8 Survival Analysis in Phase III Clinical Trials

In chronic disease clinical trials; e.g. Cancer, AIDS, Cardiovascular disease, Diabetes, etc., the primary endpoint is often time to an event, such as time to death, time to relapse of disease, etc. For such clinical trials the major focus is to compare the distribution of time to event among competing treatments.

Typically, the clinical trials occur over a finite period of time and consequently the time to event is not ascertained on all the patients in the study. This results in censored data. In addition, since patients enter the clinical trial at different calendar times (staggered entry), the length of follow-up varies by the individual. The combination of censoring and differential follow-up creates some unusual difficulties in the analysis of such data that do not allow standard statistical techniques to be used. Because of this, a whole new research area in Statistics has emerged to study such problems. This is called Survival Analysis or Censored Survival Analysis. A brief introduction of Survival Analysis will be given in this chapter, but for a more thorough study of this area I recommend the course in Applied Survival Analysis offered every Spring semester and for individuals interested in a more rigorous treatment of the subject there is a course on Advanced Survival Analysis offered every other year.

In survival analysis, the endpoint of interest is time to an event which we denote by the positive random variable $T$. Some examples include

- survival time (time from birth to death)
- time from treatment of lung cancer to death among patients with lung cancer
- among patients with an infection that are treated with an antibiotic, the time from treatment until eradication of infection

In order to be unambiguous, the start and end of the event must be clearly identified.
8.1 Describing the Distribution of Time to Event

We will describe some different, but equivalent, ways to define the distribution of the random variable, $T$, “time to event.”

- The distribution function:
  
  $$F(t) = P(T \leq t);$$

- The survival function:
  
  $$S(t) = P(T \geq t);$$

The right-continuous version of the survival function will be denoted by

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

**Remark:** For the most part, we will assume that $T$ is a continuous random variable in which case $S(t^-) = S(t) = 1 - F(t)$. We will also assume that $T$ has a density function

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

Clearly:

$$F(t) = \int_0^t f(u)du,$$

and

$$S(t) = \int_t^\infty f(u)du.$$

**Hazard rate**

The hazard rate is a useful way of defining the distribution of a survival time which can also be used to describe the aging of a population. We motivate the definition of a hazard rate by first introducing “mortality rate” or discrete hazard rate.

The mortality rate at time $t$ (where $t$ is usually taken to be an integer of some unit of time; i.e. day, week, month, year, etc.) is the proportion of the population who fail between times $t$ and $(t + 1)$ among individuals alive (who have not failed) at time $t$.

$$m(t) = P(t \leq T < t + 1|T \geq t).$$
In a human population, the mortality rate has a pattern like

The hazard rate $\lambda(t)$ is the limit of the mortality rate or the instantaneous rate of failure at time $t$ given the individual is alive at time $t$. That is,

$$\lambda(t) = \lim_{h \to 0} \left\{ \frac{P(t \leq T < t + h|T \geq t)}{h} \right\}.$$

This can be expressed as

$$\lambda(t) = \lim_{h \to 0} \left\{ \frac{P(t \leq T < t + h)/h}{P(T \geq t)} \right\} = \frac{f(t)}{S(t)} = \frac{-dS(t)}{dt} = \frac{-d\log\{S(t)\}}{dt}. $$

Integrating both sides of the equation above, we get

$$-\log\{S(t)\} = \int_0^t \lambda(u)du = \Lambda(t),$$

where $\Lambda(t)$ is defined as the cumulative hazard function. Consequently,

$$S(t) = \exp\left\{ -\int_0^t \lambda(u)du \right\} = \exp\{-\Lambda(t)\}.$$

**Note:** Although the mortality rate is a probability, the hazard rate is NOT a probability; thus it can take on any positive value unlike a mortality rate which must be bounded by 1.
The mortality rate

\[ m(t) = \frac{P(T \geq t) - P(T \geq t + 1)}{P(T \geq t)} \]
\[ = 1 - \frac{P(T \geq t + 1)}{P(T \geq t)} \]
\[ = 1 - \frac{\exp\{-\Lambda(t + 1)\}}{\exp\{-\Lambda(t)\}} \]
\[ = 1 - \exp\left\{-\int_t^{t+1} \lambda(u)du\right\}. \]

Notice that if the probability of an event occurring in a single time unit is small and the hazard rate doesn’t change quickly within that time unit, then the hazard rate is approximately the same as the mortality rate. To see this, note that

\[ m(t) = 1 - \exp\left\{-\int_t^{t+1} \lambda(u)du\right\} \approx 1 - \left\{1 - \int_t^{t+1} \lambda(u)du\right\} \]
\[ = \int_t^{t+1} \lambda(u)du \approx \lambda(t). \]

Also, by definition, the hazard rate depends on the time scale being used. Therefore, at the same point in time the hazard rate in days is 1/365 times the hazard rate in years.

Because of the one-to-one relationships that were previously derived, the distribution of a continuous survival time \( T \) can be defined by any of the following:

\[ S(t), F(t), f(t), \lambda(t). \]

**Exponential distribution**

If the hazard rate is constant over time

\[ \lambda(t) = \lambda, \quad \text{then} \]
\[ S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\} = \exp(-\lambda t). \]

This is an exponential distribution with hazard equal to \( \lambda \). Sometimes this is referred to as the negative exponential.

It is sometimes useful to plot the log survival probability over time. This is because \(- \log\{S(t)\} = \Lambda(t)\).
If $T$ follows an exponential distribution with hazard rate $\lambda$, then the median survival time

$$
\{ m : P(T \geq m) = .5 \}; \exp(-\lambda m) = .5; m = -\log(.5)/\lambda = \log(2)/\lambda = .6931/\lambda,
$$

and the mean survival time

$$
E(T) = \int_0^\infty t\lambda \exp(-\lambda t) dt = \lambda^{-1}.
$$

Other parametric models commonly used

**Weibull distribution**

The hazard function of the Weibull distribution is given by

$$
\lambda(t) = \lambda t^{\gamma - 1}; \lambda, \gamma > 0.
$$

This is a two-parameter model with the scale parameter $\lambda$ and the shape parameter $\gamma$. This model allows us to consider hazard functions which increase or decrease over time according to the choice of $\gamma$.

$$
S(t) = \exp\left(\frac{-\lambda t^\gamma}{\gamma}\right).
$$

**Gompertz-Makeham distribution**

This distribution is useful for modeling the hazard function of human populations especially later in life. The hazard function is given by

$$
\lambda(t) = \theta + \beta e^{\gamma t}.
$$
Must be careful to choose $\theta, \beta, \gamma$ so that $\lambda(t) \geq 0$ and if we also want a proper distribution, i.e. $S(t) \to 0$ as $t \to \infty$ then

$$\Lambda(\infty) = \infty.$$ 

Other popular distributions include the log normal distribution where

$$\log(T) \sim N(\mu, \sigma^2),$$

and the gamma distribution whose density

$$f(t) \text{ is proportional to } t^\rho e^{-\lambda t}.$$ 

**Remark:** In most clinical trials applications and research in survival analysis it has become common practice to use non-parametric and semi-parametric models where the shape of the distribution function is left unspecified.

### 8.2 Censoring and Life-Table Methods

Two important issues in clinical trials where the primary endpoint of interest is time to an event which are different than most studies:

1. Some individuals are still alive (event of interest has not occurred) at the time of analysis. This results in right censored data.
2. The length of follow-up varies due to staggered entry.

This is illustrated in the schematic shown in the next page.

The time to event of interest in most clinical trials is the time from entry into the study until death (right-hand panel).

In addition to censoring occurring because of insufficient follow-up, it may also occur for other reasons such as

- loss to follow-up (patient drops out of the study, stops coming to the clinic or moves away)
Figure 8.3: Illustration of censored data

- death from other causes
  (competing risks; e.g. gets run over by a bus)

The above are examples of what is called random right censoring. That is, we conceptualize a random variable (possibly unobserved) corresponding to the potential time that an individual may be censored. This censoring time is a random variable which varies by individual. Random right censoring creates unique difficulties which does not allow the use of standard inferential techniques. This is illustrated in the following example from a study of 146 patients with previous history of heart disease treated with a new anti-hypertensive drug. The study was carried out over a ten year period and the data are grouped into one year intervals.
<table>
<thead>
<tr>
<th>Year since entry into study</th>
<th>Number at risk at beginning of interval</th>
<th>Number dying in interval</th>
<th>Number censored in interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>146</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>1-2</td>
<td>116</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>2-3</td>
<td>88</td>
<td>21</td>
<td>10</td>
</tr>
<tr>
<td>3-4</td>
<td>57</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>4-5</td>
<td>45</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5-6</td>
<td>41</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>6-7</td>
<td>28</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>7-8</td>
<td>20</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>8-9</td>
<td>11</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9-10</td>
<td>8</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

Question: Estimate the five-year mortality rate? Two naive estimators are as follows:

1. \[
\frac{76 \text{ deaths in 5 years}}{146 \text{ individuals}} = .521, \quad \hat{S}(5) = .479
\]

2. \[
\frac{76 \text{ deaths in 5 years}}{146 - 29 \text{ (withdrawn)}} = .650, \quad \hat{S}(5) = .350.
\]

Estimator 1. corresponds to censoring on the right; that is, if everyone that was withdrawn in the first 5 years was withdrawn exactly at 5 years, then this approach would give an unbiased estimator. Since this isn’t what happened, this estimator is too optimistic.

In contrast, estimator 2. would be appropriate if everyone that was withdrawn in the first 5 years was withdrawn immediately at time “0”. If this were the case then this approach would yield an unbiased estimator. Since this isn’t what happened, this estimator is too pessimistic.

The more appropriate method uses life-table estimates, illustrated as follows:

Assume censoring occurs at the right of each yearly interval.


\[
R = \frac{d}{n_r} \quad \hat{S}_R = \Pi(1 - m_R)
\]

\[
0-1 \quad 146 \quad 27 \quad 3 \quad .185 \quad .815 \quad .815
\]

\[
1-2 \quad 116 \quad 18 \quad 10 \quad .155 \quad .845 \quad .689
\]

\[
2-3 \quad 88 \quad 21 \quad 10 \quad .239 \quad .761 \quad .524
\]

\[
3-4 \quad 57 \quad 9 \quad 3 \quad .158 \quad .842 \quad .441
\]

\[
4-5 \quad 45 \quad 1 \quad 3 \quad .022 \quad .978 \quad .432
\]

5 year survival estimate = .432

5 year mortality rate estimate = .568

Assume censoring occurs at the left of each interval

\[
L = \frac{d}{n_r - w} \quad \hat{S}_L = \Pi(1 - m_L)
\]

\[
0-1 \quad 146 \quad 27 \quad 3 \quad .189 \quad .811 \quad .811
\]

\[
1-2 \quad 116 \quad 18 \quad 10 \quad .170 \quad .830 \quad .673
\]

\[
2-3 \quad 88 \quad 21 \quad 10 \quad .269 \quad .731 \quad .492
\]

\[
3-4 \quad 57 \quad 9 \quad 3 \quad .167 \quad .833 \quad .410
\]

\[
4-5 \quad 45 \quad 1 \quad 3 \quad .024 \quad .976 \quad .400
\]

5 year survival estimate = .400

5 year mortality rate = .600

We note that the naive estimator for the five year survival probability ranged from .35 to .479, whereas the life-table estimates ranged from .40 to .432 depending on whether we assumed censoring occurred on the left or right of each interval.

More than likely, censoring occurred during the interval. Thus \(\hat{S}_L\) and \(\hat{S}_R\) are under and over estimates respectively. A compromise would be to use

\[
m = \frac{d}{(n_r - w/2)} \quad \text{in the tables above.}
\]

This is what is referred to as the life-table estimate and for this example leads to the estimate of the 5 year survival probability \(\hat{S}(5) = .417\).

Since the life-table estimator is an estimator for the underlying population survival probability based on a sample of data, it is subject to variability. To assess the variability of this estimator,
the standard error can be computed using Greenwood’s formulas as follows:

\[
\text{se}\{\hat{S}(t)\} = \hat{S}(t) \left\{ \sum_{j=1}^{t} \frac{d_j}{(n_{rj} - w_j/2)(n_{rj} - d_j - w_j/2)} \right\}^{1/2}.
\]

With sufficiently large sample sizes this estimator is approximately normally distributed; in which case, the \((1 - \alpha)^{th}\) confidence interval for \(S(t)\) can be approximated by

\[
\hat{S}(t) \pm Z_{\alpha/2}[\text{se}\{\hat{S}(t)\}],
\]

where \(Z_{\alpha/2}\) is the \((1 - \alpha/2)\) quantile of the standard normal distribution.

In our example

<table>
<thead>
<tr>
<th>time</th>
<th>(n_r)</th>
<th>(d)</th>
<th>(w)</th>
<th>(\hat{S})</th>
<th>(\frac{d}{(n_r-w/2)(n_r-d-w/2)})</th>
<th>(\frac{d}{(n-w/2)(n-d-w/2)})</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>146</td>
<td>27</td>
<td>3</td>
<td>.813</td>
<td>.00159</td>
<td>.00159</td>
<td>.032</td>
</tr>
<tr>
<td>1-2</td>
<td>118</td>
<td>18</td>
<td>10</td>
<td>.681</td>
<td>.00168</td>
<td>.00327</td>
<td>.039</td>
</tr>
<tr>
<td>2-3</td>
<td>88</td>
<td>21</td>
<td>10</td>
<td>.509</td>
<td>.00408</td>
<td>.00735</td>
<td>.044</td>
</tr>
<tr>
<td>3-4</td>
<td>57</td>
<td>9</td>
<td>3</td>
<td>.426</td>
<td>.00345</td>
<td>.01084</td>
<td>.044</td>
</tr>
<tr>
<td>4-5</td>
<td>45</td>
<td>1</td>
<td>3</td>
<td>.417</td>
<td>.00054</td>
<td>.01138</td>
<td>.044</td>
</tr>
</tbody>
</table>

The 95% confidence interval for \(S(5)\) is given as .417 ± 1.96(.044) = (.331 – .503).

### 8.3 Kaplan-Meier or Product-Limit Estimator

We notice in our previous example that the bias that occurs in estimating the survival distribution by incorrectly assuming that censoring occurs at the left or right of each interval was decreased when the interval was taken to be smaller (i.e. 1 year intervals as opposed to 5 year intervals). If the data were not grouped (i.e. we know the exact times to death or censoring), then this suggests that we may want to apply the life-table estimator using many intervals with small interval widths. The limit of the life-table estimator when the intervals are taken so small that at most only one observation occurs within any interval is called the product-limit estimator which is also the same as the Kaplan-Meier estimator. Kaplan and Meier (1958) derived the estimator based on likelihood principles. We believe it is more instructive and intuitive to consider this estimator as the limit of the life-table estimator.
To illustrate how this estimator is constructed, consider the following example:

**Figure 8.4: An illustrative example of Kaplan-Meier estimator**

![Kaplan-Meier estimator diagram](image)

\[
1 - \hat{m}(x) : \quad 1 \quad 1 \quad 1 \quad 1 \quad \frac{9}{10} \quad 1 \quad 1 \quad \frac{8}{9} \quad 1 \quad 1 \quad 1 \quad \frac{6}{7} \quad 1 \quad 1 \quad 1 \quad \frac{4}{5} \quad \frac{3}{4} \quad 1 \quad 1 \quad \frac{1}{2} \quad 1 \quad 1
\]

\[
\hat{S}(t) : \quad 1 \quad 1 \quad 1 \quad 1 \quad \frac{9}{10} \quad \ldots \quad \frac{8}{10} \quad \ldots \quad \frac{48}{70} \quad \ldots \quad \frac{192}{350} \quad \frac{144}{350} \quad \ldots \quad \frac{144}{700} \quad \ldots
\]

\[
m = \frac{d}{n_r} = \frac{\text{number of deaths in an interval}}{\text{number at risk at beginning of interval}}
\]

\[
= \left(\frac{1}{n_r} \text{ or 0 depending on whether or not a death occurred in interval}\right)
\]

\[
(1 - m) = (1 - d/n_r) = ((1 - 1/n_r) \text{ or 1}).
\]

In the limit, the Kaplan-Meier (product-limit) estimator will be a step function taking jumps at times where a failure occurs. Therefore at any time \(t\), the product-limit estimator of the survival distribution is computed as the product

\[
\prod_{\text{all deaths}} \left(1 - \frac{1}{\text{number at risk}}\right)
\]

over all death times occurring up to and including time \(t\).

By convention, the Kaplan-Meier estimator is taken to be right-continuous.

**Non-informative Censoring**
In order that life-table estimators give unbiased results, there is an implicit assumption that individuals who are censored have the same risk of subsequent failure as those who are alive and uncensored. The risk set at any point in time (individuals still alive and uncensored) should be representative of the entire population alive at the same time in order that the estimated mortality rates reflect the true population mortality rates.

**Some notation and software**

In describing censored survival data, it is useful to conceptualize two latent random variables (possibly unobserved) corresponding to the failure time and censoring time. For the $i$-th individual we will denote these by $T_i$ and $C_i$ respectively. Specifically, $T_i$ represents the survival time if that individual was followed until death; whereas, $C_i$ corresponds to the time measured from their entry into the study until they were censored in the hypothetical situation that they could not die. For example, $C_i$ may be the time from entry into the study until the time the final analysis was conducted. However, if the individual could be lost to follow-up, then the variable $C_i$ would have to account for that possibility as well. In any case, $C_i$ corresponds to the time that an individual would have been censored in a study if their death could be prevented.

In contrast to these latent variables, the variables we actually get to observe for the $i$-th individual are denoted by $(U_i, \Delta_i)$, where $U_i$ denotes the observed time on study (i.e. the time to death or censoring, and $\Delta_i$ denotes the failure indicator taking on the value 1 if the patient is observed to die and the value 0 if the patient is censored. In terms of the latent variables we assume $U_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$.

The main objective of a clinical trial is to make inference about the probability distribution of the latent survival time $T_i$ even though this variable is not always observed due to censoring. In the one-sample problem we are interested in estimating the survival distribution $S(t) = P(T_i \geq t)$ using a sample of observable data

$$(U_i, \Delta_i), \ i = 1, \ldots, n.$$  

If we define the number of individuals at risk at any time $t$ by

$$n(t) = \sum_{i=1}^{n} I(U_i \geq t),$$

that is, the number of individuals in our sample who neither died or were censored by time $t$,
then the Kaplan-Meier estimator is given by

$$KM(t) = \prod_{i: U_i \leq t} \left\{ \frac{n(U_i) - 1}{n(U_i)} \right\}^{\Delta_i}.$$  

This is the Kaplan-Meier estimator when there are no tied survival times in our sample. More generally, if we denote by

$$d(t) = \sum_{i=1}^{n} I(U_i = t, \Delta_i = 1),$$

the number of observed deaths in our sample at time $t$, thus allowing the possibility that $d(t) \geq 2$ in cases where survival times are tied, then we can write the Kaplan-Meier estimator as

$$KM(t) = \prod_{\text{death times } u \leq t} \left\{ 1 - \frac{d(u)}{n(u)} \right\}.$$  

The standard error of the Kaplan-Meier estimator is also taken as the limit in Greenwood’s formula, Namely,

$$se\{KM(t)\} = KM(t) \left\{ \sum_{\text{death times } u \leq t} \frac{d(u)}{n(u)\{n(u) - d(u)\}} \right\}^{1/2}.$$  

**Proc lifetest in SAS**

Many statistical packages, including SAS, have software available for censored survival analysis. For example, the above Kaplan-Meier estimator can be obtained using the following *SAS* program:

```
Data example;
  input survtime censcode;
  cards;
  4.5 1
  7.5 1
  8.5 0
  11.5 1
  13.5 0
  15.5 1
  16.5 1
  17.5 0
  19.5 1
  21.5 0;

Proc lifetest;
  time survtime*censcode(0);
run;
```

And part of the output from the above program is
8.4 Two-sample Tests

The major objective of many Phase III clinical trials is to compare two or more treatments with respect to some primary endpoint. If the primary endpoint is time to an event (e.g. survival time), then interest will focus on whether one treatment will increase or decrease the distribution of this time as compared to some standard or control treatment. Let us begin by considering the comparison of two treatments. Let the variable \( A \) denote treatment group, where we take \( A = 0 \) to denote the control group or standard treatment and \( A = 1 \) the new treatment.

The problem of comparing two treatments is often cast as a hypothesis testing question. The null hypothesis being that the distribution of time to death (event) is the same for both treatments. Letting \( T \) denote a patient’s underlying survival time, we define the treatment specific survival distributions by \( S_1(t) = P(T \geq t| A = 1) \) and \( S_0(t) = P(T \geq t| A = 0) \). The null hypothesis is given as

\[
H_0 : S_1(t) = S_0(t) = S(t), \ t > 0,
\]

or equivalently

\[
H_0 : \lambda_1(t) = \lambda_0(t) = \lambda(t),
\]

where \( \lambda_j(t), j = 0, 1 \) denote the treatment-specific hazard rates.
The alternative hypothesis of most interest in such trials is that the survival time for one treatment is stochastically larger than the survival time for the other treatment. Specifically, we say the survival time for treatment 1 is stochastically larger than the survival time for treatment 0 if \( S_1(t) \geq S_0(t) \) for all \( t > 0 \) with strict inequality for at least one value of \( t \).

It has become standard practice in clinical trials to use nonparametric tests; that is, tests based on statistics whose distribution under the null hypothesis does not depend on the underlying survival distribution \( S(t) \) (At least asymptotically). The most widely used test with censored survival data is the logrank test which we now describe.

Data from a clinical trial comparing the survival distribution between two treatments can be viewed as realizations of the random triplets

\[(U_i, \Delta_i, A_i), i = 1, \ldots, n,\]

where

- \( U_i = \min(T_i, C_i) \)
  - \( T_i \) denotes the latent failure time
  - \( C_i \) denotes the latent censoring time
- \( \Delta_i = I(T_i \leq C_i) \) denotes failure indicator
- \( A_i \) denotes treatment indicator

We also define the following notation:

- \( n_j = \sum_{i=1}^{n} I(A_i = j) \) denotes the number of patients assigned treatment \( j = 0, 1; n = n_0 + n_1 \)
- \( n_j(u) = \sum_{i=1}^{n} I(U_i \geq u, A_i = j) \) denotes the number at risk at time \( u \) from treatment \( j = 0, 1 \)
- \( n(u) = n_0(u) + n_1(u) \) denotes the total number at risk at time \( u \) from both treatments
- \( d_j(u) = \sum_{i=1}^{n} I(U_i = u, \Delta_i = 1, A_i = j) \) denotes the number of observed deaths at time \( u \) from treatment \( j = 0, 1 \)
• $d(u) = d_0(u) + d_1(u)$ denotes the number of observed deaths at time $u$ from both samples.

The notation above allows the possibility of more than one death occurring at the same time (tied survival times).

The logrank test is based on the statistic

$$\sum_{\text{all death times } u} \left\{ d_1(u) - \frac{n_1(u)}{n(u)} d(u) \right\}.$$  

This statistic can be viewed as the sum over the distinct death times of the observed number of deaths from treatment 1 minus the expected number of deaths from treatment 1 if the null hypothesis were true.

Thus at any point in time $u$ corresponding to a time where a death was observed, i.e. $d(u) \geq 1$, the data at that point in time can be viewed as a $2 \times 2$ table; namely,

<table>
<thead>
<tr>
<th></th>
<th>treatment</th>
<th>0</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of deaths</td>
<td>$d_1(u)$</td>
<td>$d_0(u)$</td>
<td>$d(u)$</td>
</tr>
<tr>
<td>number alive</td>
<td>$n_1(u) - d_1(u)$</td>
<td>$n_0(u) - d_0(u)$</td>
<td>$n(u) - d(u)$</td>
</tr>
<tr>
<td>number at risk</td>
<td>$n_1(u)$</td>
<td>$n_0(u)$</td>
<td>$n(u)$</td>
</tr>
</tbody>
</table>

• The observed number of deaths at time $u$ from treatment 1 is $d_1(u)$

• The expected number of deaths from treatment 1 at time $u$ if the null hypothesis were true is $\frac{d(u)}{n(u)} n_1(u)$

• Thus the observed minus expected number of deaths at time $u$ is \(d_1(u) - \frac{d(u)}{n(u)} n_1(u)\)

From this point of view, the survival data from a clinical trial can be summarized as $k$ $2 \times 2$ tables, where $k$ denotes the number of distinct death times. If the null hypothesis is true, then we would expect \(d_1(u) - \frac{d(u)}{n(u)} n_1(u)\) to be about zero on average for all $\{u : d(u) \geq 1\}$. However, if the hazard rate for treatment 0 is greater than the hazard rate for treatment 1 consistently over all $u$, then, on average, we would expect \(d_1(u) - \frac{d(u)}{n(u)} n_1(u)\) to be negative. The opposite would be expected if the hazard rate for treatment 1 was greater than the hazard rate for treatment 0 for all $u$. 
This suggests that the null hypothesis of treatment equality should be rejected if the test statistic is sufficiently large or small depending on the alternative of interest for one-sided tests or if the absolute value of the test statistic is sufficiently large for two-sided tests. In order to gauge the strength of evidence against the null hypothesis we must be able to evaluate the distribution of the test statistic (at least approximately) under the null hypothesis. Therefore, the test statistic has to be standardized appropriately. Specifically, the logrank test is given by

\[ T_n = \frac{\sum \left\{ d_1(u) - \frac{d(u)}{n(u)} n_1(u) \right\}}{\left[ \frac{\sum n_1(u) n_0(u) d(u) \{ n(u) - d(u) \}}{n^2(u) \{ n(u) - 1 \}} \right]^{1/2}}. \]  

**Remark:** In a $2 \times 2$ contingency table

\[
\begin{array}{ccc|c}
  d_1(u) & \cdot & d(u) \\
  \cdot & \cdot & n(u) - d(u) \\
  n_1(u) & n_0(u) & n(u) \\
\end{array}
\]

The value $d_1(u)$, under the null hypothesis, conditional on the marginal totals, has a hypergeometric distribution with mean

\[ \frac{d(u)}{n(u)} n_1(u) \]

and variance

\[ \left[ \sum \frac{n_1(u) n_0(u) d(u) \{ n(u) - d(u) \}}{n^2(u) \{ n(u) - 1 \}} \right]. \]

The sum of the hypergeometric variances of these $2 \times 2$ tables, summed over the distinct death times, is the estimator of the variance of the test statistic under the null hypothesis. Therefore, the logrank test $T_n$ given by (8.1) is distributed as a standard normal under $H_0$; i.e.

\[ T_n \overset{H_0}{\sim} N(0, 1). \]

Consequently, a level $\alpha$ test (two-sided) would reject the null hypothesis when $|T_n| \geq Z_{\alpha/2}$. One sided level $\alpha$ tests would reject whenever $T_n \geq Z_\alpha$ or $-T_n \geq Z_\alpha$ depending on the question. For example, if we were interested in showing that treatment 1 is better (longer survival times) than treatment 0, then we would reject $H_0$ when $-T_n \geq Z_\alpha$ because under the alternative hypothesis we would expect the observed number of deaths from treatment 1 to be less than that expected under the null hypothesis.
Note: All the arguments made above were based on summarizing the data as $2 \times 2$ tables at distinct death times. Nowhere did we have to make any assumptions (other than the null hypothesis) about the actual shape of the underlying survival distribution in deriving the numerator of the logrank test statistic or its variance. This, intuitively, explains why this test is nonparametric.

If censored survival data are organized as $(U_i, \Delta_i, A_i), i = 1, \ldots, n$, where $U_i$ denotes time to failure or censoring, $\Delta_i$ denotes failure indicator, and $A_i$ denotes treatment indicator, then the logrank test can be computed using SAS. To illustrate, we again use the data from CALGB 8541 (clinical trial on breast cancer).

Recall that CALGB 8541 was a randomized three arm clinical trial for patients with stage II node positive breast cancer. Although there were three treatments, the major focus was comparing treatment 1 (Intensive CAF) to treatment 2 (Low dose CAF), where CAF is the combination of the drugs Cyclophosphamide, Adriamycin and 5 Fluorouracil. For the purpose of this illustration we will restrict attention to the comparison of these two treatments. Later we will discuss the comparison of all three treatments.

data trt12; set bcancer;
  if (trt=1) or (trt=2);
run;

%title "Log-rank test comparing treatments 1 and 2";
%proc lifetest data=trt12 notable;
  time years*censor(0);
  strata trt;
run;

Part of the output from the above SAS program:

The LIFETEST Procedure
Testing Homogeneity of Survival Curves for years over Strata

<table>
<thead>
<tr>
<th>trt</th>
<th>Log-Rank</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-30.030</td>
<td>-23695</td>
</tr>
<tr>
<td>2</td>
<td>30.030</td>
<td>23695</td>
</tr>
</tbody>
</table>
8.5 Power and Sample Size

Thus far, we have only considered the properties of the logrank test under the null hypothesis. In order to assess the statistical sensitivity of this test, we must also consider the power to detect clinically meaningful alternatives. A useful way to define alternative hypotheses is through the proportional hazards assumption. That is, letting $\lambda_1(t)$ and $\lambda_0(t)$ denote the hazard functions at time $t$, for treatments 1 and 0 respectively, the proportional hazards assumption assumes that

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma), \text{ for all } t \geq 0. \quad (8.2)$$

We use $\exp(\gamma)$ here because a hazard ratio must be positive and because $\gamma = 0$ will correspond to a hazard ratio of one which would imply that both treatments have the same hazard function (i.e. the null hypothesis). The proportional hazards assumption, if true, also has a nice interpretation. The hazard ratio $\exp(\gamma)$ can be viewed as a relative risk and for purposes of testing the null hypothesis of no treatment difference
• \( \gamma > 0 \) implies that individuals on treatment 1 have worse survival (i.e. die faster)

• \( \gamma = 0 \) implies the null hypothesis

• \( \gamma < 0 \) implies that individuals on treatment 1 have better survival (i.e. live longer)

If the proportional hazards assumption were true; that is,

\[
\lambda_1(t) = \lambda_0(t) \exp(\gamma),
\]

then this would imply that

\[
-\frac{d \log S_1(t)}{dt} = -\frac{d \log S_0(t)}{dt} \exp(\gamma),
\]

or

\[
-\log S_1(t) = -\log S_0(t) \exp(\gamma).
\]

Consequently,

\[
S_1(t) = \{S_0(t)\}^{\exp(\gamma)},
\]

and

\[
\log\{-\log S_1(t)\} = \log\{-\log S_0(t)\} + \gamma.
\]

This last relationship can be useful if we want to assess whether a proportional hazards assumption is a reasonable representation of the data. By plotting the two treatment-specific Kaplan-Meier curves on a \( \log\{-\log\} \) scale we can visually inspect whether these two curves differ from each other by a constant over time.

Also, in the special case where we feel comfortable in assuming that the survival distributions follow an exponential distribution; i.e. constant hazards, the proportional hazards assumption is guaranteed to hold. That is,

\[
\frac{\lambda_1(t)}{\lambda_0(t)} = \frac{\lambda_1}{\lambda_0}.
\]

In section 9.1 we showed that the median survival time for an exponential distribution with hazard \( \lambda \) is equal to \( m = \log(2)/\lambda \). Therefore, the ratio of the median survival times for two treatments whose survival distributions are exponentially distributed with hazard rates \( \lambda_1 \) and \( \lambda_0 \) is

\[
\frac{m_1}{m_0} = \left\{ \frac{\log(2)}{\lambda_1} \right\} = \frac{\lambda_0}{\lambda_1}.
\]
That is, the ratio of the medians of two exponentially distributed random variables is inversely proportional to the ratio of the hazards. This relationship may be useful when one is trying to illicit clinically important differences from medical collaborators during the design stage of an experiment. Clinical investigators generally have a good sense of the median survival for various treatments and can more easily relate to the question of determining an important increase in median survival. However, as we just illustrated, if the survival distributions are well approximated by exponential distributions then the differences in median survival can be easily translated to a hazard ratio through the inverse relationship derived above.

The reason we focus on proportional hazards alternatives is that, in addition to having “nice” interpretation, theory has been developed that shows that the logrank test is the most powerful nonparametric test to detect proportional hazards alternatives. Moreover, it has also been shown that the distribution of the logrank test under the alternative

\[ H_A : \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma_A), \ t \geq 0 \]

is approximately distributed as a normal distribution

\[ T_n \overset{H_A}{\sim} N \left( \{d\theta(1 - \theta)\}^{1/2} \gamma_A, 1 \right), \]

where \( d \) denotes the total number of deaths (events), and \( \theta \) denotes the proportion randomized to treatment 1 (generally .5). That is, under a proportional hazards alternative, the logrank
test is distributed approximately as a normal random variable with variance 1 and noncentrality parameter

\[ \{d\theta(1 - \theta)\}^{1/2} \gamma_A. \]

When \( \theta = .5 \), the noncentrality parameter is

\[ \gamma_A d^{1/2}/2. \]

In order that a level \( \alpha \) test (say, two-sided) have power \( 1 - \beta \) to detect the alternative

\[ \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\gamma_A), \]

then the noncentrality parameter must equal \( Z_{\alpha/2} + Z_\beta \). That is,

\[ \gamma_A d^{1/2}/2 = Z_{\alpha/2} + Z_\beta, \]

or

\[ d = \frac{4(Z_{\alpha/2} + Z_\beta)^2}{\gamma_A^2}. \]  \hspace{1cm} (8.3)

This means that the power of the logrank test to detect a proportional hazards alternative is directly related to the number of events (deaths) \( d \) during the course of the clinical trial.

**Remark:** This very nice and simple relationship would not apply if we didn’t use the logrank test in conjunction with a proportional hazards alternative.

If we take \( \alpha = .05 \) (two-sided), power \( 1 - \beta = .90 \), and \( \theta = .5 \), then

\[ d = \frac{4(1.96 + 1.28)^2}{\gamma_A^2}. \]  \hspace{1cm} (8.4)

Some examples of the number of deaths necessary to detect an alternative where the hazard ratio equals \( \exp(\gamma_A) \) with 90% power using the logrank test at the .05 (two-sided) level of significance is given in the following table.

<table>
<thead>
<tr>
<th>Hazard Ratio ( \exp(\gamma_A) )</th>
<th>Number of deaths ( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>88</td>
</tr>
<tr>
<td>1.50</td>
<td>256</td>
</tr>
<tr>
<td>1.25</td>
<td>844</td>
</tr>
<tr>
<td>1.10</td>
<td>4623</td>
</tr>
</tbody>
</table>
During the design stage we must ensure that a sufficient number of patients are entered into the trial and followed long enough so that the requisite number of events are attained.

**Sample Size Considerations**

One straightforward approach to ensure the desired power of a trial to detect a clinically important alternative is to continue a clinical trial until we obtain the required number of failures.

**Example:** Suppose patients with advanced lung cancer historically have a median survival of six months and the survival distribution is approximately exponentially distributed. An increase in the median survival from six months to nine months is considered clinically important and we would like to have at least 90% power to detect such a difference if it can be achieved with a new therapy using a logrank test at the .05 (two-sided) level of significance. How should we design a clinical trial comparing the standard treatment to a promising new treatment?

If the survival distributions for both treatments were exponentially distributed, then the clinically important hazard ratio would be

\[
\frac{\lambda_1}{\lambda_0} = \frac{m_0}{m_1} = \frac{6}{9} = 2/3.
\]

Hence \(\gamma_A = \log(2/3) = -0.4055\). Using the formula given by equation (8.4) we determine that we need to carry out a study that will ultimately require a total of 256 deaths.

Since, for this example, patients do not survive for very long, one strategy is to enter some number of patients greater than 256 and continue the study until 256 deaths occur. For example, we may enter, say 350 patients, whom we randomize with equal probability to the two treatments and then analyze the data after 256 deaths.

**Note** The 350 patients was chosen arbitrarily. We now give more careful consideration to some of the design aspects of a clinical trial with a survival endpoint.

**Design Specifications**

In most clinical trials, arbitrarily picking a number of patients and waiting for the requisite number of events to occur will not be adequate for the proper planning of the trial. More often we need to consider the following in the proper design of a clinical trial with a time to event endpoint.
• number of patients

• accrual period

• follow-up time

It was shown by Schoenfeld that to obtain reasonable approximations for the power, we need the expected number of events (deaths), computed under the alternative, to equal the number given by equation (8.3). Specifically, the expected number of deaths need to be computed separately for each treatment arm, under the assumption that the alternative hypothesis were true, and the sum from both treatments should equal (8.3) to achieve the desired power.

In order to compute the expected number of deaths for each treatment during the design stage, we must consider the following. Also, see figure below.

Figure 8.6: Illustration of accrual and follow-up

\[
0 \quad E_i \quad \text{(Accrual)} \quad \text{Acc} \quad \text{(Follow-up)} \quad L = \text{Acc} + F
\]

Remark: We will assume that any censoring that may occur is due to end of study censoring resulting from staggered entry. We will sometimes refer to this as “administrative censoring”. If, in addition, we wanted to consider other forms of censoring, such as lost to follow-up due to withdrawal or censoring due to death from other causes, then the formulas given below would have to be modified.

We define the following notation:

• Acc denotes the accrual period; that is, the calendar period of time that patients are entering the study

• \( F \) denotes the calendar period of time after accrual has ended before the final analysis is conducted

• \( L = \text{Acc} + F \) denotes the total calendar time of the study from the time the study opens until the final analysis
• Denote the accrual rate at calendar time $u$ by $a(u)$; more precisely as

$$a(u) = \lim_{h \to 0} \left\{ \frac{\text{Expected no. of patients entering between } [u, u + h]}{h} \right\}.$$ 

The total expected number of patients in the clinical trial is given by

$$\int_0^{\text{Acc}} a(u) du.$$ 

If we have a constant accrual rate (this is the most common assumption made during the design stage), then $a(u) = a$ and the expected number of patients is $a \times \text{Acc}$.

In a randomized study where patients are assigned with equal probability to each of two treatments the accrual rate to each treatment would be $\frac{a(u)}{2}$. If we denote the distribution function of the survival time for each treatment (0 or 1) by

$$F_j(u) = P(T \leq u | A = j) = 1 - S_j(u), j = 0, 1,$$

then the expected number of observed deaths from treatment $j$ is

$$d_j = \int_0^{\text{Acc}} \frac{a(u)}{2} F_j(L - u) du, j = 0, 1.$$ (8.5)

In words, we expect $\frac{a(u)}{2} du$ patients to enter between times $[u, u + du]$, of which the proportion $F_j(L - u)$ are expected to die by the end of the study (at time $L$). This number summed (integrated) over $u$, for values of $u$ during the accrual period $[0, \text{Acc}]$, will yield the expected number of deaths on treatment $j = 0, 1$. To get the desired power we want

$$d_1 + d_0 = \frac{4(Z_{\alpha/2} + Z_\beta)^2}{\gamma^2_A},$$

Note: The number of failures can be affected by

• the accrual rate
• accrual period (sample size)
• follow-up period
• the failure (hazard) rate (survival distribution)
Some, or all, of these factors can be controlled by the investigator and have to be considered during the design stage.

**Example:**

Assume the accrual rate is constant $a$ patients per year and we randomize equally to two treatments so that the accrual rate is $a/2$ patients per year on each treatment. Also assume that the treatment specific survival distributions are exponentially distributed with hazards $\lambda_j, j = 0, 1$. Then

$$d_j = \int_0^{Acc} \frac{a}{2} \left[ 1 - \exp\{-\lambda_j(L - u)\} \right] du$$

$$= \frac{a}{2} \left[ Acc - \frac{\exp(-\lambda_j L)}{\lambda_j} \{\exp(\lambda_j Acc) - 1\} \right].$$

During the design stage, we expect 100 patients per year to be recruited into the study. The median survival for treatment 0 is expected to be about 4 years and assumed to follow an exponential distribution ($\lambda_0 = .173$). We want the power to be 90% if the new treatment (treatment 1) can increase the median survival to 6 years ($\lambda_1 = .116$) using a logrank test at the .05 (two-sided) level of significance.

For this problem the clinically important hazard ratio is $2/3$ corresponding to $\gamma_A = \log(2/3)$. Using equation (8.3), the total number of deaths necessary is 256. Hence, we need

$$d_1(Acc, L) + d_0(Acc, L) = 256.$$

**Note:** We use the notation $d_1(Acc, L)$ to emphasize the fact that the number of deaths will depend on our choice of accrual period $Acc$ and length of study $L$.

According to the formulas we derived, we need $Acc$ and $L$ to satisfy the equation

$$50 \left[ Acc - \frac{\exp(-.116 L)}{.116} \{\exp(.116 Acc) - 1\} \right]$$

$$+ 50 \left[ Acc - \frac{\exp(-.173 L)}{.173} \{\exp(.173 Acc) - 1\} \right]$$

$$= 256.$$

There are many combinations of $Acc$ and $L$ that would satisfy the equation above. Clearly, the minimum length of the study would occur if we continued accruing patients continuously until
we obtained the desired number of events. This corresponds to the case where $Acc = L$. When $Acc = L$ there is only one unknown above and solving the equation above yields $Acc = L = 7$ years. Such a design would require a total of 700 patients.

**Note:** This equation does not have a closed form solution and the solution must be obtained numerically by iterative techniques.

In contrast, if we accrued for five years, $Acc = 5$, or a total of 500 patients, then solving for $L$ yields $L = 7.65$.

Clearly, to obtain 256 deaths, we need to accrue at least 256 patients. Therefore, $Acc$ must be greater than 2.56 years. However, choosing $L$ that small in order to accrue as few patients as possible would result in a trial that would take an exceedingly long time to complete.

A good strategy is to experiment with a variety of combinations of $Acc$ and $L$ that can be used to present to the clinical investigators and then the choice which best suits the needs of the study can be made.

Other factors that may affect power that we will not discuss are

- loss to follow-up (withdrawal)
- competing risks
- non-compliance

An excellent account on how to deal with these issues during the design stage is given by Lakatos (1988), *Biometrics*.

### 8.6 K-Sample Tests

We now consider the case where we randomize to $K > 2$ treatments and are interested in testing the null hypothesis that the survival distributions are the same for all treatments versus the alternative that there is some difference. With right censoring, the data from such a clinical trial can be represented as realizations of the iid triplets $(U_i, \Delta_i, A_i), i = 1, \ldots, n,$ where for the $i$-th
individual, $U_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and $A_i = (1, \ldots, K)$ denotes the treatment group that the individual was randomized to. Letting $S_j(t) = P(T \geq t | A = j), j = 1, \ldots, K$ denote the treatment-specific survival distributions, the null hypothesis of no treatment difference is

$$H_0 : S_1(t) = \ldots = S_K(t), \ t \geq 0,$$

or equivalently

$$H_0 : \lambda_1(t) = \ldots = \lambda_K(t), \ t \geq 0,$$

where $\lambda_j(t), j = 1, \ldots, K$ denote the treatment-specific hazard rates.

The test of the null hypothesis for this $K$-sample problem will be a direct generalization of the logrank test used for the two-sample test. Using the same notation as in the previous section where $n_j(u)$ denotes the number of individuals at risk at time $u$ in treatment group $j$ and $d_j(u)$ the number of observed deaths in treatment group $j$ at time $u$, we view the data as a series of $2 \times K$ tables at the distinct failure times $u$ where $d(u) \geq 1$. For example,

<table>
<thead>
<tr>
<th>Treatments</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of deaths</td>
<td>$d_1(u)$</td>
<td>$d_2(u)$</td>
<td>...</td>
<td>$d_K(u)$</td>
<td>$d(u)$</td>
</tr>
<tr>
<td>No. alive</td>
<td>$n_1(u) - d_1(u)$</td>
<td>$n_2(u) - d_2(u)$</td>
<td>...</td>
<td>$n_K(u) - d_K(u)$</td>
<td>$n(u) - d(u)$</td>
</tr>
<tr>
<td>No. at risk</td>
<td>$n_1(u)$</td>
<td>$n_2(u)$</td>
<td>...</td>
<td>$n_K(u)$</td>
<td>$n(u)$</td>
</tr>
</tbody>
</table>

**Generalization of the two-sample test**

At each time $u$ where $d(u) \geq 1$, we now consider a vector of observed minus expected number of deaths under the null hypothesis for each treatment group $j$. Namely,

$$\begin{pmatrix}
    d_1(u) - \frac{n_1(u)}{n(u)}d(u) \\
    \vdots \\
    d_K(u) - \frac{n_K(u)}{n(u)}d(u)
\end{pmatrix}^{K \times 1}$$

**Note:** The sum of the elements in this vector is equal to zero. Because of this, such vectors can only vary in $(K - 1)$-dimensional space. Consequently, it suffices to use only $K - 1$ of these
vectors for further consideration in constructing test statistics. It doesn’t matter which $K - 1$ of these we use, thus, for convention, we will take from now on $j = 1, \ldots, K - 1$.

Analogous to the two-sample problem, if we condition on the marginal counts of these $2 \times K$ contingency tables, then, under the null hypothesis, the distribution of the counts $\{d_1(u), \ldots, d_K(u)\}$ is a multivariate version of the hypergeometric distribution. Specifically, this implies that, conditional on the marginal counts, the conditional expectation of $d_j(u)$ is

$$E_C\{d_j(u)\} = \frac{n_j(u)}{n(u)} d(u), j = 1, \ldots, K.$$ 

**Note:** We use the notation $E_C(\cdot)$ to denote conditional expectation.

Also the conditional variance-covariance matrix of the $d_j(u)$’s is given by

$$\text{var}_C\{d_j(u)\} = \frac{d(u)\{n(u) - d(u)\}n_j(u)\{n(u) - n_j(u)\}}{n^2(u)\{n(u) - 1\}}, j = 1, \ldots, K,$$

and for $j \neq j'$

$$\text{cov}_C\{d_j(u), d_{j'}(u)\} = -\left[\frac{d(u)\{n(u) - d(u)\}n_j(u)n_{j'}(u)}{n^2(u)\{n(u) - 1\}}\right].$$

Again, analogous to the construction of the logrank test in two-samples, consider the $K - 1$ dimensional vector $T_n$, made up by the sum of the observed minus expected counts for treatments $j = 1, \ldots, K - 1$, summed over distinct death times $u$ such that $d(u) \geq 1$. That is

$$T_n = \begin{pmatrix}
\sum\text{death times } u \left\{d_1(u) - \frac{n_1(u)}{n(u)} d(u)\right\} \\
\vdots \\
\sum\text{death times } u \left\{d_{K-1}(u) - \frac{n_{K-1}(u)}{n(u)} d(u)\right\}
\end{pmatrix}^{(K-1)\times 1}$$

The corresponding $(K - 1) \times (K - 1)$ covariance matrix of the vector $T_n$ is given by $V_n^{(K-1)\times(K-1)}$, where for $j = 1, \ldots, K - 1$ the diagonal terms are given by

$$V_{njj} = \sum_{\text{death times } u} \left[\frac{d(u)\{n(u) - d(u)\}n_j(u)\{n(u) - n_j(u)\}}{n^2(u)\{n(u) - 1\}}\right],$$

and for $j \neq j'$, the off-diagonal terms are

$$V_{njj'} = -\sum_{\text{death times } u} \left[\frac{d(u)\{n(u) - d(u)\}n_j(u)n_{j'}(u)}{n^2(u)\{n(u) - 1\}}\right].$$
The test statistic used to test the null hypothesis is given by the quadratic form

\[ T_n = T_n^T [V_n]^{-1} T_n, \]

where we use the notation \((\cdot)^T\) to denote the transpose of a vector. Under \(H_0\), the statistic \(T_n\) is distributed asymptotically as a central chi-square distribution with \(K - 1\) degrees of freedom. This test statistic is called the \(K\)-sample logrank test statistic. If the null hypothesis is true, we would expect the elements in the vector \(T_n\) to be near zero. Hence the quadratic form, \(T_n\), should also be near zero. If however, there were treatment differences, then we would expect some of the elements in the vector \(T_n\) to deviate from zero and thus expect the quadratic form, \(T_n\), to be larger. Thus, we will reject the null hypothesis if \(T_n\) is sufficiently large. For a level \(\alpha\) test, we would reject \(H_0\) when

\[ T_n = T_n^T [V_n]^{-1} T_n \geq \chi^2_{\alpha, K-1}. \]

Proc lifetest in SAS implements this test and we will illustrate shortly using the data from CALGB 8541. But first let us end this section with a short discussion on how sample size calculations can be implemented during the design stage when one is comparing the survival distribution of more than two treatments with the logrank test.

### 8.7 Sample-size considerations for the \(K\)-sample logrank test

The computations of the sample size necessary to attain the desired power to detect clinically important treatment differences using the logrank test generalize from the two-sample problem to the \(K\)-sample problem in a manner similar to that considered for the comparison of proportions or means discussed in Chapter 7.

Specifically, if we randomize with equal probability to \(K\) treatments, then the total number of deaths necessary for the \(K\)-sample logrank test at the \(\alpha\) level of significance to have power at least \(1 - \beta\) to detect a hazard ratio (assumed constant over time) between any two treatments greater than equal to \(\exp(\gamma_A)\) is given by

\[ d = \frac{2K \phi^2(\alpha, \beta, K - 1)}{\gamma_A^2}, \]

where

\[ \phi^2(\alpha, \beta, K - 1), \]
is the value of the non-centrality parameter necessary so that a non-central chi-square distributed random variable with $K - 1$ degrees of freedom and non-centrality parameter $\phi^2(\alpha, \beta, K - 1)$ will exceed the value $\chi^2_{\alpha; K-1}$ with probability $(1 - \beta)$. Tables of $\phi^2(\alpha, \beta, K - 1)$ for $\alpha = .05$ were provided in chapter 7.

For example, if we take $K = 3$, then in order to ensure that we have at least 90% power to detect a hazard ratio between any two treatments that may exceed 1.5, using a logrank test at the .05 level of significance, we would need the total number of deaths to exceed

$$d = \frac{2 \times 3 \times 12.654}{(\log(1.5))^2} = 462.$$  

We can contrast this to a two-sample comparison which needs 256 events. As in the two-sample problem, the computations during the design stage will involve the best guesses for the accrual rate, accrual period, follow-up period, and underlying treatment-specific survival distributions which can be translated to the desired number of failures. Thus we can experiment with different values of

- accrual rate $a(u)$
- underlying treatment-specific failure time distributions $F_j(t) = P(T \leq t | A = j) = 1 - S_j(t), j = 1, \ldots, K$ under the alternative hypothesis of interest (we may take these at the least favorable configuration)
- the accrual period $A$
- the length of study $L$

so that

$$\sum_{j=1}^{K} d_j\{a(\cdot), F_j(\cdot), Acc, L\} = \frac{2K\phi^2(\alpha, \beta, K - 1)}{\gamma_A^2},$$

where $d_j\{a(\cdot), F_j(\cdot), Acc, L\}$ denotes the expected number of deaths in treatment group $j$ as a function of $a(\cdot), F_j(\cdot), Acc, L$, computed using equation (8.5).

**CALGB 8541 Example**

We now return to the data from CALGB 8541 which compared three treatments in a randomized study of node positive stage II breast cancer patients. The three treatments were
• treatment 1 (Intensive CAF)

• treatment 2 (Low dose CAF)

• treatment 3 (Standard dose CAF)

where CAF denotes the combination of Cyclophosphamide, Adriamycin and 5-Fluourouracil. As well as testing for overall differences in the three treatments, we shall also look at the three pairwise comparisons.

**SAS program:**

```sas
title "Log-rank test comparing all three treatments";
proc lifetest data=bcancer notable;
   time years*censor(0);
   strata trt;
run;
```

Part of the output:

```
Log-rank test comparing all three treatments

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for years over Strata

Rank Statistics

<table>
<thead>
<tr>
<th>trt</th>
<th>Log-Rank</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-21.245</td>
<td>-27171</td>
</tr>
<tr>
<td>2</td>
<td>37.653</td>
<td>43166</td>
</tr>
<tr>
<td>3</td>
<td>-16.408</td>
<td>-15995</td>
</tr>
</tbody>
</table>

Covariance Matrix for the Log-Rank Statistics

<table>
<thead>
<tr>
<th>trt</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120.132</td>
<td>-57.761</td>
<td>-62.371</td>
</tr>
<tr>
<td>2</td>
<td>-57.761</td>
<td>114.004</td>
<td>-56.243</td>
</tr>
<tr>
<td>3</td>
<td>-62.371</td>
<td>-56.243</td>
<td>118.615</td>
</tr>
</tbody>
</table>

Covariance Matrix for the Wilcoxon Statistics

<table>
<thead>
<tr>
<th>trt</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
1.6295E8 -7.94E7 -8.355E7
-7.94E7 1.5675E8 -7.734E7
-8.355E7 -7.734E7 1.6089E8

Test of Equality over Strata

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Rank</td>
<td>12.4876</td>
<td>2</td>
<td>0.0019</td>
</tr>
<tr>
<td>Wilcoxon</td>
<td>12.1167</td>
<td>2</td>
<td>0.0023</td>
</tr>
<tr>
<td>-2Log(LR)</td>
<td>11.3987</td>
<td>2</td>
<td>0.0033</td>
</tr>
</tbody>
</table>