What is survival data?

Survival data arise when the time from a defined time origin until the occurrence of a predetermined event, often termed failure, is measured for each subject.

Examples:

1. Time to death from cancer after diagnosis
2. Length of time in remission for leukaemia patients
3. Time to recurrence of a cancerous tumour after removal

How is survival data characterised and why it is not amenable to standard statistical techniques:

Survival data are characterised by three things:

1. having at most one event occurring per subject. That is, failure can occur only once
2. by times being highly skewed and positive
3. by having some subjects potentially being unobserved for the full time to failure, that is, censoring may occur

The fact that survival time data are highly skewed (usually positively skewed) means that standard statistical techniques, based on the assumption of normality, cannot be validly applied. This difficulty may sometimes be overcome by performing a transformation (e.g. log-transformation) to the times to obtain a more "symmetric/normal" distribution. However a more satisfactory solution is to use an alternative distributional model for the original survival time data. For example, we could assume an exponential, Weibull or a gamma distribution. (Note: We will not be discussing these parametric models in this lecture.)

However, it is censoring which makes standard statistical techniques unsuitable for analysing survival data. As mentioned before, censoring is said to occur when the end-point of interest has not been observed by the end of data collection. It occurs, for example, when some patients survive to the end of the trial investigating time to death; when a certain type of cancer does not recur after surgical removal; when a patient has died from an unrelated cause to the one being investigated; and when a patient is lost to follow-up.
In each of the situations described above, the censoring patterns were of the same form. That is, the individuals were observed failure-free for certain lengths of time, after which no more information about these individuals was known. Thus all that can be said about the true survival times of these individuals were that they exceeded their observed censored survival times. That is, the true survival times have known minima. This type of censoring is known as right censoring and it will be the only type of censoring considered here although a brief mention of left censoring will be made at the end of these notes.

Survival and hazard functions:

In survival analysis, there are two functions that are very important when looking at survival data. They are the survival or survivor function and the hazard function.

Let $T$ be a non-negative random variable ($T \geq 0$) with density $f(t)$ and cumulative distribution function $F(t)$. $T$ represents the survival time of an individual from a homogeneous population and $F(t)$ is interpreted as the probability of dying, before or at time $t$.

Now define the survivor function $S(t)$ to be

$$S(t) = 1 - F(t) = P(T > t)$$

That is, $S(t)$ is the probability of an individual surviving longer than time $t$. An equivalent way of describing $S(t)$ is the proportion of individuals, from a homogeneous population, whom survive after time $t$.

Define the hazard function, $h(t)$, to be the rate of failure at time $t$, given survival up to time $t$. That is,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

In other words, the hazard function is the probability of a death in the next small interval $\Delta t$ given that the individual has survived up to time $t$ divided by $\Delta t$. That is, it is a function that measures the instantaneous risk of failing per unit time - a rate not a probability.

Formally, let $T_1, \ldots, T_n$ be the times to failure of $n$ individuals, and assume that they are iid random variables each with distribution function $F(t)$. Let $C_1, \ldots, C_n$ be the right-censoring times associated with the $T_i$'s. We assume that they are iid with distribution function $G(t)$. We also assume that the censoring is uninformative. That
is, the $T_i$’s and $C_i$’s are independent. (Note: In practice this assumption should be checked carefully.)

Now in a clinical trial, say, we are only in a position to observe the pairs

$$(X_1, \delta_1), \ldots, (X_n, \delta_n)$$

where

$$X_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = 1(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq C_i, \text{ that is, } T_i \text{ is not censored} \\ 0 & \text{if } T_i > C_i, \text{ that is, } T_i \text{ is censored} \end{cases}$$

The $X_i$’s are the observed survival times and the $\delta_i$’s are indicator random variables that indicate whether the individuals were censored or not.

Now, interest lies in estimating the survival probabilities.

**Kaplan-Meier or product-limit estimator:**

Kaplan and Meier (1958) introduced the product-limit estimator (also known as Kaplan-Meier estimator):

$$\hat{S}(t) = \prod_{\text{distinct } X_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where $d_i$ is the number of observed failures (deaths) that occurred at $X_i$, $n_i$ is the number of observations (patients) in the risk set at time $X_i$ (i.e., the number of subjects whom are still alive just before time $X_i$) and the product is taken over all distinct observed survival times. In the case of a tie between censored and uncensored observations, we assume that the uncensored observation occurs before the censored observation.

Equivalently, to determine the Kaplan-Meier estimate of the survivor function from the survival data, a sequence of time bands is formed, where each of these time bands is constructed so that only one uncensored survival (death) time is contained in that interval, and this death time is taken to occur at the start of the interval.

The Kaplan-Meier estimate at time $t$ is then given by the product of terms of the form

$$\left(1 - \frac{d_i}{n_i}\right)$$

over all time intervals before or including $t$, where $d_i$ now represents the number of events (deaths) that occurred in the $i$th interval and $n_i$ represents the number of subjects at risk (alive) at the beginning of the $i$th interval. That is, the Kaplan-Meier estimate for the survivor function is given by

$$\hat{S}(t) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right).$$
where $t$ lies in the $j$th interval.

The Kaplan-Meier estimate can be represented in tabular or graphical form. More commonly displayed graphically. The curve is a graph in which the horizontal axis represents the survival times and the vertical axis represents the survival probabilities. The curve is a right-continuous step function, which will start from 1 (100% of patients alive) at time 0, and decline towards 0 (all patients have died) with increasing time. It is plotted as a step function, since the survival probability remains constant between successive uncensored times and only drops instantaneously at the time when an event (death) occurs. The graph will only reach 0 if the patient with the longest observed time has died.

**Interpreting a survival curve:**

The Kaplan-Meier curve is the best description of the survival times of patients in the study. The reading of the curve is straightforward. For example, if the estimated median survival time is to be reported (i.e., the time beyond which 50% of the patients in the study are expected to survive - $\hat{S}^{-1}(1/2)$) then this can be done by extending a horizontal line from the survival probability 0.5 on the vertical axis, to the point where the curve and the line intersects and then dropping a vertical line downwards to the time axis. The estimated median survival time is the time point where the vertical line cuts the time axis.

Note that the survival curve shows the *pattern of mortality* (if death is the outcome of interest) over time and *not the details*. Therefore any conclusions made based on the fine details of a curve are likely to be inaccurate. In particular, if the tail of the survival curve is flat, this does not mean that the risk to patients still alive is non-existent. In fact, this may occur because the number of patients still alive and being followed-up along the tail is small and therefore reliable estimates of the survival probabilities along the tail are not obtained. Also, drastic drops and large flat sections in other parts of the curve may be due to a large proportion of censored observations.

Hence the survival curve is unreliable if based on small numbers at risk. Also, there is a natural tendency for the eye to be drawn to the right-hand side of the curve where it is least reliable. Therefore it is wise not to place too much confidence in the fine details of the curve, unless there is a valid reason (based possibly on prior knowledge) to do so. The overall picture is more reliable! It is worthwhile to calculate the confidence bands for the whole survival curve (using Greenwood's variance formula) so as to help assess the uncertainty in the estimates of the survival probabilities.
Testing survival data - Comparing two or more groups (log-rank test):

In many clinical studies, the main aim of the study is to compare the survival times in two or more groups to determine, for example, whether the survival times of patients are systematically longer in the second group than in the first.

The simplest way of comparing the survival times obtained from two or more groups is to plot the Kaplan-Meier curves for these groups on the same graph. However, this graph does not allow us to say, with any confidence, whether or not there is a real difference between the groups. The observed difference may be a true difference, but equally, it could also be due merely to chance variation. Assessing whether or not there is a real difference between groups can only be done, with any degree of confidence, by utilising statistical tests. Below we describe briefly the log-rank test for comparing the differences between groups.

However before describing the log-rank test, we mention that it is not good practice to compare survival curves at specific time points. This is because of the following reasons:

1. we are usually interested in comparing the curves over the whole time spectrum rather than at particular points
2. comparing curves at specific time points may lead to a proliferation of hypothesis test and therefore problems of multiple testing, and
3. comparing curves at specific time points is open to abuse and potentially misleading. That is, a person may choose a particular time point (a posteriori) so as to support a particular hypothesis

Log-rank test:

The log-rank test, developed by Mantel and Haenszel, is a non-parametric test for comparing two or more independent survival curves. Since it is a non-parametric test, no assumptions about the data’s distributional form need be made. This test is however most powerful when used for non-overlapping survival curves.

The test is based on (assuming k groups)

1. the constructing of a 2×k table for each uncensored (death) time from the pooled data,

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>…</th>
<th>k</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death at $x_i$</td>
<td>$n_{i1}$ - $d_{i1}$</td>
<td>$d_{i1}$</td>
<td>…</td>
<td>$d_{ik}$</td>
<td>$d_i$</td>
</tr>
<tr>
<td>Surviving beyond $x_i$</td>
<td>$n_{i2} - d_{i2}$</td>
<td>$d_{i2}$</td>
<td>…</td>
<td>$n_{ik} - d_{ki}$</td>
<td>$n_i - d_i$</td>
</tr>
<tr>
<td>At risk just before $x_i$</td>
<td>$n_{i1}$</td>
<td>$n_{i2}$</td>
<td>…</td>
<td>$n_{ik}$</td>
<td>$n_i$</td>
</tr>
</tbody>
</table>
where \( d_g \) and \( n_g \) are the number of deaths that occurred in group \( g \) at \( x_i \) and the number at risk just before \( x_i \). \( d_i \) and \( n_i \) are the total observed deaths at \( x_i \) and the total number at risk just before \( x_i \).

2. calculating the expected number of deaths in each group, under the null hypothesis of no differences in the survival experiences between the \( k \) groups

3. the combining of the information from these tables to provide a test statistic, which is of a similar form to the familiar chi-squared test statistic

This test-statistic, under the null hypothesis, is asymptotically chi-squared with \((k-1)\) degrees of freedom. The null hypothesis is rejected if the test statistic is found to be "too" large.

Modelling survival data - Cox proportional hazards model:

The log-rank test just described is a useful way of comparing the survival experiences of two or more groups. However, it requires the assumption that within each level of the primary or exposure variable (e.g. treatment), the populations are homogeneous in survival experiences. This assumption is rarely realistic in medical studies, as populations are made heterogeneous (in survival experiences) by demographic variables and possible risk or prognostic factors. Furthermore the log-rank test can only detect differences, not quantify or describe the differences among the groups. Thus some way of quantifying these differences, and also including supplementary information on subjects into any analysis is required. The regression method introduced by D.R. Cox in 1972 is widely used to adjust for, as well as to assess the effects of several variables simultaneously on survival. This method is known as Cox regression or Cox proportional hazards analysis.

Cox's method is a semi-parametric approach to modelling the survival times. It is semi-parametric because it does not assume any particular distribution for the survival times; but it does make an assumption that the effects of the different prognostic variables on survival are constant over time and that the hazards are proportional. The fact that it does not make any distributional assumptions makes it very appealing.

Cox regression is conceptually similar to linear and logistic regression. Thus the strategies and methods employed in fitting the Cox proportional hazards model are analogous to those for linear and logistic regression. Whilst in linear regression we model the mean (or expected value) of the response variable, and in logistic regression we model the probability of having a certain value from a binary response (on the log odds or logit scale), in Cox regression we model the hazard function, \( h(t) \), as this is of clinical interest.

With linear regression it is assumed that the response is continuous and is normally distributed:
\( y \sim N(E(y), \sigma^2), \)

where \( E(y) \) is the mean (or expected or predicted) response.

The explanatory variables enter into the linear regression model through the predicted part:

\[ E(y) = \beta^T z \]

With logistic regression for binary data, e.g. classifying success and failure, it is assumed that the observed response (taking values 0 and 1) is distributed as a Bernoulli trial with parameter \( p \),

\[ y \sim \text{Bernoulli}(p), \]

where \( p \) is the predicted probability of responding 1.

The explanatory variables contribute by explaining the predicted part of the model, but on the log odds (logit) scale:

\[ \log \frac{p}{1-p} = \beta^T z \]

With Cox regression, by contrast, no assumptions about the distribution of the survival times are made. However the explanatory variables are assumed to enter the model through the hazard function:

\[ h(t) = h_0(t) \exp(\beta^T z) \quad (1) \]

where \( h_0(t) \) is called the baseline or underlying hazard function. Note that the adjustment for baseline measurement is achieved through the exponential of the linear combination of the predictors. Also note that in a Cox model an intercept parameter is not included in the model as it can be subsumed into the baseline hazard function. Finally, note that the exponential function is used so as to ensure that the hazard function is positive.

As in logistic regression where the \( \beta \)'s are interpreted as log odds ratios, in Cox regression the \( \beta \)'s are interpreted as log hazard ratios. The exponential of the \( \beta \)'s are therefore interpreted as hazard ratios and the standard way of explaining the effects of different types of covariates (e.g. binary, categorical or continuous) apply to Cox models also.
Proportional hazards assumption:

From (1), it is easily shown that the hazard ratio will depend only on the values of the explanatory variables and not on time. That is, the hazard ratio is constant over time. This is the proportional hazards assumption. Equivalently thought of as requiring that the hazard for one individual be proportional to the hazard for any other individual, where the constant of proportionality is independent of time. Note that the proportional hazards assumption will not be satisfied if the hazards (or survival curves) cross.

More formally, the proportional hazards assumption is achieved in the Cox model by defining the hazard to be a product of two terms. The first is the baseline hazard, $h_0(t)$, which is independent of the explanatory variables. The second term is the exponential of the linear predictors and is independent of $t$. The independence of the second term from $t$ ensures that the proportional hazards assumption is satisfied. Note, that also implicit in the proportional hazards assumption is that the $z$’s are independent of $t$. The $z$’s are called time-independent variables.

Comments:

1. The actual method of estimating the parameters and standard errors will not be described here, except to say that it is based on the maximising of the partial likelihood. This partial likelihood can be thought of in the same way as an ordinary likelihood. It is written as

$$\prod_{i=1}^{n} \left[ \frac{e^{\beta^T z_i}}{\sum_{j \in R(x_i)} e^{\beta^T z_j}} \right]^{d_i},$$

where the summation in the denominator is over all subjects in the risk set at $x_i$, denoted by $R(x_i) = \{ j : x_j \geq x_i \}$ and the product is over all $n$ subjects.

2. The baseline hazard function need not be estimated in order to estimate the parameters. Therefore inferences can be made about the effect of parameters without knowledge of the baseline hazard function.

3. The estimates obtained from a Cox model are usually quite robust. That is, the estimates obtained from a Cox model will be close to those obtained if the true distribution of the survival times followed a certain parametric distribution (e.g. Weibull or exponential).
Checking the proportional hazard assumption:

After the Cox model has been fitted to the data, the adequacy of the fitted model (in terms of the proportional hazards assumption being satisfied) needs to be evaluated. This is important as valid interpretation of regression coefficients in the Cox model depend on the proportional hazards assumption being valid.

There are a number of methods for checking the proportional hazards assumption. We will briefly mention the graphical method based on the complementary log survival curves and the hypothesis testing method based on assessing whether the re-scaled Schoenfeld residuals and some function of time are correlated.

Graphical assessment:

A popular graphical approach for checking the proportional hazards (PH) assumption is based on the comparison of complementary log (or log-log) survival curves for each of the categories of the variable of interest. If the curves are approximately parallel to one another then the PH assumption is considered valid for this variable. Otherwise the variable does not satisfy the PH assumption.

A complementary log survival curve for a certain category of the variable of interest is obtained by plotting a graph of \( \log(\log(1 - S(t))) \) against either time or some transformation of time (e.g. the natural logarithm of time). Here \( S(t) \) is either the Kaplan-Meier curve estimated for those subjects falling into the appropriate category of the variable of interest or it is the estimated survival curve obtained from a stratified Cox model (that adjust for variables where the PH assumption is known to be valid) fitted to the data, where the variable of interest is the stratifying variable.

There are a number of problems with this approach. They are

1. some categories of the variable of interest may have too few subjects in them to obtain reliable estimates of the survival curves
2. continuous variables need to be categorised
3. deciding how parallel the curves need to be for the PH assumption to be valid

Analytical approach:

The analytical approach to assessing the proportional hazards assumption is based on assessing whether the re-scaled Schoenfeld residuals are correlated with some transformation of time (e.g. logarithmic, the Kaplan-Meier, the rank or even the identity transformation). If the null hypothesis of no association is rejected then the proportional hazards assumption is not valid. Note that you may get different results by using different transformations of time. I recommend use of the Kaplan-Meier or rank transformation.
PH assumption invalid:

When the proportional hazards assumption is not satisfied for a particular variable one way of coping with this problem is to include this variable into the model as a stratifying variable. That is, we assume different baseline hazards for the different levels of the stratifying variable:

\[ h(t) = h_0(t) \exp(\beta^T z) \] \[ (2) \]

where \( s \) denotes the different levels of the stratifying variable. Note that the stratifying variable is not adjusted for in the model (i.e. it is not in the exponential part) and therefore an estimate of the effect of this stratifying variable is not obtained. We do not discuss here ways of obtaining an estimate for the effect of the stratifying variable on survival, as this would lead on to time-dependent Cox models. Note also that in (2) we have assumed that the effects of the adjusted variables are the same in the different strata. However, by including an interaction between the strata variable and any of the adjusted variables, we can allow for different effects of the adjusted variable in the different strata.

Observation and Selection Schemes: (Not Examinable)

Up to now, we have discussed the analysis of time to event data when right censoring takes place. There are, however, other possible forms of restriction on the observation of the failure times for subjects in a study. Examples are left censoring and left and right truncation.

Left censoring occurs when the true survival time, \( T_i \), of a subject is less than the actual time, \( C_i \), observed. For example, in a study of time to recurrence of leukaemia after “successful” bone marrow transplantation, some subjects may relapse before their first post-surgery medical visit. Thus, only the incomplete information that the true recurrence times of these subjects are less than the times to their first medical visit is available.

Truncation of survival data occurs when individuals may be observed only within a certain observation period/window, say \((S_L, S_R)\). A subject whose time to event is not in this window is not observed and no information on this subject is available to the investigator. This is in contrast to censoring, where there is at least partial information on each subject (that is, for example, the true survival time is larger than the observed time (right censoring), or the true survival time is smaller than the observed time (left censoring)).

Right truncation arises when \( S_L = 0 \). That is, an individual will come under observation only if their true survival time is less than \( S_R \). An example of when right truncation occurs is in estimating the distribution of the distance of stars from earth.
In this case, stars within a certain threshold distance can be detected, whilst stars beyond this distance are not detectable and are therefore right truncated. Another area where right truncation is common is in retrospective studies of AIDS cases. For example, the 1986 study of blood “transfusion-related” AIDS was based on all AIDS cases reported to the Centre for Disease Control and Prevention (CDC) with transfusion as the only known mode of infection and who were diagnosed before 1986. The dates of infection were assumed to be the dates of transfusion with contaminated blood. Any analyses performed on the time to developing AIDS from transfusion (i.e. the incubation time) must take account the fact that the true survival times must be less than the time between the earliest calendar date of infection and the 1st January 1986. Any AIDS cases occurring after 1st January 1986 of patients transfused before that date will not be observed and are therefore right truncated.

It may happen that subjects with a time to event less than some threshold, \( S_L \), may not be observed at all. This is known as left truncation, and here individuals come under observation only some known time after the “natural”/defined time origin of the phenomenon/event under study (and \( S_R = \infty \)). It is different from left censoring in that, had an individual experienced the event of interest before the truncation time in question, that individual would not have been known about. Thus any contribution to the likelihood must be conditional on the truncation threshold, \( S_L \), having been exceeded.

An example of where left truncation arises is in estimating the distribution of the diameters of microscopic particles. Only particles large enough to be seen based on the resolution of the microscope are observed, whilst smaller particles will not be observed and we will know nothing about them. Another example is in studies where subjects suffering from a particular disease come under observation at the time of entry into the clinic, but interest lies in the time from diagnosis of the disease (or maybe even from birth) to death. Therefore individuals diagnosed with the disease who die before they were able to enter the clinic would not be known about and are left truncated. Thus we do not know how many diagnosed patients died before entering the clinic and inference must be made conditional on survival until entry into the clinic.

Left truncation and right censoring:

In what follows, we concentrate on the situation where we may have left truncation (or delayed entry) in addition to right censoring. Under the assumptions that the right censoring is non-informative and that left truncated times, \( A_i \)'s (some possibly zero), for individuals are independent of the true survival times, \( T_i \)'s, and therefore those who enter the study at time \( A \) (from the defined time origin) are a random sample of those in the population still at risk at \( A \), suitably modified Kaplan-Meier curves and semi-parametric Cox regression models can be used to describe and model the observed survival data.

The modifications needed to the Kaplan-Meier estimator and the partial likelihood presented earlier, in order to be applicable when there is a combination of left truncation and right censoring of the survival data, are in the definitions of the risk
sets. That is, how the \( n_i \)'s are defined for the Kaplan-Meier estimator and how the \( R(x_i) \)'s are defined in the partial likelihood.

Let \( A_1, \ldots, A_n \) be the entry times (some possibly zero), from the defined time origin, for the \( n \) subjects in the study. Let \( A_{\text{min}} = \min(A_1, \ldots, A_n) \), and as before, let \((X_1, \delta_1), \ldots, (X_n, \delta_n)\) be the observed survival time and event indicator pairs for the subjects. Then the “modified” Kaplan-Meier estimator, \( \hat{S}_{A_{\text{min}}}(.) \), is

\[
\hat{S}_{A_{\text{min}}}(t) = \Pr(T > t \mid T \geq A_{\text{min}}) = \prod_{i}^{\text{distinct } X_i < t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq A_{\text{min}}
\]

where \( d_i \) is the number of observed failures (deaths) that occurred at \( X_i \), \( n_i \) is the number of observations (patients) in the risk set at time \( X_i \) who have already entered the study by \( X_i \) (i.e. the count of subjects whom are still alive and under observation just before time \( X_i \)). Mathematically, \( n_i = \#\{ j : A_j \leq X_i, \delta_j = 0 \} \). As before the product is taken over all distinct observed survival times.

This estimator can be further extended to estimate the conditional probability of surviving beyond a time \( t \), given survival to a time \( a(\geq A_{\text{min}}) \). That is,

\[
\hat{S}_{a}(t) = \Pr(T > t \mid T \geq a) = \prod_{a \leq X_i < t} \left(1 - \frac{d_i}{n_i}\right), \quad t \geq a
\]

with \( d_i \) and \( n_i \) defined as in the previous paragraph.

Some care in directly applying these estimators is required. For left truncated data, it is possible for the number at risk to be quite small for small values of \( x_i \) (i.e. the realised value of \( X_i \)). If, for some \( x_i \), \( d_i \) and \( n_i \) are equal, then the Kaplan-Meier estimator will be zero for all \( t \) beyond this point, even though we may observe “survivors” and “deaths” beyond this \( x_i \). In such cases, it is common to estimate the survivor function conditional on survival to a time, \( a \), where this will not happen, by considering only those event times beyond this point.

Under non-informative right censoring and independent left truncation (given the covariates \( Z \)), the “modified” partial likelihood is
\[
\prod_{i=1}^{n} \left( \frac{\prod_{j \in R(x_i)} e^{\beta^T z_i}}{\sum_{j \in R(x_i)} e^{\beta^T z_j}} \right)^{d_i},
\]

where left truncation is accounted for by the redefining of the risk set to
\[ R(x_i) = \{ j : x_j \geq x_i, a_j \} \]
and \( a_j \) is the realised value of \( A_j \). The interpretation of the parameters, \( \beta \), are as before (i.e. as log hazard ratios).

References:
