17 Survival analysis

Armitage, Berry, Matthews
Statistical Methods in Medical
Research 4th edition

17.1 Introduction

In many studies the variable of direct interest is the length of time that elapses before some event occurs. This event may be death, or death due to a particular disease, and for this reason the analysis of such data is often referred to as *survival analysis*.

An example of such a study is a clinical trial for the treatment of a malignant tumour where the prognosis is poor; death or remission of the tumour would be the end-point. Such studies usually include individuals for whom the event has not occurred at the time of the analysis. Although the time to the event for such a patient is unknown, there is some information on its value since it is known that it must exceed the current survival time; an observation of this type is referred to as a *censored* value. Methods of analysis must be able to cope with censored values. Often a number of variables are observed at the commencement of a trial, and survival is related to the values of these variables; that is, the variables are prognostic. Methods of analysis must be able to take account of the distribution of prognostic variables in the groups under study.

The number of studies of the above type has increased during the last three decades and statistical methods have been developed to analyse them; many of these methods were developed during the 1970s. Some of the methods will be described in this chapter; readers interested in more details are referred to a review by Andersen and Keiding (1998), to the books by Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984), Collett (1994), Marubini and Valsecchi (1995), Parmar and Machin (1995), Kleinbaum (1996) and Klein and Moeschberger (1997), and to the two papers by Peto *et al.* (1976, 1977). Software for performing the computations are reviewed by Goldstein and Harrell (1998).

A second situation where survival analysis has been used occurs in the study of occupational mortality where it is required to assess if a group of workers who are exposed to a pollutant are experiencing excess mortality. Subjects enter the study when healthy and, for this reason, a common method of analysis has been the comparison of observed mortality, both in timing and cause, with what would be expected if the study group were subject to a similar mortality to that

of the population of which it is a part. Discussion of this situation is deferred until §19.7.

Many of the methods of analysis are based on a life-table approach and in the next section the life-table is described.

17.2 Life-tables

The *life-table*, first developed adequately by E. Halley (1656–1742), is one of the basic tools of vital statistics and actuarial science. Standardization is introduced in §19.3 as a method of summarizing a set of age-specific death rates, thus providing a composite measure of the mortality experience of a community at all ages and permitting useful comparison with the experience of other groups of people. The life-table is an alternative summarizing procedure with rather similar attributes. Its purpose is to exhibit the pattern of survival of a group of individuals subject, throughout life, to the age-specific rates in question.

There are two distinct ways in which a life-table may be constructed from mortality data for a large community; the two forms are usually called the *current life-table* and the *cohort* or *generation life-table*. The current life-table describes the survival pattern of a group of individuals, subject throughout life to the age-specific death rates currently observed in a particular community. This group is necessarily hypothetical. A group of individuals now aged 60 years will next year experience approximately the current mortality rate specific to ages 60–61; but those who survive another 10 years will, in the 11th year, experience not the *current* rate for ages 70–71 but the rate prevailing 10 years hence. The current life-table, then, is a convenient summary of current mortality rather than a description of the actual mortality experience of any group.

The method of constructing the current life-tables published in national sources of vital statistics or in those used in life assurance offices is rather complex (Chiang, 1984). A simplified approach is described by Hill and Hill (1991). The main features of the life-table can be seen from Table 17.1, the left side of which summarizes the English Life Table No. 10 based on the mortality of males in England and Wales in 1930–32. The second column gives q_x, the probability that an individual, alive at age x years exactly, will die before his or her next birthday. The third column shows l_x, the number of individuals out of an arbitrary 1000 born alive who would survive to their xth birthday. To survive for this period an individual must survive the first year, then the second, and so on. Consequently.

$$l_x = l_0\, p_0\, p_1 \ldots p_{x-1}, \tag{17.1}$$

where $p_x = 1 - q_x$. This formula can be checked from Table 17.1 for $x = 1$, but not subsequently because values of q_x are given here only for selected values of x; such a table is called an *abridged life-table*.

Table 17.1 Current and cohort abridged life-tables for men in England and Wales born around 1931.

Age (years) x	Current life-tables 1930–32			Cohort life-table, 1931 cohort
	Probability of death between age x and $x+1$ q_x	Life-table survivors l_x	Expectation of life $\overset{\circ}{e}_x$	Life-table survivors l_x
0	0·0719	1000	58·7	1000
1	0·0153	928·1	62·2	927·8
5	0·0034	900·7	60·1	903·6
10	0·0015	890·2	55·8	894·8
20	0·0032	872·4	46·8	884·2
30	0·0034	844·2	38·2	874·1
40	0·0056	809·4	29·6	861·8
50	0·0113	747·9	21·6	829·7
60	0·0242	636·2	14·4	––
70	0·0604	433·6	8·6	––
80	0·1450	162·0	4·7	––

The fourth column shows $\overset{\circ}{e}_x$, the expectation of life at age x. This is the mean length of additional life beyond age x of all the l_x people alive at age x. In a complete table $\overset{\circ}{e}_x$ can be calculated approximately as

$$\overset{\circ}{e}_x = (l_{x+1} + l_{x+2} + \ldots)/l_x + \frac{1}{2}, \qquad (17.2)$$

since the term in parentheses is the total number of years lived beyond age x by the l_x individuals if those dying between age y and age $y+1$ did so immediately after their yth birthday, and the $\frac{1}{2}$ is a correction to allow for the fact that deaths take place throughout each year of age, which very roughly adds half a year to the mean survival time.

The cohort life-table describes the actual survival experience of a group, or 'cohort', of individuals born at about the same time. Those born in 1900, for instance, are subject during their first year to the mortality under 1 year of age prevailing in 1900–01; if they survive to 10 years of age they are subject to the mortality at that age in 1910–11; and so on. Cohort life-tables summarize the mortality at different ages at the times when the cohort would have been at these ages. The right-hand side of Table 17.1 summarizes the l_x column from the cohort life-table for men in England and Wales born in the 5 years centred around 1931. As would be expected, the values of l_1 in the two life-tables are very similar, being dependent on infant mortality in about the same calendar years. At higher ages the values of l_x are greater for the cohort table because this is based on mortality rates at the higher ages which were experienced since 193. and which are lower than the 1931 rates.

Both forms of life-table are useful for vital statistical and epidemiological studies. Current life-tables summarize current mortality and may be used as an alternative to methods of standardization for comparisons between the mortality patterns of different communities. Cohort life-tables are particularly useful in studies of occupational mortality, where a group may be followed up over a long period of time (§19.7).

17.3 Follow-up studies

Many medical investigations are concerned with the survival pattern of special groups of patients—for example, those suffering from a particular form of malignant disease. Survival may be on average much shorter than for members of the general population. Since age is likely to be a less important factor than the progress of the disease, it is natural to measure survival from a particular stage in the history of the disease, such as the date when symptoms were first reported or the date on which a particular operation took place.

The application of life-table methods to data from follow-up studies of this kind will now be considered in some detail. In principle the methods are applicable to situations in which the critical end-point is not death, but some non-fatal event, such as the recurrence of symptoms and signs after a remission, although it may not be possible to determine the precise time of recurrence, whereas the time of death can usually be determined accurately. Indeed, the event may be favourable rather than unfavourable; the disappearance of symptoms after the start of treatment is an example. The discussion below is in terms of survival after an operation.

At the time of analysis of such a follow-up study patients are likely to have been observed for varying lengths of time, some having had the operation a long time before, others having been operated on recently. Some patients will have died, at times which can usually be ascertained relatively accurately; others are known to be alive at the time of analysis; others may be lost to follow-up for various reasons between one examination and the next; others may have had to be withdrawn from the study for medical reasons—perhaps by the intervention of some other disease or an accidental death.

If there were no complications like those just referred to, and if every patient were followed until the time of death, the construction of a life-table in terms of time after operation would be a simple matter. The life-table survival rate, l_x, is l_0 times the proportion of survival times greater than x. The problem would be merely that of obtaining the distribution of survival time—a very elementary task. To overcome the complications of incomplete data, a table like Table 17.2 is constructed.

This table is adapted from that given by Berkson and Gage (1950) in one of the first papers describing the method. In the original data, the time intervals

Table 17.2 Life-table calculations for patients with a particular form of malignant disease, adapted from Berkson and Gage (1950).

(1) Interval since operation (years) x to $x+1$	(2) Last reported during this interval Died d_x	(3) Last reported during this interval Withdrawn w_x	(4) Living at start of interval n_x	(5) Adjusted number at risk n'_x	(6) Estimated probability of death q_x	(7) Estimated probability of survival p_x	(8) Percentage of survivors after x years l_x
0–1	90	0	374	374·0	0·2406	0·7594	100·0
1–2	76	0	284	284·0	0·2676	0·7324	75·9
2–3	51	0	208	208·0	0·2452	0·7548	55·6
3–4	25	12	157	151·0	0·1656	0·8344	42·0
4–5	20	5	120	117·5	0·1702	0·8298	35·0
5–6	7	9	95	90·5	0·0773	0·9227	29·1
6–7	4	9	79	74·5	0·0537	0·9463	26·8
7–8	1	3	66	64·5	0·0155	0·9845	25·4
8–9	3	5	62	59·5	0·0504	0·9496	25·0
9–10	2	5	54	51·5	0·0388	0·9612	23·7
10	21	26	47	—	—	—	22·8

were measured from the time of hospital discharge, but for purposes of exposition we have changed these to intervals following operation. The columns (1)–(8) are formed as follows.

(1) The choice of time intervals will depend on the nature of the data. In the present study estimates were needed of survival rates for integral numbers of years, to 10, after operation. If survival after 10 years had been of particular interest, the intervals could easily have been extended beyond 10 years. In that case, to avoid the table becoming too cumbersome it might have been useful to use 2-year intervals for at least some of the groups. Unequal intervals cause no problem; for an example, see Merrell and Shulman (1955).

(2) and (3) The patients in the study are now classified according to the time interval during which their condition was last reported. If the report was of a death, the patient is counted in column (2); patients who were alive at the last report are counted in column (3). The term 'withdrawn' thus includes patients recently reported as alive, who would continue to be observed at future follow-up examinations, and those who have been lost to follow-up for some reason.

(4) The numbers of patients living at the start of the intervals are obtained by cumulating columns (2) and (3) from the foot. Thus, the number alive at 10 years is $21 + 26 = 47$. The number alive at 9 years includes these 47 and also the $2 + 5 = 7$ died or withdrawn in the interval 9–10 years; the entry is therefore $47 + 7 = 54$.

(5) The adjusted number at risk during the interval x to $x + 1$ is

$$n'_x = n_x - \tfrac{1}{2}w_x. \tag{17.3}$$

The purpose of this formula is to provide a denominator for the next column. The rationale is discussed below.

(6) The estimated probability of death during the interval x to $x + 1$ is

$$q_x = d_x/n'_x. \tag{17.4}$$

For example, in the first line,

$$q_0 = 90/374\cdot0 = 0\cdot2406.$$

The adjustment from n_x to n'_x is needed because the w_x withdrawals are necessarily at risk for only part of the interval. It is possible to make rather more sophisticated allowance for the withdrawals, particularly if the point of withdrawal during the interval is known. However, it is usually quite adequate to assume that the withdrawals have the same effect as if half of them were at risk for the whole period; hence the adjustment (17.3). An alternative argument is that, if the w_x patients had *not* withdrawn, we might have expected about $\tfrac{1}{2}q_x w_x$ extra deaths. The total number of deaths would then have been $d_x + \tfrac{1}{2}q_x w_x$ and we should have had an estimated death rate

$$q_x = \frac{d_x + \tfrac{1}{2}q_x w_x}{n_x}. \tag{17.5}$$

(17.5) is equivalent to (17.3) and (17.4).

(7) $p_x = 1 - q_x$.

(8) The estimated probability of survival to, say, 3 years after the operation is $p_0\, p_1\, p_2$. The entries in the last column, often called the *life-table survival rates*, are thus obtained by successive multiplication of those in column (7), with an arbitrary multiplier $l_0 = 100$. Formally,

$$l_x = l_0 p_0 p_1 \ldots p_{x-1}, \tag{17.6}$$

as in (17.1).

Two important assumptions underlie these calculations. First, it is assumed that the withdrawals are subject to the same probabilities of death as the non-withdrawals. This is a reasonable assumption for withdrawals who are still in the study and will be available for future follow-up. It may be a dangerous assumption for patients who were lost to follow-up, since failure to examine a patient for any reason may be related to the patient's health. Secondly, the various values of p_x are obtained from patients who entered the study at different points of time. It must be assumed that these probabilities remain reasonably constant over time;

otherwise the life-table calculations represent quantities with no simple interpretation.

In Table 17.2 the calculations could have been continued beyond 10 years. Suppose, however, that d_{10} and w_{10} had both been zero, as they would have been if no patients had been observed for more than 10 years. Then n_{10} would have been zero, no values of q_{10} and p_{10} could have been calculated and, in general, no value of l_{11} would have been available unless l_{10} were zero (as it would be if any one of $p_0, p_1, \ldots, p_9$ were zero), in which case l_{11} would also be zero. This point can be put more obviously by saying that no survival information is available for periods of follow-up longer than the maximum observed in the study. This means that the expectation of life (which implies an indefinitely long follow-up) cannot be calculated from follow-up studies unless the period of follow-up, at least for some patients, is sufficiently long to cover virtually the complete span of survival. For this reason the life-table survival rate (column (8) of Table 17.2) is a more generally useful measure of survival. Note that the value of x for which $l_x = 50\%$ is the *median* survival time; for a symmetric distribution this would be equal to the expectation of life.

For further discussion of life-table methods in follow-up studies, see Berkson and Gage (1950), Merrell and Shulman (1955), Cutler and Ederer (1958) and Newell *et al.* (1961).

17.4 **Sampling errors in the life-table**

Each of the values of p_x in a life-table calculation is subject to sampling variation. Were it not for the withdrawals the variation could be regarded as binomial, with a sample size n_x. The effect of withdrawals is approximately the same as that of reducing the sample size to n'_x. The variance of l_x is given approximately by the following formula due to Greenwood (1926), which can be obtained by taking logarithms in (17.6) and using an extension of (5.20).

$$\text{var}(l_x) = l_x^2 \sum_{i=0}^{x-1} \frac{d_i}{n'_i(n'_i - d_i)}. \tag{17.7}$$

In Table 17.2, for instance, where $l_4 = 35 \cdot 0\%$,

$$\text{var}(l_4) = (35 \cdot 0)^2 \left[\frac{90}{(374)(284)} + \frac{76}{(284)(208)} + \frac{51}{(208)(157)} + \frac{25}{(151)(126)} \right]$$
$$= 6 \cdot 14$$

so that $\text{SE}(l_4) = \sqrt{6 \cdot 14} = 2 \cdot 48$, and approximate 95% confidence limits for l_4 are

$$35 \cdot 0 \pm (1 \cdot 96)(2 \cdot 48) = 30 \cdot 1 \text{ and } 39 \cdot 9.$$

Application of (17.7) can lead to impossible values for confidence limits outside the range 0 to 100%. An alternative that avoids this is to apply the double-log transformation, $\ln(-\ln l_x)$, to (17.6), with $l_0 = 1$, so that l_x is a proportion with permissible range 0 to 1 (Kalbfleisch & Prentice, 1980). Then Greenwood's formula is modified to give 95% confidence limits for l_x of

$$l_x^{\exp(\pm 1 \cdot 96s)}, \tag{17.8}$$

where

$$s = \text{SE}(l_x)/(-l_x \ln l_x).$$

For the above example, $l_4 = 0 \cdot 35, \text{SE}(l_4) = 0 \cdot 0248, s = 0 \cdot 0675, \exp(1 \cdot 96s) = 1 \cdot 14, \exp(-1 \cdot 96s) = 0 \cdot 876$, and the limits are $0 \cdot 35^{1 \cdot 14}$ and $0 \cdot 35^{0 \cdot 876}$, which equal $0 \cdot 302$ and $0 \cdot 399$. In this case, where the limits using (17.7) are not near either end of the permissible range, (17.8) gives almost identical values to (17.7).

Peto *et al.* (1977) give a formula for $\text{SE}(l_x)$ that is easier to calculate than (17.7):

$$\text{SE}(l_x) = l_x \sqrt{[(1 - l_x)/n'_x]}. \tag{17.9}$$

As in (17.8), it is essential to work with l_x as a proportion. In the example, (17.9) gives $\text{SE}(l_4) = 0 \cdot 0258$. Formula (17.9) is conservative but may be more appropriate for the period of increasing uncertainty at the end of life-tables when there are few survivors still being followed.

Methods for calculating the sampling variance of the various entries in the life-table, including the expectation of life, are given by Chiang (1984, Chapter 8).

17.5 **The Kaplan–Meier estimator**

The estimated life-table given in Table 17.2 was calculated after dividing the period of follow-up into time intervals. In some cases the data may only be available in group form and often it is convenient to summarize the data into groups. Forming groups does, however, involve an arbitrary choice of time intervals and this can be avoided by using a method due to Kaplan and Meier (1958). In this method the data are, effectively, regarded as grouped into a large number of short time intervals, with each interval as short as the accuracy of recording permits. Thus, if survival is recorded to an accuracy of 1 day then time intervals of 1-day width would be used. Suppose that at time t_j there are d_j deaths and that just before the deaths occurred there were n'_j subjects surviving. Then the estimated probability of death at time t_i is

$$q_{t_j} = d_j/n'_j. \tag{17.10}$$

This is equivalent to (17.4). By convention, if any subjects are censored at time t_j, then they are considered to have survived for longer than the deaths at time t_j and adjustments of the form of (17.3) are not applied. For most of the time intervals $d_j = 0$ and hence $q_{t_j} = 0$ and the survival probability $p_{t_j} (= 1 - q_{t_j}) = 1$. These intervals may be ignored in calculating the life-table survival using (17.6). The survival at time t, l_t, is then estimated by

$$l_t = \prod_j p_{t_j} = \prod_j \frac{n'_j - d_j}{n'_j}, \tag{17.11}$$

where the product is taken over all time intervals in which a death occurred, up to and including t. This estimator is termed the *product-limit* estimator because it is the limiting form of the product in (17.6) as the time intervals are reduced towards zero. The estimator is also the maximum likelihood estimator. The estimates obtained are invariably expressed in graphical form. The survival curve consists of horizontal lines with vertical steps each time a death occurred (see Fig. 17.1 on p. 580). The calculations are illustrated in Table 17.4 (p. 579).

17.6 **The logrank test**

The test described in this section is used for the comparison of two or more groups of survival data. The first step is to arrange the survival times, both observed and censored, in rank order. Suppose, for illustration, that there are two groups, A and B. If at time t_j there were d_j deaths and there were n'_{jA} and n'_{jB} subjects alive just before t_j in groups A and B, respectively, then the data can be arranged in a 2×2 table:

	Died	Survived	Total
Group A	d_{jA}	$n'_{jA} - d_{jA}$	n'_{jA}
Group B	d_{jB}	$n'_{jB} - d_{jB}$	n'_{jB}
Total	d_j	$n'_j - d_j$	n'_j

Except for tied survival times, $d_j = 1$ and each of d_{jA} and d_{jB} is 0 or 1. Note also that if a subject is censored at t_j then that subject is considered at risk at that time and so included in n'_j.

On the null hypothesis that the risk of death is the same in the two groups, then we would expect the number of deaths at any time to be distributed between the two groups in proportion to the numbers at risk. That is,

$$\left. \begin{aligned} \mathrm{E}(d_{jA}) &= n'_{jA} d_j/n'_j, \\ \mathrm{var}(d_{jA}) &= \frac{d_j(n'_j - d_j)n'_{jA}n'_{jB}}{n'^2_j(n'_j - 1)} \end{aligned} \right\}. \tag{17.12}$$

In the case of $d_j = 1$, (17.12) simplifies to

$$\mathrm{E}(d_{jA}) = p'_{jA},$$
$$\mathrm{var}(d_{jA}) = p'_{jA}(1 - p'_{jA}),$$

where $p'_{jA} = n'_{jA}/n'_j$, the proportion of survivors who are in group A.

The difference between d_{jA} and $\mathrm{E}(d_{jA})$ is evidence against the null hypothesis. The logrank test is the combination of these differences over all the times at which deaths occurred. It is analogous to the Mantel–Haenszel test for combining data over strata (see §15.6) and was first introduced in this way (Mantel, 1966).

Summing over all times of death, t_j, gives

$$\left. \begin{aligned} O_A &= \sum d_{jA} \\ E_A &= \sum \mathrm{E}(d_{jA}) \\ V_A &= \sum \mathrm{var}(d_{jA}) \end{aligned} \right\}. \tag{17.13}$$

Similar sums can be obtained for group B and it follows from (17.12) that $E_A + E_B = O_A + O_B$.

E_A may be referred to as the 'expected' number of deaths in group A but since, in some circumstances, E_A may exceed the number of individuals starting in the group, a more accurate description is the *extent of exposure to risk of death* (Peto *et al.*, 1977). A test statistic for the equivalence of the death rates in the two groups is

$$X_1^2 = \frac{(O_A - E_A)^2}{V_A}, \tag{17.14}$$

which is approximately a $\chi^2_{(1)}$. An alternative and simpler test statistic, which does not require the calculation of the variance terms, is

$$X_2^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}. \tag{17.15}$$

This statistic is also approximately a $\chi^2_{(1)}$. In practice (17.15) is usually adequate, but it errs on the conservative side (Peto & Pike, 1973).

The logrank test may be generalized to more than two groups. The extension of (17.14) involves the inverse of the variance–covariance matrix of the $O - E$ over the groups (Peto & Pike, 1973), but the extension of (17.15) is straightforward. The summation in (17.15) is extended to cover all the groups, with the

quantities in (17.13) calculated for each group in the same way as for two groups. The test statistic would have $k - 1$ degrees of freedom (DF) if there were k groups.

The ratios O_A/E_A and O_B/E_B are referred to as the relative death rates and estimate the ratio of the death rate in each group to the death rate among both groups combined. The ratio of these two relative rates estimates the death rate in Group A relative to that in Group B, sometimes referred to as the *hazard ratio*. The hazard ratio and sampling variability are given by

$$\left.\begin{aligned} h &= \frac{O_A/E_A}{O_B/E_B} \\ \mathrm{SE}[\ln(h)] &= \sqrt{\left(\frac{1}{E_A} + \frac{1}{E_B}\right)} \end{aligned}\right\}. \tag{17.16}$$

An alternative estimate is

$$\left.\begin{aligned} h &= \exp\left(\frac{O_A - E_A}{V_A}\right) \\ \mathrm{SE}[\ln(h)] &= \sqrt{\frac{1}{V_A}} \end{aligned}\right\} \tag{17.17}$$

(Machin & Gardner, 1989). Formula (17.17) is similar to (4.33). Both (17.16) and (17.17) are biased, and confidence intervals based on the standard errors (SE) will have less than the nominal coverage, when the hazard ratio is not close to unity. Formula (17.16) is less biased and is adequate for h less than 3, but for larger hazard ratios an adjusted standard error may be calculated (Berry *et al.*, 1991) or a more complex analysis might be advisable (§17.8).

Example 17.1

In Table 17.3 data are given of the survival of patients with diffuse histiocytic lymphoma according to stage of tumour. Survival is measured in days after entry to a clinical trial. There was little difference in survival between the two treatment groups, which are not considered in this example.

The calculations of the product-limit estimate of the life-table are given in Table 17.4 for the stage 3 group and the comparison of the survival for the two stages is shown in Fig. 17.1. It is apparent that survival is longer, on average, for patients with a stage 3 tumour than for those with stage 4. This difference may be formally tested using the logrank test.

The basic calculations necessary for the logrank test are given in Table 17.5. For brevity, only deaths occurring at the beginning and end of the observation period are shown. The two groups are indicated by subscripts 3 and 4, instead of A and B used in the general description

Table 17.3 Survival of patients with diffuse hystiocytic lymphoma according to stage of tumour (data abstracted from McKelvey *et al.*, 1976).

	Survival (days)							
Stage 3	6	19	32	42	42	43*	94	126*
	169*	207	211*	227*	253	255*	270*	310*
	316*	335*	346*					
Stage 4	4	6	10	11	11	11	13	17
	20	20	21	22	24	24	29	30
	30	31	33	34	35	39	40	41*
	43*	45	46	50	56	61*	61*	63
	68	82	85	88	89	90	93	104
	110	134	137	160*	169	171	173	175
	184	201	222	235*	247*	260*	284*	290*
	291*	302*	304*	341*	345*			

*Still alive (censored value).

Table 17.4 Calculation of product-limit estimate of life-table for stage 3 tumour data of Table 17.3.

Time (days) t_j	Died d_j	Living at start of day n_j'	Estimated probability of: Death q_{t_j}	Estimated probability of: Survival p_{t_j}	Percentage of survivors at end of day l_{t_j}
0	—	19	—	—	100·0
6	1	19	0·0526	0·9474	94·7
19	1	18	0·0556	0·9444	89·5
32	1	17	0·0588	0·9412	84·2
42	2	16	0·1250	0·8750	73·7
94	1	13	0·0769	0·9231	68·0
207	1	10	0·1000	0·9000	61·2
253	1	7	0·1429	0·8571	52·5

Applying (17.14) gives

$$X_1^2 = (8 - 16 \cdot 6870)^2 / 11 \cdot 2471$$
$$= 8 \cdot 6870^2 / 11 \cdot 2471$$
$$= 6 \cdot 71 \quad (P = 0 \cdot 010).$$

To calculate (17.15) we first calculate E_4, using the relationship $O_3 + O_4 = E_3 + E_4$. Thus $E_4 = 37 \cdot 3130$ and

$$X_2^2 = 8 \cdot 6870^2 (1/16 \cdot 6870 + 1/37 \cdot 3130)$$
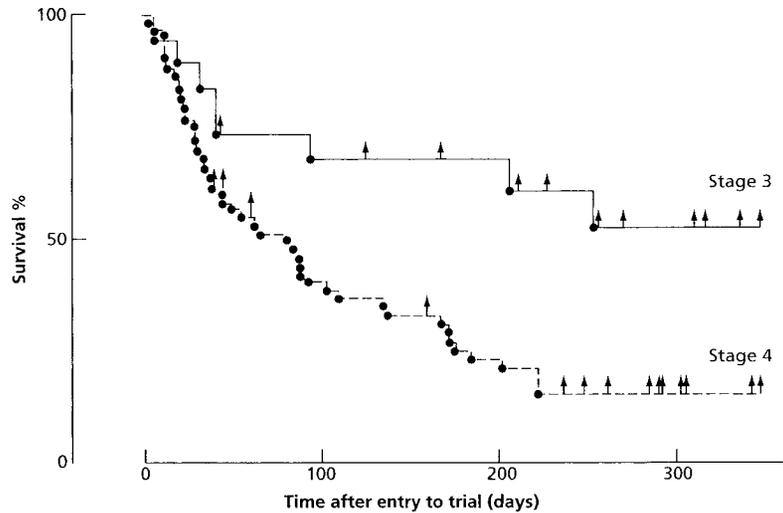$$= 6 \cdot 54 \quad (P = 0 \cdot 010).$$

Fig. 17.1 Plots of Kaplan–Meier product-limit estimates of survival for patients with stage 3 or stage 4 lymphoma. ● times of death. ⊬ censored times of survivors.

Table 17.5 Calculation of logrank test (data of Table 17.3) to compare survival of patients with tumours of stages 3 and 4.

Days when deaths occurred	Numbers at risk		Deaths			
	n'_3	n'_4	d_3	d_4	$E(d_3)$	$var(d_3)$
4	19	61	0	1	0·2375	0·1811
6	19	60	1	1	0·4810	0·3606
10	18	59	0	1	0·2338	0·1791
11	18	58	0	3	0·7105	0·5278
13	18	55	0	1	0·2466	0·1858
17	18	54	0	1	0·2500	0·1875
19	18	53	1	0	0·2535	0·1892
20	17	53	0	2	0·4857	0·3624
⋮						
201	10	12	0	1	0·4545	0·2479
207	10	11	1	0	0·4762	0·2494
222	8	11	0	1	0·4211	0·2438
253	7	8	1	0	0·4667	0·2489
Total			8	46	16·6870	11·2471
			O_3	O_4	E_3	V_3

Thus it is demonstrated that the difference shown in Fig. 17.1 is unlikely to be due to chance.

The relative death rates are $8/16·6870 = 0·48$ for the stage 3 group and $46/37·3130 = 1·23$ for the stage 4 group. The ratio of these rates estimates the death rate of stage 4 relative to that of stage 3 as $1·23/0·48 = 2·57$. Using (17.16), $\text{SE}[\ln(h)] = 0·2945$ and the 95% confidence interval for the hazard ratio is $\exp[\ln(2.57) \pm 1·96 \times 0·2945] = 1·44$ to $4·58$. Using (17.17), the hazard ratio is $2·16$ (95% confidence interval $1·21$ to $3·88$).

The logrank test can be extended to take account of a covariate that divides the total group into strata. The rationale is similar to that discussed in §§15.6 and 15.7 (see (15.20) to (15.23)). That is, the quantities in (17.13) are summed over the strata before applying (17.14) or (17.15). Thus, denoting the strata by h, (17.14) becomes

$$X_1^2 = \frac{\left(\sum_h O_A - \sum_h E_A\right)^2}{\sum_h V_A}. \tag{17.18}$$

As in analogous situations in Chapter 15 (see discussion after (15.23)), stratification is usually only an option when the covariate structure can be represented by just a few strata. When there are several variables to take into account, or a continuous variable which it is not convenient to categorize, then methods based on stratification become cumbersome and inefficient, and it is much preferable to use regression methods (§17.8).

The logrank test is a non-parametric test. Other tests can be obtained by modifying Wilcoxon's rank sum test (§10.3) so that it can be applied to compare survival times for two groups in the case where some survival times are censored (Cox & Oakes, 1984, p. 124). The generalized Wilcoxon test was originally proposed by Gehan (1965) and is constructed by using weights in the summations of (17.13). Gehan's proposal was that the weight is the total number of survivors in each group. These weights are dependent on the censoring and an alternative avoiding this is to use an estimator of the combined survivor function (Prentice, 1978). If none of the observations were censored, then this test is identical to the Wilcoxon rank sum test. The logrank test is unweighted—that is, the weights are the same for every death. Consequently the logrank test puts more weight on deaths towards the end of follow-up when few individuals are surviving, and the generalized Wilcoxon test tends to be more sensitive than the logrank test in situations where the ratio of hazards is higher at early survival times than at late ones. The logrank test is optimal under the proportional-hazards assumption, that is, where the ratio of hazards is constant at all survival times (§17.8). Intermediate systems of weights have been proposed, in particular that the weight is a power, ξ, between 0 and 1, of the number of survivors or the combined survivor function. For the generalized Wilcoxon test $\xi = 1$, for the

logrank test $\xi = 0$, and the square root, $\xi = \frac{1}{2}$, is intermediate (Tarone & Ware, 1977).

17.7 **Parametric methods**

In mortality studies the variable of interest is the survival time. A possible approach to the analysis is to postulate a distribution for survival time and to estimate the parameters of this distribution from the data. This approach is usually applied by starting with a model for the death rate and determining the form of the resulting survival time distribution.

The death rate will usually vary with time since entry to the study, t, and will be denoted by $\lambda(t)$; sometimes $\lambda(t)$ is referred to as the *hazard function*. Suppose the probability density of survival time is $f(t)$ and the corresponding distribution function is $F(t)$. Then, since the death rate is the rate at which deaths occur divided by the proportion of the population surviving, we have

$$\left. \begin{aligned} \lambda(t) &= \frac{f(t)}{1 - F(t)} \\ &= f(t)/S(t) \end{aligned} \right\} , \qquad (17.19)$$

where $S(t) = 1 - F(t)$ is the proportion surviving and is referred to as the *survivor function*.

Equation (17.19) enables $f(t)$ and $S(t)$ to be specified in terms of $\lambda(t)$. The general solution is obtained by integrating (17.19) with respect to t and noting that $f(t)$ is the derivative of $F(t)$ (§3.4). We shall consider certain cases. The simplest form is that the death rate is a constant, i.e. $\lambda(t) = \lambda$ for all t. Then

$$\lambda t = -\ln[S(t)]. \qquad (17.20)$$

That is,

$$S(t) = \exp(-\lambda t).$$

The survival time has an exponential distribution with mean $1/\lambda$. If this distribution is appropriate, then, from (17.20), a plot of the logarithm of the survivor function against time should give a straight line through the origin.

Data from a group of subjects consist of a number of deaths with known survival times and a number of survivors for whom the censored length of survival is known. These data can be used to estimate λ, using the method of maximum likelihood (§14.2). For a particular value of λ, the likelihood consists of the product of terms $f(t)$ for the deaths and $S(t)$ for the survivors. The maximum likelihood estimate of λ, the standard error of the estimate and a

significance test against any hypothesized value are obtained, using the general method of maximum likelihood, although, in this simple case, the solution can be obtained directly without iteration.

The main restriction of the exponential model is the assumption that the death rate is independent of time. It would usually be unreasonable to expect this assumption to hold except over short time intervals. One way of overcoming this restriction is to divide the period of follow-up into a number of shorter intervals, and assume that the hazard rate is constant within each interval but that it is different for the different intervals (Holford, 1976).

Another method of avoiding the assumption that the hazard is constant is to use a different parametric model of the hazard rate. One model is the *Weibull*, defined by

$$\lambda(t) = \alpha \gamma t^{\gamma - 1}, \qquad (17.21)$$

where γ is greater than 1. This model has proved applicable to the incidence of cancer by age in humans (Cook *et al.*, 1969) and by time after exposure to a carcinogen in animal experiments (Pike, 1966). A third model is that the hazard increases exponentially with age, that is,

$$\lambda(t) = \alpha \exp(\beta t). \qquad (17.22)$$

This is the *Gompertz* hazard and describes the death rate from all causes in adults fairly well. A model in which the times of death are log-normally distributed has also been used but has the disadvantage that the associated hazard rate starts to decrease at some time.

17.8 **Regression and proportional-hazards models**

It would be unusual to analyse a single group of homogeneous subjects but the basic method may be extended to cope with more realistic situations by modelling the hazard rate to represent dependence on variables recorded for each subject as well as on time. For example, in a clinical trial it would be postulated that the hazard rate was dependent on treatment, which could be represented by one or more dummy variables (§11.7). Again, if a number of prognostic variables were known, then the hazard rate could be expressed as a function of these variables. In general, the hazard rate could be written as a function of both time and the covariates, that is, as $\lambda(t, x)$, where x represents the set of covariates $(x_1, x_2, \ldots, x_p)$.

Zippin and Armitage (1966) considered one prognostic variable, x, the logarithm of white blood count, and an exponential survival distribution, with

$$\lambda(t, x) = (\alpha + \beta x)^{-1}; \qquad (17.23)$$