4. Comparison of Two (K) Samples

K=2

Problem: compare the *survival distributions* between two groups.

Ex: competing treatments on patients with a particular disease.

Z: Treatment indicator, i.e. $Z = 1$ for treatment 1 (new treatment); $Z = 0$ for treatment 0 (standard treatment or placebo)

Null Hypothesis:

H_0: no treatment (group) difference
H_0: $S_0(t) = S_1(t)$, for $t \geq 0$
H_0: $\lambda_0(t) = \lambda_1(t)$, for $t \geq 0$

Alternative Hypothesis:

H_a: the survival time for one treatment is **stochastically larger or smaller** than the survival time for the other treatment.
H_a: $S_1(t) \geq S_0(t)$, for $t \geq 0$ with strict inequality for some t (one-sided)
H_a: either $S_1(t) \geq S_0(t)$, or $S_0(t) \geq S_1(t)$, for $t \geq 0$ with strict inequality for some t

Solution: 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, with the logrank test being the most commonly used.

Notations

A sample of triplets $(X_i, \Delta_i, Z_i), i = 1, 2, \ldots, n$, where

$$X_i = \min(T_i, C_i) \qquad \Delta_i = I(T_i \leq C_i) \qquad Z_i = \begin{cases} 1 & new\ treatment \\ 0 & standard\ Treatment \end{cases}$$

$T_i =$ latent failure time; $C_i =$ latent censoring time

Also, define,

$n_1 =$ number of individuals in group 1

$n_0 =$ number of individuals in group 0

$$n_j = \sum_{i=1}^{n} I(Z_j = j)\,, j = 0, 1$$

$$n = n_0 + n_1$$

$Y_1(x) =$ number of individuals at risk at time x from trt $1 = \sum_{i=1}^{n} I(X_i \geq x, Z_i = 0)$

$Y_0(x) =$ number of individuals at risk at time x from trt $0 = \sum_{i=1}^{n} I(X_i \geq x, Z_i = 1)$

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

$dN_1(x) =$ # of deaths observed at time x from trt $1 = \sum_{i=1}^{n} I(X_i = x, \Delta_i = 1, Z_i = 1)$

$dN_0(x) =$ # of deaths observed at time x from trt $0 = \sum_{i=1}^{n} I(X_i = x, \Delta_i = 1, Z_i = 0)$

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

Note: $dN(x)$ 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.

Weighted logrank Test Statistic

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

Where,

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

$se\big(U(w)\big)$ will be given later.

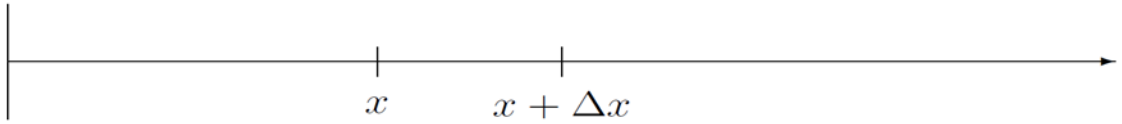
The null hypothesis of treatment equality will be rejected if $T(w)$ is sufficiently different from zero.

Note:

1. 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.
2. 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.
3. When $w(x) = 1$, logrank test statistic

Motivation

Take a slice of time $[x, x + \Delta x)$:



The following 2×2 table can be formulated:

	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	$Y_0(x)$	$Y_1(x)$	$Y(x)$

Under H_0:

$$dN_1(x)|Y_1(x), Y(x), dN(x) \sim Hypergeometric(Y_1(x), dN(x), Y(x))$$

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

$dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)}$ is the observed number of deaths minus expected number of deaths due to treatment 1. Hence,

- if H_0 is true, sum of $dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)}$ over x is expected to be near zero.
- If the hazard rate for treatment 1 were **lower** than that for treatment 0 **consistently** over x, then on average, we expect $dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)}$ to be negative.
- If the hazard rate for treatment 1 were **higher** than that for treatment 0 **consistently** over x, then on average, we expect $dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)}$ to be positive.

Specifically, the weighted logrank test statistic is given by

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

Under H_0: $T(w) \overset{a}{\sim} N(0, 1)$

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

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

Remarks:

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

2. 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.

3. 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.

4. 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.

5. $w(x) = 1$: Logrank test

6. $w(x) = Y(x)$: Gehan's generalization of wilcoxon test

7. $w(x) = KM(x)$: Peto−Prentice's generalization of wilcoxon test

Note: Since both $Y(x)$ and $KM(x)$ are non-increasing functions of x, both Gehan's and Peto−Prentice's tests emphasize the difference early in the survival curves.

A Heuristic Proof

Define a set of random variables:

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

Assume H_0 is true. Knowing $F(x)$ 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 is $dN_1(x)$.

Conditional on $F(x)$, 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 count are fixed conditional on $F(x)$). Therefore, the conditional distribution of one of the counts, say, $dN_1(x)$, in the cell of the table, given $F(x)$ follows a hypergeometric distribution.

$$P[dN_1(x) = c | Y_1(x), Y(x), dN(x)] = \frac{\binom{dN(x)}{c}\binom{Y(x)-dN(x)}{Y_1(x)-c}}{\binom{Y(x)}{Y_1(x)}}$$

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

$$Var[dN_1(x)|F(x)] = \frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}$$

The numerator of the weighted logrank test statistic is:

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

Notice that under H_0 :

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

Next, we will find an unbiased estimator for the variance of $U(w)$. Let

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

Then,

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

Notice that we already show: $E[A(x)] = E[A(y)] = 0$. WOLG, suppose y $< x$, then,

$$Cov\big(A(x), A(y)\big) = E[A(x) * A(y)] = E[A(x) * A(y)|F(x)] = E\big[A(y)E[A(x)|F(x)]\big] = 0$$

Now, $$Var[U(w)] = \sum_x Var[A(x)] = \sum_x E[A^2(x)] = \sum_x E\big[E[A^2(x)|F(x)]\big]$$

$$= \sum_x E\left\{E\left[w^2(x)\left[dN_1(x) - \frac{Y_1(x) \times dN(x)}{Y(x)}\right]^2 \middle| F(x)\right]\right\}$$

$$= \sum_x E\left\{w^2(x)E\left(\big[dN_1(x) - E[dN_1(x)]\big]^2 \middle| F(x)\right)\right\}$$

$$= \sum_x E\{w^2(x)Var[dN_1(x)|F(x)]\}$$

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

This means:

$\sum_x w^2(x)\left[\frac{Y_1(x)Y_0(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}\right]$ is an unbiased estimator for $Var[U(w)]$.

Recapping:

Under $H_0 : S_0(t) = S_1(t)$

1. The Statistics $U(w) = \sum_x A(x)$ has expectation equal to zero, i.e. $\mathrm{E}[U(w)] = 0$.

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:

$$T(w) = \frac{U(w)}{se\big(U(w)\big)} \overset{a}{\sim} N(0, 1)$$

3. An unbiased estimate of 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,

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

#

An Example

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
…			
1296	1	2	0

dur is the patient's survival or censored time,
status is the censoring indicator,
trt is the treatment indicator,
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. $H_0 : S_0(t) = S_1(t)$

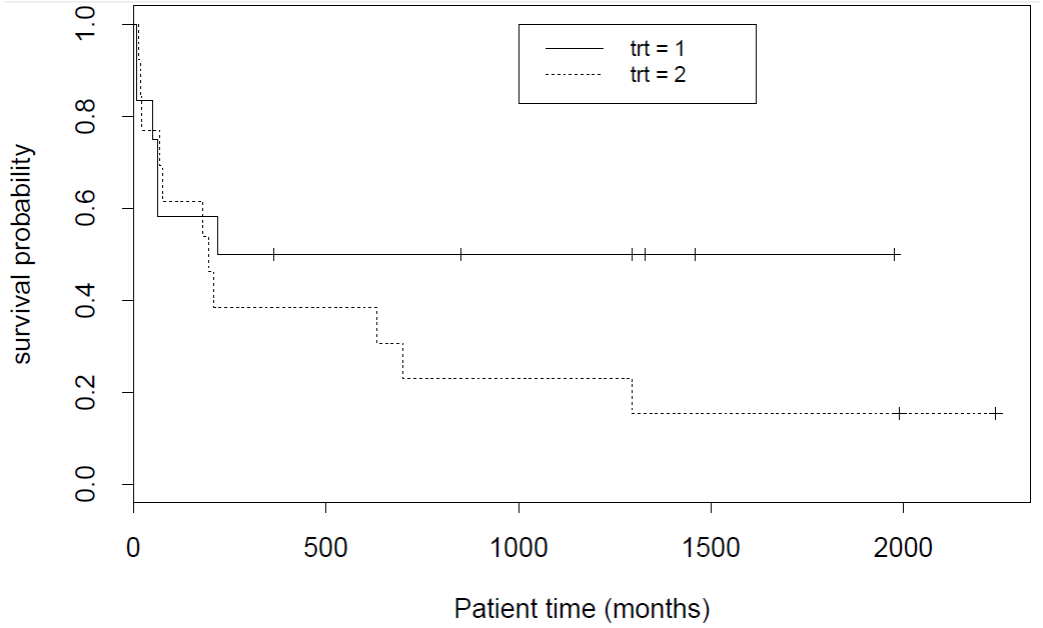
SAS & R codes

Note:

1. 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.
2. The likelihood ratio test is based on exponential model.
3. 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?

The treatment specific Kaplan-Meier survival estimates were generated using the following R functions:

```
pdf(file="fig_myel.pdf", horizontal = F, height=6, width=8.5, pointsize=14)
# par(mfrow=c(1,2))
example <- read.table(file="chap4_myel.txt", header=T);
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()
```



Kaplan-Meier estimates for two treatments

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

logrank test in R

Peto-Prentice's Wilcoxon test

Power and Sample Size

Since a survival curve is infinite dimensional, describing departures from the null as differences at every point in time over the survival curve would be complicated. Clearly, some simplifying conditions must be given. In clinical trials, proportional hazards alternatives have become very popular. That is,

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

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).

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta) \Leftrightarrow -\frac{dlog\{S_1(t)\}}{dt} = -\frac{dlog\{S_1(t)\}}{dt}\exp(\beta)$$

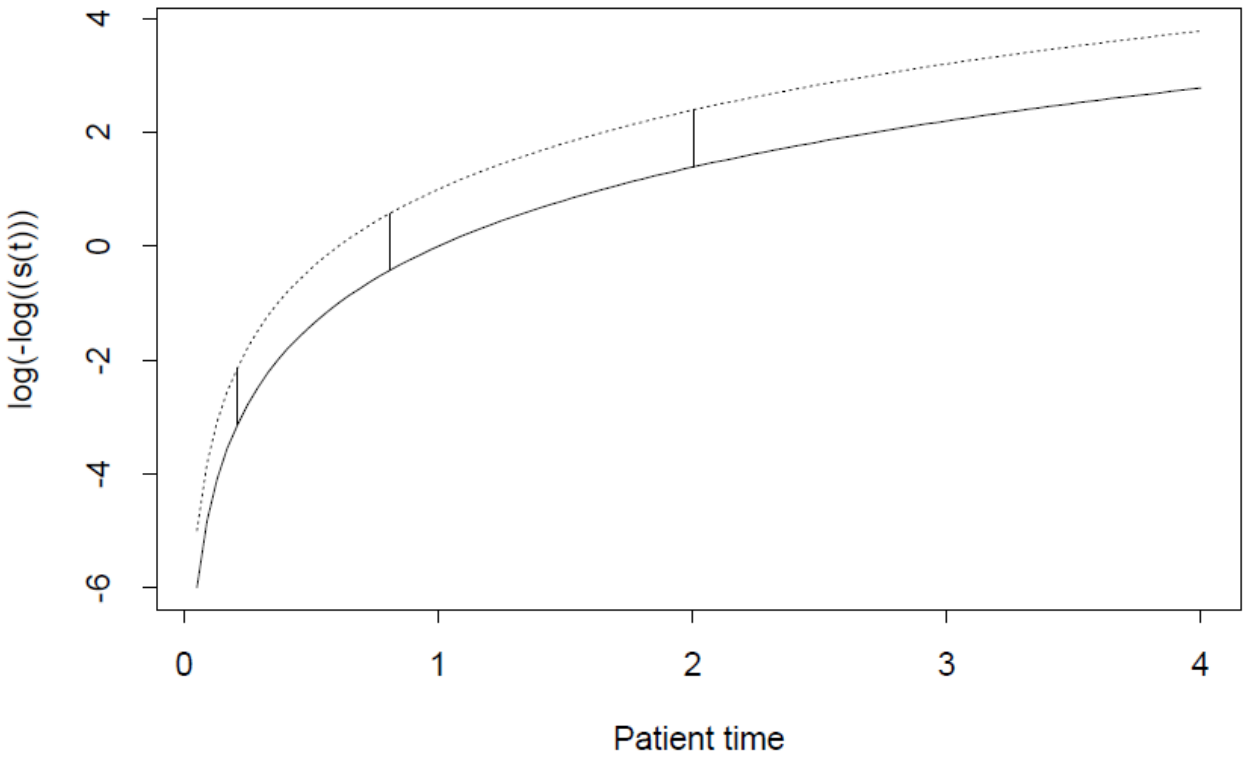
$$\Leftrightarrow log\{S_1(t)\} = log\{S_0(t)\}\exp(\beta) + \mathrm{C} \quad (\mathrm{t} = 0 \Rightarrow \mathrm{C} = 0)$$

$$\Leftrightarrow S_1(t) = S_1^{\gamma}(t), \qquad \gamma = \exp(\beta)$$

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

Based on the last equation, 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.

EX:



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

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

The ratio of median (m) or mean (μ) survival times for two groups having exponential distributions is

$$\frac{m_1}{m_0} = \frac{\log(2)/\lambda_1}{\log(2)/\lambda_0} = \frac{\lambda_0}{\lambda_1} = \frac{1/\lambda_1}{1/\lambda_0} = \frac{\mu_1}{\mu_0}$$

logrank Test & Proportional Hazards

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 logrank tests (unweighted) .

For H_a: $\frac{\lambda_1(t)}{\lambda_0(t)} = \exp(\beta_A)$; $\beta_A \neq 0$, When censoring does not depend on treatment (*e.g.*, randomized experiments), the logrank test has distribution approximated by

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

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.

Let $\mu = \beta_A\sqrt{d\theta(1-\theta)}$ $\quad$ $T_n \overset{a}{\sim} N(\mu, 1)$

Sample Size Formula

Recall that our test procedure is that: Reject H_0 when $|T_n| > z_{\alpha/2}$

under $H_0, T_n \overset{a}{\sim} N(0,1)$ and under $H_a, T_n \overset{a}{\sim} N(\mu, 1)$

By the definition of power, we have

$$P\left[|T_n| > z_{\alpha/2} \middle| H_a\right] = 1 - \gamma \qquad (1-\gamma) \text{ is the desired power.}$$

$$\Leftrightarrow P\left[T_n > z_{\alpha/2} \middle| H_a\right] + P\left[T_n < -z_{\alpha/2} \middle| H_a\right] = 1 - \gamma$$

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

$$P\left[T_n < -z_{\alpha/2} \middle| H_a\right] = P\left[T_n - \mu < -z_{\alpha/2} - \mu \middle| H_a\right] = P\left[Z < -z_{\alpha/2} - \mu\right]$$
$$= P\left[Z > z_{\alpha/2} + \mu\right] \approx 0 \qquad (Z \sim N(0,1))$$

$$P\left[T_n > z_{\alpha/2} \middle| H_a\right] = P\left[T_n - \mu < z_{\alpha/2} - \mu \middle| H_a\right] = P\left[Z > z_{\alpha/2} - \mu\right]$$

$$P\left[Z > z_{\alpha/2} - \mu\right] \approx 1 - \gamma \Leftrightarrow P\left[Z < z_{\alpha/2} - \mu\right] \approx \gamma \Leftrightarrow P\left[Z > -z_{\alpha/2} + \mu\right] \approx \gamma$$

$$-z_{\alpha/2} + \mu = z_\gamma \Leftrightarrow \mu = z_\gamma + z_{\alpha/2} \Leftrightarrow \beta_A\sqrt{d\theta(1-\theta)} = z_\gamma + z_{\alpha/2} \Leftrightarrow d = \frac{\left(z_\gamma + z_{\alpha/2}\right)^2}{(\beta_A)^2 * \theta(1-\theta)}$$

- Exactly the **same** formula for d can be derived if $\beta_A < 0$.
- d acts as the sample size.

Take a two-sided logrank test with level $\alpha = 0.05$, power $1 - \gamma = 0.90, \theta = 0.5$. 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

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

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

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)

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

That is, 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 $E(d)$ in One Sample

Suppose (X_i, Δ_i), $i = 1, \ldots, n$ represents a sample of possibly censored survival data, with $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, and the following notations:

T		C
$f(t)$	Density function	$g(t)$
$F(t)$	C.D.F	$G(t)$
$S(t)$	Survival function	$H(t)$
$\lambda(t)$	Hazard function	$\mu(t)$

The expected number of deaths is

$$E(d) = n * P[\Delta = 1] = \int_0^\infty f(x, \Delta = 1)dx = \int_0^\infty f(x)H(x)dx$$

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

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

Design with Censoring Due To Staggered Entry

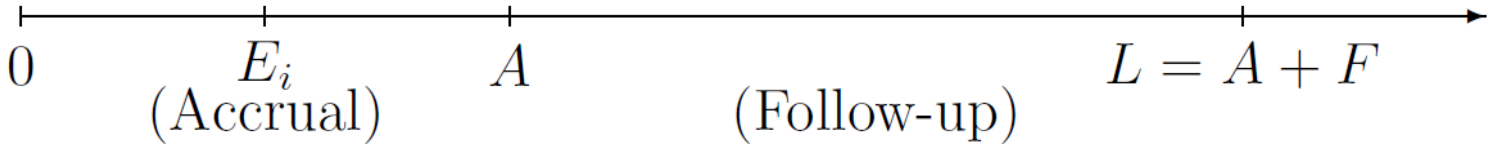
Suppose n patients enter the study at times $E_1, E_2, \ldots, E_n$ assumed to be independent and identically distributed (*i.i.d.*) with distribution function $Q_E(u) = P[E \leq u]$. If there was no other loss to follow-up or competing risk, the censoring random variable would be $C = L - E$. Hence,

$$H_C(u) = P[L - E \geq u] = P[E \leq L - u] = Q_E(L - u), \qquad u \in [0, L].$$

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

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

Ex: 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.



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

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

Consequently,

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

Hence,

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

Therefore, if we accrue n patients uniformly over A years, who fail according to an exponential distribution with hazard ¸, and follow them for an additional F years, then the expected number of deaths in the sample is $n * \left\{ 1 - \frac{e^{-\lambda L}}{\lambda A} \left(e^{\lambda A} - 1 \right) \right\}$

Lung cancer example (continued)

$$m_0 = 4\ years; \lambda_0 = \frac{\log(2)}{m_0} = 0.173; m_1 = 6\ years; \lambda_0 = \frac{\log(2)}{m_1} = 0.116;$$

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

Suppose we decide to accrue patients for $A = 5$ years and then follow them for an additional $F = 3$ years, so $\mathrm{L} = A + F = 8$ 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

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

For this problem, the expected number of deaths is

$$D_1 + D_0 = \frac{n}{2} * \left\{1 - \frac{e^{-0.173*8}}{0.173 * 5}(e^{0.173*5} - 1)\right\} + \frac{n}{2} * \left\{1 - \frac{e^{-0.116*8}}{0.116 * 5}(e^{0.116*5} - 1)\right\}$$

$$= \frac{n}{2} * 0.6017 + \frac{n}{2} * 0.4642 = \frac{n}{2} * 1.0658$$

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

$$\frac{n}{2} * 1.0658 = 256 \quad \Leftrightarrow \quad n = 480$$

Note:

1. Different combinations of sample sizes, accrual periods and follow-up periods can be experimented to give the desired answer and best suits the needs of the experiment being conducted.
2. 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.
3. 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.
4. Other issues that affect power and may have to be considered are: a). loss to follow-up; b). competing risks; c). non-compliance.
5. Originally, we introduced a class of weighted logrank tests to test $\mathrm{H_0}: \mathrm{S_1(t)} = \mathrm{S_0(t)}, \text{for } \mathrm{t} \geq 0$. The weighted logrank test with weight function $\mathrm{w(t)}$ is optimal to detect the following alternative hypothesis

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

K sample weighted logrank test

Testing the null hypothesis that the survival distributions are the same for $K > 2$ groups.

Notations:

A sample of triplets $(X_i, \Delta_i, Z_i), i = 1, 2, \dots, n$, where

$X_i = \min(T_i, C_i)$ $\quad$ $\Delta_i = I(T_i \leq C_i)$

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

Denote $S_j(t) = P[T_j \geq t]$ as the survival function for the jth group. The null hypothesis can then be represented as:

$$\mathrm{H}_0\colon S_1(t) = S_2(t) = \cdots = S_K(t), \text{ for } t \geq 0, \quad \text{or equivalently}$$
$$\mathrm{H}_0\colon \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_K(t), \text{ for } t \geq 0$$

$dN_j(x) =$ # of deaths observed at time x $([x + \Delta x))$ from group $j = 1,2, \dots, K$

$Y_j(x) =$ # at risk at time x from group j

$dN(x) = \sum_{j=1}^{K} dN_j(x)$, total # of observed deaths at time x

$Y(x) = \sum_{j=1}^{K} Y_j(x)$, total # at risk at time x

$F(x) = \{dN_j(u), Y_j(x); j = 1,2, \dots, K \text{ for all grid points } u < x, and\ dN(x)\}$

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

	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	$Y_1(x)$	$Y_2(x)$	$\cdots$	$Y_K(x)$	$Y(x)$

Conditioning on $F(x)$, we know the marginal counts of this $2 \times K$ table, in which case the vector $\left(dN_1(x), dN_2(x), \ldots, dN_K(x)\right)^{\mathrm{T}}$ is distributed as a multivariate version of a hypergeometric distribution. Particularly,

$$E\left[dN_j(x)|F(x)\right] = \frac{Y_j(x)dN(x)}{Y(x)}, \qquad j = 1,2,\ldots,K$$

$$Var\left[dN_j(x)|F(x)\right] = \frac{Y_j(x)\left[Y(x) - Y_j(x)\right]dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}$$

$$\mathrm{Cov}\left[dN_j(x), dN_{j'}(x)|F(x)\right] = -\frac{Y_j(x)Y_{j'}(x)dN(x)[Y(x) - dN(x)]}{Y^2(x)[Y(x) - 1]}$$

Consider a $(K-1)$ dimensional vector $U(w)$, made up by the weighted sum of observed number of deaths minus their expected number of deaths for each treatment group $j = 1,2,\dots,K$, summer over x

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

Note:

1. The $(K-1)$ dimensional vector is considered here 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.
2. Using arguments similar to the two-sample test, it can be shown 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 $V = \left[V_{jj'}\right],\ j, j' = 1,2,\dots,K-1$, where

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

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

Test Statistic: K sample weighted logrank test

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

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

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

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

Where $\chi^2_{\alpha;K-1}$ is the quantity that satisfies $\mathrm{P}\left[\chi^2_{K-1} \geq \chi^2_{\alpha;K-1}\right] = \alpha$

Remark:

1. 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.
2. 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, S+, etc. For example, the SAS code is exactly the same the that for two sample tests.

Stratified Test: Do We Need it?

- When comparing survival distributions among groups, especially in non-randomized studies, confounding effect, i.e., other factors that may affect the interpretation of the relationship between survival and groups, is a concern.
- For example, suppose we extract hospital records for patients who were treated after a myocardial infarction (heart attack) with either bypass surgery or angioplasty, and wish to 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*, a logrank or weighted logrank test can be used.
- However, since this study was not randomized, there is no guarantee that the patients being compared are prognostically similar, e.g., it may be that the group of patients receiving angioplasty are younger on average or prognostically better in other ways. 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.
- The effect of these prognostic factors can be adjusted either through stratification or through regression modeling (discussed later .
- To adjust by stratification, strata of our population were definde 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.

Stratified logrank Test

Consider a 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. Consider two-sample comparisons (treatments 0 vs. 1), and let j index the strata $j = 1,2, \dots, p$. The null hypothesis being tested in a stratified test is

$$\mathrm{H}_0\colon\ S_{1j}(t) = S_{0j}(t), \text{ for } t \geq 0, j = 1,2, \dots, p$$

The stratified logrank test consists of computing two-sample test statistic within each strata and then combining these results across strata. For example,

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

Since within each of the strata there was no additional balance being forced between two groups 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 H_0, *i.e.*,

$$T(w) \overset{\text{a}}{\sim} N(0,1) \qquad \text{or} \qquad [T(w)]^2 \overset{\text{a}}{\sim} \chi_1^2$$

Remarks:

- Stratified tests can be constructed for K samples as well. We 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.
- 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 $w(x) \equiv 1$ 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

$$\mathrm{H_a}: \lambda_{1j}(x) = \lambda_{0j}(x)\exp(\beta_A)$$

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

$$d = \frac{4 * \left(z_\alpha + z_{1-\gamma}\right)^2}{\beta_A^2}$$

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.

SAS Example