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

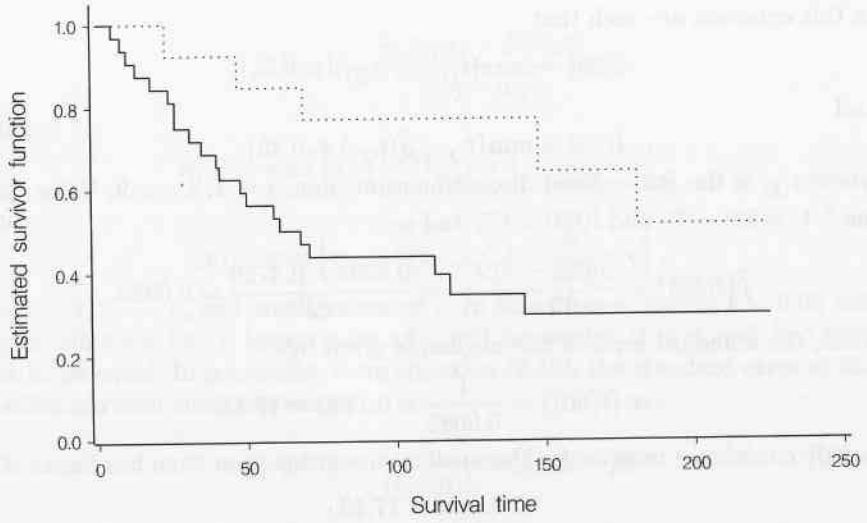

Figure 2.9 *Kaplan-Meier estimate of the survivor functions for women with tumours that were positively stained (—) and negatively stained (· · ·).*

from those in the other. An alternative explanation is that there are no real differences between the survival times in each group, and that the difference that has been observed is merely the result of chance variation. To help distinguish between these two possible explanations, we use a procedure known as the *hypothesis test*. Because the concept of the hypothesis test has a central role in the analysis of survival data, the underlying basis for this procedure is described in detail in the following section.

2.6.1 Hypothesis testing

The hypothesis test is a procedure that enables us to assess the extent to which an observed set of data are consistent with a particular hypothesis, known as the *working* or *null hypothesis*. A null hypothesis generally represents a simplified view of the data-generating process, and is typified by hypotheses that specify that there is no difference between two groups of survival data, or that there is no relationship between survival time and explanatory variables such as age or serum cholesterol level. The null hypothesis is then the hypothesis that will be adopted, and subsequently acted upon, unless the data indicate that it is untenable.

The next step is to formulate a *test statistic* that measures the extent to which the observed data depart from the null hypothesis. In general, the test statistic is so constructed that the larger the value of the statistic, the greater the departure from the null hypothesis. Hence, if the null hypothesis is that there is no difference between two groups, relatively large values of the test statistic will be interpreted as evidence against this null hypothesis.

Once the value of the test statistic has been obtained from the observed data, we calculate the probability of obtaining a value as extreme or more extreme than the observed value, when the null hypothesis is true. This quantity summarises the strength of the evidence in the sample data against the null hypothesis, and is known as the *probability value*, or *P-value* for short. If the P-value is large, we would conclude that it is quite likely that the observed data would have been obtained when the null hypothesis was true, and that there is no evidence to reject the null hypothesis. On the other hand, if the P-value is small, this would be interpreted as evidence against the null hypothesis; the smaller the P-value, the stronger the evidence.

In order to obtain the P-value for a hypothesis test, the test statistic must have a probability distribution that is known, or at least approximately known, when the null hypothesis is true. This probability distribution is referred to as the *null distribution* of the test statistic. More specifically, consider a test statistic, W, which is such that the larger the observed value of the test statistic, w, the greater the deviation of the observed data from that expected under the null hypothesis. If W has a continuous probability distribution, the P-value is then $\mathrm{P}(W \geqslant w) = 1 - F(w)$, where $F(w)$ is the distribution function of W, under the null hypothesis, evaluated at w.

In some applications, the most natural test statistic is one for which large positive values correspond to departures from the null hypothesis in one direction, while large negative values correspond to departures in the opposite direction. For example, suppose that patients suffering from a particular illness have been randomised to receive either a standard treatment or a new treatment, and their survival times are recorded. In this situation, a null hypothesis of interest will be that there is no difference in the survival experience of the patients in the two treatment groups. The extent to which the data are consistent with this null hypothesis might then be summarised by a test statistic for which positive values indicate that the new treatment is superior to the standard, while negative values indicate that the standard treatment is superior. When departures from the null hypothesis in either direction are equally important, the null hypothesis is said to have a *two-sided alternative*, and the hypothesis test itself is referred to as a two-sided test.

If W is a test statistic for which large positive or large negative observed values lead to rejection of the null hypothesis, a new test statistic, such as $|W|$ or W^2, can be defined, so that only large positive values of the new statistic indicate that there is evidence against the null hypothesis. For example, suppose that W is a test statistic that under the null hypothesis has a standard normal distribution. If w is the observed value of W, the appropriate P-value is $\mathrm{P}(W \leqslant -|w|) + \mathrm{P}(W \geqslant |w|)$, which in view of the symmetry of the standard normal distribution, is $2\,\mathrm{P}(W \geqslant |w|)$. Alternatively, we can make use of the result that if W has a standard normal distribution, W^2 has a chi-squared distribution on one degree of freedom, written χ^2_1. Thus a P-value for the two-sided hypothesis test based on the statistic W is the probability that a χ^2_1 random variable exceeds w^2. The required P-value can therefore be found using tables of the standard normal or chi-squared distribution functions.

When interest centres on departures in a particular direction, the hypothesis test is said to be *one-sided*. For example, in comparing the survival times of two groups of patients where one group receives a standard treatment and the other group a new treatment, it might be argued that the new treatment cannot possibly be inferior to the standard. Then, the only relevant alternative to the null hypothesis of no treatment difference is that the new treatment is superior. If positive values of the test statistic W reflect the superiority of the new treatment, the P-value is then $\mathrm{P}(W \geqslant w)$. If W has a standard normal distribution, this P-value is half of that which would have been obtained for the corresponding two-sided alternative hypothesis.

A one-sided hypothesis test can only be appropriate when there is no interest whatsoever in departures from the null hypothesis in the opposite direction to that specified in the one-sided alternative. For example, consider again the comparison of a new treatment with a standard treatment, and suppose that the observed value of the test statistic is either positive or negative, depending on whether the new treatment is superior or inferior to the standard. If the alternative to the null hypothesis of no treatment difference is that the new treatment is superior, a large negative value of the test statistic would not be regarded as evidence against the null hypothesis. Instead, it would be assumed that this large negative value is simply the result of chance variation. Generally speaking, the use of one-sided tests can rarely be justified in medical research, and so two-sided tests will be used throughout this book.

If a P-value is smaller than some value α, we say that the hypothesis is rejected at the $100\alpha\%$ *level of significance*. The observed value of the test statistic is then said to be significant at this level. But how do we decide on the basis of the P-value whether or not a null hypothesis should actually be rejected? Traditionally, P-values of 0.05 or 0.01 have been used in reaching a decision about whether or not a null hypothesis should be rejected, so that if $P < 0.05$, for example, the null hypothesis is rejected at the 5% significance level. Guidelines such as these are not hard-and-fast rules and should not be interpreted rigidly. For example, there is no practical difference between a P-value of 0.046 and 0.056, even though only the former indicates that the observed value of the test statistic is significant at the 5% level.

Instead of reporting that a null hypothesis is rejected or not rejected at some specified significance level, a more satisfactory policy is to report the actual P-value. This P-value can then be interpreted as a measure of the strength of evidence against the null hypothesis, using a vocabulary that depends on the range within which the P-value lies. Thus, if $P > 0.1$, there is said to be no evidence to reject the null hypothesis; if $0.05 < P \leqslant 0.1$, there is slight evidence against the null hypothesis; if $0.01 < P \leqslant 0.05$, there is moderate evidence against the null hypothesis; if $0.001 < P \leqslant 0.01$, there is strong evidence against the null hypothesis, and if $P \leqslant 0.001$, the evidence against the null hypothesis is overwhelming.

An alternative to quoting the exact P-value associated with a hypothesis test, is to compare the observed value of the test statistic with those values that would correspond to particular P-values, when the null hypothesis is

true. Values of the test statistic that lead to rejection of the null hypothesis at particular levels of significance can be found from tables of the percentage points of the null distribution of that statistic. In particular, if W is a test statistic that has a standard normal distribution, for a two-sided test, the upper $\alpha/2$-point of the distribution, depicted in Figure 2.5, is the value of the test statistic for which the P-value is α. For example, values of the test statistic of 1.96, 2.58 and 3.29 correspond to P-values of 0.05, 0.01 and 0.001. Thus, if the observed value of W were between 1.96 and 2.58, we would declare that $0.01 < P < 0.05$. On the other hand, if the null distribution of W is chi-squared on one degree of freedom, the upper α-point of the distribution is the value of the test statistic which would give a P-value of α. Then, values of the test statistic of 3.84, 6.64 and 10.83 correspond to P-values of 0.05, 0.01 and 0.001, respectively. Notice that these values are simply the squares of those for the standard normal distribution, which they must be in view of the fact that the square of a standard normal random variable has a chi-squared distribution on one degree of freedom.

For commonly encountered probability distributions, such as the normal and chi-squared, percentage points are tabulated in many introductory text books on statistics, or in statistical tables such as those of Lindley and Scott (1984). Statistical software packages used in computer-based statistical analyses of survival data usually provide the exact P-values associated with hypothesis tests as a matter of course. Note that when these are rounded off to, say, three decimal places, a P-value of 0.000 should be interpreted as $P < 0.001$.

In deciding on a course of action, such as whether or not to reject the hypothesis that there is no difference between two treatments, the statistical evidence summarised in the P-value for the hypothesis test will be just one ingredient of the decision-making process. In addition to the statistical evidence, there will also be scientific evidence to consider. This may, for example, concern whether the size of the treatment effect is clinically important. In particular, in a large trial, a difference between two treatments that is significant at, say, the 5% level may be found when the magnitude of the treatment effect is so small that it does not indicate a major scientific breakthrough. On the other hand, a new formulation of a treatment may prolong life by a factor of two, and yet, because of small sample sizes used in the study, may not appear to be significantly different from the standard.

Rather than report findings in terms of the results of a hypothesis testing procedure, it is more informative to provide an estimate of the size of any treatment difference, supported by a confidence interval for this difference. Unfortunately, the non-parametric approaches to the analysis of survival data being considered in this chapter do not lend themselves to this approach. We will therefore return to this theme in subsequent chapters when we consider models for survival data.

In the comparison of two groups of survival data, there are a number of methods that can be used to quantify the extent of between-group differences. Two non-parametric procedures will now be considered, namely the log-rank test and the Wilcoxon test.

2.6.2 *The log-rank test*

In order to construct the log-rank test, we begin by considering separately each death time in two groups of survival data. These groups will be labelled Group I and Group II. Suppose that there are r distinct death times, $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, across the two groups, and that at time $t_{(j)}$, d_{1j} individuals in Group I and d_{2j} individuals in Group II die, for $j = 1, 2, \ldots, r$. Unless two or more individuals in a group have the same recorded death time, the values of d_{1j} and d_{2j} will either be zero or unity. Suppose further that there are n_{1j} individuals at risk of death in the first group just before time $t_{(j)}$, and that there are n_{2j} at risk in the second group. Consequently, at time $t_{(j)}$, there are $d_j = d_{1j} + d_{2j}$ deaths in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk. The situation is summarised in Table 2.7.

Table 2.7 *Number of deaths at the jth death time in each of two groups of individuals.*

Group	Number of deaths at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Number at risk just before $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Now consider the null hypothesis that there is no difference in the survival experiences of the individuals in the two groups. One way of assessing the validity of this hypothesis is to consider the extent of the difference between the observed number of individuals in the two groups who die at each of the death times, and the numbers expected under the null hypothesis. Information about the extent of these differences can then be combined over each of the death times.

If the marginal totals in Table 2.7 are regarded as fixed, and the null hypothesis that survival is independent of group is true, the four entries in this table are solely determined by the value of d_{1j}, the number of deaths at $t_{(j)}$ in Group I. We can therefore regard d_{1j} as a random variable, which can take any value in the range from 0 to the minimum of d_j and n_{1j}. In fact, d_{1j} has a distribution known as the *hypergeometric distribution*, according to which the probability that the random variable associated with the number of deaths in the first group takes the value d_{1j} is

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}. \tag{2.21}$$

In this formula, the expression

$$\binom{d_j}{d_{1j}}$$

represents the number of different ways in which d_{1j} times can be chosen from d_j times and is read as "d_{1j} C d_j". It is given by

$$\binom{d_j}{d_{1j}} = \frac{d_j!}{d_{1j}!(d_j - d_{1j})!},$$

where $d_j!$, read as "d_j *factorial*", is such that

$$d_j! = d_j \times (d_j - 1) \times \cdots \times 2 \times 1.$$

The other two terms in expression (2.21) are interpreted in a similar manner.

The mean of the hypergeometric random variable d_{1j} is given by

$$e_{1j} = n_{1j} d_j / n_j, \tag{2.22}$$

so that e_{1j} is the expected number of individuals who die at time $t_{(j)}$ in Group I. This value is intuitively appealing, since under the null hypothesis that the probability of death at time $t_{(j)}$ does not depend on the group that an individual is in, the probability of death at $t_{(j)}$ is d_j/n_j. Multiplying this by n_{1j}, gives e_{1j} as the expected number of deaths in Group I at $t_{(j)}$.

The next step is to combine the information from the individual 2×2 tables for each death time to give an overall measure of the deviation of the observed values of d_{1j} from their expected values. The most straightforward way of doing this is to sum the differences $d_{1j} - e_{1j}$ over the total number of death times, r, in the two groups. The resulting statistic is given by

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}). \tag{2.23}$$

Notice that this is $\sum d_{1j} - \sum e_{1j}$, which is the difference between the total observed and expected numbers of deaths in Group I. This statistic will have zero mean, since $\mathrm{E}(d_{1j}) = e_{1j}$. Moreover, since the death times are independent of one another, the variance of U_L is simply the sum of the variances of the d_{1j}. Now, since d_{1j} has a hypergeometric distribution, the variance of d_{1j} is given by

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, \tag{2.24}$$

so that the variance of U_L is

$$\mathrm{var}(U_L) = \sum_{j=1}^{r} v_{1j} = V_L, \tag{2.25}$$

say. Furthermore, it can be shown that U_L has an approximate normal distribution, when the number of death times is not too small. It then follows that $U_L/\sqrt{V_L}$ has a normal distribution with zero mean and unit variance,

denoted $N(0,1)$. We therefore write

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1),$$

where the symbol "$\sim$" is read as "is distributed as". The square of a standard normal random variable has a chi-squared distribution on one degree of freedom, denoted χ^2_1, and so we have that

$$\frac{U_L^2}{V_L} \sim \chi^2_1. \tag{2.26}$$

This method of combining information over a number of 2×2 tables was proposed by Mantel and Haenszel (1959), and is known as the Mantel-Haenszel procedure. In fact, the test based on this statistic has various names, including *Mantel-Cox* and *Peto-Mantel-Haenszel*, but it is probably best known as the log-rank test. The reason for this name is that the test statistic can be derived from the ranks of the survival times in the two groups, and the resulting rank test statistic is based on the logarithm of the Nelson-Aalen estimate of the survivor function.

The statistic $W_L = U_L^2/V_L$ summarises the extent to which the observed survival times in the two groups of data deviate from those expected under the null hypothesis of no group differences. The larger the value of this statistic, the greater the evidence against the null hypothesis. Because the null distribution of W is approximately chi-squared with one degree of freedom, the P-value associated with the test statistic can be obtained from the distribution function of a chi-squared random variable. Alternatively, percentage points of the chi-squared distribution can be used to identify a range within which the P-value lies. An illustration of the log-rank test is presented below in Example 2.12.

Example 2.12 Prognosis for women with breast cancer

In this example, we return to the data on the survival times of women with breast cancer, grouped according to whether a section of the tumour was positively or negatively stained. In particular the null hypothesis that there is no difference in the survival experience of the two groups will be examined using the log-rank test. The required calculations are laid out in Table 2.8.

We begin by ordering the observed death times across the two groups of women; these times are given in column 1 of Table 2.8. The numbers of women in each group who die at each death time and the numbers who are at risk at each time are then calculated. These values are d_{1j}, n_{1j}, d_{2j} and n_{2j} given in columns 2 to 5 of the table. Columns 6 and 7 contain the total numbers of deaths and the total numbers of women at risk over the two groups, at each death time. The final two columns give the values of e_{1j} and v_{1j}, computed from equations (2.22) and (2.24) respectively. Summing the entries in columns 2 and 8 gives $\sum d_{1j}$ and $\sum e_{1j}$, from which the log-rank statistic can be calculated from $U_L = \sum d_{1j} - \sum e_{1j}$. The value of $V_L = \sum v_{1j}$ can be obtained by summing the entries in the final column. We find that

Table 2.8 *Calculation of the log-rank statistic for the data from Example 1.2.*

Death time	d_{1j}	n_{1j}	d_{2j}	n_{2j}	d_j	n_j	e_{1j}	v_{1j}
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2344
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Total	5						9.5652	5.9289

$U_L = 5 - 9.565 = -4.565$ and $V_L = 5.929$, and so the value of the log-rank test statistic is $W_L = (-4.565)^2/5.929 = 3.515$.

The corresponding P-value is calculated from the probability that a chi-squared variate on one degree of freedom is greater than or equal to 3.515, and is 0.061, written $P = 0.061$. This P-value is sufficiently small to cast doubt on the null hypothesis that there is no difference between the survivor functions for the two groups of women. In fact, the evidence against the null hypothesis is nearly significant at the 6% level. We therefore conclude that the data do provide some evidence that the prognosis of a breast cancer patient is dependent on the result of the staining procedure.

2.6.3 The Wilcoxon test

The Wilcoxon test, sometimes known as the *Breslow test*, is also used to test the null hypothesis that there is no difference in the survivor functions for two

groups of survival data. The Wilcoxon test is based on the statistic

$$U_W = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j}),$$

where, as in the previous section, d_{1j} is the number of deaths at time $t_{(j)}$ in the first group and e_{1j} is as defined in equation (2.22). The difference between U_W and U_L is that in the Wilcoxon test, each difference $d_{1j} - e_{1j}$ is weighted by n_j, the total number of individuals at risk at time $t_{(j)}$. The effect of this is to give less weight to differences between d_{1j} and e_{1j} at those times when the total number of individuals who are still alive is small, that is, at the longest survival times. This statistic is therefore less sensitive than the log-rank statistic to deviations of d_{1j} from e_{1j} in the tail of the distribution of survival times.

The variance of the Wilcoxon statistic U_W is given by

$$V_W = \sum_{j=1}^{r} n_j^2 v_{1j},$$

where v_{1j} is given in equation (2.24), and so the Wilcoxon test statistic is

$$W_W = U_W^2/V_W, \tag{2.27}$$

which has a chi-squared distribution on one degree of freedom when the null hypothesis is true. The Wilcoxon test is therefore conducted in the same manner as the log-rank test.

Example 2.13 Prognosis for women with breast cancer
For the data on the survival times of women with tumours that were positively or negatively stained, the value of the Wilcoxon statistic is $U_W = -159$, and the variance of the statistic is $V_W = 6048.136$. The value of the chi-squared statistic, U_W^2/V_W, is 4.180, and the corresponding P-value is 0.041. This is slightly smaller than the P-value for the log-rank test, and on the basis of this result, we would declare that the difference between the two groups is significant at the 5% level.

2.6.4 Comparison of the log-rank and Wilcoxon tests

Of the two tests, the log-rank test is the more suitable when the alternative to the null hypothesis of no difference between two groups of survival times is that the hazard of death at any given time for an individual in one group is proportional to the hazard at that time for a similar individual in the other group. This is the assumption of *proportional hazards*, which underlies a number of methods for analysing survival data. For other types of departure from the null hypothesis, the Wilcoxon test is more appropriate than the log-rank test for comparing the two survivor functions.

In order to help decide which test is the more suitable in any given situation, we make use of the result that if the hazard functions are proportional,

the survivor functions for the two groups of survival data do not cross one another. To show this, suppose that $h_1(t)$ is the hazard of death at time t for an individual in Group I, and $h_2(t)$ is the hazard at that same time for an individual in Group II. If these two hazards are proportional, then we can write $h_1(t) = \psi h_2(t)$, where ψ is a constant that does not depend on the time t. Integrating both sides of this expression, multiplying by -1 and exponentiating gives

$$\exp\left\{-\int_0^t h_1(u)\,\mathrm{d}u\right\} = \exp\left\{-\int_0^t \psi h_2(u)\,\mathrm{d}u\right\}. \tag{2.28}$$

Now, from equation (1.5),

$$S(t) = \exp\left\{-\int_0^t h(u)\,\mathrm{d}u\right\},$$

and so if $S_1(t)$ and $S_2(t)$ are the survivor functions for the two groups of survival data, from equation (2.28),

$$S_1(t) = \{S_2(t)\}^{\psi}.$$

Since the survivor function takes values between zero and unity, this result shows that $S_1(t)$ is greater than or less than $S_2(t)$, according to whether ψ is less than or greater than unity, at any time t. This means that if two hazard functions are proportional, the true survivor functions do not cross. This is a necessary, but not a sufficient condition for proportional hazards.

An informal assessment of the likely validity of the proportional hazards assumption can be made from a plot of the estimated survivor functions for two groups of survival data, such as that shown in Figure 2.9. If the two estimated survivor functions do not cross, the assumption of proportional hazards may be justified, and the log-rank test is appropriate. Of course, sample-based estimates of survivor functions may cross even though the corresponding true hazard functions are proportional, and so some care is needed in the interpretation of such graphs. A more satisfactory graphical method for assessing the validity of the proportional hazards assumption is described in Section 4.4.1 of Chapter 4.

In summary, unless a plot of the estimated survival functions, or previous data, indicate that there is good reason to doubt the proportional hazards assumption, the log-rank test should be used to test the hypothesis of equality of two survivor functions.

Example 2.14 Prognosis for women with breast cancer
From the graph of the two estimated survivor functions in Figure 2.9, we see that the survivor function for the negatively stained women always lies above that for the positively stained women. This suggests that the proportional hazards assumption is appropriate, and that the log-rank test is more appropriate than the Wilcoxon test. However, in this example, there is very little difference between the results of the two hypothesis tests.