Reliability Data and Survival Analysis
review A

Chenghua Liu
liuch18@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

目录

1 Introduction and Basic Quantities

1.1 Describe the Distribution of Time-to-Event

The distribution of the random variable T can be described in a number of equivalent ways. There is of course the usual (cumulative) distribution function

$$F(t) = P\left[T \leq t\right], \quad t \geq 0$$

which is right continuous, i.e., $\lim_{u \to t^+} F(u) = F(t)$. When T is a survival time, $F(t)$ is the probability that a randomly selected subject from the population will die before time t.

If T is a continuous random variable, then it has a density function $f(t)$, which is related to $F(t)$ through following equations

$$f(t) = \frac{dF(t)}{dt}, F(t) = \int_0^t f(u)du$$

In biomedical applications, it is often common to use the survival function

$$S(t) = P[T \geq t] = 1 - F\left(t^-\right)$$

where $F(t^-) = \lim_{u \to t^-} F(u)$).If T is continuous random variable ,obviously we have

$$f(t) = -\frac{dS(t)}{dt}, S(t) = \int_t^\infty f(u)du$$

Now we define some label as follow.

1. **Mean Survival Time**: $\mu = \mathrm{E}(T)$. Due to censoring, sample mean of observed survival times is no longer an unbiased estimate of $\mu = \mathrm{E}(T)$. If we can estimate $S(t)$ well, then we can estimate $\mu = \mathrm{E}(T)$ using the following fact:

$$\mathrm{E}(T) = \int_0^\infty S(t)dt$$

2. **Median Survival Time**: Median survival time m is defined as the quantity m satisfying $S(m) = 0.5$. Sometimes denoted by $t_{0.5}$. If S(t) is not strictly decreasing, m is the smallest one such that $S(m) \leq 0.5$

3. p **th quantile of Survival Time** ($100p$ th percentile) : t_p such that $S\left(t_p\right) = 1 - p(0 < p < 1)$ If S(t) is not strictly decreasing, t_p is the smallest one such that $S\left(t_p\right) \leq 1 - p$

4. **Mean Residual Life Time**(mrl):

$$mrl\left(t_0\right) = \mathrm{E}\left[T - t_0 \mid T \geq t_0\right]$$

i.e., $mrl\left(t_0\right)$ = average remaining survival time given the population has survived beyond t_0. It can be shown that

$$mrl\left(t_0\right) = \frac{\int_{t_0}^\infty S(t)dt}{S\left(t_0\right)}$$

For convenience,we first define "Mortality rate at time ": the proportion of the population who fail (die) between times t and t + 1 among individuals alive at time t, i.e.,

$$m(t) = P\left[t \leq T < t + h \mid T \geq t\right]$$

Hazard rate is the instantaneous rate of failure at time t give that an individual is alive at time t

$$\lambda(t) = \lim_{h \to 0} \frac{P[t \leq T < t + h \mid T \geq t]}{h}$$

If h is very small,

$$P[t \leq T < t + h \mid T \geq t] \approx \lambda(t)h$$

The hazard function $\lambda(t)$ or $h(t)$

$$\lambda(t) = \frac{\lim_{h \to 0} \frac{P[t \leq T < t+h]}{h}}{P[T \geq t]} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d \log\{S(t)\}}{dt}$$

From this, we can integrate both sides to get

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

where $\Lambda(t)$ is referred to as the cumulative hazard function. Here we used the fact that $S(0) = 1$. Hence,

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

Node: The hazard rate is NOT a probability, it is a probability rate. Therefore it is possible that a hazard rate can exceed one in the same fashion as a density function f(t) may exceed one.

1.2 Censoring and Likelihood Function

Firstly we introduce two type censoring.

1. Type I censoring: observe only

$$(U_i, \delta_i) = \{min(T_i, c), I(T_i \leq c)\} \quad i = 1, \cdots, n$$

   i.e, we only have the survival information up to a fixed time c.

2. Type II censoring: observe only

$$T_{(1,n)}, T_{(2,n)}, \cdots, T_{(r,n)}$$

   where $T_{(i,n)}$ is the ith smallest survival time, i.e, we only observe the first r smallest survival times.

3. Random censoring (The most common type of censoring): $C_1, C_2, \cdots, C_n$ are potential censoring times for n subjects, observe only

$$(U_i, \delta_i) = \{\min(T_i, C_i), I(T_i \leq C_i)\}, i = 1, \cdots, n$$

We often treat the censoring time C_i as i.i.d. random variables in statistical inferences.

4. Interval censoring: observe only $(L_i, U_i), i = 1, \cdots, n$ such that $T_i \in [L_i, U_i)$

In the last, we give likelihood function for life table data: Assuming that the censoring is at t_i

Type of Censoring	Characteristic	Number of Cases	Likelihood of Responses $L_i(\boldsymbol{\pi};$ data $_i)$
Left at t_i	$T \leq t_i$	ℓ_i	$[F(t_i)]^{\ell_i}$
Interval	$t_{i-1} < T \leq t_i$	d_i	$[F(t_i) - F(t_{i-1})]^{d_i}$
Right at t_i	$T > t_i$	r_i	$[1 - F(t_i)]^{r_i}$

2 Parametric Survival Data and Likelihood for Censoring and Truncation Data

2.1 Parametric Survival Data

For common parametric models$(\alpha, \lambda > 0)$,we have table:

D	$\Lambda(t)$	$\lambda(t)$	$S(t)$	density $f(t)$	$\mathrm{E}(T)$	$\mathrm{Var}(T)$	$t_{0.5}$
Exp	λt	λ	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\log(2)}{\lambda}$
Weibull	$(\lambda t)^p$	$p\lambda(\lambda t)^{p-1}$	$e^{-(\lambda t)^p}$	$p\lambda(\lambda t)^{p-1}e^{-(\lambda t)^p}$	$\frac{\Gamma(1+1/p)}{\lambda}$	$\frac{\left[\Gamma(1+\frac{2}{p})-\{\Gamma(1+\frac{1}{p})\}^2\right]}{\lambda^2}$	$\left[\frac{\log(2)}{\lambda^p}\right]^{1/p}$
Gamma	/	$\frac{f(t)}{S(t)}$	$1 - I(\lambda t, \beta)$	$\frac{\lambda^\beta t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$	$\frac{\beta}{\lambda^2}$	/

where $I(t, \beta) = \int_0^t \frac{u^{\beta-1}e^{-u}}{\Gamma(\beta)}du$, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ is the Gamma function. For integer $\alpha, \Gamma(\alpha) = (\alpha - 1)!$

We can find the relationship between different distribution.

$$G(1, \lambda) \sim W(1, \lambda) \sim \mathrm{EXP}(\lambda)$$
$$W(p, \lambda) \sim \{\mathrm{EXP}(\lambda^p)\}^{1/p}$$

When we analysis survival data by parametric model, we need to cheak our assumption for distribution.

1. Exponential assumption

   (a) Obtain the non-parametric estimate of $S(t) : \hat{S}(t)$

   Plot $\log\{\hat{S}(t)\}$ vs. t and see if it is approximately a straight line

(b) Obtain the non-parametric estimate of $S(t)$: $\hat{S}_{KM}(t)$ Obtain the estimate of $S(t)$: $\hat{S}_1(t)$ by assuming the exponential distribution of the data

Plot $\hat{S}_1(t)$ vs. $\hat{S}_{KM}(t)$ and see if it is approximately a straight line

2. Weibull Assumption

(a) Since $\log\{S(t)\} = -(\lambda t)^p$, then we have $\log[-\log\{S(t)\}] = \log(\lambda) + p\log(t)$

Plot $\log[-\log\{S(t)\}]$ vs. log(t), is it a straight line

(b) Plot the Weibull estimate of$\hat{S}_1(t)$ vs. $\hat{S}_{KM}(t)$ and see if it is approximately a straight line

Moreover, we discuss about inference for exponential distribution. Observed fisher information $\hat{I}(\lambda) = r/\lambda^2$ Fisher (expected) information

$$I(\lambda) = E\{\hat{I}(\lambda)\} = \frac{E(r)}{\lambda^2} = \frac{nP\left(C_i > T_i\right)}{\lambda^2} = np/\lambda^2$$

It follows from the property of MLE

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda^2/np}} \to N(0,1)$$

in distribution as $n \to \infty$.

$$\widehat{\mathrm{Var}}(\hat{\lambda}) = -\left[\frac{d^2\ell(\lambda; X, \Delta)}{d\lambda^2}\bigg|_{\lambda=\hat{\lambda}}\right]^{-1} = \left(\frac{r}{(r/W)^2}\right)^{-1} = r/W^2$$

$\hat{\lambda}$ approximately follows $N\left(\lambda, r/W^2\right)$ for large n ,where $r = \sum_{i=1}^n \delta_i$ = # failures $W = \sum_{i=1}^n \tilde{t}_i$ = total follow up time
Using delta method,

$$\log(\hat{\lambda}) \sim N\left(\log(\lambda), r^{-1}\right)$$

Hypothesis testing for $H_0 : \lambda = \lambda_0$

$$Z = \sqrt{r}\left\{\log(\hat{\lambda}) - \log\left(\lambda_0\right)\right\} \sim N(0,1)$$

under H_0. We will reject the null hypothesis if $|Z|$ is too big. Confidence interval

$$\mathrm{pr}[\log(\hat{\lambda}) - 1.96\sqrt{1/r} < \log(\lambda) < \log(\hat{\lambda}) + 1.96\sqrt{1/r}] \approx 0.95$$

which suggests that the .95 confidence interval for λ is

$$\left(\hat{\lambda}e^{-1.96/\sqrt{r}}, \hat{\lambda}e^{1.96/\sqrt{r}}\right).$$

2.2 Likelihood for Censoring and Truncation Data

Statistically, we have observed data $\left(\tilde{T}_i, \Delta_i\right), i = 1, 2, \ldots, n$, where

$$\tilde{T}_i = \min\left(T_i, C_i\right)$$
$$\Delta_i = I\left(T_i \leq C_i\right) = \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (observed failure)} \\ 0 & \text{if } T_i > C_i \text{ (observed censoring)} \end{cases}$$

Namely, the potential data are $\{(T_1, C_1), (T_2, C_2), \ldots, (T_n, C_n)\}$, but the actual observed data are $\left\{\left(\tilde{T}_1, \Delta_1\right), \left(\tilde{T}_2, \Delta_2\right), \ldots, \left(\tilde{T}_n, \Delta_n\right)\right\}$ Of course we are interested in making inference on the random variable T, i.e., any one of the following functions

$$\begin{aligned} f(t) &= \text{ density function} \\ F(t) &= \text{ distribution function} \\ S(t) &= \text{ survival function} \\ \lambda(t) &= \text{ hazard function} \end{aligned}$$

Since we need to work with our data: $\left\{\left(\tilde{T}_1, \Delta_1\right), \left(\tilde{T}_2, \Delta_2\right), \ldots, \left(\tilde{T}_n, \Delta_n\right)\right\}$, we define the following corresponding functions for the censoring time C :

$$\begin{aligned} g(t) &= \text{ density function} \\ G(t) &= \text{ distribution function } = P[C \leq t] \\ H(t) &= \text{ survival function } = P[C \geq t] = 1 - G(t) \\ \mu(t) &= \text{ hazard function } = \frac{g(t)}{H(t)} \end{aligned}$$

Usually, the density function $f(t)$ of T may be governed by some parameters θ and $g(t)$ by some other parameters ϕ. In these cases, we are interested in making inference on θ.

In order to derive the density of $(\tilde{T}, \Delta)$, we assume independent censoring, i.e., random variables T and C are independent. The density function of $(\tilde{T}, \Delta)$ is defined as

$$f(t, \delta) = \lim_{h \to 0} \frac{P[t \leq \tilde{T} < t + h, \Delta = \delta]}{h}, t \geq 0, \delta = \{0, 1\}$$

Do not mix up the density $f(t)$ of T and $f(t, \delta)$ of $(\tilde{T}, \Delta)$. If we want to be more specific, we will use $f_T(t)$ for T and $f_{\tilde{T},\Delta}(t, \delta)$ for $(\tilde{T}, \Delta)$. But when there is no ambiguity, we will suppress the subscripts.

1. Case 1: $\delta = 1$, i.e., $T \leq C, \tilde{T} = \min(T, C) = T$, we have

$$
\begin{aligned}
&P[t \le \tilde{T} < t+h, \Delta = 1] \\
&= P[t \le T < t+h, C \ge T] \\
&\approx P[t \le T < t+h, C \ge t] \text{(Note: t is a fixed number)} \\
&= P[t \le T < t+h] \cdot P[C \ge t] \text{(by independence of T and C)} \\
&= f_T(\xi) h H_C(t), \quad \xi \in [t, t+h), \quad \text{(Note: H(t) is the survival function of } C)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
f(t, \delta = 1) &= \lim_{h \to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = 1]}{h} \\
&= \lim_{h \to 0} \frac{f_T(\xi) h H_C(t)}{h} \\
&= f_T(t) H_C(t)
\end{aligned}
$$

2. Case 2 : $\delta = 0$, i.e., $T > C, \tilde{T} = \min(T, C) = C$, we have

$$
\begin{aligned}
&P[t \le \tilde{T} < t+h, \Delta = 0] \\
&= P[t \le C < t+h, T > C] \\
&\approx P[t \le C < t+h, T \ge t] \text{(Note: t is a fixed number)} \\
&= P[t \le C < t+h] \cdot P[T \ge t] \text{(by independence of T and C)} \\
&= g_C(\xi) h S_T(t), \quad \xi \in [t, t+h)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
f(t, \delta = 0) &= \lim_{h \to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = 0]}{h} \\
&= \lim_{h \to 0} \frac{g_C(\xi) h S_T(t)}{h} \\
&= g_C(t) S_T(t)
\end{aligned}
$$

Combining these two cases, we have the density function of $(\tilde{T}, \Delta)$:

$$
\begin{aligned}
f(t, \delta) &= [f_T(t) H_C(t)]^{\delta} [g_C(t) S_T(t)]^{1-\delta} \\
&= \left\{ [f_T(t)]^{\delta} [S_T(t)]^{1-\delta} \right\} \left\{ [g_C(t)]^{1-\delta} [H_C(t)]^{\delta} \right\}
\end{aligned}
$$

Sometimes it may be useful to use hazard functions. Recalling that the hazard function

$$
\lambda_T(t) = \frac{f_T(t)}{S_T(t)}, \text{ or } f_T(t) = \lambda_T(t) S_T(t)
$$

we can write $[f_T(t)]^{\delta} [S_T(t)]^{1-\delta}$ as

$$
[f_T(t)]^{\delta} [S_T(t)]^{1-\delta} = [\lambda_T(t) S_T(t)]^{\delta} [S_T(t)]^{1-\delta} = [\lambda_T(t)]^{\delta} [S_T(t)]
$$

Another useful way of defining the distribution of the random variable $(\tilde{T}, \Delta)$ is through the **cause-specific hazard function.** Definition: The cause-specific hazard function is defined as

$$\lambda(t, \delta) = \lim_{h \to 0} \frac{P[t \leq \tilde{T} < t + h, \Delta = \delta \mid \tilde{T} \geq t]}{h}.$$

For example, $\lambda(t, \delta = 1)$ corresponds to the probability rate of observing a failure at time t given an individual is at risk at time t(*i.e.*, neither failed nor was censored prior to time t). If T and C are statistically independent, then through the following calculations, we obtain

$$\begin{aligned} P[t \leq \tilde{T} < t + h, \Delta = \delta \mid \tilde{T} \geq t] &= \frac{P[(t \leq \tilde{T} < t + h, \Delta = \delta) \cap (\tilde{T} \geq t)]}{P[\tilde{T} \geq t]} \\ &= \frac{P[t \leq \tilde{T} < t + h, \Delta = \delta]}{P[\tilde{T} \geq t]} \end{aligned}$$

Hence

$$\lambda(t, \delta = 1) = \frac{\lim_{h \to 0} \frac{P[t \leq \tilde{T} < t + h, \Delta = 1]}{h}}{P[\tilde{T} \geq t]} = \frac{f(t, \delta = 1)}{P[\tilde{T} \geq t]}$$

Since $f(t, \delta = 1) = f_T(t) H_C(t)$ and

$$\begin{aligned} &P[\tilde{T} \geq t] = P[\min(T, C) \geq t] = P[(T \geq t) \cap (C \geq t)] \\ &= P[T \geq t] \cdot P[C \geq t] (\text{by independence of T and C }) \\ &= S_T(t) H_C(t) \end{aligned}$$

Therefore,

$$\lambda(t, \delta = 1) = \frac{f_T(t) H_C(t)}{S_T(t) H_C(t)} = \frac{f_T(t)}{S_T(t)} = \lambda_T(t)$$

Note: This last statement is very important. It says that if T and C are independent then the cause-specific hazard for failing (of the observed data) is the same as the underlying hazard of failing for the variable T we are interested in. This result was used implicitly when constructing the life-table, Kaplan-Meier and Nelson-Aalen estimators in later lectures.

We define that D: Index set of death time R: Index set of right censored time L: Index set of left censored time I: Index set of interval censored time ,then we have

Likelihood for Censored Data

$$L(\theta; \tilde{t}, \delta) = \prod_{d \in D} f\left(\tilde{t}_d\right) \prod_{r \in R} S\left(\tilde{t}_r\right) \prod_{l \in L} \left[1 - S\left(\tilde{t}_l\right)\right] \prod_{i \in I} \left[S\left(u_i\right) - S\left(v_i\right)\right]$$

Likelihood for Truncated Data

$$L(\theta; \tilde{t}, y, \delta) = \prod_{d \in D} \frac{f\left(\tilde{t}_d\right)}{S\left(y_d\right)} \prod_{r \in R} \frac{S\left(\tilde{t}_r\right)}{S\left(y_r\right)} \prod_{l \in L} \frac{\left[S\left(y_l\right) - S\left(\tilde{t}_l\right)\right]}{S\left(y_l\right)} \prod_{i \in I} \frac{\left[S\left(u_i\right) - S\left(v_i\right)\right]}{S\left(y_i\right)}$$

3 Nonparametric Estimation

3.1 Kaplan-Meier

Kaplan and Meier (1958) proposed an idea of conditional probability to estimate survival function.The estimator of the survival function $S(t)$ (the probability that life is longer than t) is given by:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i.

The plug-in approach

By elementary calculations,

$$\begin{aligned} S(t) &= \text{Prob}(\tau > t \mid \tau > t-1)\,\text{Prob}(\tau > t-1) \\ &= (1 - \text{Prob}(\tau \leq t \mid \tau > t-1))\,\text{Prob}(\tau > t-1) \\ &= (1 - \text{Prob}(\tau = t \mid \tau \geq t))\,\text{Prob}(\tau > t-1) \\ &= q(t)S(t-1) \end{aligned}$$

where the one but last equality used that τ is integer valued and for the last line we introduced

$$q(t) = 1 - \text{Prob}(\tau = t \mid \tau \geq t)$$

By a recursive expansion of the equality $S(t) = q(t)S(t-1)$, we get

$$S(t) = q(t)q(t-1)\cdots q(0)$$

Note that here $q(0) = 1 - \text{Prob}(\tau = 0 \mid \tau > -1) = 1 - \text{Prob}(\tau = 0)$. The Kaplan-Meier estimator can be seen as a "plug-in estimator" where each $q(s)$ is estimated based on the data and the estimator of $S(t)$ is obtained as a product of these estimates.

Derivation as a maximum likelihood estimator

Kaplan-Meier estimator can be derived from maximum likelihood estimation of hazard function. More specifically given d_i as the number of events and n_i the total individuals at risk at time t_i, discrete hazard rate h_i can be defined as the probability of an individual with an event at time t_i. Then survival rate can be defined as:

$$S(t) = \prod_{i:t_i \leq t} (1 - h_i)$$

and the likelihood function for the hazard function up to time t_i is:

$$\mathcal{L}\left(h_{j:j\leq i} \mid d_{j:j\leq i}, n_{j:j\leq i}\right) = \prod_{j=1}^{i} h_j^{d_j} \left(1 - h_j\right)^{n_j - d_j}$$

therefore the log likelihood will be:

$$\log(\mathcal{L}) = \sum_{j=1}^{i} \left(d_j \log\left(h_j\right) + \left(n_j - d_j\right) \log\left(1 - h_j\right)\right)$$

finding the maximum of log likelihood with respect to h_i yields:

$$\frac{\partial \log(\mathcal{L})}{\partial h_i} = \frac{d_i}{\widehat{h_i}} - \frac{n_i - d_i}{1 - \widehat{h_i}} = 0 \Rightarrow \widehat{h}_i = \frac{d_i}{n_i}$$

where hat is used to denote maximum likelihood estimation. Given this result, we can write:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \widehat{h}_i\right) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Greenwood's formula

The Kaplan-Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common estimators is Greenwood's formula:

$$\widehat{\mathrm{Var}}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i\left(n_i - d_i\right)}$$

where d_i is the number of cases and n_i is the total number of observations, for $t_i < t$.

We assume that d_i is binomially distributed with parameters p_i .Thus $E\left(d_i\right) = n_i p_i$ and $\mathrm{Var}\left(d_i\right) = n_i p_i \left(1 - p_i\right)$.According $f(X) \approx f(c) + f'(c)(X - c)$ with $c = p_i$

$$\begin{aligned}
\log \widehat{S}(t) &= \sum_{t_i \leq t} \log(1 - \frac{d_i}{n_i}) \\
&\approx \sum_{t_i \leq t} \left(\log\left(1 - p_i\right) - \frac{1}{1 - p_i}\left(\frac{d_i}{n_i} - p_i\right)\right) \\
&= C(p) - \sum_{t_i \leq t} \frac{1}{1 - p_i}\left(\frac{d_i}{n_i} - p_i\right)
\end{aligned}$$

The terms in the sum in above formula are not independent,but the fact the each term in the sum has mean zero conditional on the earlier terms can be used to show that the variance of the sum in formula is the sum of the variances, even though the terms are not independent (martingale). Since $\mathrm{Var}\left(d_i/n_i \mid n_i\right) = p_i\left(1 - p_i\right)/n_i$, using delta method

$$\begin{aligned}
\mathrm{Var}(\log \widehat{S}(t)) &\approx \sum_{t_i \leq t} \frac{1}{\left(1 - p_i\right)^2} \frac{p_i\left(1 - p_i\right)}{n_i} = \sum_{t_i \leq t} \frac{1}{n_i} \frac{p_i}{1 - p_i} \\
&\approx \sum_{t_i \leq t} \frac{d_i}{n_i\left(n_i - d_i\right)}
\end{aligned}$$

by setting $p_i = \widehat{p}_i = d_i/n_i$.Using delta method ,we can drive the Greenwood's formula.

Confidence Interval and Confidence Band

1. Confidence Interval

   (a) Plain approach: $\hat{S}(t) \pm z_{1-\alpha/2}\,\text{se}[\hat{S}(t)]$

   (b) Log-log approach: Get a 95% confidence interval for $\text{L(t)} = \log(-\log(\text{S(t)}))$, and then transform back by $\text{S(t)} = \exp(-\exp(\text{L(t)})$
   The CI: $\left([\hat{S}(t)]^{e^A}, [\hat{S}(t)]^{e^{-A}}\right)$
   where $\text{A} = 1.96 * \text{se}(\widehat{L}(t))$ and $\text{se}(\hat{L}(t)) = \text{sqrt}\left(\frac{1}{[\log \hat{S}(t)]^2} \sum_{j:\tau_j \le t} \frac{d_j}{(r_j - d_j) r_j}\right)$

2. Confidence Band*

   (a) Hall and Wellner (1980)

   (b) Nair (1984), "equal precision"
   Let $\sigma_S^2(t) = \text{var}(\hat{S}(\text{t}))/\hat{S}(\text{t}) = \text{var}(\widehat{\Lambda}(\text{t}))$. The EP bands are proportional to pointwise CIs.

$$\hat{S}(t) \pm c_\alpha\,(a_L, a_U)\,\sigma_S(t)\hat{S}(t)$$

   where $c_\alpha\,(a_L, a_U)$ is obtained using Table C.3 in the reference book mentioned on the previous slide.

$$a_L = \frac{n\sigma_S^2\,(t_L)}{1 + n\sigma_S^2\,(t_L)}, \quad a_U = \frac{n\sigma_S^2\,(t_U)}{1 + n\sigma_S^2\,(t_U)}$$

   where n is the sample size, $t_L < t_U$ and t_L and t_U are selected such that

   $t_L \ge$ the smallest observed event time

   $t_U \le$ the largest observed event time

3.2 Nelson-Aalen

NA estimator is used to estimate the cumulative hazard $\Lambda(t)$, and then estimate $\text{S}(t) = e^{-\Lambda(t)}$.Define the NA Estimator as

$$\widetilde{\Lambda}(t) = \sum_{j:\tau_j \le t} d_j/r_j$$

Denote:

- $\tau_1, \cdots, \tau_K$: K distinct event times observed in the sample;
- d_j : # deaths at event τ_j
- r_j :# individuals 'at risk' right before the j -th death time (everyone dead or censored at or after that time);
- $\mathcal{H}_j$: historical knowledge before τ_j.

Prove that the estimated variance of NA estimator is

$$\begin{aligned}\widehat{\mathrm{Var}}(\widetilde{\Lambda}(t)) &= \sum_{j:\tau_j \le t} \frac{\frac{d_j}{r_j}\left(\frac{r_j - d_j}{r_j}\right)}{r_j - 1} \\ &= \sum_{j:\tau_j \le t} d_j / r_j^2 \quad \text{if no ties}\end{aligned}$$

Assume that d_j are modelled as binomial variables, i.e.,

$$d_j \mid \mathcal{H}_j \sim \mathrm{B}\left(r_j, \pi_j\right), \quad j = 1, \cdots, K$$

Intuitively, this assumption consider $d_j \approx$ # deaths in interval $[\tau_j, \tau_{j+1}\ )$. We can estimate the expectation and variance of NA estimator as follows.

$$\begin{aligned}\mathrm{E}[\widetilde{\Lambda}(t)] &= \mathrm{E}\left[\sum_{j:\tau_j \le t} \frac{d_j}{r_j}\right] = \sum_{j:\tau_j \le t} \mathrm{E}\left[\frac{d_j}{r_j}\right] \\ &= \sum_{j:\tau_j \le t} \mathrm{E}\left[\mathrm{E}\left(\frac{d_j}{r_j} \mid \mathcal{H}_j\right)\right] \\ &= \sum_{j:\tau_j \le t} \pi_j \\ &\approx \sum_{j:\tau_j \le t} \lambda\left(\tau_j\right)\left(\tau_{j+1} - \tau_j\right) \approx \int_0^t \lambda(x)dx \\ &= \Lambda(t).\end{aligned}$$

When number of sample is large enough, $\max\left\{\tau_{j+1} - \tau_j\right\}$ is small, we may assume that hazard λ is near constant in each of the intervals $[\tau_j, \tau_{j+1})$, and the first approximation holds. If we further assume that the hazard function is continuous, then the second approximation holds, too.

The definition of variance is given by

$$\begin{aligned}\mathrm{Var}(\widetilde{\Lambda}(t)) &= \mathrm{E}[\widetilde{\Lambda}(t) - \mathrm{E}(\widetilde{\Lambda}(t))]^2 \\ &= \mathrm{E}\left[\sum_{j:\tau_j \le t} \frac{d_j}{r_j} - \sum_{j:\tau_j \le t} \pi_j\right]^2 = \mathrm{E}\left[\sum_{j:\tau_j \le t} \left\{\frac{d_j}{r_j} - \pi_j\right\}\right]^2 \\ &= \mathrm{E}\left[\sum_{j:\tau_j \le t} \left\{\frac{d_j}{r_j} - \pi_j\right\}^2 + 2\sum_{j,k:} \left\{\frac{d_j}{r_j} - \pi_j\right\}\left\{\frac{d_k}{r_k} - \pi_k\right\}\right] \\ &= \sum_{j:\tau_j \le t} \mathrm{E}\left[\frac{d_j}{r_j} - \pi_j\right]^2 + 2 \sum_{\substack{j,k: \\ \tau_j \le t_k \le t}} \mathrm{E}\left[\left\{\frac{d_j}{r_j} - \pi_j\right\}\left\{\frac{d_k}{r_k} - \pi_k\right\}\right] \\ &= \sum_{j:\tau_j \le t} \mathbf{I}_j + 2 \sum_{\substack{j,k: \\ \tau_j < \tau_k \le t}} \mathbf{II}_j.\end{aligned}$$

Notice that the cross product terms have expectation equal to zero:

$$\begin{aligned}
&\mathrm{E}\left[\left\{\frac{d_j}{r_j}-\pi_j\right\}\left\{\frac{d_k}{r_k}-\pi_k\right\}\right] \\
=&\mathrm{E}\left[\mathrm{E}\left[\left\{\frac{d_j}{r_j}-\pi_j\right\}\left\{\frac{d_k}{r_k}-\pi_k\right\} \mid \mathcal{H}_k\right]\right] \\
=&\mathrm{E}\left[\left\{\frac{d_j}{r_j}-\pi_j\right\}\cdot \mathrm{E}\left[\left\{\frac{d_k}{r_k}-\pi_k\right\} \mid \mathcal{H}_k\right]\right] \\
=&0.
\end{aligned}$$

And for $\mathbf{I}$, we have

$$\begin{aligned}
\mathbf{I}_j &= \mathrm{E}\left[\frac{d_j}{r_j}-\pi_j\right]^2 = \mathrm{E}\left[\mathrm{E}\left(\left\{\frac{d_j}{r_j}-\pi_j\right\}^2 \mid \mathcal{H}_j\right)\right] \\
&= \mathrm{E}\left[\frac{1}{r_j^2}\operatorname{Var}\left(d_j \mid \mathcal{H}_j\right)\right] \\
&= \frac{\pi_j\left(1-\pi_j\right)}{r_j}
\end{aligned}$$

Summing up we get

$$\begin{aligned}
&\mathrm{E}\left[\sum_{j:\tau_j\le t}\frac{\frac{d_j}{r_j}\left[\frac{r_j-d_j}{r_j}\right]}{r_j-1}\right] = \sum_{j:\tau_j\le t}\mathrm{E}\left[\frac{\frac{d_j}{r_j}\left[\frac{r_j-d_j}{r_j}\right]}{r_j-1}\right] \\
&= \sum_{j:\tau_j\le t}\mathrm{E}\left[\mathrm{E}\left[\frac{\frac{d_j}{r_j}\left\{\frac{r_j-d_j}{r_j}\right\}}{r_j-1} \mid \mathcal{H}_j\right]\right] \\
&= \sum_{j:\tau_j\le t}\mathrm{E}\left[\frac{\pi_j\left[1-\pi_j\right]}{r_j}\right] \\
&= \operatorname{Var}[\widetilde{\Lambda}(t)].
\end{aligned}$$

In the last we comepare NA with KM.NA estimator performs better when the sample size is small. NA estimator is asymptotically equivalent to KM estimator and is also NPMLE.NA estimator is commonly used to get crude estimation of the hazard function.

3.3 Life-table (Actuarial Estimator)

Widely used by actuaries, demographers, etc.Life-table is one of the oldest techniques around, and useful for large sample size.We applies it when the data are grouped.

Define the effective # subjects at risk in each interval r'_j. If we assume censoring occur:

- at the beginning of intervals: $r'_j = r_j - c_j$
- at the end of intervals: $r'_j = r_j$
- during the interval (on average halfway): $r'_j = r_j - c_j/2$

Quantities estimated

- Conditional probability of dying/events $\hat{p}_j = \frac{d_j}{r'_j}$
- Conditional probability of surviving $\hat{q}_j = 1 - \frac{d_j}{r'_j}$
- Cumulative probability of surviving at t_j, $\hat{S}(t_j) = \prod_{l \leq j} \hat{q}_j = \prod_{l \leq j} \left(1 - \frac{d_j}{r'_j}\right)$

Find variance of each $\log(\widehat{q_j})$ and use delta method. Similar approach as calculating variance of KM estimator.(Greenwood's formula)

$$\widehat{\text{var}}\left(\hat{S}(t_j)\right) = \left(\hat{S}(t_j)\right)^2 \sum_{l \leq j} \frac{d_j}{\left(r'_j - d_j\right) r'_j}$$

Assuming censoring times and death times are uniformly distributed within each interval Hazard in the j -th interval ($t_{mj} = 0.5 \times (t_{j-1} + t_j)$, midpoint)

$$\hat{\lambda}(t_{mj}) = \frac{\hat{f}(t_{mj})}{\hat{S}(t_{mj})} = \frac{d_j}{\left[(t_j - t_{j-1})\left(r'_j - d_j/2\right)\right]}$$

$r'_j - d_j/2$ is taken to be the number exposed (or at risk) in the j -th interval. Denominator = person-units in j-th interval and $\hat{\lambda}(t_{mj})$ = Number of events per person-units.

4 Nonparametric Test to Compare Groups

4.1 Pearson's Chi-square Test

	Event		
Group	Yes	No	Total
0	d_0	$n_0 - d_0$	n_0
1	d_1	$n_1 - d_1$	n_1
Total	d	$n - d$	n

How to measure the difference of the observation and the theoretical distribution?

$$\sum_{i=1}^{r} c_i \left(v_i/n - p_i\right)^2$$

K. Pearson 1900, take $c_i = n/p_i$

$$K_n = \sum_{i=1}^{r} \frac{\left(v_i - np_i\right)^2}{np_i}$$

(K. Pearson) Under H_0, as the sample size $n \to \infty, K_n \xrightarrow{d} \chi^2_{r-1}$. Reject H_0 if $K_n > \chi^2_{r-1,\alpha}$ Let k_0 be the observed value of K_n, goodness-of-fit

$$p(k_0) = P(K_n \geq k_0 \mid H_0) \approx P\left(\chi^2_{r-1} \geq k_0\right)$$

Pearson's chi-square statistic in this case is:(we assume the row margins were considered fixed)

$$\chi_P^2 = \frac{\left[d_0 - \frac{n_0 d}{n}\right]^2}{n_0 n_1 d(n-d)/n^3} \sim \chi_1^2$$

One approvement is Mental-Haenszel Test

$$\chi_{MH}^2 = \frac{\left[d_0 - n_0 d/n\right]^2}{\frac{n_0 n_1 d(n-d)}{n^2(n-1)}} \overset{\text{approx.}}{\sim} \chi_1^2$$

And the Mental-Haenszel Test can be expand to Cochran-Mantel-Haenszel test which can handle K independent 2x2 tables to test for a common group effect.

$$\chi_{CMH}^2 = \frac{\left\{\sum_{j=1}^K \left(d_{0j} - n_{0j} * d_j/n_j\right)\right\}^2}{\sum_{j=1}^K n_{1j} n_{0j} d_j \left(n_j - d_j\right) / \left[n_j^2 \left(n_j - 1\right)\right]} \sim \chi_1^2$$

4.2 Log-rank Test

The idea of Logrank test is

1. Logrank test is obtained by constructing a 2x2 table at each distinct failure time.
2. Comparing failure rates between 2 groups conditional on number of subjects at risk in groups.
3. Combine the table using the Cochran-Mantel-Haenszel test.

Let $\tau_1 < \ldots < \tau_K$ represent the K ordered, distinct failure times. At j -th failure time τ_j,

	Die/Fail		
Group	Yes	No	Total
0	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Total	d_j	$r_j - d_j$	r_j

where d_{0j} and d_{1j} are the number of failures in group 0 and 1 . respectively at the j -th failure time, and r_{0j} and r_{1j} are the number at risk at that time, in groups 0 and 1 .
The logrank test is

$$\chi_{\text{logrank}}^2 = \frac{\left[\sum_{j=1}^K \left(d_{0j} - r_{0j} * d_j/r_j\right)\right]^2}{\sum_{j=1}^K \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{\left[r_j^2 (r_j - 1)\right]}}$$

K 2×2 tables are treated as independent, the logrank test has an approximate χ_1^2.

The power of the logrank test depends on the number of observed failures rather than the sample sizes.Logrank test is the most powerful for detecting the alternatives

$$H_1 : S_1(t) = S_0(t)^{\exp(\beta)} \Leftrightarrow \lambda_1(t) = \lambda_0(t) e^{\beta}, \beta \neq 0$$

If censoring process independent of treatment group, log-rank test statistics T has distribution (Schoenfeld, 1981 Biometrika)

$$\mathrm{T} \overset{\mathrm{a}}{\sim} \mathrm{N}\left(|\beta|\sqrt{D\pi_0\left(1-\pi_0\right)}, 1\right)$$

The power of log-rank test under the alterative $\lambda_1(t) = \lambda_0(t)e^{\beta}$ is approximately

$$\Phi\left(|\beta|\sqrt{D\pi_0\left(1-\pi_0\right)} - 1.96\right)$$

where D= expected total number of failures (under Ha) from both groups, π_0= the allocation proportion of patient in groups 0.

4.3 Some Extensions of Log-rank test

1. **Weighted Log-rank test** Note that in the logrank test, $O_j - E_j$ measure of how $\lambda_0\left(t_j\right) = \lambda_{0j}$ and $\lambda_1\left(t_j\right) = \lambda_{1j}$ differ. Suppose we wanted to compare groups, but in a way that "emphasized" certain times more than others. Let $w_1 \geq 0, w_2 \geq 0, \ldots, w_K \geq 0$ be known constants,

   $$Z_w = \frac{\sum_j w_j(O_j - E_j)}{\sqrt{\sum_j w_j^2 V_j}} = \frac{\left\{\sum_{j=1}^K w_j(d_{1j} - r_{1j}\cdot d_j/r_j)\right\}}{\sqrt{\sum_{j=1}^K \frac{w_j^2 r_{1j} r_{jj} d_j (r_j - d_j)}{r_j^2(r_j - 1)}}}$$

   under $H_0, Z_w \overset{\text{apx}}{\approx} N(0,1)$.We will proof it in last of the section.

   Generalized Wilcoxon test is one case of weighted log-rank test, which set $\mathrm{w}_j = r_j$ in the numerator

   $$\sum_j r_j\left(O_j - E_j\right) = \sum_j r_j\left(d_{1j} - r_{1j}\frac{d_j}{r_j}\right) = \sum_j \left(r_{0j}d_{1j} - r_{1j}d_{0j}\right)$$

2. **Stratified Log-rank test** Mantel-Haenszel Test (1959) is a special case of stratified Cox model. Separate data into L groups, where $L = \#$ levels of the categorical covariates on which you want to stratify (e.g., $L = 2$ when stratifying by gender)

   Compute O, E, V (say $, O^{(l)}, E^{(l)}, V^{(l)}\big)$ within each group, just as with the ordinary logrank

   $$Z = \frac{\sum_{l=1}^L \left(O^{(l)} - E^{(l)}\right)}{\sqrt{\sum_{l=1}^L V^{(l)}}} \overset{\text{apx}}{\approx} N(0,1) under H_0$$

3. **Log-rank test for $>$ 2 groups**

   Roughly, generating for K tables, the test statistics:

   $$\left(\sum_j \left(\mathbf{O}_j - \mathbf{E}_j\right)\right)' \left(\sum_j \mathbf{V}_j\right)^{-1} \left(\sum_j \left(\mathbf{O}_j - \mathbf{E}_j\right)\right) \sim \chi^2_{P-1} \quad \text{under } H_0$$

4. **Log-rank trend test**

$$Z_{tr} = \frac{\mathbf{c}^T(\mathbf{O} - \mathbf{E}.)}{\sqrt{\mathbf{c}^T \mathbf{V} \mathbf{c}}} \overset{\text{apx}}{\sim} N(0,1) \text{ under } H_0$$

We can understand some test through Weighted Log-rank test.and all these weighted tests will lack power if the hazards "cross".

Test	Weight w_j
Logrank	$w_j = 1$
Gehan's Wilcoxon	$w_j = r_j$
Peto/Prentice	$w_j = \hat{S}(t_j)$
Fleming-Harrington	$w_j = \left[\hat{S}(t_j)\right]^{\rho} (\rho \geq 0)$
Tarone-Ware	$w_j = \sqrt{r_j}$

When we choose test to use,we have suggestions as follows.

- The Gehan's Wilcoxon is sensitive to early differences between survival.
- The logrank is most powerful under proportional hazards.
- The Wilcoxon also has high power when the survival times are log-normally distributed, with equal variance in both groups but a different mean.
- The Peto/Prentice is most powerful under the alternative hypothesis of log-logistic model.

4.4 Heuristic Proof

Define the numerator to be $U(w) = \sum_{j=1}^{K} w_j (d_{1j} - r_{1j} \cdot d_j / r_j) = \sum A_j$, then

$$\begin{aligned}
\mathrm{E}(U(w)) &= \sum_{j=1}^{K} \mathrm{E}(A_j) = \sum_{j=1}^{K} \mathrm{E}\left[\mathrm{E}(A_j) \mid \mathcal{H}_j\right] \\
&= \sum_{j=1}^{K} \mathrm{E}\left[w_j \mathrm{E}\left(d_{1j} - \frac{r_{1j} d_j}{r_j} \mid \mathcal{H}_j\right)\right] \\
&= \sum_{j=1}^{K} \mathrm{E}\left[w_j \left(\mathrm{E}(d_{1j}) - \frac{r_{1j} d_j}{r_j}\right)\right] \\
&= 0.
\end{aligned}$$

So $\mathrm{E}(U(w)) = 0$, and

$$\mathrm{Var}(U(w)) = \mathrm{Var}\left(\sum_{j=1}^{K} A_j\right) = \sum_{j=1}^{K} \mathrm{Var}(A_j) + 2 \sum_{i<j} \mathrm{Cov}(A_i, A_j)$$

Since $\mathrm{E}(A_i) = \mathrm{E}(A_j) = 0$, we have

$$\mathrm{Cov}(A_i, A_j) = \mathrm{E}(A_i A_j) = \mathrm{E}[\mathrm{E}(A_i A_j) \mid \mathcal{H}_j] = \mathrm{E}[A_i \mathrm{E}(A_j \mid \mathcal{H}_j)] = 0,$$

and thus

$$\mathrm{Var}(U(w)) = \sum_{j=1}^{K} \mathrm{Var}(A_j) = \sum_{j=1}^{K} \mathrm{E}\left(A_j^2\right) = \sum_{j=1}^{K} \mathrm{E}\left(\mathrm{E}\left(A_j^2 \mid \mathcal{H}_j\right)\right).$$

For its j -th term, notice that

$$\mathrm{E}\left(A_j^2 \mid \mathcal{H}_j\right) = w_j^2 \mathrm{E}\left(\left(\mathrm{d}_{1j} - \frac{\mathrm{r}_{1j}\ \mathrm{d}_j}{\mathrm{r}_j}\right)^2 \mid \mathcal{H}_j\right) = w_j^2 \mathrm{Var}(\mathrm{d}_{1j} \mid \mathcal{H}_j)$$

And $\widehat{\mathrm{Var}}(\mathrm{d}_{1j}) = \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$ (since Hypergeometric dist.) So

$$\sum_{j=1}^{K} \frac{w_j^2 r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$$

is unbiased estimator of $\mathrm{Var}(U(w))$.

By the central limit theory, $\frac{U(w)}{\mathrm{se}(U(w))} \overset{H_0 \text{ asymp.}}{\sim} \mathrm{N}(0,1)$

$$Z_w = \frac{\sum_{j=1}^{K} w_j (O_j - E_j)}{\sqrt{\sum_{j=1}^{K} w_j^2 V_j}} = \frac{\left\{\sum_{j=1}^{K} w_j (d_{1j} - r_{1j} \cdot d_j / r_j)\right\}}{\sqrt{\sum_{j=1}^{K} \frac{w_j^2 r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}} \overset{H_0 \text{ asymp.}}{\sim} \mathrm{N}(0,1)$$

5 Design of Survival Study

5.1 Basic Design Theorem

In clinical trials proportional hazards alternatives have become very popular.That is

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

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

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

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

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

$$\log\{S_1(t)\} = \log\{S_0(t)\}\exp(\beta)$$
$$\Leftrightarrow S_1(t) = S_0^{\gamma}(t)$$

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

If we multiply both sides of the above formula by -1 and then take log, we will have:

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

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

The 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 3. In this case, we say two curves are parallel. Remember that do not be misled by the visual impression of the curves near the origin.

Exponentially distribution

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

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

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

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

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

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

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.

Distribution of log-rank statistic

When censoring does not depend on treatment, suppose the hazard rates in the two groups are $\lambda_0(t)$ and $\lambda_1(t)$ with hazard ratio

$$H_A: \quad \lambda_1(t)/\lambda_0(t) = e^{\beta_A}, \beta_A \neq 0$$
$$H_0: \quad \lambda_1(t)/\lambda_0(t) = 1$$

Under H_A, the log-rank test has distribution 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. (Schoenfeld, 1981)

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 H_a. Recall that our test procedure:

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

and $T_n \overset{a}{\sim} \mathrm{N}(0,1)$ under H_0 and $T_n \overset{a}{\sim} \mathrm{N}(\mu,1)$ under H_A.
By the definition of power, we have

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

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

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

and

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

Therefore,

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

Consequently,

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

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 observe 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

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

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

$$\tilde{T}_i = \min\left(T_i, C_i\right)$$
$$\Delta_i = I\left(T_i \leq C_i\right)$$

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 previous derivation, we know that the density 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 $f(x)$ is the probability density function of the survival time T and $H(x) = P[C \geq x]$ is the survival function of 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:

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

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

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

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

$$\begin{aligned} nP[\Delta = 1] &= n\int_0^L f(u)Q_E(L-u)du \\ &= n\int_0^L \lambda_T(u)S_T(u)Q_E(L-u)du \end{aligned}$$

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.

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,

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

Hence,

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

After some straightforward algebra, we get

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

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

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

5.3 Design in Reality

- Accrual rate is not constant
- Loss to follow-up

  To increase sample size

  - Method 1:simple inflation (commonly applied). Let $l \times 100\%$ be drop-off rate,

    $$N^* = \left(\frac{1}{1-\ell} \right) \cdot N$$

  - Method 2: exponential loss assumption. By assuming that time to loss also follows an exponential distribution, and modify θ and $1-\theta$

- Cross-overs
- Stratification
- Sequential Design

  A few techniques to protest against type I inflation:

  - Pocock Approach: Pick a smaller significance level (say α ') at each IA so that overall all type I error stays at α. Always conservative at final analysis.
  - O'Brien and Fleming Approach: most popular approach. By varying the alpha levels used for each of the K interim analyses, and try to keep final analysis "close" to the desired overall significance level.
  - A small portion to the required sample size will be added.

- Equivalence hypotheses