend to be larger than they ought to be. For this reason, the Greenwood estimate is recommended for general use.

Example 2.5 Time to discontinuation of the use of an IUD

The standard error of the estimated survivor function, and 95% confidence limits for the corresponding true value of the function, for the data from Example 1.1 on the times to discontinuation of use of an IUD, are given in Table 2.4. In this table, confidence limits outside the range $(0, 1)$ have been replaced by zero or unity.

From this table we see that in general the standard error of the estimated survivor function increases with the discontinuation time. The reason for this is that estimates of the survivor function at later times are based on fewer individuals. A graph of the estimated survivor function, with the 95% confidence limits shown as dashed lines, is given in Figure 2.6.

It is important to observe that the confidence limits plotted on such a graph are only valid for any given time. Different methods are needed to

Table 2.4 *Standard error of $\hat{S}(t)$ and confidence intervals for $S(t)$ for the data from Example 1.1.*

Time interval	$\hat{S}(t)$	se $\{\hat{S}(t)\}$	95% confidence interval
0–	1.0000	0.0000	
10–	0.9444	0.0540	(0.839, 1.000)
19–	0.8815	0.0790	(0.727, 1.000)
30–	0.8137	0.0978	(0.622, 1.000)
36–	0.7459	0.1107	(0.529, 0.963)
59–	0.6526	0.1303	(0.397, 0.908)
75–	0.5594	0.1412	(0.283, 0.836)
93–	0.4662	0.1452	(0.182, 0.751)
97–	0.3729	0.1430	(0.093, 0.653)
107	0.2486	0.1392	(0.000, 0.522)

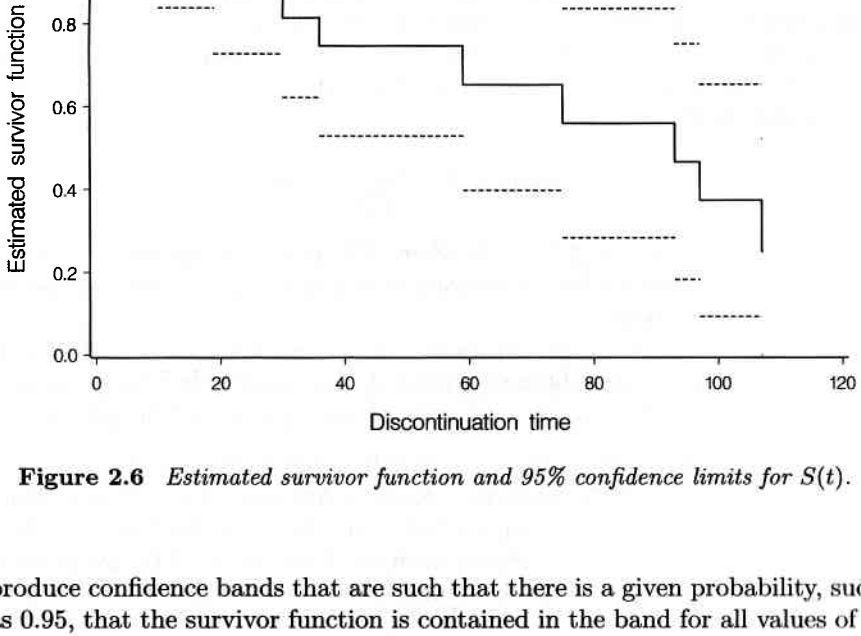

Figure 2.6 *Estimated survivor function and 95% confidence limits for $S(t)$.*

produce confidence bands that are such that there is a given probability, such as 0.95, that the survivor function is contained in the band for all values of t. These bands will tend to be wider than the band formed from the pointwise confidence limits. Details will not be included, but references to these methods are given in the final section of this chapter. Notice also that the width of these intervals is very much greater than the difference between the Kaplan-Meier and Nelson-Aalen estimates of the survivor function, shown in Tables 2.2 and 2.3. Similar calculations lead to confidence limits based on life-table and Nelson-Aalen estimates of the survivor function.