Survival Analysis

Introduction to Survival Data

In many studies the outcome for a patient, volunteer or experimental unit is some event. Examples would be:

a) clearance of psoriasis in a trial of alternative therapies;

b) metastatic spread in a study of the natural history of a particular cancer;

c) death, from any cause, of patients in a trial of different treatments for leukaemia;

d) death, due to leukaemia, in the above trial;

e) destruction of a sample in a study of the strength of dental material.

The events in these examples differ in the degree of certainty that can be ascribed to their ascertainment. Those in c) and e) are unequivocal, d) might be surprisingly uncertain, b) could be very difficult whereas a) would depend on the method used to determine 'clearance'. In what follows it will be taken for granted that the occurrence of the event can be accurately ascertained, although in practice this may not be the case and it may be necessary to take account of this.

In each of these studies an outcome that may be of considerable clinical or scientific importance is the time, measured from some appropriate origin, that elapses before the event occurs. It is this quantity that is known as the *survival time* or in some contexts, the *failure time*. Such times are either continuous variables or can be treated as continuous for all practical purposes. However, special methods are needed to analyse such data for two reasons, the first empirical, the second structural.

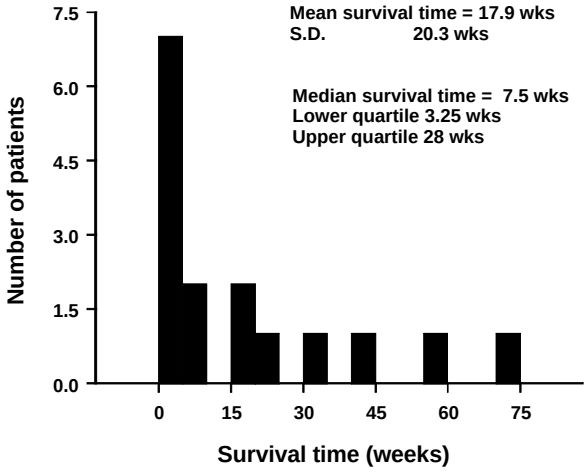

Figure 1: histogram of a typical sample of survival times

Skewness: although it is not inevitable, it is a matter of common experience that most sets of survival times are skewed and not at all suitable for the application of the usual methods based on Normal data. A typical graph of some survival times is given in figure 1. As often happens with highly skewed data, the SD exceeds the mean and clearly a Normality assumption is untenable; it predicts over 16% of the population would have survival time less than 0.

Censoring: the second problem is that during the period or locality of observation the event which defines the survival time will not happen to some of the patients or units in

the study. When the analysis is performed the survival times of such individuals are not fully observed; on the other hand these times are not totally unknown because they are known to be in excess of the time for which the individual has been observed. Such times are said to be *censored* and it is an important component of survival analysis that the information in censored times is taken into account fully.

The Survival Curve

In a *t*-test or regression the analysis is based around the estimation of and testing hypotheses about population parameters, which are numbers such as means, standard deviations or regression slopes. In a survival analysis the underlying population quantity is a curve rather than a single number, namely the *survival curve*. The survival function is denoted by $S(t)$, which is defined as:

$S(t)$ is the probability an individual survives more than time t

The survival curve is the plot of $S(t)$ (vertical axis) against t (horizontal axis). The definition has two general consequences: first, as an individual is certain to survive

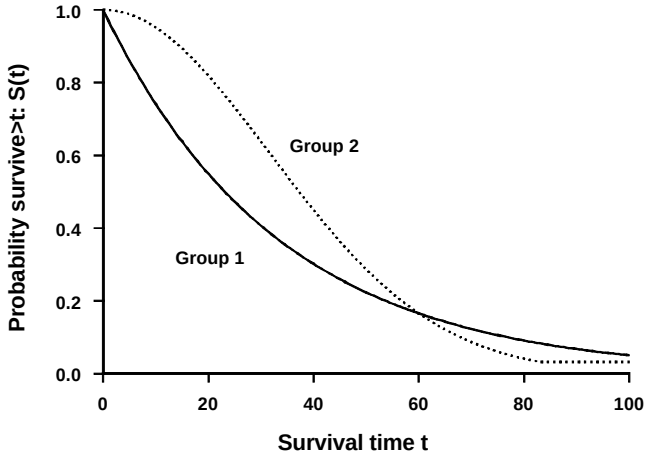

Figure 2: hypothetical examples of two population survival curves

more than zero time, $S(0) = 1$; second, as it is impossible for an individual to survive more than time t_2 without surviving all times $t_1 < t_2$, $S(t)$ must get smaller as t gets bigger. Figure 2 shows two examples of population survival curves.

The aims of the survival analyses to be described in the following sections are centred on the estimation of the survival curve and on the assessment of the equality of two survival curves. The methods will not make any assumptions about the distribution of the survival times, so the comparison of two curves can be thought of as the survival analogue of the Mann-Whitney test. More complicated methods are available, including parametric survival models (generally rather underused in medical applications) and *Cox regression models*, which feature in many parts of the literature. Both these are beyond the scope of this note: further details on Cox regression can be found in Altman (1991, p. 387-393).

Estimation of $S(t)$

The first requirement of all survival analyses is to produce an estimate of the survival curve. The method, known as the Kaplan-Meier or Product-Limit estimator, is explained in broad terms and then the procedures for its calculation are set out in more detail. Various packages, such as Stata or SAS, will produce this estimate for you but unfortunately it is not available in Minitab. It is actually easy to program in a spreadsheet such as Excel.

Background to the method.

The approach of the method is to compute the estimate sequentially. Suppose we have an estimate of $S(t)$ for times up to t_1 and we want to extend the estimate to time $t_2 > t_1$. A unit can survive beyond the later time only if it survives until the earlier time and then survives the interval between the two times, in symbols this is:

$$S(t_2) = \text{Prob(survive from } t_1 \text{ to } t_2 \text{)} \times S(t_1).$$

A simple estimate of the probability of surviving from t_1 to t_2 is $1-\frac{d}{n}$, where

d is the number failing between t_1 and t_2 and n is the number that could have failed in the interval. An important step in estimating $S(t)$ is to realise that if the interval between t_1 and t_2 contains no failures (d=0) then the estimate of $S(t)$ does not change. Therefore, we only need to update $S(t)$ at the times at which a failure is observed to occur, and then the formula displayed above is used.

The only complication to resolve is what to do with censored observations. It is assumed that if, e.g., an individual whose survival time was censored at 100 months had in fact failed at 50 months, then we would have recorded that failure. Thus, at any failure time, n in the formula includes any individual censored at that or a later time but excludes those censored earlier.

This is all there is to the Kaplan-Meier estimate. Its implementation is essentially straightforward and relies on accurate book-keeping of when individuals failed or were censored.

Computational Details

The computation of the Kaplan-Meier estimate is demonstrated table 1 using the following example data set of survival times (in days from entry to a trial) for patients with stage 3 diffuse hystiocytic lymphoma (from McKelvey *et al.*, 1976, *Cancer*, **38**, 1484-1493).

6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316*, 335*, 346*.

The times marked * are censored. It is useful to start the estimation of $S(t)$ by writing the data in ascending order, including the censored times in the ranking but ensuring that they are distinguishable from times that are not censored.

The table comprises five columns, the final two being found easily once the first three have been completed. The first column contains all the times at which at least one failure was observed, it does not contain the times at which only censoring occurred: it also contain the time 0. The second column indicates how many failures occurred at the corresponding failure time: in many tables this column predominantly

Failure time	No. failing d	No. at risk n	$1-\frac{d}{n}$	$S(t)$
0	-	19	-	1
6	1	19	0.9474	0.947
19	1	18	0.9444	0.895
32	1	17	0.9412	0.842
42	2	16	0.8750	0.737
94	1	13	0.9231	0.680
207	1	10	0.9000	0.612
253	1	7	0.8571	0.525

Table 1: calculations for Kaplan Meier estimator of $S(t)$

contains 1s. The third column contains the number that could have been seen to fail at the corresponding time.

The entries in the first two columns are easily found. The first row is also easily found: there are no failures and the number at risk is the total sample size: as $S(0) = 1$ for all survival functions, the value in the final column is known.

Subsequent rows for the third column are found by subtracting the value in the second column from that in the third column of the row above, that is the number at risk is the number that were at risk at the last failure time minus the number failing at that time. Thus in row 3, n=18 = 19 (n in previous row) - 1 (d in previous row). However, there is a complication, namely the number of subjects censored at any time from (and including) the previous failure time up to (but excluding) the present failure time must also be subtracted from the number at risk. So, in row 6 n is not 16 - 2 from the row above but 16 -2 -1 = 13, because a patient was censored at 43 days.

Once the first three columns have been found it is a matter of simple arithmetic to calculate the ratio in column 4. The entry at a given row of column 5 is the value in the previous row of column 5 multiplied by the value in the present row of column 4. It is because the evaluation is dependent on the previous row that it is necessary to set the process going by including the value 1 in the first row of column 5.

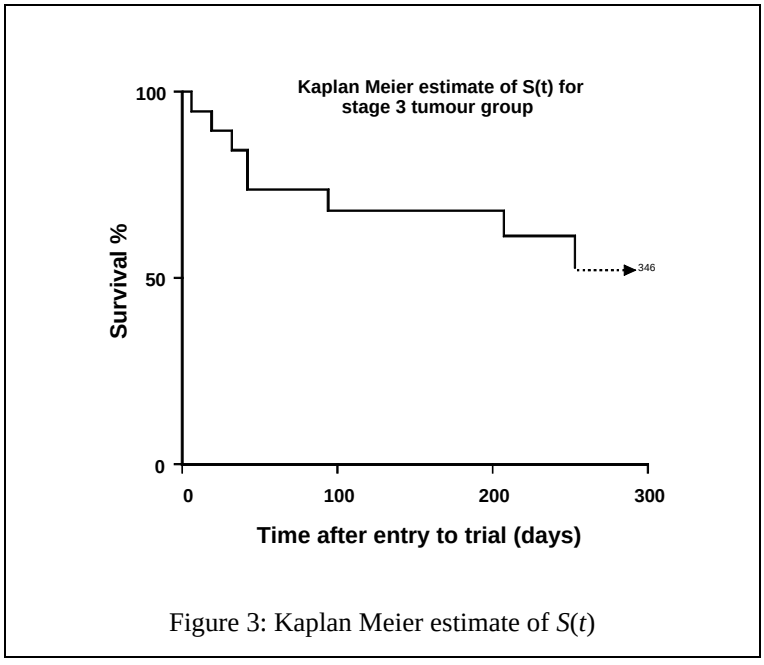

Figure 3: Kaplan Meier estimate of $S(t)$

The entries in column 5 form the Kaplan-Meier estimator; it is usual to plot $S(t)$ against t, remembering that the value of $S(t)$ at times intermediate between successive failures is the value at the earlier failure time (figure 3). Although the dataset contained times greater than 253 days, they were all censored. There is no general agreement about whether these points should be included in any plot, some authors extend the graph as a horizontal line until the

maximum censored time, others finish the graph at the largest failure time. In principle the former is more informative, although some judgment should be exercised because a large outlying censored time can distort a graph greatly.

Comparing two Survival Curves: the Log-rank test

There are many circumstances when it is required to ascertain whether or not there are differences in the survival experiences of two groups, perhaps patients in treatment groups after a clinical trial or with different prognoses, such as tumour stages. A naive approach would be to choose a time, e.g. 2 years, work out the proportions surviving beyond this time in the two groups and compare them using a χ^2 test. This has two disadvantages: first it is difficult to define the proportions in a way that is both efficient and unbiassed, largely due to the presence of censoring: see Altman (1991, section 13.5) for details. Second the choice of a time (such as 2 years) is arbitrary and it is potentially misleading to compare two survival curves at a single time.

Although the χ^2 test is incorrect, the associated 2×2 table is the basis of the correct method. The method overcomes the problems of censoring by using methods that are similar to those used in the Kaplan-Meier estimator of $S(t)$. It overcomes the problem of comparison at a single point by comparing at many time points simultaneously: again the thinking behind the Kaplan-Meier estimator is used as the times chosen are those a which a failure occurred, in either group.

The resulting test, called the log-rank test, tests the null hypothesis that the survival curves in the two groups are the same. If this null hypothesis is true, then the numbers failing in each group in each 2×2 table will be in the ratio of the numbers at risk in the groups. The expected numbers of failures so calculated in each group can be added across all times: the log-rank test proceeds by comparing these values with the actual numbers of failures in the two groups.

To be specific, the 2×2 table that is constructed at each time is given below:

	Failed	Surviving	At risk
Group A	d_A	n_A - d_A	n_A
Group B	d_B	n_B - d_B	n_B
Combined	d	n - d	n

If the null hypothesis is true then you expect the total number of failures d to be distributed between the two groups in the ratio of the group sizes. That is, the expected number of failures in group A at this time is

$$E_{At} = d \times \frac{n_A}{n}$$

The subscript t is added to make it clear that this is just the expected number of failures in group A *at one particular failure time*. The expected number of failures in group B is defined in a similar fashion but there is no need to compute it in the same way because it can be found simply by noting:

$$E_{Bt} = d \times \frac{n_B}{n} = d \times \frac{n - n_A}{n} = d - E_{At}:$$

from which it follows that the expected number in group B can be found as the number which ensures that the expected numbers of failures in both groups add up to the observed number.

From this table at a given failure time, the table at the next failure time is computed in a way analogous to the updating of the table used in the calculation of the Kaplan-Meier estimate. Updating is done separately in each group: the number at risk in the next table is the number at risk in the present table minus the number failing in the present table *minus the number of individuals censored between the present table and before the next table*.

Once tables have been found for each failure time the method continues by computing four quantities:

1. the total number of observed failures in group A, i.e. the sum of the d_A over all failure times, call this O_A
2. the total number of observed failures in group B, called O_B
3. the expected number of failures in group A, i.e. the sum of the E_{At} over all failure times *in either group*, called E_A.
4. the expected number of failures in group B, which is easily found from the above as $E_B = O_A + O_B - E_A$.

If the null hypothesis is true then the *O*s and the *E*s would be expected to be similar. The log rank test based on the above computes

$$\frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}, \qquad (*)$$

and the P-value is found by referring this to a χ^2 distribution with 1 degree of freedom.

An example: trial of 6-MP in treatment of leukaemia.

Table 2 contains the survival times (in weeks) of patients with leukaemia given 6-MP (group A) or control (group B), (Freireich, EJ *et al*. 1963, *Blood*, **21**, 699-716): as before * denotes a censored time

Group A (6-MP)	6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Group B (control)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Table 2: survival times, see text for details.

Although it is conceptually easier to think of the process in terms of 2×2 tables, it is cumbersome and unnecessary to carry this into the calculations. All that is needed is to record the numbers failing in each group at each failure time, together with the numbers at risk in each group at the same times. The calculation can then be set out in tabular form, as in table 3, with columns 1 & 2 and 4 & 5 containing the essential the data from the experiment. Column 7 can then be found by adding 1 & 4

and column 8 by adding 2 & 5; columns 3 and 6 can be found by multiplying respectively columns 2 and 5 by the ratio of column 7 to column 8.

The first row is at the time of the first failure, that is at week 1. At the second failure time, namely week 2, the number at risk in group B is 19, which comprises the 21 at risk at week 1 minus the two deaths at week 1. The first failure in group A occurred at week 6, when three deaths occurred. The next failure time is 7 weeks and the number at risk is 17, which is the 21 at risk in week 6 minus the 3 deaths at week 6, *minus the patient whose survival time was censored at 6 weeks.* The table is compiled in this way, stopping at the largest failure time which is not censored. The

Failure	d_{At}	n_{At}	E_{At}	d_{Bt}	n_{Bt}	E_{Bt}	d_t	n_t
Times	1	2	3	4	5	6	7	8
1	0	21	1.0	2	21	1.0	2	42
2	0	21	1.05	2	19	0.95	2	40
3	0	21	0.553	1	17	0.447	1	38
4	0	21	1.135	2	16	0.865	2	37
5	0	21	1.2	2	14	0.8	2	35
6	3	21	1.909	0	12	1.091	3	33
7	1	17	0.586	0	12	0.414	1	29
8	0	16	2.286	4	12	1.714	4	28
10	1	15	0.652	0	8	0.348	1	23
11	0	13	1.238	2	8	0.762	2	21
12	0	12	1.333	2	6	0.667	2	18
13	1	12	0.75	0	4	0.25	1	16
15	0	11	0.733	1	4	0.267	1	15
16	1	11	0.786	0	3	0.214	1	14
17	0	10	0.769	1	3	0.231	1	13
22	1	7	1.556	1	2	0.444	2	9
23	1	6	1.714	1	1	0.286	2	7

$O_A = 9$ $E_A = 19.251$ $O_B=21$ $E_B = 10.749$

$O_A/E_A = 0.468$ $O_B/E_B = 1.954$

Table 3: log rank test.

total number of observed and expected failures in each group is found as in table 3. The log-rank statistic can then be found from an application of the formula given above:

$$\frac{(9-19.25)^2}{19.25}+\frac{(21-10.75)^2}{10.75}=15.23,\ \mathrm{P} < 0.001,$$

so there is clear evidence that the survival curves for the two groups are different. This can be seen in figure 4, where the Kaplan-Meier estimators for the two groups are shown. In any survival analysis it is insufficient to supply just the results of a log-rank

test. It is necessary to provide the Kaplan-Meier estimators for all appropriate groups, and it is useful to provide some numerical estimate of the difference between the curves.

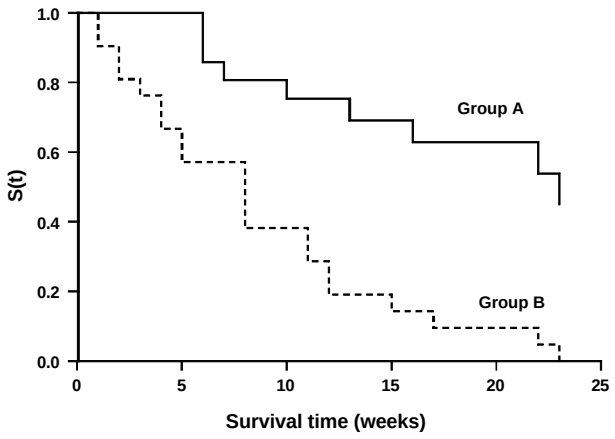

Figure 4: Kaplan-Meier estimators for two groups in table 2

The hazard ratio

The commonest estimate of the difference between two survival curves is known as the hazard ratio. The ratio of the number of failures observed to that expected assuming the null hypothesis, that is O/E, is called the relative failure rate. In this analysis the relative failure rates are 0.46 in group A and 1.95 in group B. The ratio of these two quantities is:

$$h = \frac{O_A / E_A}{O_B / E_B}$$

which is the hazard ratio, equal in this analysis to 0.46/1.95=0.24. This quantity expresses how much larger the failure rate is in group A than group B and, being less than 1, shows that group A has the more favourable survival rate.

An approximate 95% confidence interval for h can be found. As with odds ratios, the confidence interval is the back transformation of a confidence interval for $\log_e(h)$. The standard error of this quantity is approximately:

$$se(\log_e(h)) = \sqrt{\frac{1}{E_A} + \frac{1}{E_B}} = \sqrt{\frac{1}{19.25} + \frac{1}{10.75}} = 0.381,$$

so a 95% confidence interval for $\log_e(h)$ = -1.427 is $-1.427 \pm 1.96 \times 0.381 = (-2.173, -0.680)$, and back-transforming gives the confidence interval for h as (0.11, 0.51). This procedure is only approximate but gives reasonable answers if h is between a ⅓ and 3.

Extensions and variants of the Log-rank test

The log-rank test outlined above is the simplest version, designed to compare two groups as a whole. Even to do exactly the same job there are at least two other forms of the statistic which may be encountered in the literature. However, in most cases the difference between the above method and these variants is small and can be ignored.

There are several extensions of the log-rank test to accommodate different circumstances. The most obvious extension is to compare more than two groups: the method described works by computing the number of failures expected in each group under the null hypothesis and there is no reason why exactly the same method should not be extended to find the expected number of failures in any number of groups. The final test statistic is the obvious extension of (*), which is referred to a χ^2 distribution on g-1 degrees of freedom, where g is the number of groups. This extension can itself be extended to allow for ordering in the groups.

An extension of a different kind is to stratify the analysis: this may be desirable when the failure rates are expected to differ between different subgroups of patients, e.g. those with stage II or stage III tumours. This is a straightforward extension: the computations in table 3 are performed separately in each stratum and then the observed and expected numbers in each group are summed across the strata before being entered in the formula (*).

A good description of these extensions can be found in Altman (1991, section 13.4) A much more ambitious extension is to *Cox Regression*, which allows numerous groups to be compared, using several strata and also can allow for continuous variables which might influence survival, such as white blood cell count, and variables which change over time. This technique is widely used in medical research but is beyond the scope of this note, the interested reader is referred to section 13.6 of Altman (1991).

Reference

Altman, DG, (1991) *Practical Statistics for Medical Research: London*, Chapman and Hall