Research Methods 2

Week 14: Document 1

Survival Analysis.

Background

In several areas of medicine, including oncology and palliative care, the time to some event is an important variable. For example, in many clinical trials in oncology the principal outcome is the time a patient survives from the date they were enrolled. The event which defines the time need not be death – it could be the time to when metastatic disease is first detected, or it could be something more positive, such as the time it takes for the plaques to clear in a patient with psoriasis. However, regardless of the type of event used to define the time, this kind of data are habitually referred to as 'survival data', or as 'failure time data', and they are subjected to survival analysis.

General feature of survival data

In general survival data need specially developed statistical methods. Why is this? The main reasons are outlined below.

- One feature of survival data is that they are often very skewed and methods based on Normal distributions cannot be applied. This is not inevitably the case but something that experience suggests usually obtains; an example is shown below

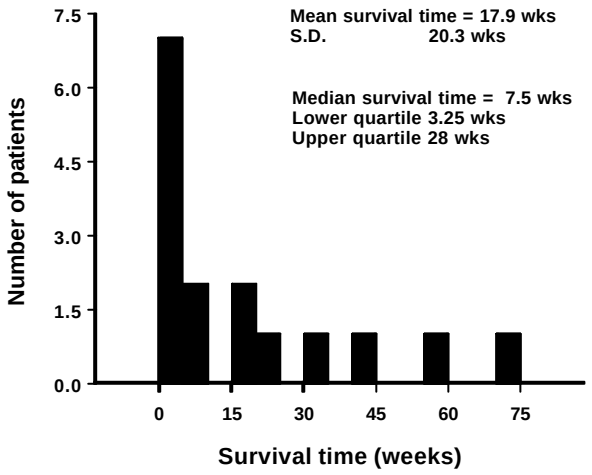

Figure 1: histogram of typical set of survival data.

- Perhaps the feature of survival data that is most obviously beyond the reach of general statistical methods is that some survival times are only partially observed. If you want to measure a patient's haemoglobin concentration at the next clinic visit, then you might obtain a sample but on the other hand you might not. The patient might not attend as planned (or you might drop the tube containing the sample). Whatever happened you will or will not get a haemoglobin concentration – there is no half-way house.
  This is not the case with survival data. If at the end of your study you observe a patient is still alive you do not know the survival time but you do know *something* about it. You know it exceeds the length of time you have already observed on the patient. So, for example, if a patient presented in January and was seen at clinic the following December but you closed the study at the end of the year, then you do not know the survival time of that patient, as they are still alive. However, you

do know, e.g., that the survival time was not 2 months or 6 months. You know that it was something in excess of 11 months. Times such as this, where the final event has not been observed, are known as *censored* observations.

- Not only do you encounter censored observations, you need your analysis to take them into account appropriately. If you omitted them then you would be losing information and if you treated them as ordinary observations, i.e. as if the event occurred just after you last saw them then you would almost certainly be underestimating the length of survival of your patients.

How to describe survival data

Another difference with survival data is that one of the most widely used ways to describe the data is through the survival curve. This contrasts with the situations you have met before in the course, where the aim has usually been to estimate a parameter, i.e. a single number.

The survival curve plots time *t* on the horizontal axis and a quantity denoted by *S*(*t*) on the vertical axis. The latter is defined as the probability an individual survives more than time *t*. An example of what the curves for two different populations of patients could look like is in figure 2.

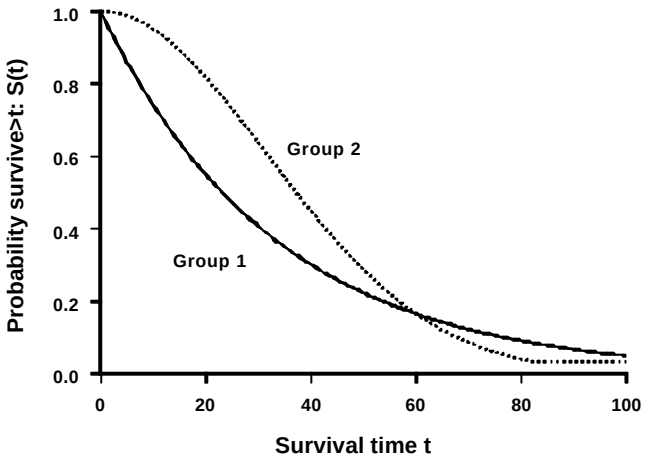

Figure 2: hypothetical examples of population survival curves.

Another way to think of the survival curve is that at time *t* you plot the proportion of your population that is still alive at time *t*.

If you knew the survival curve for the population under study, then you would be able to look up the proportion of patients who survive any time you chose to consider, such as a 12 month survival proportion, a three-year survival proportion or a five-year version.

There are some consequences of the definition of the survival curve.

i) The proportion of the population that survives beyond time 0 is 1 (or 100%). Therefore all survival curves start from 0 on the horizontal axis and 1 in the vertical axis (cf. Figure 2).

ii) All patients who survive six months have necessarily survived all other times less than six months. Consequently the proportion of patients surviving a given time must get smaller as the time increases. The curves in figure 2 illustrate this: they go down as the time increases. It is impossible for a survival curve to increase as time proceeds.

How to estimate the survival curve

Several methods for estimating survival curves are available, with different methods being pertinent to different forms of data. The most widely used method is called the *Kaplan-Meier estimator* (an alternative name is the product-limit estimator).

The method is most easily illustrated by means of an example. The following data set are 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, so, e.g., 43* means that when these data were collated this patient was still alive 43 days after entry to the trial.

The above times are entered into a column in Minitab, say C1, ignoring the *s. A second column, say C2, containing two values, say 0 and 1, with a 1 if the corresponding entry in C1 is a fully observed survival time and a 0 if it is a censored observation.

The estimated survival curve is found in Minitab by clicking on **Stat** -> **Reliability/Survival** and then on **Distribution Analysis (Right Censoring)**[†] and then select **Nonparametric Distribution Analysis ...** . The column containing the survival times needs to be entered into the **Variables:** box. Next the **Censor...** box should be clicked and the column containing the censoring indicator should be entered into the **Use censoring columns:** box and the value which indicates censored times should be entered in **Censoring value:**. Click on **OK** and then on **Graphs...** and in the resulting dialogue box check the **Survival plot** box and click **OK**. Then click on **OK** in the main box and you will obtain the following plot.

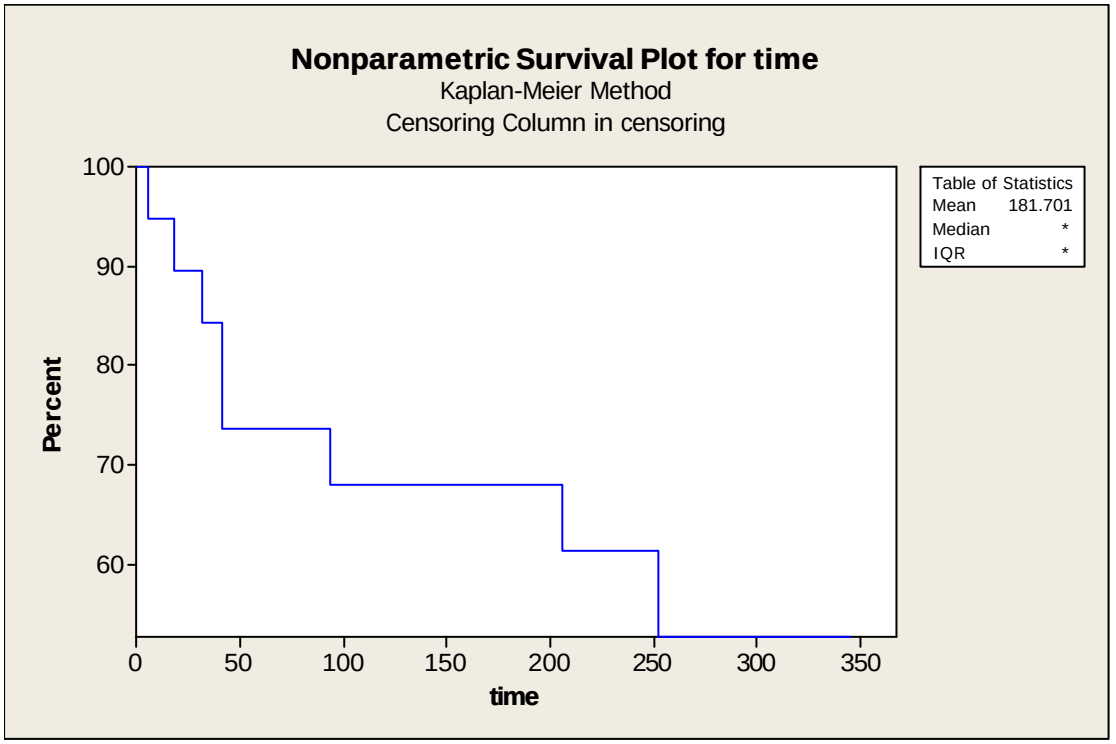

Figure 3: Kaplan-Meier survival plot of the hystiocytic lymphoma data.

This kind of curve is frequently seen in reports of oncology trials. It is worth making a few general comments.

[†] The form of censoring we have just met is more fully referred to as *Right Censoring*: Left Censoring, where we only know that a value is less than some observation will not concern us in this course.

i) It follows the basic requirements of a survival curve – at time 0 it is 1 and as you move to the right on the horizontal axis the plotted value never gets bigger.

ii) The curve has the strange step-like pattern because of the method underlying its calculation. The survival probability only changes when a death is observed. The curve changes level at times 6, 19, 32, 42 etc. It does not change level at censored values.

iii) The final change occurs when the last failure is observed. As there is no information about deaths after 253 days, the curve stops at this time.

iv) When censored survival times occurred cannot readily be seen on this plot. In some programs, small tick marks are placed on the curve at times corresponding to censored observations; see Bland, figure 15.7 for an example.

The curve can be used in the obvious way – the probability of surviving 100 days can be read off the curve as approximately 0.68. However, you can get the value precisely because in addition to the graph, Minitab produces the following output in the Session window.

```
Kaplan-Meier Estimates

      Number
          at  Number     Survival  Standard   95.0% Normal CI
Time    Risk  Failed  Probability     Error     Lower    Upper
   6      19       1     0.947368  0.051228  0.846964  1.00000
  19      18       1     0.894737  0.070406  0.756744  1.00000
  32      17       1     0.842105  0.083655  0.678145  1.00000
  42      16       2     0.736842  0.101023  0.538841  0.93484
  94      13       1     0.680162  0.107988  0.468510  0.89181
 207      10       1     0.612146  0.116659  0.383498  0.84079
 253       7       1     0.524696  0.128661  0.272526  0.77687
```

All the details of the table need not concern us but it is useful to consider the first and fourth columns. The first column is just the times at which deaths occur and the corresponding value in the fourth column is the survival probability shown on the Kaplan-Meier plot. So the value at 94 days is 0.6802, which is the value of the survival probability until the next death at 207 days, hence the proportion of patients surviving 100 days is estimated as 0.6802.

Comparing survival curves – the log-rank test and Cox regression

The Kaplan-Meier estimator is the most widely used estimator of the survival curve. Unlike the situation with Normally distributed data or binary data, where estimation of parameters is straightforward, we have just seen that the main estimator for survival analysis is quite complicated to compute. Partly for this reason, and partly because of the increasing technical complexity involved, we will not dwell long on further aspects of survival analysis, such as the survival analogues of hypothesis tests. However, one of these is frequently encountered in the literature and a brief description is in order.

If the survival experience of a group of patients is measured by the survival curve, then it is natural to compare the survival of groups of patients by comparing their respective survival curves. A technique which tests the null hypothesis that the survival curves in the different groups are the same is known as the *log-rank* test. If two survival curves are compared, then the log-rank test will provide a P-value to test the null hypothesis that the two population survival curves are the same.

Figure 4 shows the Kaplan-Meier plots for two groups of patients with leukaemia, those that are classed as AG+ and those that are AG-. The time is the time to death in weeks: data from Feigl and Zelen, 1965, *Biometrics*, 21, 826-838.

The log-rank test gives P= 0.0037. This indicates that whatever the reason for difference between the survival curves, it is very unlikely that the difference is simply due to chance.

The difference between the curves can also be described by a quantity called the *hazard ratio*, which is essentially the ratio between the risk of death in one group relative to the other. For the example below the hazard ratio is 2.4, i.e. the risk of death is 2.4 times larger in the AG- group than in the AG+ group.

It is also possible to compute a 95% confidence interval for this quantity, here it is 1.1 to 5.1. However the use of a hazard ratio embodies assumptions and a discussion of these is beyond our scope. See Bland, section 15.6 and references therein for a fuller discussion.

You will also see the term 'Cox regression' in the literature. This is a way of assessing how a variety of variables affect survival times. However, it leads to very deep water and we will not consider it here: see Bland section 17.9 for some details.

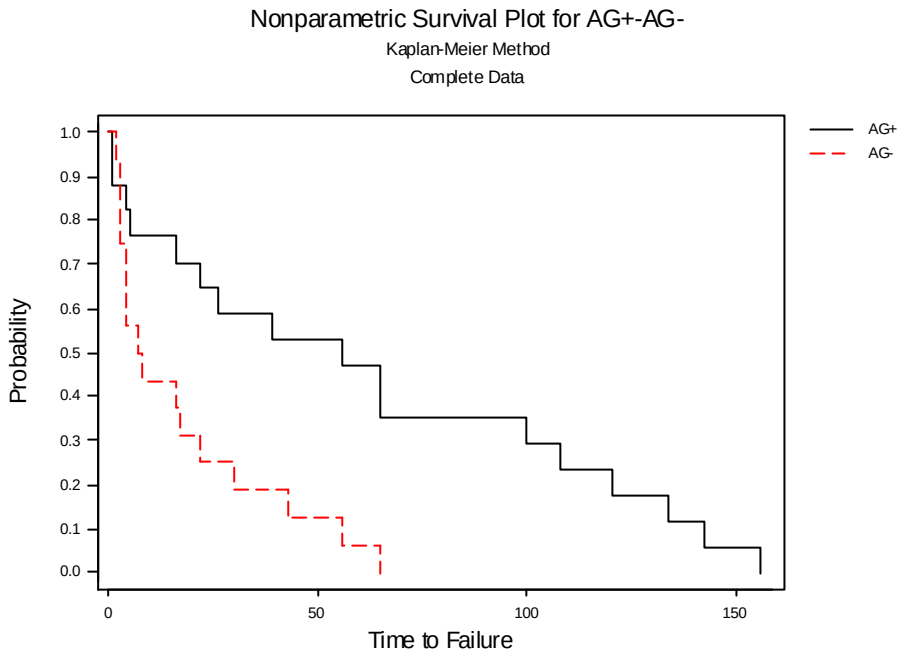

Figure 4: Kaplan-Meier plots for two groups (AG+ and AG-) with leukaemia (times in weeks)