4 Two (K) Sample Problems

In many biomedical experiments we are interested in comparing the survival distributions between two or more groups. For example, in phase III clinical trials we may be interested in comparing the survival distributions between two or more competing treatments on patients with a particular disease. For the time being, we will consider two sample comparisons and later extend to $k > 2$ sample comparisons.

The problem of comparing two treatments with respect to a time to event endpoint can be posed as a hypothesis testing problem. Let Z denote the treatment indicator. That is, $Z = 1$ for treatment 1 and $Z = 0$ for treatment 0.

In general, we will use treatment 0 (or group 0) to mean the standard treatment or placebo comparator, and treatment 1 (or group 1) to denote the new treatment that is to be compared to the standard or the placebo. This of course does not have to be the case; for example, we may be comparing two new promising treatments to each other in a disease for which there is no agreement what treatment is standard.

The null hypothesis is generally that of no treatment (group) difference; that is, the distribution of the time to event is the same for both treatments. If we denote by $S_0(t)$ and $S_1(t)$ the survival functions for treatments 0 and 1 respectively, then the null hypothesis can be expressed as

$$H_0 : S_0(t) = S_1(t), \quad \text{for } t \geq 0,$$

or equivalently,

$$H_0 : \lambda_0(t) = \lambda_1(t), \quad \text{for } t \geq 0,$$

where $\lambda_0(t)$ and $\lambda_1(t)$ are the hazard functions for treatments 0 and 1 respectively. Recall that

$$\lambda_j(t) = -\frac{d\log\{S_j(t)\}}{dt}, \quad j = 0, 1.$$

The alternative hypothesis we are most interested in is that the survival time for one treatment is **stochastically larger or smaller** than the survival time for the other treatment.

For example, we may be interested in the alternative that the new treatment is better than the standard one.

$$H_a : S_1(t) \geq S_0(t), \quad \text{for } t \geq 0, \text{ with strict inequality for some } t.$$

This is an example of a one-sided alternative . Most often, we are interested in declaring a difference from the null hypothesis if either treatment is better than the other. If this is the case, we use a two sided alternative:

$$H_a : \quad \text{either} \quad S_1(t) \geq S_0(t), \quad \text{or} \quad S_0(t) \geq S_1(t), \quad \text{with strict inequality for some } t.$$

In biomedical applications, it has become common practice to use nonparametric tests; that is, using test statistics whose distribution under the null hypothesis does not depend on specific parametric assumptions on the shape of the probability distribution. With censored survival data, the class of weighted logrank tests are mostly used to test the null hypothesis of treatment equality, with the logrank test being the most commonly used.

Censored survival data for comparing two groups are given as a sample of triplets (X_i, Δ_i, Z_i), $i = 1, 2, ..., n$, where

$$\begin{aligned} X_i &= \min(T_i, C_i), \\ &\quad T_i = \text{latent failure time} \\ &\quad C_i = \text{latent censoring time} \\ \Delta_i &= I(T_i \leq C_i) \\ Z_i &= \begin{cases} 1 & \text{new treatment} \\ 0 & \text{standard treatment} \end{cases} \end{aligned}$$

We now define the following notation:

$$n_1 = \text{number of individuals in group 1}$$

$$n_0 = \text{number of individuals in group 0}$$

Obviously,

$$\begin{aligned} n_j &= \sum_{i=1}^{n} I(Z_i = j), \quad j = 0, 1 \\ n &= n_0 + n_1. \end{aligned}$$

The number of individuals at risk at time x from treatments 0 and 1 is denoted by $Y_0(x)$ and $Y_1(x)$ respectively, where

$$\begin{aligned} Y_0(x) &= \sum_{i=1}^{n} I(X_i \geq x, Z_i = 0), \\ Y_1(x) &= \sum_{i=1}^{n} I(X_i \geq x, Z_i = 1), \end{aligned}$$

and the total number at risk at time x is denoted by

$$Y(x) = Y_0(x) + Y_1(x).$$

Similarly, let $dN_0(x)$ and $dN_1(x)$ denote the number of deaths observed at time x from treatments 0 and 1 respectively,

$$\begin{aligned} dN_0(x) &= \sum_{i=1}^{n} I(X_i = x, \Delta_i = 1, Z_i = 0), \\ dN_1(x) &= \sum_{i=1}^{n} I(X_i = x, \Delta_i = 1, Z_i = 1), \end{aligned}$$

and

$$dN(x) = \sum_{i=1}^{n} I(X_i = x, \Delta_i = 1) = dN_0(x) + dN_1(x).$$

Note: In some applications, $dN(x)$ will actually correspond to the observed number of deaths in time window $[x, x + \Delta x)$ for some partition of the time axis into intervals of length Δx. If the partition is sufficiently fine then thinking of the number of deaths occurring exactly at x or in $[x, x + \Delta x)$ makes little difference, and in the limit makes no difference at all. When we are dealing with data, we can view $dN(x)$ as the number of deaths observed at time x. In theory

the probability that we can observe a death at time x is always zero. So we understand $dN(x)$ as the number of deaths observed in $[x, x + \Delta x)$ in our theoretical arguments later.

The weighted logrank test statistic is given by

$$T(w) = \frac{U(w)}{\text{se}(U(w))}$$

where

$$U(w) = \sum_x w(x) \left\{ dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)} \right\},$$

and $\text{se}(U(w))$ will be given later. The null hypothesis of treatment equality will be rejected if $T(w)$ is sufficiently different from zero.

Note: At any time x for which there is no observed death

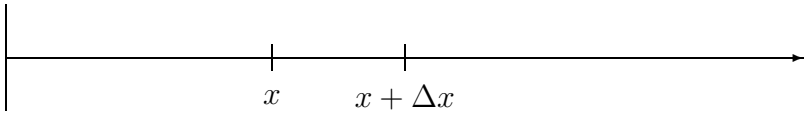
$$dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)} = 0.$$

This means that the sum above is only over distinct failure times.

Motivation of the test

These two-sample tests can be viewed as a weighted sum over the distinct failure times of observed number of deaths from treatment 1 minus the expected number of deaths from treatment 1 if the null hypothesis were true.

Figure 4.1: *A slice of time*



Take a slice of time as shown in Figure 4.1 and consider the following resulting 2×2 table

If H_0 is true, then conditional on $Y_1(x), Y(x)$ and $dN(x)$,

$$dN_1(x) | (Y_1(x), Y(x), dN(x)) \sim \text{Hypergeometric}\left(Y_1(x), dN(x), Y(x)\right).$$

Table 4.1: 2×2 *table from* $[x, x + \Delta x)$

	Treatment		
	0	1	total
# of death	$dN_0(x)$	$dN_1(x)$	$dN(x)$
# of not dying	$Y_0(x) - dN_0(x)$	$Y_1(x) - dN_1(x)$	$Y(x) - dN(x)$
# at risk	$Y_0(x)$	$Y_1(x)$	$Y(x)$

This is analogous to assuming that there are $dN(x)$ black balls and $Y(x) - dN(x)$ white balls. Randomly draw $Y_1(x)$ balls from these $Y(x)$ balls using sampling without replacement. Then the number of black balls $dN_1(x)$ in the sample has the above hypergeometric distribution.

Obviously,

$$\mathrm{E}[dN_1(x)|(Y_1(x), Y(x), dN(x))] = \frac{dN(x)}{Y(x)} Y_1(x).$$

Since the observed number of deaths at x from treatment 1 is $dN_1(x)$, so the observed minus expected is equal to

$$\left\{ dN_1(x) - \frac{dN(x) \times Y_1(x)}{Y(x)} \right\}.$$

From this point of view, the censored survival data can be viewed as k such 2×2 tables, where k corresponds to the total number of distinct failure times from two groups combined.

If the null hypothesis were true, we would expect

$$\left\{ dN_1(x) - \frac{dN(x) \times Y_1(x)}{Y(x)} \right\}$$

to be equal to zero on the average, and hence so should the sum over all x. If however, the hazard rate for treatment 1 were lower than that for treatment 0 **consistently** over x, then on

average we would expect

$$\left\{dN_1(x) - \frac{dN(x) \times Y_1(x)}{Y(x)}\right\}$$

to be negative. The opposite should be true if the hazard rate for treatment 1 were consistently higher than that for treatment 0.

This suggests that we should reject the null hypothesis if our test statistics T_w is sufficiently far from zero, positive, negative, or in absolute value depending on the alternative hypothesis.

In order to measure the strength of the evidence against the null hypothesis, we must be able to evaluate the distribution of the test statistic (at least approximately) under the null hypothesis. Specifically, the weighted logrank test statistic is given by

$$T(w) = \frac{\sum_x w(x)\left[dN_1(x) - \frac{dN(x)\times Y_1(x)}{Y(x)}\right]}{\left\{\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x)-dN(x)]}{Y^2(x)[Y(x)-1]}\right]\right\}^{1/2}}.$$

Under the null hypothesis, this test statistic is approximately distributed as a standard normal

$$T(w) \overset{a}{\sim} \mathrm{N}(0,1).$$

Therefore, a level α test (two-sided) will reject $H_0 : S_0(t) = S_1(t)$ whenever

$$|T(w)| \geq z_{\alpha/2},$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of a standard normal distribution.

A heuristic justification for this result will be given shortly. We want to mention however, that the most commonly used test statistic is the log-rank test, where $w(x) = 1$ for all x.

$$\text{logrank test stat} = \frac{\sum_x \left[dN_1(x) - \frac{dN(x)\times Y_1(x)}{Y(x)}\right]}{\left\{\sum_x \left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x)-dN(x)]}{Y^2(x)[Y(x)-1]}\right]\right\}^{1/2}}.$$

Remark: The statistic in the numerator is a weighted sum of observed minus the expected over the k 2×2 tables, where k is the number of distinct failure times.

The weight function $w(x)$ can be used to emphasize differences in the hazard rates over time according to their relative values. For example, if the weight early in time is larger and later becomes smaller, then such test statistic would emphasize early differences in the survival curves. The weights to be chosen depends on the type of alternative difference we wish to detect.

Note: If the weights $w(x)$ are stochastic (functions of data), then they need to be a function of the censoring and survival information *prior to* time x.

The most commonly used test is the logrank test where $w(x) = 1$ for all x. Other tests given in the literature are:

1. Gehan's generalization of Wilcoxon test that uses $w(x) = Y(x)$.

2. Peto-Prentice's generalization of Wilcoxon test that uses $w(x) = KM(x)$, where $KM(x)$ is the Kaplan-Meier estimator using the combined sample; *i.e.*,

$$w(x) = \prod_{u \leq x} \left[1 - \frac{dN(u)}{Y(u)}\right].$$

Since both $Y(x)$ and $KM(x)$ are non-increasing functions of x, both tests emphasize the difference early in the survival curves.

Heuristic proof of the statistical properties for the weighted logrank test

As you will soon see, the proofs are similar to those used to find the variance of the Nelson-Aalen estimator, and will rely heavily on the double expectation theorem (or iterative expectation (variance) theorem).

Toward that end, we define a set of random variables

$$\mathcal{F}(x) = \{dN_0(u), dN_1(u), Y_0(u), Y_1(u), w_0(u), w_1(u), dN(x) \text{ for all grid points } u < x\}.$$

That is, when we define $\mathcal{F}(x)$, then we know all the failure and censoring that has occurred prior to time x form either treatment, the number of individuals at risk at time x as well as the

number of total deaths ($dN(x)$) that occurs in $[x, x+\Delta x)$. What we don't know is the number of deaths from each treatment group in $[x, x+\Delta x)$.

Let us consider the 2×2 table (Table 4.1) that is created using the slice of time $[x, x+\Delta x)$.

We already have argued that given an individual is at risk at time x, and is in treatment group 1, then (assuming independent censoring) the probability of dying in $[x, x+\Delta x)$ is equal to $\lambda_1(x)\Delta x$, where $\lambda_1(x)$ is the hazard function for treatment group 1.

Similarly, the probability is equal to $\lambda_0(x)\Delta x$ for an individual at risk at time x from treatment group 0. Under the null hypothesis

$$H_0 : \lambda_1(x) = \lambda_0(x),$$

the conditional probability of dying in $[x, x+\Delta x)$, given being at risk at time x, is the same for both treatment groups.

Assume the null hypothesis is true. Knowing $\mathcal{F}(x)$ would imply (with respect to the 2×2 table) that:

> We know $Y_1(x), Y_0(x)$ (*i.e.*, the number at risk at time x from either treatment group), and, in addition, we know $dN(x)$ (*i.e.*, the number of deaths (total from both treatment groups) occurring in $[x, x+\Delta x)$).

The only thing we don't know about the 2×2 table is $dN_1(x)$ (Note: knowing this would complete the knowledge of the counts in the 2×2 table).

Fact: In a 2×2 table, under the assumption of independence, the count in one cell of the table, conditional on the marginal counts, follows a hypergeometric distribution. (This is the basis of Fisher's exact test for independence in a 2×2 table).

Conditional on $\mathcal{F}(x)$, we have a 2×2 table which under the null hypothesis follows independence and we have the knowledge of the marginal counts of the table (*i.e.*, the marginal counts are fixed conditional on $\mathcal{F}(x)$).

Therefore, the conditional distribution of one of the counts, say, $dN_1(x)$, in the cell of the table, given $\mathcal{F}(x)$ follows a hypergeometric distribution.

This is equivalent to imaging that there are $Y(x) = dN(x) + (Y(x) - dN(x))$ balls in a urn, of which $dN(x)$ are black balls, and $(Y(x) - dN(x))$ are white balls. We then randomly draw $Y_1(x)$ balls from the urn without replacement. Let $dN_1(x)$ be the number of black balls in this sample. Then $dN_1(x)$ has a hypergeometric distribution, *i.e.*,

$$P[dN_1(x) = c|Y_1(x), Y_0(x), dN(x)] = \frac{\begin{pmatrix} dN_1(x) \\ c \end{pmatrix} \begin{pmatrix} Y(x) - dN(x) \\ Y_1(x) - c \end{pmatrix}}{\begin{pmatrix} Y(x) \\ Y_1(x) \end{pmatrix}}.$$

From the properties of a hypergeometric distribution, we know that conditional on $\mathcal{F}(x)$, $dN_1(x)$ has the following mean and variance

$$\begin{aligned} \mathrm{E}[dN_1(x)|\mathcal{F}(x)] &= \frac{dN(x)Y_1(x)}{Y(x)}, \\ \mathrm{Var}[dN_1(x)|\mathcal{F}(x)] &= \frac{dN(x)Y_1(x)Y_0(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}. \end{aligned}$$

Note that $Y_1(x)$ is the sample size,

$$\begin{aligned} &\frac{dN(x)}{Y(x)} = \text{proportion of black balls}, \\ &\frac{Y(x) - dN(x)}{Y(x)} = \text{proportion of white balls}, \\ &\frac{Y(x) - Y_1(x)(= Y_0(x))}{Y(x) - 1} = \text{variance correction factor}. \end{aligned}$$

Now that we have taken care of some preliminaries, let us go back to our weighted logrank test. The first thing I want to demonstrate is that under the null hypothesis the weighted logrank test statistic has a numerator with mean zero.

The numerator of the weighted logrank test statistic is

$$U(w) = \sum_x w(x) \left[dN_1(x) - \frac{dN(x)Y_1(x)}{Y(x)} \right],$$

which has the expectation

$$\begin{aligned} \mathrm{E}[U(w)] &= \sum_x \mathrm{E}\left[w(x)\left[dN_1(x) - \frac{dN(x)Y_1(x)}{Y(x)}\right]\right] \\ &= \sum_x \mathrm{E}\left\{\mathrm{E}\left[w(x)\left[dN_1(x) - \frac{dN(x)Y_1(x)}{Y(x)}\right]\middle| \mathcal{F}(x)\right]\right\}. \end{aligned}$$

By the assumption we made about $w(x)$, we know that $w(x)$ is a function of data prior to x. That is to say, conditional on $\mathcal{F}(x)$, $w(x)$ is a known value. Again given $\mathcal{F}(x)$, $Y_1(x)$, $dN(x)$ and $Y(x)$ are known. Therefore, the inner expectation in the above sum can be written as

$$\begin{aligned} & \mathrm{E}\left[w(x)\left[dN_1(x) - \frac{dN(x)Y_1(x)}{Y(x)}\right]\middle| \mathcal{F}(x)\right] \\ = \; & w(x)\left[\mathrm{E}[dN_1(x)|\mathcal{F}(x)] - \frac{dN(x)Y_1(x)}{Y(x)}\right] \\ = \; & 0. \end{aligned}$$

So th inner expectation is equal to zero. Consequently, so is the total expectation as the sum of the expectations. Therefore, under the null hypothesis, we have

$$\mathrm{E}[U(w)] = 0.$$

Finding an unbiased estimator for the variance of $U(w)$

For ease of notation, let us define

$$U(w) = \sum_x A(x),$$

where

$$A(x) = w(x)\left[dN_1(x) - \frac{dN(x)Y_1(x)}{Y(x)}\right].$$

The variance of $U(w)$ is

$$\begin{aligned} \mathrm{Var}[U(w)] &= \mathrm{Var}\left[\sum_x A(x)\right] \\ &= \sum_x \mathrm{Var}(A(x)) + \sum_{x \neq y} \mathrm{Cov}(A(x), A(y)). \end{aligned}$$

By using a conditioning argument, we will now show that each covariance term (*i.e.*, the cross product term) is equal to zero. Let us take one arbitrary covariance term $\text{Cov}(A(x), A(y))$ for $y < x$, where x and y are the grid points of the partition of the time axis.

Remember that we have already shown that

$$\text{E}(A(x)) = 0 \text{ and } \text{E}(A(y)) = 0.$$

Therefore,

$$\text{Cov}(A(x), A(y)) = \text{E}[A(x) * A(y)].$$

By the double expectation theorem, we have

$$\begin{aligned} \text{Cov}(A(x), A(y)) &= \text{E}[A(x) * A(y)] \\ &= \text{E}[\text{E}[A(x) * A(y)|\mathcal{F}(x)]]. \end{aligned}$$

Since $y < x$, this implies that all the elements which make up $A(y)$ (, *i.e.*, $dN_1(y), Y_1(y), dN(y), Y(y)$) are known when we condition on $\mathcal{F}(x)$. Hence the inner expectation is

$$\text{E}[A(x) * A(y)|\mathcal{F}(x)]] = A(y) * \text{E}[A(x)|\mathcal{F}(x)]] = 0.$$

Therefore we showed that

$$\text{Cov}(A(x), A(y)) = 0.$$

Since $y < x$ are arbitrary, hence all the covariance terms are equal to zero. Therefore,

$$\begin{aligned} \text{Var}[U(w)] &= \sum_x \text{Var}(A(x)) \\ &= \sum_x \text{E}[A^2(x)] \\ &= \sum_x \text{E}[\text{E}[A^2(x)|\mathcal{F}(x)]]. \end{aligned}$$

Let us examine the inner expectation more closely:

$$\begin{aligned}
\mathrm{E}[A^2(x)|\mathcal{F}(x)] &= \mathrm{E}\left[w^2(x)\left[dN_1(x) - \frac{Y_1(x)dN(x)}{Y(x)}\right]^2 \middle| \mathcal{F}(x)\right] \\
&= w^2(x) * \mathrm{E}[\{dN_1(x) - \mathrm{E}[dN_1(x)|\mathcal{F}(x)]\}^2|\mathcal{F}(x)] \\
&= w^2(x) * \mathrm{Var}[dN_1(x)|\mathcal{F}(x)] \\
&= w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x)-1]}\right].
\end{aligned}$$

Therefore, we have shown that

$$\begin{aligned}
\mathrm{Var}[U(w)] &= \sum_x \mathrm{E}[\mathrm{E}[A^2(x)|\mathcal{F}(x)]] \\
&= \sum_x \mathrm{E}\left[w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x)-1]}\right]\right],
\end{aligned}$$

which means that the following statistic

$$\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x)-1]}\right]$$

is an unbiased estimate for $\mathrm{Var}[U(w)]$.

Recapping: Under the null hypothesis $H_0 : S_0(t) = S_1(t)$,

1. The statistic $U(w) = \sum_x A(x)$ has expectation equal to zero.

$$\mathrm{E}(U(w)) = 0.$$

2. $U(w) = \sum_x A(x)$ is made up of a sum of conditionally uncorrelated terms each with mean zero. By the central limit theory for such martingale structures, $U(w)$ properly normalized will be approximately a standard normal random variable. That is

$$\frac{U(w)}{\mathrm{se}(U(w))} \overset{a}{\sim} \mathrm{N}(0,1), \quad \text{under } H_0$$

3. We showed that an unbiased estimate for the variance of $U(w)$ was given by

$$\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x)-1]}\right].$$

Therefore

$$T(w)=\frac{U(w)}{\text{se}(U(w))}=\frac{\sum_x w(x)\left[dN_1(x)-\frac{dN(x)*Y_1(x)}{Y(x)}\right]}{\left\{\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x)-dN(x)]}{Y^2(x)[Y(x)-1]}\right]\right\}^{1/2}}\overset{a}{\sim}\text{N}(0,1).$$

This ends the heuristic proof.

We will illustrate the tests (logrank test, Gehan's Wilcoxons and Peto-Prentice's Wilcoxon test) using the following data set (data file = `myel.dat`) taken from Paul Allison's book. The data give the survival times for 25 myelomatosis patients randomized to two treatments (1 or 2):

```
dur status trt renal
8    1 1 1
180  1 2 0
632  1 2 0
852  0 1 0
52   1 1 1
2240 0 2 0
220  1 1 0
63   1 1 1
195  1 2 0
76   1 2 0
70   1 2 0
8    1 1 0
13   1 2 1
1990 0 2 0
1976 0 1 0
18   1 2 1
700  1 2 0
1296 0 1 0
1460 0 1 0
210  1 2 0
63   1 1 1
1328 0 1 0
1296 1 2 0
365  0 1 0
23   1 2 1
```

where `dur` is the patient's survival or censored time, `status` is the censoring indicator, `trt` is the treatment indicator and `renal` is the indicator of impaired renal function (0 = normal; 1 = impaired). To test the null hypothesis the treatment `trt` has no effect (*i.e.*, $H_0: S_0(t)=S_1(t)$), we used the following `SAS` program to perform logrank and Gehan's Wilcoxon tests:

```
options ls=78 ps=60;
```

```
data myel;
  infile "myel.dat" firstobs=2;
  input dur status trt renal;
run;

proc lifetest data=myel;
  time dur*status(0);
  strata trt;
run;
```

Part of the output from this program gives logrank and Gehan's Wilcoxon tests:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                        09:21 Tuesday, January 11, 2000   3

                        The LIFETEST Procedure

                Testing Homogeneity of Survival Curves over Strata
                              Time Variable DUR


                          Rank Statistics

                TRT          Log-Rank     Wilcoxon

                1             -2.3376      -18.000
                2              2.3376       18.000


              Covariance Matrix for the Log-Rank Statistics

                TRT                  1             2

                1              4.16301      -4.16301
                2             -4.16301       4.16301


              Covariance Matrix for the Wilcoxon Statistics

                TRT                  1             2

                1               1301.00      -1301.00
                2              -1301.00       1301.00


                    Test of Equality over Strata

                                                  Pr >
            Test      Chi-Square    DF  Chi-Square

            Log-Rank      1.3126     1      0.2519
            Wilcoxon      0.2490     1      0.6178
            -2Log(LR)     1.5240     1      0.2170
```

This output gives the numerators of logrank and Gehan's Wilcoxon tests and their estimated variances for each group. So for example, the numerator of logrank test for treatment 1 is -2.3376 with its estimated variance 4.16301 (negative means on average treatment 1 is better than treatment 2; but we have to judge this statement using the p-value from our test). So $(-2.3376)^2/4.16301 = 1.3126$ and the p-value of this test is $P[\chi^2 > 1.3126] = 0.2519$. Similarly the p-value for Gehan's Wilcoxon test is 0.6178. Note that the numerator of Gehan's Wilcoxon test is much larger than that of logrank test since Gehan's Wilcoxon test uses the number at risk as the weight and logrank test uses identity weight. (The last test is likelihood ratio test based on exponential model. See next chapter for more detail).

In this example, logrank test gives a more significant result than Gehan's Wilcoxon test (although none of them provides strong evidence against the null hypothesis). Why is that? Recall that Gehan's Wilcoxon test puts more weight to early times than to later times (the number at risk is a decreasing function of time) and logrank test puts equal weight to all times. So if the true survival distributions for treatments 1 and 2 differ less early on than later (of course, there eventually will be no difference in their survival functions when the time is sufficiently large), then logrank test is more powerful (more sensitive) than Gehan's Wilcoxon test. This is the case if $\lambda_1(t)$ and $\lambda_0(t)$ are related by

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

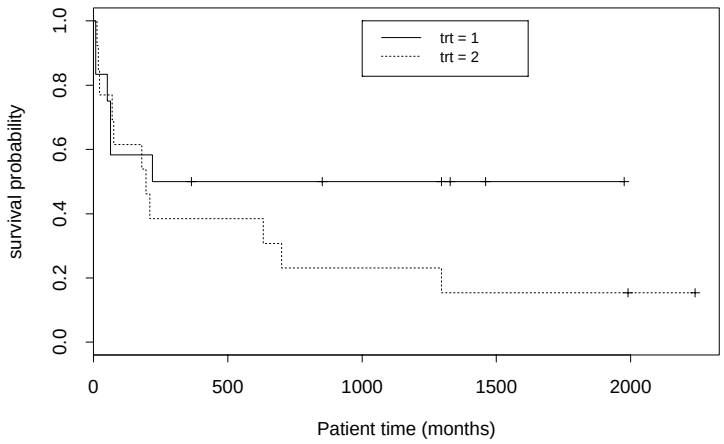
where $\gamma > 0$ is a constant. This means that the hazard for group 1 is *proportional to* that for group 0. This is the *proportional hazards model* proposed by D.R. Cox. We will discuss this more when we talk about the power of these tests.

The treatment specific Kaplan-Meier survival estimates were generated using the following splus functions and were presented in Figure 4.2.

```
  postscript(file="fig4.2.ps", horizontal = F,
    height=6, width=8.5, pointsize=14)
 # par(mfrow=c(1,2))

  example <- read.table(file="myel.dat", header=T)
```

Figure 4.2: *Kaplan-Meier estimates for two treatments*




```
  fit <- survfit(Surv(dur, status) ~ trt, example)
  plot(fit, xlab="Patient time (months)", ylab="survival probability",
       lty=c(1,2))

  legend(1000,1, c("trt = 1", "trt = 2"),
lty=c(1,2), cex=0.8)

  dev.off()
```

Figure 4.2 shows that there is less difference in the estimated survival functions early on than later. So logrank test gives a more significant result.

We can also use the following splus functions to do logrank test:

```
> survdiff(Surv(dur, status) ~ trt, example)
 survdiff(Surv(dur, status) ~ trt, example)
Call:
survdiff(formula = Surv(dur, status) ~ trt, data = example)

       N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 12        6     8.34     0.655      1.31
trt=2 13       11     8.66     0.631      1.31

 Chisq= 1.3  on 1 degrees of freedom, p= 0.252
```

If we want to perform Peto-Prentice's Wilcoxon test, we need to specify `rho=1` in the above splus functions:

```
> survdiff(Surv(dur, status) ~ trt, rho=1, example)
```

```
survdiff(Surv(dur, status) ~ trt, rho=1, example)
Call:
survdiff(formula = Surv(dur, status) ~ trt, data = example, rho = 1)

       N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 12     4.80     5.60     0.115     0.304
trt=2 13     6.83     6.03     0.106     0.304

 Chisq= 0.3  on 1 degrees of freedom, p= 0.581
```

This test has the similar p-value to Gehan's Wilcoxon test since both tests put more weights to the earlier time than to later time.

Power and Sample Size

Focus so far has been on the null hypothesis. We showed that weighted logrank test statistics (after properly normalized) are asymptotically distributed as a standard normal if the null hypothesis is true, enabling us to use these test statistics to compute p-value to assess the strength of evidence against the null hypothesis (in favor of treatment difference)

In considering the sensitivity of the tests, we must also assess the power, or the probability of rejecting the null when "in truth" we have departures from the null. Describing departures from the null hypothesis that we feel are important to detect is complicated. That is because a survival curve is infinite dimensional and departures from the null have to be described as differences at every point in time over the survival curve. Clearly, some simplifying conditions must be given. In clinical trials proportional hazards alternatives have become very popular. That is

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

We use $\exp(\beta)$, since by necessity, hazard ratios have to be positive and that $\beta = 0$ would correspond to **no** treatment difference.

Note:

1. $\beta > 0 \Rightarrow$ individuals on treatment 1 have worse survival (*i.e.*, die faster).

2. $\beta = 0 \Rightarrow$ no treatment difference (null is true)

3. $\beta < 0 \Rightarrow$ individuals on treatment 1 have better survival (*i.e.*, live longer).

Other ways of representing proportional hazards follow from the following relationship

$$\begin{aligned} & \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta) \\ \Longleftrightarrow \quad & -\frac{d\log\{S_1(t)\}}{dt} = -\frac{d\log\{S_0(t)\}}{dt}\exp(\beta) \\ \Longleftrightarrow \quad & \frac{d\log\{S_1(t)\}}{dt} = \frac{d\log\{S_0(t)\}}{dt}\exp(\beta) \\ \Longleftrightarrow \quad & \log\{S_1(t)\} = \log\{S_0(t)\}\exp(\beta) + C, \end{aligned} \tag{4.1}$$

where C is a constant to be determined. In the above identity, take $t = 0$, we get $C = 0$. Therefore we get

$$\begin{aligned} & \log\{S_1(t)\} = \log\{S_0(t)\}\exp(\beta) \\ \Longleftrightarrow \quad & S_1(t) = S_0^{\gamma}(t), \end{aligned} \tag{4.2}$$

where $\gamma = \exp(\beta)$.

If we multiply both sides of (4.3) by -1 and then take log, we will have:

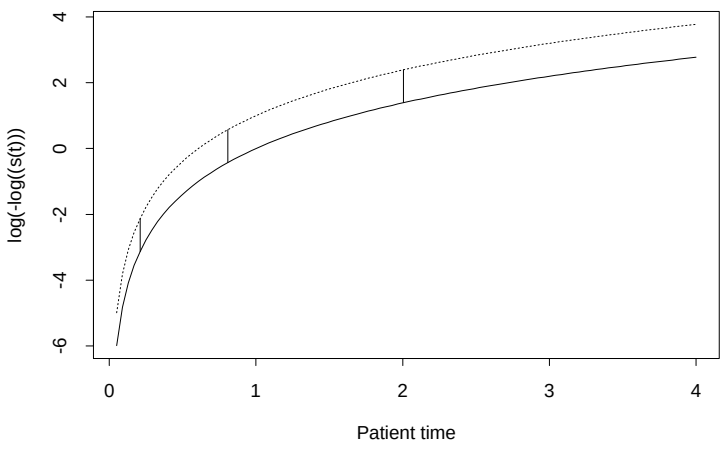
$$\log[-\log\{S_1(t)\}] = \log[-\log\{S_0(t)\}] + \beta.$$

(Since $0 \leq S_j(t) \leq 1$, $\log\{S_j(t)\} < 0$. So we need to multiply $\log\{S_j(t)\}$ by -1 before we can take log)

This last relationship is very useful to help us identify situations where we may have proportional hazards. By plotting estimated survival curves (say, Kaplan-Meier estimates) for two treatments (groups) on a log[-log] scale, we would see constant vertical shift of the two curves if the hazards are proportional. The situation is illustrated in Figure 4.3. In this case, we say two curves are parallel. Do not be misled by the visual impression of the curves near the origin.

For the specific case where the survival curves for the two groups are exponentially distributed

Figure 4.3: *Two survival functions with proportional hazards on log[-log] scale*



(*i.e.*, constant hazard), we automatically have proportional hazards, since

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

The median survival time for an exponentially distributed random variable is given by m

$$S(m) = \mathrm{e}^{-\lambda m} = 0.5, \quad \text{or} \quad m = \log(2)/\lambda.$$

The ratio of median survival times for two groups having exponential distributions is

$$\frac{m_1}{m_0} = \frac{\log(2)/\lambda_1}{\log(2)/\lambda_0} = \frac{\lambda_0}{\lambda_1}, \tag{4.3}$$

i.e., the ratio of median survival times is inversely proportional to the ratio of hazard rates. This result may be useful when trying to illicit clinically important differences from your collaborators. If survival times are exponentially distributed (or approximately so) then the desired increase in median survival times can be easily translated to the desired difference in hazard ratio.

Note: If the survival times have exponential distributions, then the ratio of mean survival times is also inversely proportional to the ratio of hazard rates. Therefore, the clinically important difference in the survival times can also be specified via the ratio of mean survival times.

The logrank test is the **most powerful** test among the weighted logrank tests to detect proportional hazards alternatives. In fact, it is the most powerful test among all nonparametric tests for detecting proportional hazards alternatives.

Therefore, the proportional hazards alternative has not only a nice interpretation but also nice statistical properties. These features leads to the natural use of (unweighted) logrank tests.

In order to compute power to detect the difference of interest and corresponding sample sizes, we must be able to derive the distribution of our test statistic under the alternative hypothesis (here proportional hazards alternative). When censoring does not depend on treatment (*e.g.*, randomized experiments), the logrank test has distribution under the following alternative

$$H_A : \frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta_A); \quad \beta_A \neq 0, \tag{4.4}$$

approximated by

$$T_n \overset{a}{\sim} \mathrm{N}\left(\beta_A\sqrt{d\theta(1-\theta)}, 1\right),$$

where d is the total number of deaths (events), θ is the proportion in group 1, β_A is the log hazard ratio under the alternative. That is, under the proportional hazards alternative, the logrank test statistic is distributed approximated as a normal distribution with mean $\beta_A\sqrt{d\theta(1-\theta)}$ and variance 1.

For the common case where $\theta = 1/2$ (randomization with equal probability), the mean equals

$$\frac{\beta_A}{2}\sqrt{d}.$$

Such a result will also be useful to aid us in determining sample size during the design stage of an experiment. It is fairly easy to show that in order that a level α test (say, two-sided) has power $1-\gamma$ to detect the alternative of interest, the mean of the test statistic (under the alternative) must be equal to $(z_{\alpha/2} + z_\gamma)$.

Remark: We are using γ here to describe the type II error probability since we already used β to describe the log hazard ratio. "β_A" is used to denote the log hazard ratio that is felt to be clinically important to detect.

Let $\mu = \beta_A\sqrt{d\theta(1-\theta)}$, the mean of the log rank test statistic T_n under the alternative 4.4. Recall that our test procedure:

$$\text{reject } H_0 \text{ when } |T_n| > z_{\alpha/2},$$

and

$$T_n \overset{a}{\sim} \mathrm{N}(0,1) \text{ under } H_0 \text{ and } T_n \overset{a}{\sim} \mathrm{N}(\mu,1) \text{ under } H_A.$$

By the definition of power, we have

$$\begin{aligned} & P[|T_n| > z_{\alpha/2}|H_A] = 1-\gamma \\ \iff\ & P[T_n > z_{\alpha/2}|H_A] + P[T_n < -z_{\alpha/2}|H_A] = 1-\gamma \end{aligned}$$

Assume $\beta_A > 0$ at this moment, then $\mu > 0$. In this case,

$$\begin{aligned} P[T_n < -z_{\alpha/2}|H_A] &= P[T_n - \mu < -z_{\alpha/2} - \mu|H_A] \\ &= P[Z < -z_{\alpha/2} - \mu] \ \ (Z \sim \mathrm{N}(0,1)) \\ &= P[Z > z_{\alpha/2} + \mu] \\ &\approx 0 \ \ (\text{at least less than } \alpha/2, \text{ since } P[Z > z_{\alpha/2}] = \alpha/2), \end{aligned}$$

and

$$\begin{aligned} P[T_n > z_{\alpha/2}|H_A] &= P[T_n - \mu > z_{\alpha/2} - \mu|H_A] \\ &= P[Z > z_{\alpha/2} - \mu] \ \ (Z \sim \mathrm{N}(0,1)) \end{aligned}$$

Therefore,

$$\begin{aligned} & P[Z > z_{\alpha/2} - \mu] \approx 1-\gamma \\ \iff\ & P[Z < z_{\alpha/2} - \mu] \approx \gamma \\ \iff\ & P[Z > -z_{\alpha/2} + \mu] \approx \gamma \\ \iff\ & -z_{\alpha/2} + \mu \approx z_\gamma \ \ (\text{since } P[Z > z_\gamma] = \gamma \text{ by definition}) \\ \iff\ & \mu = z_{\alpha/2} + z_\gamma. \end{aligned}$$

Consequently,

$$\begin{aligned} & \sqrt{d}\beta_A\sqrt{\theta(1-\theta)} = z_{\alpha/2} + z_\gamma \\ \iff \quad & d = \frac{(z_{\alpha/2} + z_\gamma)^2}{(\beta_A)^2 * \theta(1-\theta)}. \end{aligned}$$

Exactly the **same** formula for d can be derived if $\beta_A < 0$. This is the requirement for number of events "d" we have to observ in order for our level α logrank test to have a power $1-\gamma$. In this sense, "d" acts as the sample size.

For the case where $\theta = 1/2$, we have

$$d = \frac{4(z_{\alpha/2} + z_\gamma)^2}{(\beta_A)^2}.$$

An Example

Take a two-sided logrank test with level $\alpha = 0.05$, power $1-\gamma = 0.90$, $\theta = 1/2$. Then

$$d = \frac{4(1.96 + 1.28)^2}{(\beta_A)^2}.$$

The following table gives some required number of events for different hazard ratio $\exp(\beta_A)$.

Hazard ratio ($\exp(\beta_A)$)	d
2.00	88
1.50	256
1.25	844
1.10	4623

Therefore one has to answer during the design stage that a sufficient number of patients are entered and followed long enough so that the required number of events are attained.

Sample size (number of patients) calculations

One simple approach is to continue the experiment (*i.e.*, keep recruiting patients) until the required number of failures is obtained.

Example: Suppose patients with advanced lung cancer have a median survival time of 6 months. We have a new treatment which we hope will increase the median survival time to 9 months. If the survival time follows exponential distributions, then this difference would correspond to a hazard ratio of

$$\exp(\beta_A) = \frac{\lambda_1(t)}{\lambda_0(t)} = \frac{\lambda_1}{\lambda_0} = \frac{m_0}{m_1} = \frac{6}{9} = \frac{2}{3}.$$

Then the log hazard ratio is $\beta_A = \log(2/3)$.

Suppose we were asked to help design a clinical trial where these two treatments were going to be compared in a randomized experiment on newly diagnosed lung cancer patients. If patients were randomized with equal probability to the two treatments, and we desired 90% power to detect the above difference using a level $\alpha = 0.05$ two-sided log rank test, then the number of failures (deaths) necessary is given by

$$d = \frac{4(1.96 + 1.28)^2}{(\log(2/3))^2} = 256 \text{ (always rounding up).}$$

One strategy is to enter some larger number of patients, say 350 patients (about 175 patients on each treatment arm) and then continue following until we have 256 deaths.

Design Specification

More often in survival studies we need to be able to specify to the investigators the following:

1. number of patients;
2. accrual period;
3. follow-up time.

It was shown by Schoenfeld that reasonable approximations for obtaining the desired power can be made by ensuring that the total expected number of deaths (events) from both groups, computed under the alternative, should equal (assuming equal probability of assigning treatments)

$$\mathrm{E}(d) = \frac{4(z_{\alpha/2} + z_{\gamma})^2}{(\beta_A)^2}.$$

Namely, we compute the expected number of deaths for both groups "0" and "1" separately under the alternative hypothesis, the sum of these should be equal to the above formula.

Computing expected number of deaths when we have censoring

We only need to consider the one-sample problem here since expected number of deaths needs to be computed separately for each treatment group.

Suppose $(X_i, \Delta_i), i = 1, 2, ..., n$ represents a sample of possibly censored survival data, with the usual kind of assumption we have been making, *i.e.*,

$$\begin{aligned} X_i &= \min(T_i, C_i) \\ \Delta_i &= I(T_i \leq C_i). \end{aligned}$$

T is the underlying survival time having density $f(t)$, distribution function $F(t)$, survival function $S(t)$ and hazard function $\lambda(t)$. (We may want to subscribe by T to denote that these functions refer to the survival time T, such as $\lambda_T(t)$) C is the underlying censoring time having density $g(t)$, distribution function $G(t)$, survival function $H(t)$ and hazard function $\mu(t)$.

The expected number of deaths is equal to

$$n * P[\Delta = 1].$$

From the derivation in Chapter 3, we know that the desity for the pair of random variables (X, Δ):

$$f(x, \delta) = [f(x)]^{\delta}[S(x)]^{1-\delta} * [g(x)]^{1-\delta}[H(x)]^{\delta}.$$

So

$$f(x, \delta = 1) \;\; = \;\; f(x)H(x)$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard for the survival time T and $M(t) = \int_0^t \mu(u)du$ is the cumulative hazard for the censoring time C. Therefore,

$$\begin{aligned} P[\Delta = 1] &= \int_0^\infty f(x, \delta = 1)dx \\ &= \int_0^\infty f(x)H(x)dx, \end{aligned}$$

or integrating any of the above equivalent relationships.

Alternatively, the probability $P[\Delta = 1]$ can be calculated in another way:

$$\begin{aligned} P[\Delta = 1] &= P[T \le C] = \iint_D f(t, c)dtdc \;\; (\text{Here } D = \{(t, c) : t \le c\}) \\ &= \iint_D f(t)g(c)dtdc = \int_0^\infty \left[\int_t^\infty f(t)g(c)dc\right] dt \\ &= \int_0^\infty f(t)H(t)dt. \end{aligned}$$

Example: Suppose T is exponential with hazard λ and C is exponential with hazard μ, then

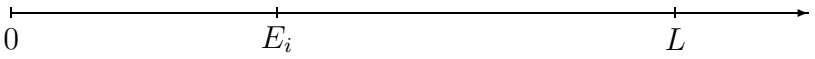
$$\begin{aligned} P[\Delta = 1] &= \int_0^\infty f(x)H(x)dx \\ &= \int_0^\infty \lambda \mathrm{e}^{-\lambda x}\mathrm{e}^{-\mu x}dx \\ &= \lambda \int_0^\infty \mathrm{e}^{-(\lambda+\mu)x}dx \\ &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

How to use these results for designing survival experiments

End of study censoring due to staggered entry: Suppose the only censoring we expect to see in a clinical trial is due to incomplete follow-up resulting at the time of analysis, as illustrated by Figure 4.4.

n patients enter the study at times $E_1, E_2, ..., E_n$ assumed to be independent and identicaly distributed (*i.i.d.*) with distribution function $Q_E(u) = P[E \le u]$. The censoring random

Figure 4.4: *Censoring due to staggered entry*



variable, if there was no other loss to follow-up or competing risk, would be $C = L - E$. Hence,

$$\begin{aligned} H_C(u) &= P[L - E \geq u] \\ &= P[E \leq L - u] \\ &= Q_E(L - u), \quad u \in [0, L]. \end{aligned}$$

Therefore, for such an experiment, the expected number of deaths in a sample of size n would be equal to

$$nP[\Delta = 1] = n \int_0^L \lambda_T(u) S_T(u) Q_E(L - u) du.$$

Example

Suppose the underlying survival of a population follows an exponential distribution. A study will accrue patients for A years uniformly during that time and then analysis will be conducted after an additional F years of follow-up. What is the expected number of deaths for a sample of n patients.

Figure 4.5: *Illustration of accrual and follow-up*

The entry rate follows a uniform distribution in $[0, A]$. That is

$$Q_E(u) = P[E \leq u] = \begin{cases} 0 & \text{if } u \leq 0 \\ \frac{u}{A} & \text{if } 0 < u \leq A \\ 1 & \text{if } u > A \end{cases}$$

Consequently,

$$H_C(u) = Q_E(L-u) = \begin{cases} 1 & \text{if } u \le L-A \\ \frac{L-u}{A} & \text{if } L-A < u \le L \\ 0 & \text{if } u > L \end{cases}$$

Hence,

$$\begin{aligned} P[\Delta = 1] &= \int_0^L \lambda_T(u) S_T(u) H_C(u) du \\ &= \int_0^{L-A} \lambda e^{-\lambda u} du + \int_{L-A}^{L} \lambda e^{-\lambda u} \frac{L-u}{A} du \\ &= \int_0^{L-A} \lambda e^{-\lambda u} du + \frac{L}{A} \int_{L-A}^{L} \lambda e^{-\lambda u} du - \frac{1}{A} \int_{L-A}^{L} u \lambda e^{-\lambda u} du. \end{aligned}$$

After some straightforward algebra, we get

$$P[\Delta = 1] = \left\{ 1 - \frac{e^{-\lambda L}}{\lambda A} \left(e^{\lambda A} - 1 \right) \right\}.$$

Therefore, if we accrue n patients uniformly over A years, who fail according to an exponential distribution with hazard λ, and follow them for an additional F years, then the expected number of deaths in the sample is

$$n * \left\{ 1 - \frac{e^{-\lambda L}}{\lambda A} \left(e^{\lambda A} - 1 \right) \right\}.$$

Example of a designed experiment

The survival time for treatment 0 is assumed to follow an exponential distribution with the median survival time equal to m_0 =4 years (so the hazard rate is $\lambda_0 = \log(2)/m_0 = 0.173$). We hope the new treatment "1" will increase the median survival time to $m_1 = 6$ years (assume exponential distribution, $\lambda_1 = \log(2)/m_1 = 0.116$), which we want to have 90% power to detect using a (two-sided) logrank test at the 0.05 level of significance. The hazard ratio is 2/3 and $\beta_A = \log(2/3)$. The total number of deaths from both treatments must be equal to

$$d = \frac{4(1.96 + 1.28)^2}{(\log(2/3))^2} = 256.$$

Suppose we decide to accrue patients for $A = 5$ years and then follow them for an additional $F = 3$ years, so $L = A + F = 8$ years. How large a sample size is necessary?

In a randomized trial where we randomize the patients to the two treatments with equal probability, the expected number of deaths would be equal to $D_1 + D_0$,where

$$D_j = \frac{n}{2}\left\{1 - \frac{\mathrm{e}^{-\lambda_j L}}{\lambda_j A}\left(\mathrm{e}^{\lambda_j A} - 1\right)\right\}, \quad j = 0, 1.$$

For our problem, the expected number of deaths is

$$\begin{aligned} D_1 + D_0 &= \frac{n}{2}\left\{1 - \frac{\mathrm{e}^{-0.173*8}}{0.173 * 5}\left(\mathrm{e}^{0.173*5} - 1\right)\right\} + \frac{n}{2}\left\{1 - \frac{\mathrm{e}^{-0.116*8}}{0.116 * 5}\left(\mathrm{e}^{0.116*5} - 1\right)\right\} \\ &= \frac{n}{2} * 0.6017 + \frac{n}{2} * 0.4642 = \frac{n}{2} * 1.0659. \end{aligned}$$

Thus if we want the expected number of deaths to equal 256, then

$$\frac{n}{2} * 1.0659 = 256 \Longleftrightarrow n = 480.$$

We can, of course, experiment with different combinations of sample sizes, accrual periods and follow-up periods, that gives us the desired answer and best suits the needs of the experiment being conducted.

The above calculation for the sample size requires that we are able to get $n = 480$ patients within $A = 5$ years. If this is not the case, we will be underpowered to detect the difference of interest. In fact, the sample size n and the accrual period A are tied by the accrual rate R (number of patients available per year) by

$$n = AR.$$

If we have information on R, the above calculation has to be modified.

Other issues that affect power and may have to be considered are

1. loss to follow-up

2. competing risks

3. non-compliance.

Remark: Originally, we introduced a class of weighted logrank tests to test $H_0 : S_1(t) = S_0(t)$, for $t \geq 0$. The weighted logrank test with weight function $w(t)$ is optimal to detect the following alternative hypothesis

$$\begin{aligned} & \lambda_1(t) = \lambda_0(t)e^{\beta w(t)}, \\ \text{or} \quad & \log\left[\frac{\lambda_1(t)}{\lambda_0(t)}\right] = \beta w(t); \quad \beta \neq 0. \end{aligned}$$

Hence, for proportional hazards, *i.e.*, $w(t) = 1$, the logrank test is the most powerful.

K sample weighted logrank test

Suppose we are interested in testing the null hypothesis that the survival distributions are the same for $K > 2$ groups. For example, we may be evaluating $K > 2$ treatments in a randomized clinical trial.

With right censoring, the data from such a clinical trial can be represented as (X_i, Δ_i, Z_i), $i = 1, 2, ..., n$, where for the ith individual

$$\begin{aligned} X_i &= \min(T_i, C_i) \\ \Delta_i &= I(T_i \leq C_i) \end{aligned}$$

and $Z_i = \{1, 2, ..., K\}$ corresponding to group membership in one of the K groups.

Denote by $S_j(t) = P[T_j \geq t]$, the survival distribution for the jth group, where T_j is the survival time for this group. The null hypothesis of no treatment difference can be represented as

$$H_0 : S_1(t) = S_2(t) = ... = S_K(t); \quad t \geq 0,$$

or equivalently,

$$H_0 : \lambda_1(t) = \lambda_2(t) = ... = \lambda_K(t); \quad t \geq 0,$$

where $\lambda_j(t)$ is the hazard function for group j.

No assumptions will be made regarding the distribution of the censoring time C or its relationship to Z. However, we will need to assume that conditional on Z, censoring is non-informative; *i.e.*, that T and C are conditionally independent given Z.

The test of the null hypothesis will be a direct generalization of the two-sample weighted logrank tests. Towards that end, we define the following quantities for a grid point x

$$\begin{aligned}
dN_j(x) &= \text{number of observed deaths at time } x \text{ } ([x, x+\Delta x)) \text{ from group } j = 1, 2, ..., K. \\
Y_j(x) &= \text{number at risk at time } x \text{ from group } j. \\
dN(x) &= \sum_{j=1}^{K} dN_j(x), \quad \text{total number of observed deaths at time } x \\
Y(x) &= \sum_{j=1}^{K} Y_j(x), \quad \text{total number at risk at time } x \\
\mathcal{F}(x) &= \{dN_j(u), Y_j(u); j = 1, 2, ..., K, \text{for all grid points } u < x, \text{ and } dN(x)\},
\end{aligned}$$

That is, $\mathcal{F}(x)$ is the information available at time x.

At a slice of time $[x, x + \Delta x)$, the data can be viewed as a $K \times 2$ contingency table shown in Table 4.2

Table 4.2: $K \times 2$ *table from* $[x, x + \Delta x)$

	Treatments				
	1	2	$\cdots$	K	total
# of death	$dN_1(x)$	$dN_2(x)$	$\cdots$	$dN_K(x)$	$dN(x)$
# alive	$Y_1(x) - dN_1(x)$	$Y_2(x) - dN_2(x)$	$\cdots$	$Y_K(x) - dN_K(x)$	$Y(x) - dN(x)$
# at risk	$Y_1(x)$	$Y_2(x)$	$\cdots$	$Y_K(x)$	$Y(x)$

We now consider a vector of observed number of deaths minus their expected number of deaths under the null hypothesis for each treatment group $j = 1, 2, ..., K$

$$\begin{pmatrix} dN_1(x) - \frac{Y_1(x)*dN(x)}{Y(x)} \\ dN_2(x) - \frac{Y_2(x)*dN(x)}{Y(x)} \\ \vdots \\ dN_K(x) - \frac{Y_K(x)*dN(x)}{Y(x)} \end{pmatrix}_{K\times 1}.$$

Note: The sum of the elements in this vector is equal to zero, which means one element is redundant.

If we condition on $\mathcal{F}(x)$, then we know the marginal counts of this $K \times 2$ table, in which case the vector $(dN_1(x), dN_2(x), ..., dN_K(x))^T$ is distributed as a multivariate version of a hypergeometric distribution.

Particularly, conditional on $\mathcal{F}(x)$, we know the following conditional means, variances and covariances:

$$\begin{aligned} \mathrm{E}[dN_j(x)|\mathcal{F}(x)] &= \frac{Y_j(x) * dN(x)}{Y(x)}, \quad j = 1, 2, ..., K. \\ \mathrm{Var}[dN_j(x)|\mathcal{F}(x)] &= \frac{dN(x)[Y(x) - dN(x)]Y_j(x)[Y(x) - Y_j(x)]}{Y^2(x)[Y(x) - 1]}. \\ \mathrm{Cov}[dN_j(x), dN_{j'}(x)|\mathcal{F}(x)] &= -\frac{dN(x)[Y(x) - dN(x)]Y_j(x) * Y_{j'}(x)}{Y^2(x)[Y(x) - 1]}. \end{aligned}$$

Consider the $(K - 1)$ dimensional vector $U(w)$, made up by the weighted sum of observed minus expected deaths in group $j = 1, 2, ..., K - 1$, summed over time x

$$U(w) = \begin{pmatrix} \sum_x w(x) \left[dN_1(x) - \frac{Y_1(x)*dN(x)}{Y(x)}\right] \\ \sum_x w(x) \left[dN_2(x) - \frac{Y_2(x)*dN(x)}{Y(x)}\right] \\ \vdots \\ \sum_x w(x) \left[dN_{K-1}(x) - \frac{Y_{K-1}(x)*dN(x)}{Y(x)}\right] \end{pmatrix}.$$

Note: We take the $(K-1)$ dimensional vector since the sum of all K elements is equal to zero and hence we have redundancy. If we included all K elements then the resulting vector would have a singular variance matrix.

Using arguments similar to the two-sample test, we can show that the vector of observed minus expected counts computed at different times, x and x' are uncorrelated.

Consequently, the corresponding $(K-1)\times(K-1)$ covariance matrix of the vector $T_n(w)$ is given by

$$V=[V_{jj'}],\quad j,j'=1,2,...,K-1,$$

where

$$V_{jj}=\sum_x w^2(x)\left[\frac{dN(x)[Y(x)-dN(x)]Y_j(x)[Y(x)-Y_j(x)]}{Y^2(x)[Y(x)-1]}\right],$$

and

$$V_{jj'}=-\sum_x w^2(x)\left[\frac{dN(x)[Y(x)-dN(x)]Y_j(x)*Y_{j'}(x)}{Y^2(x)[Y(x)-1]}\right],\quad \text{for } j\neq j\prime=1,2,...,K-1.$$

The test statistic used to test the null hypothesis is given by the quadratic form

$$T(w)=[U(w)]^TV^{-1}U(w).$$

Note: This statistic would be numerically identical regardless which of the $K-1$ groups were included to avoid redundancy.

Under H_0, this is distributed asymptotically as a χ^2 distribution with $(K-1)$ degrees of freedom.

Hence, a level α test would reject the null hypothesis whenever

$$T(w)=[U(w)]^TV^{-1}U(w)\geq\chi^2_{\alpha;K-1},$$

where $\chi^2_{\alpha;K-1}$ is the quantity that satisfies $P[\chi^2_{K-1}\geq\chi^2_{\alpha;K-1}]=\alpha$.

Remark: As with the two-sample tests, if the weight function $w(x)$ is stochastic, then it must be a function of the survival and censoring data prior to time x.

The most popular test was a weight $w(x) \equiv 1$ and is referred to as the $K-$sample logrank test. These tests are available on most major software packages such as `SAS`, $\mathtt{S}^+$, etc. For example, the `SAS` code is exactly the same the that for two sample tests.

Stratified logrank test

When comparing survival distributions among groups, especially in non-randomized studies, we may be concerned about the confounding effects that other factors may have on the interpretation of the relationship between survival and groups. For example, suppose we extract hospital records to obtain information on patients who were treated after a myocardial infarction (heart attack) with either bypass surgery or angioplasty. We wish to study subsequent survival and test whether or not there is a difference in the survival distributions between these treatments.

If we believe that these two groups of patients are comparable, we might test treatment equality using a logrank test or weighted logrank test. However, since this study was not randomized, there is no guarantee that the patients being compared are prognostically similar. For example, it may be that the group of patients receiving angioplasty are younger on average or prognostically better in other ways.

If this were the case, then we wouldn't know whether significant difference in treatment groups, if they occurred, were due to treatment or other prognostic factors. Or the treatments do have different effects. But the difference was blocked by some other factors the were distributed unbalancedly between treatment groups.

In such cases, we may want to adjust for the effect of these prognostic factors either through stratification or through regression modeling. Regression modeling will be discussed later in much greater detail. Very similarly, to adjust by stratification, we define strata of our population according to combination of factors which make individuals within each strata more prognostically

similar. Comparisons of survival distribution between groups are made within each strata and then these results are combined across the strata.

In clinical trials, the use stratified tests may also be important even though balance of prognostic factors by comparison groups is obtained by randomization. Use of permuted block randomization as well as other treatment allocation schemes which balance treatment group within strata by more than would be expected by chance alone may affect the statistical properties of the usual two and $K-$sample tests.

If the strata are prognostic, this enforced balance may cause the treatment (group) difference to be less variable than would be expected by chance, since the groups are more alike than would have been obtained by chance alone. Less variability is a desirable property if we can take advantage of it. The statistical tests developed so far (*i.e.*, two-sample and $K-$sample weighted logrank tests) have distributional theory developed under the assumption of simple randomness and consequently may lead to inference that is conservative when applied to clinical trials, which used treatment balancing methods within strata. A simple remedy to this problem is also to use stratified tests.

Think of the population being sampled as consisting of p strata. The strata, for example, could be those used in balanced randomization of a clinical trial, or combination of factors that make individuals within each strata prognostically similar. For example, consider the four strata created by the combination of sex categories (M, F) and age categories (≤ 50, > 50): $[(M, \leq 50), (M, > 50), (F, \leq 50), (F, > 50)]$.

consider two-sample comparisons, say, treatments 0 and 1, and let j index the strata $j = 1, 2, ..., p$. The null hypothesis being tested in a stratified test is

$$H_0 : S_{1j}(t) = S_{0j}(t), \quad t \geq 0, j = 1, 2, ..., p.$$

That is, the survival distributions from the two treatments are the same within each of the strata. The stratified logrank test consists of computing two-sample test statistic within each

strata and then combining these results across strata. For example,

$$T(w) = \frac{\sum_{j=1}^{p} \left\{ \sum_x w_j(x) \left[dN_{1j}(x) - \frac{dN_j(x) * Y_{1j}(x)}{Y_j(x)} \right] \right\}}{\left\{ \sum_{j=1}^{p} \left[\sum_x w_j^2(x) \left[\frac{Y_{1j}(x) Y_{0j}(x) dN_j(x) [Y_j(x) - dN_j(x)]}{Y_j^2(x)[Y_j(x) - 1]} \right] \right] \right\}^{1/2}}.$$

Note: Here j indexes the strata. In the previous section of the notes, j indexed treatment for more than two treatments.

Since within each of the strata there was no additional balance being forced between two groups (or if we believe the two groups are similar prognostically within each strata other than the treatment group being compared) beyond chance, the mean and variance of the test statistics computed within strata under the null hypothesis, are correct. The combining of the statistics and their variances over independent strata is now also correct. The resulting stratified logrank test has a standard normal distribution (asymptotically) under the null hypothesis, *i.e.*,

$$T(w) \overset{a}{\sim} \mathrm{N}(0, 1),$$

or

$$[T(w)]^2 \overset{a}{\sim} \chi_1^2.$$

Remark:

1. Stratified tests can be constructed for K samples as well. You just add the vector of test statistics over strata, as well as the covariance matrices before you compute the quadratic form leading to the χ^2 statistic with $(K-1)$ degrees of freedom.

2. Sample size consideration are similar to the unstratified tests. Power is dependent on the number of observed deaths and the hazard ratio between groups within strata. For example, the stratified logrank test with $w_j(x) \equiv 1$ for all x and j, is most powerful to detect proportional hazards alternatives within strata, where the hazard ratio is also assumed constant between strata. Namely

$$H_A : \lambda_{1j}(x) = \lambda_{0j}(x) \exp(\beta_A).$$

The number of deaths total in the study necessary to obtain power $(1-\gamma)$ for detecting a difference corresponding to β_A above, using a stratified logrank test at the α level of significance (two-sided), is equal to

$$d = \frac{4 * (z_{\alpha/2} + z_\gamma)^2}{\beta_A^2}.$$

This assumes equal randomization to the two treatments and is the same value as that obtained for unstratified tests. To compute the expected number of deaths using the design stage, we must compute separately over treatments and strata and these should add up to the desired number above.

Myelomatosis data revisited: When we analyzed the myelomatosis data on page 54, we found that the two treatments do not differ on prolonging patients' survival time. One may argue that we did not see treatment effect because the patients assigned to different treatment arms do not the same renal condition (on average), so we perform stratified tests using the following `SAS` program

```
proc lifetest data=myel;
  time dur*status(0);
  strata renal;
  test trt;
run;
```

Part of the output from this program is as follows

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
                                                                             4
                                                 16:15 Wednesday, February 9, 2000

                                  The LIFETEST Procedure

                      Rank Tests for the Association of DUR with Covariates
                                    Pooled over Strata


                         Univariate Chi-Squares for the WILCOXON Test

                           Test        Standard                     Pr >
        Variable      Statistic       Deviation     Chi-Square    Chi-Square
```

TRT	-2.6352	1.2963	4.1324	0.0421

Covariance Matrix for the WILCOXON Statistics

Variable	TRT
TRT	1.68039

Forward Stepwise Sequence of Chi-Squares for the WILCOXON Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
TRT	1	4.1324	0.0421	4.1324	0.0421

Univariate Chi-Squares for the LOG RANK Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
TRT	-4.4306	1.8412	5.7908	0.0161

Covariance Matrix for the LOG RANK Statistics

Variable	TRT
TRT	3.38990

Forward Stepwise Sequence of Chi-Squares for the LOG RANK Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
TRT	1	5.7908	0.0161	5.7908	0.0161

This result tells us that after adjusting for renal effect, treatments 1 is (statistically) significantly better than treatment 2 from either logrank test or Wilcoxon test. But we have to be very cautious in interpreting this result. If the patients were stratified into two different blocks based on their renal function **before** randomization and then treatments were randomly assigned to patients within each block, then we should adjust any possible renal effect in identifying treatment effect. If this was not the case, then it is hard for people to accept the renal-adjusted treatment effect. Also due to the small sample size, a small imbalance in renal function (treat-

ment 1 has 4 out of 12 patients with impaired renal function, while treatment 2 only has 3 out of 13 patients with impaired renal function) may have a significant impact on the final result. But this secondary analysis may give us some insight about the true treatment effect.

Note: If the number of treatments in a stratified test is greater than 2, we need to define indicator variables and put them in the `test` statement in `Proc lifetest`.