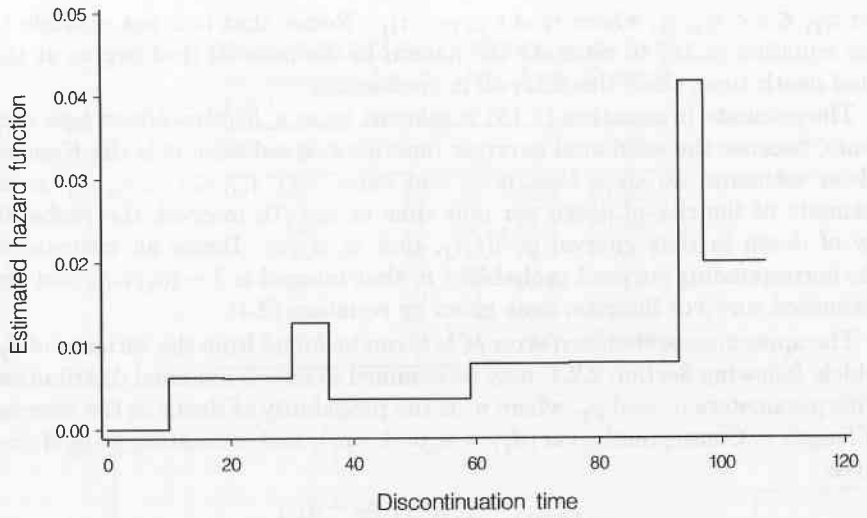

Figure 2.8 *Kaplan-Meier type estimate of the hazard function for the data from Example 1.1.*

errors of the estimated hazard function at different times are of little help in interpreting this plot.

In practice, estimates of the hazard function obtained in this way will often tend to be rather irregular. For this reason, plots of the hazard function may be "smoothed", so that any pattern can be seen more clearly. There are a number of ways of smoothing the hazard function, that lead to a weighted average of values of the estimated hazard $\hat{h}(t)$ at death times in the neighbourhood of t. For example, a *kernel smoothed* estimate of the hazard function, based on the r ordered death times, $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, with d_j deaths and n_j at risk at time $t_{(j)}$, can be found from

$$h^{\dagger}(t) = b^{-1} \sum_{j=1}^{r} 0.75 \left\{ 1 - \left(\frac{t - t_{(j)}}{b} \right)^2 \right\} \frac{d_j}{n_j},$$

where the value of b needs to be chosen. The function $h^{\dagger}(t)$ is defined for all values of t in the interval from b to $t_{(r)} - b$, where $t_{(r)}$ is the greatest death time. For any value of t in this interval, the death times in the interval $(t-b, t+b)$ will contribute to the weighted average. The parameter b is known as the *bandwidth* and its value controls the shape of the plot; the larger the value of b, the greater the degree of smoothing. There are formulae that lead to "optimal" values of b, but these tend to be rather cumbersome. Fuller details can be found in the references provided in the final section of this chapter. In this book, the use of a modelling approach to the analysis of survival data is advocated, and so model-based estimates of the hazard function will be considered in subsequent chapters.

2.3.3 Estimating the cumulative hazard function

The cumulative hazard function is important in the identification of models for survival data, as will be seen later in Sections 4.4 and 5.2. In addition, since the derivative of the cumulative hazard function is the hazard function itself, the slope of the cumulative hazard function provides information about the shape of the underlying hazard function. In particular, a linear cumulative hazard function over some time interval suggests that the hazard is constant over this interval. Accordingly, methods that can be used to estimate this function will now be described.

The cumulative hazard at time t, $H(t)$, was defined in equation (1.6) to be the integral of the hazard function, but is more conveniently found using equation (1.7). According to this result, $H(t) = -\log S(t)$, and so if $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function, $\hat{H}(t) = -\log \hat{S}(t)$ is an appropriate estimate of the cumulative hazard to time t.

Now, using equation (2.4),

$$\hat{H}(t) = -\sum_{j=1}^{k} \log \left(\frac{n_j - d_j}{n_j} \right),$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, and $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ are the r ordered death times, with $t_{(r+1)} = \infty$.

If the Nelson-Aalen estimate of the survivor function is used, the estimated cumulative hazard function, $\tilde{H}(t) = -\log \tilde{S}(t)$, is given by

$$\tilde{H}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j}.$$

This is the cumulative sum of the estimated probabilities of death from the first to the kth time interval, $k = 1, 2, \ldots, r$. This quantity therefore has immediate intuitive appeal as an estimate of the cumulative hazard.

An estimate of the cumulative hazard function also leads to an estimate of the corresponding hazard function, since the differences between adjacent values of the estimated cumulative hazard function provide estimates of the underlying hazard, after dividing by the time interval. In particular, differences in adjacent values of the Nelson-Aalen estimate of the cumulative hazard lead directly to the hazard function estimate in Section 2.3.2.

2.4 Estimating the median and percentiles of survival times

Since the distribution of survival times tends to be positively skew, the median is the preferred summary measure of the location of the distribution. Once the survivor function has been estimated, it is straightforward to obtain an estimate of the *median survival time*. This is the time beyond which 50% of the individuals in the population under study are expected to survive, and is given by that value $t(50)$ which is such that $S\{t(50)\} = 0.5$.

Because the non-parametric estimates of $S(t)$ are step-functions, it will

not usually be possible to realise an estimated survival time that makes the survivor function exactly equal to 0.5. Instead, the estimated median survival time, $\hat{t}(50)$, is defined to be the smallest observed survival time for which the value of the estimated survivor function is less than 0.5.

In mathematical terms,

$$\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) < 0.5\},$$

where t_i is the observed survival time for the ith individual, $i = 1, 2, \ldots, n$. Since the estimated survivor function only changes at a death time, this is equivalent to the definition

$$\hat{t}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) < 0.5\},$$

where $t_{(j)}$ is the jth ordered death time, $j = 1, 2, \ldots, r$.

In the particular case where the estimated survivor function is exactly equal to 0.5 for values of t in the interval from $t_{(j)}$ to $t_{(j+1)}$, the median is taken to be the half-way point in this interval, that is $(t_{(j)} + t_{(j+1)})/2$.

In the situation where there are no censored survival times, the estimated median survival time will be the smallest time beyond which 50% of the individuals in the sample survive.

Example 2.8 Time to discontinuation of the use of an IUD
The Kaplan-Meier estimate of the survivor function for the data from Example 1.1 on the time to discontinuation of the use of an IUD was given in Table 2.2. The estimated survivor function, $\hat{S}(t)$, for these data was shown in Figure 2.4. From the estimated survivor function, the smallest discontinuation time beyond which the estimated probability of discontinuation is less than 0.5 is 93 weeks. This is therefore the estimated median time to discontinuation of the IUD for this group of women.

A similar procedure to that described above can be used to estimate other *percentiles* of the distribution of survival times. The pth percentile of the distribution of survival times is defined to be the value $t(p)$ which is such that $F\{t(p)\} = p/100$. In terms of the survivor function, $t(p)$ is such that $S\{t(p)\} = 1 - (p/100)$, so that for example the 10th and 90th percentiles are given by

$$S\{t(10)\} = 0.9, \quad S\{t(90)\} = 0.1,$$

respectively. Using the estimated survivor function, the estimated pth percentile is the smallest observed survival time, $\hat{t}(p)$, for which $\hat{S}\{\hat{t}(p)\} < 1 - (p/100)$.

It sometimes happens that the estimated survivor function is greater than 0.5 for all values of t. In such cases, the median survival time cannot be estimated. It would then be natural to summarise the data in terms of other percentiles of the distribution of survival times, or the estimated survival probabilities at particular time points.

Estimates of the dispersion of a sample of survival data are not widely used, but should such an estimate be required, the *semi-interquartile range* (SIQR)

can be calculated. This is defined to be half the difference between the 75th and 25th percentiles of the distribution of survival times. Hence,

$$\text{SIQR} = \frac{1}{2}\{t(75) - t(25)\},$$

where $t(25)$ and $t(75)$ are the 25th and 75th percentiles of the survival time distribution. These two percentiles are also known as the *first* and *third quartiles*, respectively. The corresponding sample-based estimate of the SIQR is $\{\hat{t}(75) - \hat{t}(25)\}/2$. Like the variance, the larger the value of the SIQR, the more dispersed is the survival time distribution.

Example 2.9 Time to discontinuation of the use of an IUD
From the Kaplan-Meier estimate of the survivor function for the data from Example 1.1, given in Table 2.2, the 25th and 75th percentiles of the distribution of discontinuation times are 36 and 107 weeks, respectively. Hence, the SIQR of the distribution is estimated to be 35.5 weeks.

2.5* Confidence intervals for the median and percentiles

Approximate confidence intervals for the median and other percentiles of a distribution of survival times can be found once the variance of the estimated percentile has been obtained. An expression for the approximate variance of a percentile can be derived from a direct application of the general result for the variance of a function of a random variable in equation (2.9). Using this result,

$$\text{var}\,[\hat{S}\{t(p)\}] = \left(\frac{\mathrm{d}\hat{S}\{t(p)\}}{\mathrm{d}t(p)}\right)^2 \text{var}\,\{t(p)\}, \tag{2.17}$$

where $t(p)$ is the pth percentile of the distribution and $\hat{S}\{t(p)\}$ is the Kaplan-Meier estimate of the survivor function at $t(p)$. Now,

$$-\frac{\mathrm{d}\hat{S}\{t(p)\}}{\mathrm{d}t(p)} = \hat{f}\{t(p)\},$$

an estimate of the probability density function of the survival times at $t(p)$, and on rearranging equation (2.17), we get

$$\text{var}\,\{t(p)\} = \left(\frac{1}{\hat{f}\{t(p)\}}\right)^2 \text{var}\,[\hat{S}\{t(p)\}].$$

The standard error of $\hat{t}(p)$, the estimated pth percentile, is therefore given by

$$\text{se}\,\{\hat{t}(p)\} = \frac{1}{\hat{f}\{\hat{t}(p)\}}\,\text{se}\,[\hat{S}\{\hat{t}(p)\}]. \tag{2.18}$$

The standard error of $\hat{S}\{\hat{t}(p)\}$ is found using Greenwood's formula for the standard error of the Kaplan-Meier estimate of the survivor function, given in equation (2.13), while an estimate of the probability density function at $\hat{t}(p)$

is

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{S}\{\hat{u}(p)\} - \hat{S}\{\hat{l}(p)\}}{\hat{l}(p) - \hat{u}(p)},$$

where

$$\hat{u}(p) = \max\left\{t_{(j)} \mid \hat{S}(t_{(j)}) \geqslant 1 - \frac{p}{100} + \epsilon\right\},$$

and

$$\hat{l}(p) = \min\left\{t_{(j)} \mid \hat{S}(t_{(j)}) \leqslant 1 - \frac{p}{100} - \epsilon\right\},$$

for $j = 1, 2, \ldots, r$, and small values of ϵ. In many cases, taking $\epsilon = 0.05$ will be satisfactory, but a larger value of ϵ will be needed if $\hat{u}(p)$ and $\hat{l}(p)$ turn out to be equal. In particular, from equation (2.18), the standard error of the median survival time is given by

$$\text{se}\{\hat{t}(50)\} = \frac{1}{\hat{f}\{\hat{t}(50)\}}\,\text{se}[\hat{S}\{\hat{t}(50)\}], \tag{2.19}$$

where $\hat{f}\{\hat{t}(50)\}$ can be found from

$$\hat{f}\{\hat{t}(50)\} = \frac{\hat{S}\{\hat{u}(50)\} - \hat{S}\{\hat{l}(50)\}}{\hat{l}(50) - \hat{u}(50)}. \tag{2.20}$$

In this expression, $\hat{u}(50)$ is the largest survival time for which the Kaplan-Meier estimate of the survivor function exceeds 0.55, and $\hat{l}(50)$ is the smallest survival time for which the survivor function is less than or equal to 0.45.

Once the standard error of the estimated pth percentile has been found, a $100(1-\alpha)\%$ confidence interval for $t(p)$ has limits of

$$\hat{t}(p) \pm z_{\alpha/2}\,\text{se}\{\hat{t}(p)\},$$

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

This interval estimate is only approximate, in the sense that the probability that the interval includes the true percentile will not be exactly $1-\alpha$. A number of methods have been proposed for constructing confidence intervals for the median with superior properties, although these alternatives are more difficult to compute than the interval estimate derived in this section.

Example 2.10 Time to discontinuation of the use of an IUD

The data on the discontinuation times for users of an IUD, given in Example 1.1, is now used to illustrate the calculation of a confidence interval for the median discontinuation time. From Example 2.8, the estimated median discontinuation time for this group of women is given by $\hat{t}(50) = 93$ weeks. Also, from Table 2.4, the standard error of the Kaplan-Meier estimate of the survivor function at this time is given by $\text{se}[\hat{S}\{\hat{t}(50)\}] = 0.1452$.

To obtain the standard error of $\hat{t}(50)$ using equation (2.19), we need an estimate of the density function at the estimated median discontinuation time. This is obtained from equation (2.20). The quantities $\hat{u}(50)$ and $\hat{l}(50)$ needed

in this equation are such that

$$\hat{u}(50) = \max\{t_{(j)} \mid \hat{S}(t_{(j)}) \geqslant 0.55\},$$

and

$$\hat{l}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) \leqslant 0.45\},$$

where $t_{(j)}$ is the jth ordered discontinuation time, $j = 1, 2, \ldots, 9$. Using Table 2.4, $\hat{u}(50) = 75$ and $\hat{l}(50) = 97$, and so

$$\hat{f}\{\hat{t}(50)\} = \frac{\hat{S}(75) - \hat{S}(97)}{97 - 75} = \frac{0.5594 - 0.3729}{22} = 0.0085.$$

Then, the standard error of the median is given by

$$\text{se}\{\hat{t}(50)\} = \frac{1}{0.0085} \times 0.1452 = 17.13.$$

A 95% confidence interval for the median discontinuation time has limits of

$$93 \pm 1.96 \times 17.13,$$

and so the required interval estimate for the median ranges from 59 to 127 days.

2.6 Comparison of two groups of survival data

The simplest way of comparing the survival times obtained from two groups of individuals is to plot the corresponding estimates of the two survivor functions on the same axes. The resulting plot can be quite informative, as the following example illustrates.

Example 2.11 Prognosis for women with breast cancer

Data on the survival times of women with breast cancer, grouped according to whether or not sections of a tumour were positively stained with HPA, were given in Example 1.2. The Kaplan-Meier estimate of the survivor function, for each of the two groups of survival times, is plotted in Figure 2.9. Notice that in this figure, the Kaplan-Meier estimates extend to the time of the largest censored observation in each group.

This figure shows that the estimated survivor function for those women with negatively stained tumours is always greater than that for women with positively stained tumours. This means that at any time t, the estimated probability of survival beyond t is greater for women with negative staining, suggesting that the result of the HPA staining procedure might be a useful prognostic indicator. In particular, those women whose tumours are positively stained appear to have a poorer prognosis than those with negatively stained tumours.

There are two possible explanations for an observed difference between two estimated survivor functions, such as those in Example 2.11. One explanation is that there is a real difference between the survival times of the two groups of individuals, so that those in one group have a different survival experience