2.7* Comparison of three or more groups of survival data

Both the log-rank and the Wilcoxon tests can be extended to enable three or more groups of survival data to be compared. Suppose that the survival distributions of g groups of survival data are to be compared, for $g \geqslant 2$. We then define analogues of the U-statistics for comparing the observed numbers of deaths in groups $1, 2, \ldots, g-1$ with their expected values. In an obvious extension of the notation used in Section 2.6, we obtain

$$U_{Lk} = \sum_{j=1}^{r} \left(d_{kj} - \frac{n_{kj} d_j}{n_j} \right),$$

$$U_{Wk} = \sum_{j=1}^{r} n_j \left(d_{kj} - \frac{n_{kj} d_j}{n_j} \right),$$

for $k = 1, 2, \ldots, g-1$. These quantities are then expressed in the form of a vector with $(g-1)$ components, which we denote by $\boldsymbol{U}_L$ and $\boldsymbol{U}_W$.

We also need expressions for the variances of the U_{Lk} and U_{Wk}, and for the covariance between pairs of values. In particular, the covariance between U_{Lk} and $U_{Lk'}$ is given by

$$V_{Lkk'} = \sum_{j=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

for $k, k' = 1, 2, \ldots, g-1$, where $\delta_{kk'}$ is such that

$$\delta_{kk'} = \begin{cases} 1 \text{ if } k = k', \\ 0 \text{ otherwise.} \end{cases}$$

These terms are then assembled in the form of a *variance-covariance matrix*, $\boldsymbol{V}_L$, which is a symmetric matrix that has the variances of the U_{Lk} down the diagonal, and covariance terms in the off-diagonals. For example, in the comparison of three groups of survival data, this matrix would be given by

$$\boldsymbol{V}_L = \begin{pmatrix} V_{L11} & V_{L12} \\ V_{L12} & V_{L22} \end{pmatrix},$$

where V_{L11} and V_{L22} are the variances of U_{L1} and U_{L2}, respectively, and V_{L12} is their covariance.

Similarly, the variance-covariance matrix for the Wilcoxon statistic is the matrix $\boldsymbol{V}_W$, whose (k, k')th element is

$$V_{Wkk'} = \sum_{j=1}^{r} n_j^2 \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

for $k, k' = 1, 2, \ldots, g-1$.

Finally, in order to test the null hypothesis of no group differences, we make use of the result that the test statistic $\boldsymbol{U}_L' \boldsymbol{V}_L^{-1} \boldsymbol{U}_L$, or $\boldsymbol{U}_W' \boldsymbol{V}_W^{-1} \boldsymbol{U}_W$, has a chi-squared distribution on $(g-1)$ degrees of freedom, when the null hypothesis is true.

A number of well-known statistical packages for the analysis of survival data incorporate this methodology. Furthermore, because the interpretation of the resulting chi-squared statistic is straightforward, an example will not be given here.

2.8 Stratified tests

In many circumstances, there is a need to compare two or more sets of survival data, after taking account of additional variables recorded on each individual. As an illustration, consider a multicentred clinical trial in which two forms of chemotherapy are to be compared in terms of their effect on the survival times of lung cancer patients. Information on the survival times of patients in each treatment group will be available from each centre. The resulting data are then said to be *stratified* by centre.

Individual log-rank or Wilcoxon tests based on the data from each centre will be informative, but a test that combines information about the treatment difference in each centre would provide a more precise summary of the treatment effect. A similar situation would arise in attempting to test for treatment differences when patients are stratified according to variables such as age group, sex, performance status and other potential risk factors for the disease under study.

In situations such as those described above, a stratified version of the log-rank or Wilcoxon test may be employed. Essentially, this involves calculating the values of the U- and V-statistics for each *stratum*, and then combining these values over the strata. In this section, the *stratified log-rank test* will be described, but a stratified version of the Wilcoxon test can be obtained in a similar manner. An equivalent analysis, based on a model for the survival times, is described in Section 11.1.1 of Chapter 11.

Let U_{Lk} be the value of the log-rank statistic for comparing two treatment groups, computed from the kth of s strata using equation (2.23). Also, denote the variance of the statistic for the kth stratum by V_{Lk}, where V_{Lk} would be computed for each stratum using equation (2.24). The stratified log-rank test is then based on the statistic

$$W_S = \frac{\left(\sum_{k=1}^{s} U_{Lk} \right)^2}{\sum_{k=1}^{s} V_{Lk}}, \tag{2.29}$$

which has a chi-squared distribution on one degree of freedom (1 d.f.) under the null hypothesis that there is no treatment difference. Comparing the observed value of this statistic with percentage points of the chi-squared distribution enables the hypothesis of no overall treatment difference to be tested.

Example 2.15 Survival times of melanoma patients

The aim of a study carried out by the University of Oklahoma Health Sciences Center was to compare two immunotherapy treatments for their ability to prolong the life of patients suffering from melanoma, a highly malignant tumour occurring in the skin. For each patient, the tumour was surgically

removed before allocation to *Bacillus Calmette-Guérin* (BCG) vaccine or to a vaccine based on the bacterium *Corynebacterium parvum* (*C. parvum*).

The survival times of the patients in each treatment group were further classified according to the age group of the patient. The data, which were given in Lee and Wang (2003), are shown in Table 2.9. An asterisk against a survival time indicates that the observation is censored.

Table 2.9 *Survival times of melanoma patients in two treatment groups, stratified by age group.*

21–40		41–60		61–	
BCG	*C. parvum*	BCG	*C. parvum*	BCG	*C. parvum*
19	27*	34*	8	10	25*
24*	21*	4	11*	5	8
8	18*	17*	23*		11*
17*	16*		12*		
17*	7		15*		
34*	12*		8*		
	24		8*		
	8				
	8*				

These data are analysed by first computing the log-rank statistics for comparing the survival times of patients in the two treatment groups, separately for each age group. The resulting values of the U-, V- and W-statistics, found using equations (2.23), (2.25) and (2.26), are summarised in Table 2.10.

Table 2.10 *Values of the log-rank statistic for each age group.*

Age group	U_L	V_L	W_L
21–40	−0.2571	1.1921	0.055
41–60	0.4778	0.3828	0.596
61–	1.0167	0.6497	1.591
Total	1.2374	2.2246	

The values of the W_L-statistic are quite similar for the three age groups, suggesting that the treatment effect is consistent over these groups. Moreover, none of them are significantly large at the 10% level.

To carry out a stratified log-rank test on these data, we calculate the W_S-statistic defined in equation (2.29). Using the results in Table 2.10,

$$W_S = \frac{1.2374^2}{2.2246} = 0.688.$$

The observed value of W_S is not significant when compared with percentage

points of the chi-squared distribution on 1 d.f. We therefore conclude that after allowing for the different age groups, there is no significant difference between the survival times of patients treated with the BCG vaccine and those treated with *C. parvum*.

For comparison, when the division of the patients into the different age groups is ignored, the log-rank test for comparing the two groups of patients leads to $W_L = 0.756$. The fact that this is so similar to the value that allows for age group differences suggests that it is not necessary to stratify the patients by age.

The stratified log-rank test can be extended to compare more than two treatment groups The resulting formulae render it unsuitable for hand calculation, but the methodology can be implemented using computer software for survival analysis. However, this method of taking account of additional variables is not as flexible as that based on a modelling approach, introduced in the next chapter.

2.9 Log-rank test for trend

In many applications where three or more groups of survival data are to be compared, these groups are ordered in some way. For example, the groups may correspond to increasing doses of a treatment, the stage of a disease, or the age group of an individual. In comparing these groups using the log-rank test described in previous sections, it can happen that the analysis does not lead to a significant difference between the groups, even though the hazard of death increases or decreases across the groups. Indeed, a test that uses information about the ordering of the groups is more likely to lead to a trend being identified as significant than a standard log-rank test.

The log-rank test for trend across g ordered groups is based on the statistic

$$U_T = \sum_{k=1}^{g} w_k(d_{k\cdot} - e_{k\cdot}), \tag{2.30}$$

where w_k is a *code* assigned to the kth group, $k = 1, 2, \ldots, g$, and

$$d_{k\cdot} = \sum_{j=1}^{r_k} d_{kj}, \qquad e_{k\cdot} = \sum_{j=1}^{r_k} e_{kj},$$

are the observed and expected numbers of deaths in the kth group, where the summation is over the r_k death times in that group. Note that the dot subscript in the notation $d_{k\cdot}$ and $e_{k\cdot}$ stands for summation over the subscript that the dot replaces. The codes are often taken to be equally spaced to correspond to a linear trend across the groups. For example, if there are three groups, the codes might be taken to be 1, 2 and 3, although the equivalent choice of −1, 0 and 1 does simplify the calculations somewhat. The variance

of U_T is given by

$$V_T = \sum_{k=1}^{g} (w_k - \bar{w})^2 e_{k.}, \tag{2.31}$$

where $\bar{w}$ is a weighted sum of the quantities w_k, in which the expected numbers of deaths, $e_{k.}$, are the weights, that is,

$$\bar{w} = \frac{\sum_{k=1}^{g} w_k e_{k.}}{\sum_{k=1}^{g} e_{k.}}.$$

The statistic $W_T = U_T^2/V_T$ then has a chi-squared distribution on 1 d.f. under the hypothesis of no trend across the g groups.

Example 2.16 Survival times of melanoma patients
The log-rank test for trend will be illustrated using the data from Example 2.15 on the survival times of patients suffering from melanoma. For the purpose of this illustration, only the data from those patients allocated to the BCG vaccine will be used. The log-rank statistic for comparing the survival times of the patients in the three age groups turns out to be 3.739. When compared to percentage points of the chi-squared distribution on 2 d.f., this is not significant ($P = 0.154$).

We now use the log-rank test for trend to examine whether there is a linear trend over age. For this, we will take the codes, w_k, to be equally spaced, with values -1, 0 and 1. Some of the calculations required for the log-rank test for trend are summarised in Table 2.11.

Table 2.11 *Values of w_k and the observed and expected numbers of deaths in the three age groups.*

Age group	w_k	$d_{k.}$	$e_{k.}$
21–40	−1	2	3.1871
41–60	0	1	1.1949
61–	1	2	0.6179

The log-rank test for trend is based on the statistic in equation (2.30), the value of which is

$$U_T = (d_{3.} - e_{3.}) - (d_{1.} - e_{1.}) = 2.5692.$$

Using the values of the expected numbers of deaths in each group, given in Table 2.11, the weighted mean of the w_k's is given by

$$\bar{w} = \frac{e_{3.} - e_{1.}}{e_{1.} + e_{3.}} = 0.5138.$$

The three values of $(w_k - \bar{w})^2$ are 0.2364, 0.2640 and 2.2917, and, from equation (2.31), $V_T = 2.4849$. Finally, the test statistic is

$$W_T = \frac{U_T^2}{V_T} = 2.656,$$

which is just about significant at the 10% level ($P = 0.103$) when judged against a chi-squared distribution on 1 d.f. We therefore conclude that there is slight evidence of a linear trend across the age groups.

An alternative method of examining whether there is a trend across the levels of an ordered categorical variable, based on a modelling approach to the analysis of survival data, is described and illustrated in Section 3.6.2 of the next chapter.

2.10 Further reading

The life-table, which underpins the calculation of the life-table estimate of the survivor function, is widely used in the analysis of data from epidemiological studies. Fuller details of this application can be found in Armitage *et al.* (2001), and books on statistical methods in demography and epidemiology, such as Pollard *et al.* (1990) and Woodward (1999).

The product-limit estimate of the survivor function has been in use since the early 1900s. Kaplan and Meier (1958) derived the estimate using the method of maximum likelihood, which is why the estimate now bears their name. The properties of the Kaplan-Meier estimate of the survivor function have been further explored by Breslow and Crowley (1974) and Meier (1975). The Nelson-Aalen estimate is due to Altshuler (1970), Nelson (1972) and Aalen (1978), and the estimator is considered in a counting process framework by Therneau and Grambsch (2000).

The expression for the standard error of the Kaplan-Meier estimate was first given by Greenwood (1926), but an alternative expression is given by Aalen and Johansen (1978). Alternative expressions for the variance of the Nelson-Aalen estimate of the cumulative hazard function are compared by Klein (1991). Although Section 2.2.1 shows how a confidence interval for the value of the survivor function at particular times can be found using Greenwood's formula, alternative procedures are needed for the construction of confidence bands for the complete survivor function. Hall and Wellner (1980) and Efron (1981) have shown how such bands can be computed, and these procedures are also described by Harris and Albert (1991).

Methods for constructing confidence intervals for the median survival time are described by Brookmeyer and Crowley (1982), Emerson (1982), Nair (1984), Simon and Lee (1982) and Slud *et al.* (1984). Simon (1986) emphasises the importance of confidence intervals in reporting the results of clinical trials, and includes an illustration of a method described in Slud *et al.* (1984). Klein and Moeschberger (1997) include a comprehensive review of kernel-smoothed estimates of the hazard function.

The formulation of the hypothesis testing procedure in the frequentist ap-

proach to inference is covered in many statistical texts. See, for example, Altman (1991) and Armitage *et al.* (2001) for non-technical presentations of the ideas in a medical context.

The log-rank test results from the work of Mantel and Haenszel (1959), Mantel (1966) and Peto and Peto (1972). See Lawless (2002) for details of the rank test formulation. A thorough review of the hypergeometric distribution, used in the derivation of the log-rank test in Section 2.6.2, is included in Johnson and Kotz (1969).

The log-rank test for trend is derived from the test for trend in a $2 \times k$ contingency table, given in Armitage *et al.* (2001). The test is also described by Altman (1991). Peto *et al.* (1976, 1977) give a non-mathematical account of the log-rank test and its extensions.

CHAPTER 3

Modelling survival data

The non-parametric methods described in Chapter 2 can be useful in the analysis of a single sample of survival data, or in the comparison of two or more groups of survival times. However, in most medical studies that give rise to survival data, supplementary information will also be recorded on each individual. A typical example would be a clinical trial to compare the survival times of patients who receive one or other of two treatments. In such a study, demographic variables such as the age and sex of the patient, the values of physiological variables such as serum haemoglobin level and heart rate, and factors that are associated with the lifestyle of the patient, such as smoking history and dietary habits, may all have an impact on the time that the patient survives. Accordingly, the values of these variables, which are referred to as *explanatory variables*, would be recorded at the outset of the study. The resulting data set would then be more complex than those considered in Chapter 2, and the methods described in that chapter would generally be unsuitable.

In order to explore the relationship between the survival experience of a patient and explanatory variables, an approach based on statistical modelling can be used. Indeed, the particular model that is developed in this chapter both unifies and extends the non-parametric procedures of Chapter 2.

3.1 Modelling the hazard function

Through a modelling approach to the analysis of survival data, we can explore how the survival experience of a group of patients depends on the values of one or more explanatory variables, whose values have been recorded for each patient at the time origin. For example, in the study on multiple myeloma, given as Example 1.3, the aim is to determine which of seven explanatory variables have an impact on the survival time of the patients. In Example 1.4 on the survival times of patients in a clinical trial involving two treatments for prostatic cancer, the primary aim is to identify whether patients in the two treatment groups have a different survival experience. Because additional variables such as the age of the patient and the size of their tumour are likely to influence survival time, it will be important to take account of these variables when assessing the extent of any treatment difference.

In the analysis of survival data, interest centres on the risk or hazard of death at any time after the time origin of the study. As a consequence, the hazard function is modelled directly in survival analysis. The resulting models