Summary Notes for Survival Analysis

Instructor: Mei-Cheng Wang
Department of Biostatistics
Johns Hopkins University
Spring, 2006

1 Introduction

1.1 Introduction

Definition: A failure time (survival time, lifetime), T, is a nonnegative-valued random variable.

For most of the applications, the value of T is the time from a certain event to a failure event. For example,

a) in a clinical trial, time from start of treatment to a failure event

b) time from birth to death = age at death

c) to study an infectious disease, time from onset of infection to onset of disease

d) to study a genetic disease, time from birth to onset of a disease = onset age

1.2 Definitions

Definition. Cumulative distribution function $F(t)$.

$$F(t) = \Pr(T \leq t)$$

Definition. Survial function $S(t)$.

$$S(t) = \Pr(T > t) = 1 - \Pr(T \leq t)$$

Characteristics of $S(t)$:

a) $S(t) = 1$ if $t < 0$

b) $S(\infty) = \lim_{t \to \infty} S(t) = 0$

c) $S(t)$ is non-increasing in t

In general, the survival function $S(t)$ provides useful summary information, such as the median survival time, t-year survival rate, etc.

Definition. Density function $f(t)$.

a) If T is a discrete random variable,

$$f(t) = \Pr(T = t)$$

b) If T is (absolutely) continuous, the density function is

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0^+} \frac{\Pr\text{ (Failure occurring in } [t, t + \Delta t))}{\Delta t} \\ &= \text{Rate of occurrence of failure at } t. \end{aligned}$$

Note that

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

<u>Definition</u>. Hazard function $\lambda(t)$.

a) If T is discrete,

$$\lambda(t) = \mathrm{P}(T = t | T \geq t) = \frac{\mathrm{P}(T = t)}{\mathrm{P}(T \geq t)}.$$

Note that $\lambda(t) = 0$ if t is not a “mass point” of T. If T takes values at the mass points $x_1 < x_2 < x_3 \ldots$. When $x_j \leq t < x_{j+1}$,

$$S(t) = \prod_{i=1}^{j} (1 - \lambda(x_i)),$$

since

$$\begin{aligned} S(t) &= \frac{P(T \geq x_2)}{P(T \geq x_1)} \cdot \frac{P(T \geq x_3)}{P(T \geq x_2)} \cdot \ \ldots \ \frac{P(T \geq x_{j+1})}{P(T \geq x_j)} \\ &= (1 - \lambda(x_1)) \cdot (1 - \lambda(x_2)) \ \ldots \ (1 - \lambda(x_j)) \end{aligned}$$

b) If T is (absolutely) continuous,

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0^+} \frac{\Pr(\text{Failure occurring in } [t, t + \Delta t) | T \geq t)}{\Delta t} \\ &= \text{Instantaneous failure rate at } t \text{ given survival up to } t \end{aligned}$$

Here $\lambda(t)\Delta t \approx$ the proportion of individuals experiencing
failure in $[t, t+\Delta t)$ to those surviving up to t

example a. Constant hazard $\lambda(t) = \lambda_0$
b. Increasing hazard $\lambda(t_2) \geq \lambda(t_1)$ if $t_2 \geq t_1$
c. Decreasing hazard $\lambda(t_2) \leq \lambda(t_1)$ if $t_2 \geq t_1$
d. U-shape hazard (human mortality for age at death)

Remark: Modeling the hazard function is one way for parametric modeling.

Definition Cumulative hazard function (chf)$\Lambda(t)$.

a) If T is discrete, let x_i's be the mass points,

$$\Lambda(t) = \sum_{x_i \leq t} \lambda(x_i)$$

b) If T is (absolutely) continuous,

$$\Lambda(t) = \int_0^t \lambda(u)du$$

and

$$\frac{d\Lambda(t)}{dt} = \lambda(t)$$

1.3 Relationship Among Functions

a) If T is discrete,

$$\boxed{\lambda(t) = \frac{\mathrm{P}_{(T=t)}}{\mathrm{P}_{(T\geq t)}} = \frac{f(t)}{S(t^-)}}$$

b) If T is (absolutely) continuous, $S(t) = \Pr(T > t) = \Pr(T \geq t)$,

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0^+} \frac{P(T \in [t, t+\Delta t) | T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0^+} \frac{P(T \in [t, t+\Delta t))/S(t)}{\Delta t} \\
&= \frac{1}{S(t)} \cdot \lim_{\Delta t \to 0^+} \frac{P(T \in [t, t+\Delta t))}{\Delta t} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

A well known relationship among the density, hazard and survival functions is

$$\boxed{\lambda(t) = \tfrac{f(t)}{S(t)}}\ .$$

Also,

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du \\
&= \int_0^t \frac{(-\frac{dS(u)}{du})}{S(u)}du = [-\log S(u)]\,|_0^t \\
&= [-\log S(t)] - [-\log S(0)] = -\log S(t)
\end{aligned}$$

Thus

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

We now see that $\lambda(\cdot)$ is determined if and only if $f(\cdot)$ (or $S(\cdot)$) is determined, and vice versa.

When T is a continuous variable, we also have

$$\boxed{\int_0^\infty \lambda(u)du = \infty}$$

This formula is implied by $0 = S(\infty) = e^{-\int_0^\infty \lambda(u)du}$.

Example. $\lambda(t) = \lambda_0$, a positive constant, is a valid hazard function.

Example. $\lambda(t) = \lambda_0 + \lambda_1 t$, with $\lambda_0, \lambda_1 > 0$, is a valid hazard function.

Example. $\lambda(t) = e^{-\theta t}$, $\theta > 0$, is NOT a valid hazard function.

Remark: In applications, if a disease has 'cure'; that is, we assume $P(T = \infty) > 0$, then it is OK that $\Lambda(\infty) < \infty$. This is allowed since T is not a 'regular random variable'.

1.4 Censoring

Type-I Censoring Type-I censoring occurs when a failure time t_i exceeds a pre-determined censoring time c_i. The censoring time c_i is considered as a constant in the study. For example, a clinical treatment study starts at the calendar time a and ends at b. Patients could enter the study at different calendar times. The failure time is the time between the start of treatment (entry) to a certain event. Assume no loss to follow-up. In this case, c_i is the time from entry to b. The actual fialure time t_i cannot be observed if $t_i > c_i$.

Type-II Censoring This type of censoring is frequently encountered in industrial applications. From n ordered failure times, only the first $r (r \leq n)$ times are observed, others are censored.

For example, put 100 transistors on test at the same time and stop the experiment when 50 transistors burn out. In this example, $n = 100$ and $r = 50$. Let $t_{(1)}, t_{(2)}, \ldots, t_{(50)}$ be the first 50 failure times. Note that $t_{(50)}$ is an estimate of the median failure time.

Random censoring This type of censoring will be the main censoring mechanism that we deal with in this course. It occurs when the censoring time varies from individual to individual and is unknown in advance.

For example, in a follow-up study, the censoring occurs due to the end of the study, loss to follow-up, or early withdrawals.

Reasons for censoring

- patients decide to move to another hospital
- patients quit treatment because of side-effects of a drug
- failues occur after the end of study
- etc.

Theoretical setting. Suppose C is the censoring variable. Assume T and C are independent (the so-called **independent censoring**). Define

$$Y = \begin{cases} T & \text{if} \quad T \leq C \\ C & \text{if} \quad T > C \end{cases}$$

and the censoring indciators

$$\Delta = \begin{cases} 1 & \text{if data is uncensored,} \quad T \leq C \\ 0 & \text{if data is censored,} \qquad T > C \end{cases}$$

Assume $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$ are iid copies of (Y, Δ). Under random censoring, what is the actually observed data? Ideally, we would like to observe the "complete data" $t_1, t_2, \ldots, t_n$. Due to censoring, we only observe "right-censored data" $(y_1, \delta_1), (y_2, \delta_2), \ldots, (y_n, \delta_n)$ and possibly some covariate information.

Example A set of observed survival data is

y_i	25	18	17	22	27
δ_i	1	0	1	0	1

The data can also be presented as

$$25 \quad 18^+ \quad 17 \quad 22^+ \quad 27$$

1.5 Probability Properties

Intuitively, the random variable Y tends to be 'shorter' than the failure time of interest, T. This is clear upon observing $Y = \min\{T, C\}$. Under the assumption that T and C are independent, the survival function of Y is

$$\begin{aligned} S_{Y}(y) &= P(T > y, C > y) = P(T > y)P(C > y) \\ &= S_{T}(y)S_{C}(y) \leq S_{T}(y) \ . \end{aligned}$$

Thus, as compared with S_{T}, S_{Y} assigns more probability to smaller values as compared with S.

Example. Suppose the censoring time is a fixed constant, $C = c_0$, $c_0 > 0$. Then the survival function of Y is $S_{Y}(y) = S_{T}(y)$ if $y < c_0$, and $S_{Y}(y) = 0$ if $y \geq c_0$. $\diamondsuit$

Example. Suppose $T \sim \text{Exp}(\theta)$, $\theta > 0$, and $C \sim \text{Unif}(0, \beta)$, $\beta > 0$. Then the survival function of Y is

$$S_{Y}(y) = \begin{cases} 1 & \text{if} \quad y \leq 0 \\ e^{-\theta y}\left(\frac{\beta - y}{\beta}\right) & \text{if} \quad 0 < y < \beta \\ 0 & \text{if} \quad y \geq \beta \end{cases}$$

$\diamondsuit$

Hazard function is an important function for various reasons and the so-called 'risk set' plays a key role for exploring probability structure of the hazard function. The risk set at t is defined as

$$R(t) = \{y_j \ : \ y_j \geq t, \ j = 1, 2, \ldots, n\} \ , \qquad t \geq 0$$

Property. For $t \geq y$, $P(T > t \mid T \geq y) = P(T > t \mid Y \geq y)$.

Proof. For $t \geq y$,

$$\begin{aligned} P(T > t \mid Y \geq y) &= \frac{P(T > t, C \geq y)}{P(T \geq y, C \geq y)} \\ &= \frac{P(T > t)P(C \geq y)}{P(T \geq y)P(C \geq y)} \\ &= \frac{P(T > t)}{P(T \geq y)} \\ &= P(T > t \mid T \geq y) \qquad \diamondsuit \end{aligned}$$

Implication of this property. The distribution among observed survivors at y is the same as the distribution in risk population at y. Also, the hazard probability on uncensored Y at y from $R(y)$ is the same as the hazard probability of T at y:

$$
\begin{aligned}
P(Y \approx y, \Delta = 1 \mid Y \geq y) &= P(C > T \approx y \mid Y \geq y) \\
&= \frac{S_{C}(y) f(y) dy}{S_{T}(y) S_{C}(y)} \\
&= \lambda(y) dy
\end{aligned}
$$

The above fomula can be equivalently expressed by

$$f_u(y \mid Y \geq y) dy = \lambda(y) dy$$

or more directly,

$$f_u(y \mid Y \geq y) = \lambda(y) \qquad (*)$$

where the subscript 'u' represents 'uncensored'. Formula in (*) is the base for the use of risk sets in many nonparametric and semiparametric models when analyzing survival data.

Left censoring

The failure time t_i could be too small to be observed. For example, consider a study in which interest centers on the time to recurrence of a particular cancer following surgical removal of the primary tumor. A few months after the operation, the patients are examined to determine if the cancer has recurred. Let T = time from operation to the recurrence of cancer. Some of the patients at this time may be found to have a recurrence and thus the actual time is less than the time from operation to the examination. These cases are said to be left censored.

1.6 Interval Censoring and Truncation

Interval censoring

The failure time t_i falls in an interval (ℓ_i, r_i) and observe only (ℓ_i, r_i). For example, let T = time from treatment onset to disease onset. The onset of disease falls in the interval formed by two successive clinical visits.

Let ℓ_i = time from the treatment onset to the last visit when the ith patient is free of the disease.
r_i = time from the treatment onset to the first visit when the ith patient becomes diseased.

The best knowledge we have about the true failure time $T_i = t_i$ is $\ell_i < t_i \leq r_i$.

Right truncation

The failure time t_i is too large to be included in data. A well known example is the reported AIDS incidences. In this example, T = time from HIV infection to diagnosis of AIDS. An AIDS incidence is reported to a health institution only when AIDS develops. Those cases where AIDS occur after the closing date of data collection are excluded from the data set.

Left truncation (and right censoring)

The presence of left truncation is usually due to the prevalent sampling scheme, that is, drawing samples from a disease prevalent population. Right censoring is encountered for the usual reasons (loss to follow-up etc.).

Example Failure time T = time from the onset (or diagnosis) of breast cancer to death. A prevalent cohort includes a group of women who have developed breast cancer at the time of recruitment. Those with breast cancer who died before the recruiting time are excluded from the study. The study tends to recruit women with longer failure times.

Double truncation

The failure time t_i is included in the data set only if the failure event occurs in a calendar-time window. For example, $T =$ onset age of a certain disease and the data $\{t_i\}$ are observed only if the disease occurs in the calendar-time window $[a, b]$. If double truncation is adopted as the sampling scheme, those cases that the disease occurs before a or after b will not be included into the data set.

1.7 Correlated Survival Data

Univariate survival data refer to independent, possibly censored failure times. The statistical analysis for clusered or stratified failure time data is called multivariate survival analysis.

Bivariate failure times. Observe $(y_{11}, y_{12}), (y_{21}, y_{22}), \ldots, (y_{n1}, y_{n2})$ with censoring indicators $(\delta_{11}, \delta_{12}), (\delta_{21}, \delta_{22}), \ldots, (\delta_{n1}, \delta_{n2})$.

- twin data
- eyes data

c.f. Cox and Oakes (1984).

Clustered failure times.

$(y_{11}, y_{12}, \ldots, y_{im_1}), (y_{21}, y_{22}, \ldots, y_{2m_2}), \ldots, (y_{n1}, y_{n2}, \ldots, y_{n,m_n})$ with censoring indicators $(\delta_{11}, \delta_{12}, \ldots, \delta_{1m_1}), (\delta_{21}, \delta_{22}, \ldots, \delta_{2m_2}), \ldots, (\delta_{n1}, \delta_{n2}, \ldots, \delta_{n,m_n})$

- sibling data
- family data
- clustered animal data (litters)

Recurrent event data. Observe $(t_{11}, t_{12}, \ldots, t_{1,m_1}, c_1)$, $(t_{21}, t_{22}, \ldots, t_{2,m_2}, c_2)$, ..., $(t_{n1}, t_{n2}, \ldots, t_{n,m_n}, c_n)$, where $t_{i1} < t_{i2} < \ldots, t_{i,m_i} < c_i$. Examples include repeated occurrences of hospitalizations or infections.

Statistical methods have been partially developed for data described above ....

1.8 Parametric models

Parametric models assume the knowledge of the survival or density function up to K unknown parameters. In this course, $K = 1$ or 2. Assume the failure time has the density function $f(t; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is the unknown vector of parameters. Clearly, the density and survival functions are completely specified if $\boldsymbol{\theta}$ is known.

Example: Exponential distribution.

$$T \sim \exp(\theta), \quad \theta > 0.$$

The Exponential distribution with the parameter $\theta > 0$ has the density function

$$f(t) = \theta e^{-\theta t} \ ,$$

for $t > 0$. The survival function is

$$S(t) = \int_t^\infty f(u;\theta)du = \int_t^\infty \theta e^{-\theta u}du = e^{-\theta t}$$

The hazard function is

$$\lambda(t) = \frac{f(t;\theta)}{S(t;\theta)} = \theta, \text{ a constant.} \quad ////$$

Example: Weibull distribution. The Weibull distribution with the parameters $\theta > 0$ and $\beta > 0$ assumes the parameterized survival function

$$S(t) = e^{-(\theta t)^\beta},$$

for $t > 0$. The density function is

$$f(t) = -\frac{dS_{\theta,\beta}(t)}{dt} = \beta\theta(\theta t)^{\beta-1}e^{-(\theta t)^\beta}$$

The hazard function is

$$\lambda(t) = \frac{f(t;\theta,\beta)}{S_{\theta,\beta}(t)} = \beta\theta(\theta t)^{\beta-1} .$$

Note that the hazard function $\lambda(t)$ is constant if $\beta = 1$, increasing in t if $\beta > 1$, and decreasing in t if $\beta < 1$.

Example: Gamma distribution. The Gamma distribution with the parameters $\lambda > 0$ and $r > 0$ is a continuous distribution with the density function

$$f(t) = \frac{\lambda^r}{\Gamma(r)} t^{r-1}e^{-\lambda t} ,$$

for $t \geq 0$, where $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$. The survival and hazard functions can be derived from the density function. The mean of the Gamma distribution is r/λ and the variance is r/λ^2.

Example: Log-logistic distribution. The Log-logistic distribution with the parameters $\alpha > 0$ and $-\infty < \theta < \infty$ is a continuous distribution and has the hazard function

$$\lambda(t) = \frac{e^\theta \alpha t^{\alpha-1}}{1 + e^\theta t^\alpha} .$$

The hazard function decreases monotonically if $0 < \alpha \leq 1$. The hazard function has a single mode if $\alpha > 1$. The survival function is

$$S(t) = [1 + e^\theta t^\alpha]^{-1}$$

and the density function is

$$f(t) = \frac{e^{\theta}\alpha t^{\alpha-1}}{(1+e^{\theta}t^{\alpha})^2}$$

It is called the log-logistic distribution because $logT$ has a logistic distribution (a symmetric distribution with density function similar to the normal density function).

Example: Log-normal distribution. A random variable T is said to have a lognormal distribution with parameters $-\infty < \mu < \infty$ and $\sigma > 0$. The probability density function of T is

$$f(t) = \frac{1}{\sigma(2\pi)^{1/2}} t^{-1} \exp\{-(log\ t - \mu)^2/2\sigma^2\}\ ,$$

for $t \geq 0$, from which the survival and hazard functions can be derived.

The hazard functions for the gamma and lognormal distributions are less interpretable as compared with the hazard functions for the Weibull and log-logistic distributions. Thus, the Weibull and log-logistic distributions are more useful for parametric hazard modeling.

1.9 Maximum Likelihood Estimation

Suppose that we are able to observe "complete failure times" $t_1, t_2, \ldots, t_n$.

In general, for a parametric model $T \sim f(t, \theta)$, the likelihood function on the basis of identically and independently distributed failure times $\{t_1, \ldots, t_n\}$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}).$$

The maximum likelihood estimate (mle), $\hat{\theta}$, is the θ which maximizes the likelihood function $L(\theta)$. Now we consider the case when $\boldsymbol{\theta} = \theta$ is a real number. Note tht

$$\begin{aligned} \log L(\theta) &= \sum_{i=1}^{n} \log f(t_i; \theta) \\ U(\theta) = \frac{d}{d\theta} \log L(\theta) &= \sum_{i=1}^{n} \frac{d}{d\theta} \log f(t_i; \theta) \end{aligned}$$

The mle $\hat{\theta}$ satisfies $U(\hat{\theta}) = 0$. By Taylor's expansion,

$$0 = U(\hat{\theta}) = U(\theta) + U'(\theta)(\hat{\theta} - \theta)\ +\ \text{an ignorable term.}$$

Thus

$$\hat{\theta} - \theta \approx -\frac{1}{U'(\theta)} U(\theta) = -\frac{1}{U'(\theta)} \sum_{i=1}^{n} \frac{d}{d\theta} \log f(T_i; \theta)$$

By statistical theory (law of large number, central limit theorem), when n is large,

$$\begin{aligned} (\hat{\theta}) &\overset{\text{approx}}{\sim} \text{Normal}(\theta, I^{-1}(\theta)) = N(\theta, I^{-1}(\theta)) \\ I(\theta) &= \text{Fisher information} \\ &= \mathrm{E}\left[-\frac{d^2}{d\theta^2} \log L(\theta)\right] \end{aligned}$$

<u>*Example:*</u> $T \sim \exp(\theta)$. The density function is $f(t; \theta) = \theta e^{-\theta t} I(t > 0)$.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \theta e^{-\theta t_i} \\ \log L(\theta) &= \sum_{i=1}^{n} [\log \theta - \theta t_i] \\ U(\theta) &= \frac{d}{d\theta} \log L(\theta) = \sum_{i=1}^{n} \left[\frac{1}{\theta} - t_i\right] = \frac{n}{\theta} - \sum_{i=1}^{n} t_i \end{aligned}$$

Thus $\hat{\theta} = n / \sum_{i=1}^{n} t_i$ is the mle.

Note that the Fisher information is $I(\theta) = \mathrm{E}\left[-\frac{d^2}{d\theta^2} \log L(\theta)\right] = n/\theta^2$. Thus

$$\hat{\theta} - \theta \overset{\text{approx}}{\sim} N\left(0, \frac{\theta^2}{n}\right) \quad \text{when } n \text{ is large}$$

or

$$\hat{\theta} \overset{\text{approx}}{\sim} N\left(\theta, \frac{\theta^2}{n}\right)$$

Thus $\text{Prob}\left(\hat{\theta} - 1.96\frac{\theta}{\sqrt{n}} < \theta < \hat{\theta} + 1.96\frac{\theta}{\sqrt{n}}\right) \approx 95\%$. An asymptotic 95% confidence interval for θ is

$$\left(\hat{\theta} - 1.96\frac{\hat{\theta}}{\sqrt{n}},\ \hat{\theta} + 1.96\frac{\hat{\theta}}{\sqrt{n}}\right) .$$

<u>Regression extension</u> Let x_i be a $1 \times p$ vector of covariates and θ a $p \times 1$ vector of parameters for subject i. Assume the hazard function is $\lambda(t; x_i) = x_i \theta$. Assume T has the pdf $(x_i\theta)e^{-(x_i\theta)t_i}$. Based on $(x_1, t_1), \ldots, (x_n, t_n)$, the maximum likelihood techniques can still be applied to the likelihood function

$$L(\theta) = \prod_{i=1}^{n} (x_i\theta) e^{-(x_i\theta)t_i}$$

2 One Sample Estimation

2.1 Complete Failure Times: Nonparametric Models

Recall

$$\begin{aligned} S(t) &= \mathrm{P}(T > t) \\ &= \text{Population fraction surviving beyond } t \end{aligned}$$

The set of the complete data $t_1, t_2, \ldots, t_n$ reflects the structure of population failure times. Thus, we estimate $S(t)$ by the sample fraction surviving beyond t:

$$\hat{S}(t) = \frac{\# t_i > t}{n} = \frac{1}{n}\sum_{i=1}^{n} I(t_i > t)$$

$\hat{S}(t)$ is also called the empirical survival distribution. How to derive confidence interval for $S(t)$?

Define

$$\begin{aligned} B(t) &= \sum_{i=1}^{n} I(T_i > t) = a \text{ Binomial variable} \\ B(t) &\sim \text{Binomial}(n, p = S(t)) \\ \mathrm{E}[\hat{S}(t)] &= \frac{1}{n} \cdot np = p = S(t) \\ \mathrm{Var}[\hat{S}(t)] &= \frac{1}{n^2}\mathrm{Var}(B(t)) = \frac{1}{n^2} npq \\ &= \frac{S(t)(1 - S(t))}{n} \end{aligned}$$

When n is large,

$$\hat{S}(t) \overset{\text{approx}}{\sim} \text{Normal}\left(S(t), \frac{S(t)(1 - S(t))}{n}\right) .$$

A 95% confidence interval for $S(t)$ is

$$\left(\hat{S}(t) - 1.96\sqrt{\frac{\hat{S}(t)(1 - \hat{S}(t))}{n}},\ \hat{S}(t) + 1.96\sqrt{\frac{\hat{S}(t)(1 - \hat{S}(t))}{n}}\right) .$$

Remarks

- If n is small ($n < 20$), it is more appropriate to find confidence intervals using the binomial distribution tables (see Mood, Graybill and Boes, Chapter 8).

- If n is large ($n \geq 30$), use the normal approximation to derive confidence intervals.
- The normal approximation works better when $0 << S(t) << 1$ (that is, $S(t)$ is not close to 0 or 1). When $S(t)$ is close to 0 or 1, the Poisson approximation technique is better.

2.2 Right Censored Failure Times: Parametric Models

We consider only random censoring. The observed data could be right censored:

$$(y_1, \delta_1),\ (y_2, \delta_2), \ldots, (y_n, \delta_n)$$

Note that

$$\begin{aligned} y_i &= \min(t_i, c_i) = \begin{cases} t_i & \text{uncensored case} \\ c_i & \text{censored case} \end{cases} \\ \delta_i &= I(y_i = t_i) = \begin{cases} 1 & \text{uncensored case} \\ 0 & \text{censored case} \end{cases} \end{aligned}$$

where t_i is the failure time and c_i is the censoring time.

Assume T_i and C_i are independent. In this case, the censoring process is said to be uninformative (that is, independent censoring). Let $S(t;\theta) = \mathrm{pr}(T_i > t), G(c) = \mathrm{pr}(C_i > c)$, and let $f(t;\theta)$ and $g(c)$ be the corresponding density functions. The likelihood function on the basis of $(y_1, \delta_1), \ldots, (y_n, \delta_n)$ is

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \left[f(y_i;\theta)^{\delta_i} S(y_i;\theta)^{1-\delta_i} \right] \left[g(y_i)^{1-\delta_i} G(y_i)^{\delta_i} \right] \right\}$$

or simply

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i;\theta)^{\delta_i} S(y_i;\theta)^{1-\delta_i} \right] \qquad (*)$$

Note that the validity of (*) relies on the independence between the failure and censoring times. If T_i and C_i are not independent, we then have informative censoring since the value of C_i could have implication on the value of T_i.

2.3 Right Censored Failure Times: Nonparametric Models

Without parametric assumption on the distribution of T_i, how do we estimate the survival function $S(t)$? First consider a simple example.

Example. A prospective study recruited 100 patients in January, 1990 and recruited 1000 patients in January, 1991. The study ended in January, 1992. Survival time T = time from treatment (enrollment) to death. Suppose 70 patients died in year 1 and 15 patients died in year 2 from the first cohort (recruited in 90), and 750 patients died in year 1 from the second cohort. Note that T is a discrete failure time, $T = 1, 2, \ldots$; say, $T = 2$ means death during the 2nd year.

Assume the two cohorts are sampled from the same target population. When censoring is considered random, note that this assumption implicitly implies uniformative censoring (why?).

How to estimate 2-year survival rate $S(2)$?

Approach 1 Reduced sample estimate

Only use information from individuals who had been followed for at least two years. That is, use only group 1 data to derive

$$\hat{S}(2) = \frac{100 - 70 - 15}{100} = \frac{15}{100} = 0.15$$

This estimate is statistically appropriate but inefficient. It is appropriate in the sense that $\hat{S}(2)$ is very close to $S(2)$ when n_1 is large. It is inefficient because only part of the data is used. Here

$$\hat{\text{var}}(\hat{S}(2)) = \frac{\hat{S}(2)(1 - \hat{S}(2))}{100}.$$

Approach 2 (Statistically inappropriate approaches)

— Assume 250 individuals from group 2 died in year 2,

$$\hat{S}(2) = \frac{15}{1100} = 0.014$$

— Assume 250 individuals from group 2 remained alive in year 2

$$\hat{S}(2) = \frac{15 + 250}{1100} = 0.241$$

— Exclude 250 patients from the analyzed data (Watch out! A common mistake!)

$$\hat{S}(2) = \frac{15}{1100 - 250} = 0.018.$$

Approach 3 (A simple case of the Kaplan-Meier estimate). Decompose the survival function into conditional probabilities.

$$\begin{aligned} S(2) = \mathrm{P}(T > 2) &= \frac{Pr(T \geq 2)}{Pr(T \geq 1)} \cdot \frac{Pr(T \geq 3)}{Pr(T \geq 2)} \\ &= Pr(T \geq 2 | T \geq 1) \cdot Pr(T \geq 3 | T \geq 2) \\ \hat{Pr}(T \geq 2 | T \geq 1) &= \frac{30 + 250}{1100} = \frac{280}{1100} \\ \hat{Pr}(T \geq 3 | T \geq 2) &= \frac{15}{30} \end{aligned}$$

Thus

$$\hat{S}(2) = \frac{280}{1100} \cdot \frac{15}{30} = 0.127.$$

This estimator is more efficient than the reduced sample estimate. ////

Now consider the Kaplan-Meier estimator in its general form.

Kaplan-Meier Estimator

The Kaplan-Meier estimator (1958, *JASA*) is a nonparametric estimator for the survival function S. Consider now either random censoring or type-I censoring. Assume uninformative censoring. That is, assume that T_i is independent of C_i for each i. The data are

$$(y_1, \delta_1),\ (y_2, \delta_2),\ \ldots,\ (y_n, \delta_n).$$

Let $y_{(1)} < y_{(2)} < \ldots < y_{(k)},\ k \leq n$, be the distinct, uncensored and ordered failure times.

Example. Data: $3, 2^+, 0, 1, 5^+, 3, 5$

$$(y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}) = (0, 1, 3, 5).\ ////$$

Suppose $y_{(i-1)} \leq t < y_{(i)}$. A principle of nonparametric estimation of S is to assign positive probability to and only to uncensored failure times. Therefore, we try to estimate

$$S(t) \approx \frac{Pr(T \geq y_{(2)})}{Pr(T \geq y_{(1)})} \cdot \frac{Pr(T \geq y_{(3)})}{Pr(T \geq y_{(2)})} \cdots \frac{Pr(T \geq y_{(i)})}{Pr(T \geq y_{(i-1)})}.$$

How to estimate $S(t)$? Define

$$\begin{array}{rcl} R_{(j)} &=& \{y_k \ : \ y_k \geq y_{(j)}\} \\ d_{(j)} &=& \# \text{ of failures at } y_{(j)} \\ N_{(j)} &=& \# \text{ of individuals at risk at } y_{(j)} = \#R_{(j)} \end{array}$$

Example Using the previous example 3 2^+ 0 1 5^+ 3 5

$$\begin{array}{rcl} N_{(1)} &=& 7, \ N_{(2)} = 6, \ N_{(3)} = 4, \ N_{(4)} = 2 \\ d_{(1)} &=& 1, \ d_{(2)} = 1, \ d_{(3)} = 2, \ d_{(4)} = 1. \quad //// \end{array}$$

Now estimate $\frac{Pr(T \geq y_{(j+1)})}{Pr(T \geq y_{(j)})}$ by $\frac{N_{(j)} - d_{(j)}}{N_{(j)}}$, $j = 1, 2, \ldots, i-1$. The Kaplan-Meier estimate is thus

$$\begin{array}{rcl} \hat{S}(t) &=& \left(1 - \frac{d_{(1)}}{N_{(1)}}\right) \left(1 - \frac{d_{(2)}}{N_{(2)}}\right) \cdots \left(1 - \frac{d_{(i-1)}}{N_{(i-1)}}\right) \\ &=& \boxed{\prod_{y_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{N_{(j)}}\right)} \end{array}$$

Example 3, 2^+, 0, 1, 5^+, 3, 5

uncensored times	0	1	3	5
$d_{(i)}$	1	1	2	1
$N_{(i)}$	7	6	4	2

$$\begin{array}{rcl} \hat{S}(0) &=& \left(1 - \frac{1}{7}\right) = \frac{6}{7} = 0.86 \\ \hat{S}(1) &=& \frac{6}{7}\left(1 - \frac{1}{6}\right) = \frac{5}{7} = 0.71 \\ \hat{S}(3) &=& \frac{5}{7} \cdot \left(1 - \frac{2}{4}\right) = \frac{5}{14} = 0.36 \\ \hat{S}(5) &=& \frac{5}{14}\left(1 - \frac{1}{2}\right) = \frac{5}{28} = 0.18 \end{array}$$

Remark In general, if the largest observed time is uncensored, the Kaplan-Meier estimate will reach the value 0 as $t \geq$ the largest observed time. if the largest observed time is censored, the Kaplan-Meier estimate will not go down to 0 and is unreliable for $t >$ largest y_i. In this case, we say that $\hat{S}(t)$ is undetermined for $t >$ the largest uncensored time.

Greenwood's formula

The next question is how to identify the variance of the Kaplan-Meier estimate. The idea is sketched for grouped data. First group the data using the uncensored times $y_{(1)} < y_{(2)} < \ldots < y_{(k)}$.

For each risk set $R_{(j)} = \{y_i \ : \ y_i \geq y_{(j)}\}$, counting the number of failures is a binomial experiment. Thus $d_{(j)} \sim$ Binomial $(N_{(j)}, \lambda_{(j)})$, where $\lambda_{(j)}$ is the hazard at $y_{(j)}$. Let $q_{(j)} = 1 - \lambda_{(j)}$. For $y_{(i-1)} \leq t < y_{(i)}$,

$$\begin{aligned} \mathrm{var}(\log\ \hat{S}(t)) &= \mathrm{var}(\log\{\hat{q}_{(1)}\hat{q}_{(2)}, \ldots, \hat{q}_{(i-1)}\}) \\ &= \mathrm{var}(\log \hat{q}_{(1)} + \log \hat{q}_{(2)} + \ldots + \log \hat{q}_{(i-1)}) \\ &= \sum_{j=1}^{i-1} \mathrm{var}(\log \hat{q}_{(j)}) \end{aligned}$$

The variances are additive because the risk sets at $y_{(1)}, y_{(2)}, \ldots, y_{(k)}$ are nested ($R_{(1)} \supset R_{(2)} \supset \ldots$). Thus, by statistical theory, we can treat $\log \hat{q}_{(1)}, \log \hat{q}_{(2)} \ldots$ as uncorrelated terms. Use the delta method, for a transformation ϕ of an estimate $\hat{\theta}$, we have

$$\mathrm{var}(\phi(\hat{\theta})) \approx [\phi'(\theta)]^2 \mathrm{var}(\hat{\theta}).$$

Thus

$$\begin{aligned} \mathrm{var}(\log \hat{q}_{(j)}) &\approx \left[\frac{1}{q_{(j)}}\right]^2 \mathrm{var}(\hat{q}_{(j)}) = \frac{1}{q_{(j)}^2} \cdot \frac{\lambda_{(j)} q_{(j)}}{N_{(j)}} = \frac{\lambda_{(j)}}{q_{(j)} N_{(j)}}, \\ \mathrm{var}(\log \hat{S}(t)) &= \sum_{j=1}^{i-1} \mathrm{var}(\log \hat{q}_{(j)}) \approx \sum_{y_{(j)} \leq t} \left(\frac{\lambda_{(j)}}{q_{(j)} N_{(j)}}\right) \end{aligned}$$

Use the delta method again,

$$\begin{aligned} \sigma(t)^2 = \mathrm{var}(\hat{S}(t)) &= \mathrm{var}(\ \underset{\phi}{\exp}\ \ \underset{\hat{\theta}}{(\log \hat{S}(t))}\ ) \\ &\approx [S(t)]^2 \cdot \mathrm{var}(\log \hat{S}(t)) \end{aligned}$$

Plug in $\hat{\lambda}_{(j)} = d_{(j)}/N_{(j)}$ and $\hat{q}_{(j)} = \frac{N_{(j)} - d_{(j)}}{N_{(j)}}$. The Greenwood's formula, for estimating the variance of the Kaplan-Meier estimate, is

$$\boxed{\hat{\mathrm{var}}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{y_{(j)} \leq t} \frac{d_{(j)}}{N_{(j)}(N_{(j)} - d_{(j)})}}$$

Property When n is large

$$\hat{S}(t) \overset{\text{approx}}{\sim} \text{Normal}(S(t), \sigma(t)^2)$$

where $\sigma(t)^2$ can be estimated by the Greenwood's formula.

Remark 1: This general property holds also for continuous survival data.
Remark 2: A more formal approach which allows for theoretical developments of continuous survival data is through a representation of S:

$$S(t) = e^{-\int_0^t \lambda(v)dv} = e^{-\int_0^t \frac{dF^u(v)}{R(v)}}$$

where $F^u(v)$ is the cdf of uncensored Y and $R(v)$ is the cdf Y. Let $\hat{F}^u$, $\hat{R}$ be the empirical distribution estimates. Then

$$\hat{S}_{KM}(t) \approx e^{-\int_0^t \frac{\hat{F}^u(dv)}{\hat{R}(v)}} .$$

Theoretical properties can be developed based on probability theory.

Nonparametric MLE

Kaplan and Meier showed that the K-M estimate is the unique nonparametric mle from the likelihood function

$$\mathcal{L} \propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right],$$

where the likelihood maximization is subject to the class of probability distributions which assign probability to, and only to uncensored failure times. To see the Kaplan-Meier estimator is the unique mle of the likelihood function $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &\propto \prod_{i=1}^{n} \left[f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \right] = \prod_{i=1}^{n} \left\{ \frac{f(y_i)}{S(y_i)} \right\}^{\delta_i} \{S(y_i)\} \\ &= \left\{ \prod_{(i)} \lambda_{(i)}^{d_{(i)}} \right\} \left\{ \prod_{i=1}^{n} \prod_{y_{(j)}<y_i} (1-\lambda_{(j)}) \right\} = \prod_{(i)} \lambda_{(i)}^{d_{(i)}} (1-\lambda_{(i)})^{N_{(i)}-d_{(i)}} \end{aligned}$$

Thus, the unique mle of $\lambda_{(i)}$ is $d_{(i)}/N_{(i)}$ and the Kaplan-Meier estimate is the unique mle.

Reference: Kaplan & Meier *JASA*, 1958.

Remark: K-M used $S(t) = P(T \geq t)$ instead of $S(t) = P(T > t)$ for their MLE parameterization.

Example (Lee, p29) Forty-two patients with acute leukemia were randomized into a treatment group and a placebo group to assess the treatment effect to maintain remission. T : remission time.

- 6-MP (6-mercaptopurine) group, $n_1 = 21$
  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+, 17^+,$
  $19^+, 20^+, 25^+, 32^+, 32^+, 34^+, 35^+$ (months)
- Placebo group, $n_2 = 21$
  $1, 1, 2, 2, 3, 4, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15,$
  $17, 22, 23$ (months)

The empirical survival function from the placebo group is

$$\begin{aligned}
\hat{S}(0) &= \frac{21}{21} = 1 \\
\hat{S}(1) &= \frac{19}{21} \\
\hat{S}(2) &= \frac{17}{21} \\
\hat{S}(3) &= \frac{16}{21} \\
\hat{S}(4) &= \frac{14}{21} = 0.67 \\
&\vdots \\
\hat{\mathrm{var}}(\hat{S}(4)) &= \frac{(0.67)(0.33)}{21} \\
\hat{SD}(\hat{S}(4)) &= \sqrt{\frac{(0.67)(0.33)}{21}} = 0.103
\end{aligned}$$

A 95% confidence interval at $t = 4$ is

$$(0.67 - 1.96 \times 0.103, \quad 0.67 + 1.96 \times 0.103) = (0.47, \ 0.87).$$

Warning: The sample size $n_2 = 21$ may not be large enough for the normal approximation!

For the 6MP group, use the K-M estimate to derive

$$\begin{aligned}
\hat{S}(5) &= 1 \\
\hat{S}(6) &= \left(1 - \frac{3}{21}\right) \\
\hat{S}(7) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right) \\
\hat{S}(10) &= \left(1 - \frac{3}{21}\right)\left(1 - \frac{1}{17}\right)\left(1 - \frac{1}{15}\right) = 0.753
\end{aligned}$$

....

Apply the Greenwood's formula to get

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{S}(10)) &= (0.753)^2\left(\frac{3}{21\times 18}+\frac{1}{17\times 16}+\frac{1}{15\times 14}\right)\\ &= 0.0093\end{aligned}$$

A 95% confidence interval for $S(10)$ is

$$(0.753 - 1.96\sqrt{0.0093}\ ,\ 0.753 + 1.96\sqrt{0.0093}) = (0.564\ ,\ 0.942)$$

What about $\hat{S}(11)$ and $\hat{\mathrm{var}}(\hat{S}(11))$?

— Same as ($\hat{S}(10)$ and $\hat{\mathrm{var}}(\hat{S}(10))$. ////

Remark 1 The K-M estimate is a nonparametric method which can be applied to either discrete or continuous data. For a rigorous development of statistical theory, see Kalbfleisch and Prentice (1980).

Remark 2 The accuracy of the K-M estimate and Greenwood's formula relies on large sample size of uncensored data. Make sure that you have at least, say, 20 or 30 uncensored failure times in your data set before using the methods.

Remark 3 Greenwood's formula is more appropriate when $0 << S(t) << 1$. Using Greenwood's formula, the confidence interval limits could be above 1 or below 0. In these cases, we usually replace these limit points by 1 or 0. For example, a 95% confidence interval could be $(0.845, 1.130)$, we will use $(0.845, 1)$ instead.

3 Proportional Hazrds Model (PHM)

3.1 The model

Now we move to regression analysis. Assume covariates are available on each individual

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^t.$$

The PHM assumes

$$\begin{aligned}\lambda(t; \boldsymbol{x_i}) &= \lambda_0(t)e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip}} \\ &= \lambda_0(t)e^{\boldsymbol{\beta x_i}}\end{aligned}$$

where $\boldsymbol{x_i}$ is $p \times 1$ vector of covariates and $\boldsymbol{\beta}$ is a $1 \times p$ vector of parameters. Interpretation of the model:

$$\text{Hazard at } t \text{ for given } x_i = (\text{baseline hazard at } t) \times (\text{Risk factor } e^{\boldsymbol{\beta x_i}})$$

Characteristics of the model:

- The PHM is a model on the basis of hazard function
  Note: Alternatively, you might be interested in the 'accelerated failure time model':

$$T_i = T_{0i} \cdot e^{\boldsymbol{x_i \beta}} \Longleftrightarrow \begin{cases} \log T_i = \boldsymbol{\beta x_i} + \log T_{0i}, \quad T_{0i} \sim S_0 \\ \text{(a standard linear model)} \end{cases}$$

- The baseline hazard $\lambda_0(t)$ is left unspecified (nonparametric), thus the PHM is a semiparametric model: λ_0 = nonparametric component, β: parametric component.
- In most applications related to public health, the parameter β is of primary interest and $\lambda_0(t)$ is of minor interest. However, estimation of $\lambda_0(t)$ is desirable when we wish to predict the hazard for an individual with covariates $\boldsymbol{x_i}$.

3.2 PHM as Lehmann's Alternatives

The PHM can also be expressed as

$$S(t; \boldsymbol{x_i}) = S_0(t)^{e^{\boldsymbol{\beta x_i}}}$$

Proof

$$\begin{aligned} S(t;\boldsymbol{x_i}) &= e^{-\int_o^t \lambda(u;\boldsymbol{x_i})du} \\ &= e^{-\int_o^t \lambda_0(u)e^{\boldsymbol{\beta x_i}}du} \\ &= e^{[-\int_o^t \lambda_0(u)du]\cdot e^{\boldsymbol{\beta x_i}}} \\ &= S_0(t)^{e^{\boldsymbol{\beta x_i}}}. \qquad //// \end{aligned}$$

We say that a class of distributions with the form

$$S(t) = S_0(t)^\gamma$$

for some positive γ is a family of "Lehmann's alternatives". Clearly, the PHM implies that the distribution functions form a family of "Lehmann's alternatives". The PHM is a very flexible model because of its semiparametric feature, but the validity of the model is not automatic and still needs to be confirmed.

Example A two-sample case

$$x = \begin{cases} 0 & \text{represents treatment A} \\ 1 & \text{represents treatment B} \end{cases}$$

Under the PHM,

$$\lambda(t;x) = \lambda_0(t)e^{\beta x}.$$

That is

$$\lambda_1(t) = \lambda_0(t)e^{\beta}.$$

Using Lehmann's alternative expression, we derive

$$\begin{aligned} S_1(t) &= S_0(t)^{e^\beta} \\ \log S_1(t) &= e^\beta \cdot \log S_0(t) \\ &= \text{constant } \cdot \log S_0(t) \end{aligned}$$

For exploratory analysis, to examine the validity of the PHM for two-sample case, we can use the K-M estiamtes $\hat{S}_1$ and $\hat{S}_0$ to see if

$$\phi(t) = \frac{\log \hat{S}_1(t)}{\log \hat{S}_0(t)} \approx \text{ constant.}$$

The PHM is a valid model if $\phi(t)$ remains a constant over time.

3.3 Partial Likelihood Method

Assume independent censoring: Conditional on $\boldsymbol{x_i}$, T_i and C_i are independent.

Assume the PHM

$$\lambda(t;\boldsymbol{x_i}) = \lambda_0(t)e^{\beta_1 x_{i1}+\cdots+\beta_p x_{ip}} = \lambda_0(t)e^{\boldsymbol{\beta x_i}}$$

$$\begin{aligned}
\text{Data} &: (y_1,\delta_1,\boldsymbol{x_1}),\cdots,(y_n,\delta_n,\boldsymbol{x_n}) \\
y_i &= \text{observed follow-up time} \\
\delta_i &= \text{censoring indicator} \\
\boldsymbol{x_i} &= \text{covariates} \\
H_{(i)} &= \text{data history up to } y_{(i)}^-
\end{aligned}$$

Assume failure times are <u>not</u> tied. The likelihood function is

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} f(y_i;\boldsymbol{x_i})^{\delta_i} S(y_i;\boldsymbol{x_i})^{1-\delta_i} \\
&\qquad\ \uparrow \qquad\qquad\quad \nwarrow \\
&\text{density function} \quad \text{survival function} \\
&= \prod_{(i)} p(x_{(i)}|H_{(i)},y_{(i)})P(H_{(i)},y_{(i)}) \\
&= \left\{ \prod_{\substack{\text{uncensored}\\(i)}} \left[\frac{e^{\boldsymbol{x_{(i)}\beta}}}{\sum_{j\in R_{(i)}} e^{\boldsymbol{x_j\beta}}}\right] \right\} \times \{\text{something ignorable}\}
\end{aligned}$$

where $R_{(i)}$ = Risk set at $y_{(i)}$, and $\boldsymbol{x_{(i)}}$ = covariates corresponding to $y_{(i)}$.

The first likelihood is called the "partial likelihood". Cox (1972, JRSS-B; 1975, Biometrika) identified the above likelihood structure. Thus the partial likelihood method is also referred to as Cox's method.

The result is great!! Why?

- The result is derived under an attractive model. The PHM has nice interpretations in terms of hazards and it is semiparametric.
- The partial likelihood only involves β!! It does <u>not</u> involve $\lambda_0(t)$, and thus computation of $\hat{\beta}$ is manageable and inferences can be developed.

<u>How did Cox obtain the ideas of partial likelihood?</u>

Assume no ties in the uncensored failure times. Let $L_p =$ The partial likelihood.

Any "likelihood" must correspond to a probability (or density) of some kind. Note that

$$
\begin{aligned}
& \mathrm{P}\left(\text{individual } x_{(i)} \text{ fails at } y_{(i)} \middle| \begin{array}{l}\text{a failure occurring at } y_{(i)} \text{ and} \\ \text{data history before } (<) y_{(i)}\end{array}\right) \\
= \; & \mathrm{P}\left(x_{(i)} \text{ fails at } y_{(i)} \middle| \text{ a failure occurring at } y_{(i)} \text{ and} R_{(i)}\right) \\
= \; & \frac{\lambda_0(y_{(i)}) e^{\boldsymbol{\beta x_{(i)}}}}{\sum_{j \in R_{(i)}} \lambda_0(y_{(i)}) e^{\boldsymbol{\beta x_j}}} = \frac{e^{\boldsymbol{\beta x_{(i)}}}}{\sum_{j \in R_{(i)}} e^{\boldsymbol{\beta x_j}}}
\end{aligned}
$$

Thus, the "partial likelihood" is

$$
\begin{aligned}
L_p \; &= \prod_{\substack{\text{uncensored} \\ (i)}} \mathrm{P}(\boldsymbol{x_{(i)}} \text{ fails at } y_{(i)} | \text{a failure occurring at } y_{(i)}, R_{(i)}) \\
&= \prod_{(i)} \left(\frac{e^{\boldsymbol{\beta x_{(i)}}}}{\sum_{j \in R_{(i)}} e^{\boldsymbol{\beta x_j}}} \right)
\end{aligned}
$$

Derive the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ by maximizing L_p over possible values of $\boldsymbol{\beta}$.

Example Two-sample case

No treatment:	7,	9^+,	18
Treatment:	12,	19^+	

$$
x = \begin{cases} 0 & \text{no treatment} \\ \\ 1 & \quad \text{treatment} \end{cases} \qquad \underline{\text{PHM}} : \lambda(t; x) = \lambda_0(t) e^{x\beta}
$$

The partial likelihood is

$$
\begin{aligned}
L_p \; &= \left[\frac{e^{0\beta}}{e^{0\beta} + e^{0\beta} + e^{0\beta} + e^{\beta} + e^{\beta}} \right] \left[\frac{e^{\beta}}{e^{0\beta} + e^{\beta} + e^{\beta}} \right] \left[\frac{e^{0\beta}}{e^{0\beta} + e^{\beta}} \right] \\
&= \left[\frac{1}{3 + 2e^{\beta}} \right] \left[\frac{e^{\beta}}{1 + 2e^{\beta}} \right] \left[\frac{1}{1 + e^{\beta}} \right]
\end{aligned}
$$

Obtain the mle $\hat{\beta}$ by maximizing L_p. ////

3.4 Generalization to Time-Dependent Covariates

Sometimes part of the covariates could be time-dependent. For example, the time dependent covariates could be

- age at failure time t
- dosage level at failure time t
- accumulative dosage at failure time t
- treatment status (off or on) at failure time t

or a transformation of the above time-dependent measurements.

Time-dependent covariates for the i$^{\text{th}}$ individual are

$$\boldsymbol{x_i}(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$$

We shall use the general notation $\boldsymbol{x_i}(t)$ instead of $\boldsymbol{x_i}$, even though some of the covariates are time-independent. The PHM is now

$$\lambda(t; \boldsymbol{x_i}(u),\ u \le t) = \lambda_0(t) e^{\boldsymbol{\beta x_i}(t)}.$$

With time-dependent covariates, the previous partial likelihood argument still works, and the partial likelihood becomes

$$L_p = \prod_{y_{(i)}} \left(\frac{e^{\beta x_{(i)}(y_{(i)})}}{\sum_{j \in R_{(i)}} e^{\boldsymbol{\beta x_j}(y_{(i)})}} \right)$$

Example. Suppose

$$\begin{aligned}
\boldsymbol{x_i}(t) &= (x_{i1}\ ,\ x_{i2}(t),\ x_{i3}(t)) \\
x_{i1} &= \begin{cases} 1 & \text{treatment} \\ 0 & \text{no treatment} \end{cases} \\
x_{i2}(t) &= \text{the } i^{\text{th}} \text{ individual's age at } t \\
x_{i3}(t) &= (\text{the } i^{\text{th}} \text{ individual's age at } t)^2
\end{aligned}$$

T = time from entry to death.

Note that $x_{i2}(0)$ = baseline age of the i^{th} patient. The partial likelihood is

$$L_p = \prod_{y_{(i)}} \left(\frac{e^{\beta_1 x_{(i1)} + \beta_2 x_{(i2)}(y_{(i)}) + \beta_3 x_{(i3)}(y_{(i)})}}{\sum_{j \in R_{(i)}} e^{\beta_1 x_{j1} + \beta_2 x_{j2}(y_{(i)}) + \beta_3 x_{j3}(y_{(i)})}} \right)$$

Suppose the observed data are

Treatment

I.D.	001	002
age at entry	10	12
y_i	12	19^+

No treatment

I.D.	003	004	005
age at entry	4	0	11
y_i	7	9^+	18

Time-dependent age

	I.D./$y_{(i)}$	7	12	18
$x_{i1} = 1$	001	17	22	
	002	19	24	30
	003	11		
$x_{i1} = 0$	004	7		
	005	18	23	29

(Time-dependent age)^2

	I.D./$y_{(i)}$	7	12	18
$x_{i1} = 1$	001	17^2	22^2	
	002	19^2	24^2	30^2
	003	11^2		
$x_{i1} = 0$	004	7^2		
	005	18^2	23^2	29^2

Note: Computer needs the above "covariate process data" for time-dependent covariates analysis.

$$L_p = \left[\frac{e^{\beta_1 \cdot 0 + \beta_2 \cdot 11 + \beta_3 \cdot 11^2}}{e^{\beta_1 \cdot 1 + \beta_2 \cdot 17 + \beta_3 \cdot 17^2} + e^{\beta_1 \cdot 1 + \beta_2 \cdot 19 + \beta_3 \cdot 19^2} + \ldots + e^{\beta_1 \cdot 1 + \beta_2 \cdot 18 + \beta_3 \cdot 18^2}} \right]$$

$$\cdot\left[\frac{e^{\beta_1\cdot 1+\beta_2\cdot 22+\beta_3\cdot 22^2}}{e^{\beta_1 1+\beta_2\cdot 22+\beta_3\cdot 22^2}+e^{\beta_1\cdot 1+\beta_2\cdot 24+\beta_3\cdot 24^2}+e^{\beta_1\cdot 0+\beta_2\cdot 23+\beta_3\cdot 23^2}}\right]$$

$$\cdot\left[\frac{e^{\beta_1\cdot 0+\beta_2\cdot 29+\beta_3\cdot 29^2}}{e^{\beta_1\cdot 0+\beta_2\cdot 29+\beta_3\cdot 29^2}+e^{\beta_1\cdot 1+\beta_2\cdot 30+\beta_3\cdot 30^2}}\right] \quad ////$$

Remark: Using the baseline age x_{i2} or time-dependent age $x_{i2}(t)$ as a linear term in the proportional hazards model would end up with the same partial likelihood estimate $\hat{\beta}_2$ because

$$\begin{aligned}\lambda_0(t)e^{\beta_1 x_{i1}+\beta_2 x_{i2}(t)+\beta_3 x_{i3}(t)} &= \lambda_0(t)e^{\beta_1 x_{i1}+\beta_2(x_{i2}+t)+\beta_3 x_{i3}(t)}\\ &= \lambda_0^*(t)e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}(t)}\end{aligned}$$

where $\lambda_0^*(t)=\lambda_0(t)e^{\beta_2 t}$ is also a baseline hazard function.

Example T : Time from onset of treatment to AIDS
(definition before Jan. 1993)
$x_i(t)$: CD4 count for the ith individual at time t

$$\lambda(t;x_i(u),\ u\le t)=\lambda_0(t)e^{\beta x_i(t)}.$$

$$\begin{aligned}\text{Relative hazard (R.H.) at}t &= \frac{\lambda(t;x_i(u),u\le t)}{\lambda(t;x_k(u),u\le t)}\\ &= \frac{\lambda_0(t)e^{\beta\cdot x_i(t)}}{\lambda_0(t)e^{\beta\cdot x_k(t)}}\\ &= e^{\beta(x_i(t)-x_k(t))}\end{aligned}$$

If $\beta=-0.01$, $x_i(t)=250$, $x_k(t)=200$, then

$$\text{R.H.}=e^{-0.01\times(250-200)}=e^{-0.5}\approx 0.6065.$$

Note that the R.H. is determined by the covariate information defined, theoretically, at t, although in applications we could use an earlier measurement (such as the treatment received one month ago) as the current $x(t)$. So, be smart and flexible when a time-dependent covariate is used in the analysis.

3.5 Tied Survival Data

The partial likelihood methods so far do not handle tied survival data. When we analyze discrete or grouped survival data, the problem of how to analyze such data naturally arises. Consider the following simple PHM: $\lambda(t;x_i)=\lambda_0(t)e^{\beta x_i}$,

No treatment	7	9^+	18	$x_1,x_2,x_3=0$
Treatment	18	19^+		$x_4,x_5=1$

Recall the partial likelihood construction is motivated by

$$\mathrm{P}(x_{(i)} \text{ fails at } y_{(i)} |\text{ a failure occurring at } y_{(i)},\ R_{(i)}).$$

Now, at $y_{(2)}=18$, the probability becomes

$$\mathrm{P}(x_3 \text{ and } x_4 \text{ fail at } 18\,|\text{two failures at } 18,\ \text{risk set at } 18=\{x_3,x_4,x_5\})$$

$$=\frac{\lambda_0(18)e^{\beta\cdot x_3}\cdot\lambda_0(18)e^{\beta\cdot x_4}}{\lambda_0(18)e^{\beta\cdot x_3}\cdot\lambda_0(18)e^{\beta\cdot x_4}+\lambda_0(18)e^{\beta\cdot x_4}\cdot\lambda_0(18)e^{\beta\cdot x_5}+\lambda_0(18)e^{\beta\cdot x_3}\cdot\lambda_0(18}$$

$$=\frac{e^{\beta\cdot 0+\beta\cdot 1}}{(e^{\beta\cdot 0+\beta\cdot 1}+e^{\beta\cdot 1+\beta\cdot 1}+e^{\beta\cdot 0+\beta\cdot 1})}$$

The partial likelihood is

$$\begin{aligned} L_p &= \left(\frac{e^{\beta\cdot 0}}{3\cdot e^{\beta\cdot 0}+2\cdot e^{\beta\cdot 1}}\right)\left(\frac{e^{\beta\cdot 0+\beta\cdot 1}}{e^{\beta\cdot 0+\beta\cdot 1}+e^{\beta\cdot 1+\beta\cdot 1}+e^{\beta\cdot 0+\beta\cdot 1}}\right) \\ &= \left(\frac{1}{3+2e^{\beta}}\right)\left(\frac{e^{\beta}}{2e^{\beta}+e^{2\beta}}\right) \quad //// \end{aligned}$$

For the general data $(\boldsymbol{x_1},y_1,\delta_1),(\boldsymbol{x_2},y_2,\delta_2),\ldots,(\boldsymbol{x_n},y_n,\delta_n)$, the partial likelihood for tied survival data is

$$L_p=\prod_{(i)}\left(\frac{e^{\sum_{j\in D_{(i)}}\boldsymbol{\beta}\cdot\boldsymbol{x_j}(y_{(i)})}}{\sum_{\substack{\text{combinations}\\ D^*_{(i)}\subset R_{(i)}}} e^{\sum_{j\in D^*_{(i)}}\boldsymbol{\beta}\cdot\boldsymbol{x_j}(y_{(i)})}}\right)$$

Where $D_{(i)}$ is the set of "deaths" (or failures) occurring at $y_{(i)}, D^*_{(i)}$ is a a combination of deaths (or failures) from the risk set $R_{(i)}$, with the restriction $\#D^*_{(i)} = \#D_{(i)}$.

Computation of the mle from L_p for tied survial data in a big problem. Statisticians are still developing fast algorithms for calculation!

- If you have heavily tied survival data, check your computing packages to see if they handle such data.
- Some of the computing packages use the Breslow's approach (Breslow, 1972, *Biometrics*) to handle problems with tied data. The results are reasonably accurate if you have a small proportion of ties. Here the Breslow's approach refers to: Each of a set of tied failure times is sequentially treated as though it occurred just before the others.

3.6 Discrete Survival Data

In the situation that the failure times are truly discrete, we may replace the proportional hazards model by the discrete logistic regression model

$$\frac{\lambda(t_k; x(u), u \leq t_k)}{1 - \lambda(t_k; x(u), u \leq t_k)} = \frac{\lambda_0(t_k)}{1 - \lambda_0(t_k)} e^{\beta x(t_k)}$$

where t_k, $k = 1, 2, \dots, K$, are the discrete points of the failure time T. Equivalently, the logistic model can be also expressed as

$$\frac{\lambda(t_k; x(u), u \leq t_k)}{1 - \lambda(t_k; x(u), u \leq t_k)} = e^{\alpha_k + \beta x(t_k)}$$

with $e^{\alpha_k} = \lambda_0(t_k)/\{1 - \lambda_0(t_k)\}$.

There are a number of approaches developed to estimate the parameter β; see Breslow and Day (Volume 1, 1980) for details.

3.7 Estimation of $\lambda_0(t)$

Breslow (1972, *JRSS B*) gave a heuristic argument. He assumed $\hat{\lambda}_0(t)$ to be constant between uncensored survival times. Let $\hat{\lambda}_{(0)}, \hat{\lambda}_{(1)}, \hat{\lambda}_{(2)}, \dots$ be constants

$$\hat{\lambda}_0(t) = \begin{cases} \hat{\lambda}_{(0)} & 0 \leq t < y_{(1)} \\ \hat{\lambda}_{(1)} & y_{(1)} \leq t < y_{(2)} \\ \dots \end{cases} .$$

Say, we are interested in $\hat{\lambda}_{(2)}$. Tbe people in the risk set at $y_{(2)}$ are in $R_{(2)}$. Since we know one person fails at $y_{(2)}$, thus for given $(y_{(2)}, R_{(2)})$,

$$\begin{aligned} 1 &= \sum_{j \in R_{(2)}} \text{P (the } j^{\text{th}} \text{ individual fails at } y_{(2)} | y_{(2)}, R_{(2)}) \\ &= \sum_{j \in R_{(2)}} (y_{(3)} - y_{(2)}) \hat{\lambda}_{(2)} e^{\boldsymbol{\beta x_j}} \\ &= (y_{(3)} - y_{(2)}) \hat{\lambda}_{(2)} \sum_{j \in R_{(2)}} e^{\boldsymbol{\beta x_j}} \end{aligned}$$

Thus, the hazard probability between $y_{(2)}$ and $y_{(3)}$ is

$$(y_{(3)} - y_{(2)}) \hat{\lambda}_{(2)} = \frac{1}{\sum_{j \in R_{(2)}} e^{\boldsymbol{\beta x_j}}}$$

Now use $\hat{\beta}$ (the mle derived from the partial likelihood) to derive

$$\hat{\lambda}_{(2)} = \frac{1}{(y_{(3)} - y_{(2)}) \sum_{j \in R_{(2)}} e^{\hat{\beta} \boldsymbol{x_j}}}$$

Now, you may estimate an individual's hazard probability between $y_{(2)}$ and $y_{(3)}$ by

$$\begin{aligned} & (y_{(3)} - y_{(2)}) \cdot \{ \text{ hazard with } \boldsymbol{x_i} \text{ in } [y_{(2)}, y_{(3)}) \} \\ = \ & (y_{(3)} - y_{(2)}) \cdot \hat{\lambda}_{(2)} e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_i}} \\ = \ & \frac{e^{\hat{\boldsymbol{\beta}} x_i}}{\sum_{j \in R_{(2)}} e^{\hat{\beta} \boldsymbol{x_j}}}, \end{aligned}$$

where $\boldsymbol{x_i}$ is that indiv's covariates. Similary, you can also estimate an individual's hazard probability between $y_{(m)}$ and $y_{(m+1)}$ by

$$\frac{e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_i}}}{\sum_{j \in R_{(m)}} e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_j}}}$$

If you are interested in the "cumulative hazard probability" within $(0, y_{(m+1)})$, you just add up the hazard probabilities

$$\frac{e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_i}}}{\sum_{j \in R_{(1)}} e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_j}}} + \ldots + \frac{e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_i}}}{\sum_{j \in R_{(m)}} e^{\hat{\boldsymbol{\beta}} \boldsymbol{x_j}}}$$

Note: Although the estimate of the cumulative hazard probability described above is statistically accurate when the sample size is large, the Breslow's estimate of the hazard function can be greatly improved by smoothing techniques.

3.8 Goodness of Fit

Time-independent x - material from Miller's book, p168-170

Suppose we want to check on the validity of proportional hazards model. In the case that x is one-dimensional, an approach of goodness-of-fit is to partition the x-axis into K intervals, compute a separate Kaplan-Meier estimate for each interval, then apply the 2-sample goodness-of-fit procedures. When the time-independent covariate x is multi-dimensional, we consider the following approach. Define

$$\Lambda_{x_i}(T_i) = e^{\beta x_i} \int_0^{T_i} \lambda_0(u)du$$

Thus, because $\Lambda_{x_i}(T_i)$ is monotonic in T_i,

$$\begin{aligned} \mathrm{P}(\Lambda_{x_i}(T_i) > t) &= \mathrm{P}(T_i > \Lambda_{x_i}^{-1}(t)) \\ &= \exp(-\Lambda_{x_i}(\Lambda_{x_i}^{-1}(t))) \\ &= e^{-t} \end{aligned}$$

Thus, the random variable $\Lambda_{x_i}(T_i)$ follows Exponential$(\theta = 1)$ distribution. Further, $(\Lambda_{x_1}(y_1), \delta_1)$,. . ., $(\Lambda_{x_n}(y_n)), \delta_n)$ form a sample with censoring. Because $\Lambda_{x_i}(y_i)$ depends on β and $\lambda_0(t)$, substitute the corresponding estimates and define

$$\hat{\Lambda}_i = \hat{\Lambda}_{x_i}(Y_i) = e^{\hat{\beta} x_i} \int_0^{Y_i} \hat{\lambda}_0(u)du \ .$$

Let $\hat{S}(t)$ be the Kaplan-Meier estimate based on $(\hat{\Lambda}_1, \delta_1)$,. . ., $(\hat{\Lambda}_n, \delta_n)$. Under the proportional hazards model, log$S(t) = -t$ is a linear function of t. To verify the validity of the proportional hazards model, check if

$$\frac{t}{log\hat{S}(t)} = -1$$

is approximately satisfied.

Time-dependent $x(t)$

When the covariate $x(t)$ is time-dependent, the above techniques no longer work for goodness-of-fit. There is a large literature regarding how to construct tests to verify the proportional hazards model assumptions. The so-called 'Martingale residuals' are used as the fundamental statistics for constructing the tests. For continuous survival data, define a 'residual' at $y_{(i)}$ as

$$\begin{aligned} r_{(i)} &= x_{(i)}(y_{(i)}) - \frac{\sum_{j\in R_{(i)}} x_j(y_{(i)}) \exp(\beta x_j(y_{(i)}))}{\sum_{k\in R_{(i)}} e^{\beta x_k(y_{(i)})}} \\ &= x_{(i)}(y_{(i)}) - \mathrm{E}[\text{covariate at } y_{(i)} \mid R_{(i)}] \end{aligned}$$

Each residual term has 0 expectation. Thus, after replacing β by $\hat{\beta}$, the corresponding residual plot should reflect this specific feature.

4 Two-Sample Testing

Goal of testing: Determine if there is a difference between two groups.

Some of the "traditional methods" are appropriate for complete failure times but not applicable to censored data.

4.1 Complete Failure Times

Suppose there is no censoring and the data include $t_1, t_2, \ldots, t_n$. We are interested in the t-year survival rate, $S(t)$, and observe

		D	$\bar{D}$	
Treatment	A	a	b	n_A
	B	c	d	n_B
		m_D	$m_{\bar{D}}$	n

$$\begin{aligned} D: &\quad \text{Failing in } t \text{ years} \\ \bar{D}: &\quad \text{Surviving beyond } t \text{ years} \\ p_A &= \mathrm{P}(D|A) \\ p_B &= \mathrm{P}(D|B) \end{aligned}$$

Consider the following way to construct a χ^2 test statistic:

		D	$\bar{D}$	
Treatment	A	a	b	n_A
	B	c	d	n_B
		m_D	$m_{\bar{D}}$	n

Null hypothesis $H_0 : p_A = p_B$ or, equivalently, $S_A(t) = S_B(t)$.

Conditional on $n_A, n_B, m_D, m_{\bar{D}}$, the count "a" follows a hypergeometric distribution (under H_0) with

$$\mathrm{E}_0(A) = m_D \left(\frac{n_A}{n} \right)$$

$$\mathrm{Var}_0(A) = \frac{n_A n_B m_D m_{\bar{D}}}{n^2(n-1)}$$

Construct a test statistic

$$T = \left(\frac{a - m_D\left(\frac{n_A}{n}\right)}{\sqrt{\frac{n_A n_B m_D m_{\bar{D}}}{n^2(n-1)}}} \right)^2$$

when n is large, $T \sim \chi^2(1)$.

4.2 A Test for Right Censored Data

Suppose t-year survival rate is of interest

$$H_0 : S_A(t) = S_B(t).$$

Data could be censored before t. We use the K-M estimate to estimate $S_A(t)$ and $S_B(t)$, and construct a test statistic

$$T = \frac{\hat{S}_A(t) - \hat{S}_B(t)}{\widehat{SD}[\hat{S}_A(t) - \hat{S}_B(t)]} \sim N(0,1).$$

Here $SD[\hat{S}_A(t) - \hat{S}_B(t)]$ can be estimated by Greenwood's formula,

$$\begin{aligned} \mathrm{Var}[\hat{S}_A(t) - \hat{S}_B(t)] &= \mathrm{Var}(\hat{S}_A(t)) + \mathrm{Var}\hat{S}_B(t)) \\ \widehat{SD}[\hat{S}_A(t) - \hat{S}_B(t)] &= \sqrt{\widehat{\mathrm{Var}}(\hat{S}_A(t)) + \widehat{\mathrm{Var}}(\hat{S}_B(t))}, \end{aligned}$$

where $\widehat{\mathrm{Var}}$ is derived by by Greenwood's formula.

Disadvantage of test: This test only tests the survival difference at a specified time, t. It does not test the "overall" difference of two survival functions. See Pepe and Fleming for alternative approaches (1989 *Biometrics*). Is it possible to propose "global" nonparametric tests for assessing difference in survival?

4.3 Log-rank Test for Right Censored Data

Ideas: 1. Create a 2×2 table at each uncensored failure time
2. The construction of each 2×2 table is based on the corresponding risk set.
3. Combine information from tables

The nully hypothesis is

$$H_0 : \lambda_A(t) = \lambda_B(t)(or, S_A(t) = S_B(t)) \quad \text{for all } t$$

<u>Note:</u> Where "for all t" might be replaced by "for observed t".

The general concept to construct a test statistic at an uncensored time y is the following: At an uncensored time $y(y = y_{(i)}$ for some i),

	D	$\bar{D}$	
Treatment A	d	$n_A - d$	n_A
Treatment B	$m_D - d$	$n_B - (m_D - d)$	n_B
	m_D	$m_{\bar{D}}$	N

N: # individuals in the risk set at y from pooled data
d: # failures at y from group A
m_D# failures at y from pooled data
n_A: # individuals in the risk set at y from group A
n_B: # individuals in the risk set at y from group B
$m_{\bar{D}} = N - m_D$

Use the following method to construct the test statistic: conditional on $n_A, n_B, m_D, m_{\bar{D}}$, the random number d follows a hypergeometric distribution (under H_0) with probability

$$\frac{\begin{pmatrix} n_A \\ d \end{pmatrix}\begin{pmatrix} n_B \\ m_D - d \end{pmatrix}}{\begin{pmatrix} N \\ m_D \end{pmatrix}} \qquad \max(0, m_D - n_B) \leq d \leq \min(n_A, m_d).$$

Under H_0,

$$\begin{aligned} \mathrm{E}_0(D) &= m_D \left(\frac{n_A}{N}\right) \\ \mathrm{Var}_0(D) &= \frac{n_A n_B m_D m_{\bar{D}}}{N^2(N-1)} \end{aligned}$$

$$\boxed{Z = \frac{\sum_{i=1}^{k}(D_{(i)} - \mathrm{E}_0[D_{(i)}])}{\sqrt{\sum_{i=1}^{k} \mathrm{Var}_0(D_{(i)})}} \underset{n \text{ large}}{\sim} N(0,1)}$$

For the calculation at $Z = z$,

$$z = \frac{\sum_{i=1}^{k}\left(d_{(i)} - \frac{m_{D_{(i)}} \cdot n_{A_{(i)}}}{N_{(i)}}\right)}{\sqrt{\sum_{i=1}^{K} \frac{n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}^2 (N_{(i)} - 1)}}}$$

when do we reject H_0?

The null hypothesis is $H_0 : \lambda_A(t) = \lambda_B(t)$ for all t. Consider three different kinds of alternatives:

(A1)	$H_1 : \lambda_A \neq \lambda_B$	(no prior knowledge)
(A2)	$H_1 : \lambda_A < \lambda_B$	(treatment A is better)
(A3)	$H_1 : \lambda_A > \lambda_B$	(treatment B is better)

Usually the significance level of a test is set up to be 0.05.

For (A1), use

$$Z^2 = \left[\frac{\sum_1^k (D_{(i)} - \mathrm{E}_0[D_{(i)}])}{\sqrt{\sum_1^k \mathrm{Var}_0(D_{(i)})}}\right]^2 \underset{n \text{ large}}{\sim} \chi^2(1)$$

Reject H_0 when $z^2 > 3.84$ ($|z| > 1.96$)

p-value = Probability for values larger than z^2.

For (A2),

When H_1 is true, Z is likely to be negative, so reject H_0 when z is small, that is, $z < -1.645$.

P-value $=$ Probability for values smaller than z.

For (A3)

When H_1 is true, Z is likely to be positive, so reject H_0 when z is large, that is, $z > 1.645$

P-value $=$ Probability for values larger than z.

Example	Group A	3,	5,	7,	9^+,	18		
	Group B	12,	19,	20,	20^+,	33^+		
	Uncesored:	3,	5,	7,	12,	18,	19,	20
	$H_0 : \lambda_A(t) = \lambda_B(t)$							

$y_{(1)} = 3$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\
\hline
A & 1 & 4 & 5 \\
\hline
B & 0 & 5 & 5 \\
\hline
 & 1 & 9 & 10
\end{array}$$

$y_{(2)} = 5$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\
\hline
A & 1 & 3 & 4 \\
\hline
B & 0 & 5 & 5 \\
\hline
 & 1 & 8 & 9
\end{array}$$

$y_{(3)} = 7$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\
\hline
A & 1 & 2 & 3 \\
\hline
B & 0 & 5 & 5 \\
\hline
 & 1 & 7 & 8
\end{array}$$

$y_{(4)} = 12$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\ \hline
\mathrm{A} & 0 & 1 & 1 \\ \hline
\mathrm{B} & 1 & 4 & 5 \\ \hline
 & 1 & 5 & 6
\end{array}$$

$y_{(5)} =18$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\ \hline
\mathrm{A} & 1 & 0 & 1 \\ \hline
\mathrm{B} & 0 & 4 & 4 \\ \hline
 & 1 & 4 & 5
\end{array}$$

$y_{(6)} = 19$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\ \hline
\mathrm{A} & 0 & 0 & 0 \\ \hline
\mathrm{B} & 1 & 3 & 4 \\ \hline
 & 1 & 3 & 4
\end{array}$$

$y_{(7)} = 20$

$$\begin{array}{c|c|c|c}
 & D & \bar{D} & \\ \hline
\mathrm{A} & 0 & 0 & 0 \\ \hline
\mathrm{B} & 1 & 2 & 3 \\ \hline
 & 1 & 2 & 3
\end{array}$$

$y_{(i)}$	$d_{(i)}$	$\mathrm{E}_0[d_{(i)}]$	$\mathrm{Var}_0[d_{(i)}]$
3	1	$1 \times \frac{5}{10} = 0.5$	$\frac{5\times5\times1\times9}{10^2.9} = 0.25$
5	1	$1 \times \frac{4}{9} = 0.44$	$\frac{4\times5\times1\times8}{9^2.8} = 0.2469$
7	1	$1 \times \frac{3}{8} = 0.38$	0.2344
12	0	$1 \times \frac{1}{6} = 0.17$	0.1389
18	1	$1 \times \frac{1}{5} = 0.20$	0.1600
19	0	$1 \times \frac{0}{4} = 0$	0
20	0	$1 \times \frac{0}{3} = 0$	0

$$\sum_1^7 (d_{(i)} - \mathrm{E}_0(d_{(i)})) \quad = \quad (1 - 0.5) + \ldots + (0 - 0) = 2.31$$

$$\begin{aligned}
\sum_1^7 \mathrm{Var}_0(d_{(i)}) &= 0.25 + \ldots + 0 = 1.030 \\
z &= \frac{2.31}{\sqrt{1.030}} = 2.28
\end{aligned}$$

$$\begin{aligned}
\text{Now if} \quad & H_1 : \lambda_A \neq \lambda_B \ \ \text{(two-sided)} \\
& z^2 = (2.28)^2 = 5.198 > 3.84 \\
& p\text{-value} = 0.0226 \Rightarrow \ \text{reject } H_0. \\
\text{if} \quad & H_1 : \lambda_A > \lambda_B \ \ \text{(one-sided)} \\
& z = 2.28 > 1.645 \\
& p\text{-value} = 0.0113 \Rightarrow \ \text{reject } H_0.
\end{aligned}$$

Warning: Sample size might be too small for the validity of χ^2 approximation!

4.4 Generalization of Log-Rank Test

After constructing a sequence of 2×2 tables at uncensored times, we consider the statistic

$$T = \sum_{\substack{\text{uncensored} \\ (i)}} w_{(i)}(d_{(i)} - \mathrm{E}_0[d_{(i)}])$$

where $w_{(i)}$ is the "weight" on the table at $y_{(i)}$. The variance of T is

$$\sum_{(i)} w_{(i)}^2 \mathrm{Var}(d_{(i)}).$$

Define

$$\begin{aligned}
z &= \frac{\sum_{(i)} w_{(i)}(d_{(i)} - \mathrm{E}_0(d_{(i)}))}{\sqrt{\sum_{(i)} w_{(i)}^2 \mathrm{Var}_0(d_{(i)})}} \\
&= \frac{\sum_{(i)} w_{(i)} \left(d_{(i)} - \frac{m_{D_{(i)}} n_{A_{(i)}}}{N_{(i)}} \right)}{\sqrt{\sum_{(i)} \frac{w_{(i)}^2 n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}(N_{(i)}-1)}}} \quad \underset{n \text{ large}}{\overset{\text{approx}}{\sim}} \quad N(0,1)
\end{aligned}$$

Three cases of interest:

(i) $w_{(i)} = 1$ for all (i), $T =$ log-rank test

(ii) $w_{(i)} = N_{(i)}$, $T =$ Gehan's test (1965, Biometrika)

(iii) $w_{(i)} = \sqrt{N_{(i)}}$, $T =$ Tarone and Ware test

The tests of (ii) and (iii) are motivated by examining the risk set size and giving weights to tables according to the risk set sizes. In general, the log-rank test is more efficient under the proportional hazards model, and (ii) and (iii) are more efficient under other classes of models.

Reference Tarone and Ware, *Biometrika*, (1977).

For example, if the underlying model is the PHM

$$\lambda_B(t) = \lambda_A(t)e^{\beta}$$

$$\begin{cases} H_0 : \beta = 0(\lambda_A(t) = \lambda_B(t)) \\ H_1 = \beta \neq 0 \\ \text{or} \\ H_1 = \beta > 0 \\ \text{or} \\ H_1 = \beta < 0 \end{cases}$$

The log-rank test is the most powerful test. Another example, if the relative hazard is large at earlier times, then Gehan's test might be more powerful than (i). When cross-over in hazards occurs, the weighted or unweighted log-rank tests would not be good choices in general.

Gehan's test is closely related to the Wilcoxon test. It can be regarded as a generalization of the Wilcoxon test.

4.5 Wilcoxon Test for Complete Data

$$\text{Data from} \quad \begin{array}{l} \text{treatment A} = t_1, \ldots, t_m \sim S_A \\ \text{treatment B} = z_1, \ldots, z_n \sim S_B \end{array}$$

Here $t_1, \ldots, t_m, z_1, \ldots, z_n$ are failure times (uncensored). $H_0: \ S_A = S_B$.

The general idea is the following. Pool the data from treatments A and B. Rank the data. Calculate the sum of ranks from treatment-A data. If the rank-sum is large or small, then reject the null hypothesis.

Example $A: 3, \ 7, \ 2 \qquad m = 3$
$B: 1, \ 4, \qquad n = 2$

Ordered data (1, 2, 3, 4, 7)
Ranks for 3, 7, 2 are (3, 5, 2)
Rank sum is $3 + 5 + 2 = 10$. Is "10" large or small? We will discuss it.

Order the pooled data and define

$$\begin{aligned} \gamma_i &= \text{rank of } t_i, \ t = 1, \ldots, m \\ R &= \sum_{i=1}^{m} \gamma_i \end{aligned}$$

Under $H_0: \ S_A = S_B$,

$$\begin{aligned} & \qquad\qquad \text{1st rank} \qquad \text{last rank} \\ & \qquad\qquad\quad \swarrow \qquad\qquad \swarrow \\ \mathrm{E}_0[R] &= m\left(\frac{1 + (m+n)}{2}\right) \\ \mathrm{Var}_0(R) &= \frac{mn(m+n+1)}{12} \text{ from permutation theory} \end{aligned}$$

Testing statistics is

$$W = \frac{R - E_0(R)}{\sqrt{\mathrm{Var}_0(R)}}$$

When m, n are small $\Rightarrow$ Use small sample tables. Reject H_0 when W is far away from 0.

When m, n are large, use approximation result

$$W = \frac{R - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \overset{\text{approx}}{\sim} N(0,1)$$

Reject H_0 when W is very different from 0 (that is, R is very large or small).

To use the Wilcoxon test, the usual underlying models we have in mind are likely to be

- location-shift model

$$f_A(t) = f_B(t - \theta)$$

- Stochastic ordering model

$$S_A(t) \geq S_B(t) \quad \text{or} \quad S_A(t) \leq S_B(t)$$

- Proportional hazards model

$$\lambda_B(t) = \lambda_A(t)e^{\beta}$$

4.6 Extension of Wilcoxon Test: Gehan's Test for Right Censored Data

For complete and continuous data, an alternative way to write the rank sum is

$$R = \frac{m(m+n+1)}{2} \quad + \frac{1}{2}U \qquad (*)$$

and U is defined as

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} U_{ij} \text{ where } U_{ij} = \begin{cases} 1 & \text{if} \quad t_i > z_j \\ 0 & \text{if} \quad t_i = z_j \\ -1 & \text{if} \quad t_i < z_j \end{cases}$$

The statistic "U" is also called the Mann-Whitney statistic. Reject H_0 if U is away from 0. Gehan (*Biometrika*, 1965) modified U_{ij} subject to right censored data.

To see the validity of (*), consider the condition when we have the total separation

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)} < z_{(1)} < \ldots < z_{(n)},$$

then $R = \frac{m(m+1)}{2}$. For every interchange of a consecutive (t, z) pair, R is increased by 1, and the number of interchanges is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{2}\,[U_{ij}+1].$$

Thus

$$\begin{aligned} R &= \frac{m(m+1)}{2} + \sum_i \sum_j \frac{1}{2}\,[U_{ij}+1] \\ &= \frac{m(m+1)}{2} + \frac{m \cdot n}{2} + \frac{1}{2} \sum_i \sum_j U_{ij} \\ &= \frac{m(m+n+1)}{2} + \frac{1}{2}\,U. \end{aligned}$$

Now the data are

<u>A-sample</u> $(y_1, \delta_1), \ldots, (y_m, \delta_m)$
<u>B-sample</u> $(y_1^*, \delta_1^*), \ldots, (y_n^*, \delta_n^*)$ δ_i, δ_j^* = censoring indicator.

Define

$$U_{ij} = \begin{cases} 1 & \text{if} \quad t_i > z_j \\ 0 & \text{either} \quad \text{``}t_i = z_j\text{'' or ``don't know''} \\ -1 & \text{if} \quad t_i < z_j \end{cases}$$

Note: t_i and z_j may not be observable!

The Gehan statistic is

$$G = \sum_i \sum_j U_{ij} \overset{\text{approx}}{\sim} N(0, \sigma^2) \;\; \text{Reject } H_0 \text{ if } G \text{ is large or small}$$

<u>Example</u> $A =$ $3, 5, 7, 9^+, 18$
$B =$ $12, 19, 20, 20^+, 33^+$

$$G = \sum_i \sum_j U_{ij}$$

$$i = 1, \quad \sum_{j=1}^{5} U_{1j} = (-1) + (-1) + (-1) + (-1) + (-1) =$$

$$i = 2, \quad \sum_{j=1}^{5} U_{2j} = -5$$

$$i = 3, \quad \sum_{j=1}^{5} U_{3j} = -5$$

$$i = 4, \quad \sum_{j=1}^{5} u_{4j} = 0$$

$$i = 5, \quad \sum_{j=1}^{5} U_{ij} = 1 + (-1) + (-1) + (-1) + (-1) = -3$$

The Gehan statistic is

$$G = -5 \ -5 \ -5 \ +0 \ -3 = -18.$$

To get p-value, we need to estimate σ^2. Gehan provided a complicated formula (*Biometrika*, 1965). For your calculation, just use the "weighted" formula (ii) introduced earlier. Because

$$\begin{aligned} G &= -\sum_{(i)} N_{(i)} \left[d_{(i)} - \mathrm{E}_0\left(d_{(i)}\right) \right] \\ &= -\sum_{(i)} N_{(i)} \left[d_{(i)} - \frac{m_{D_{(i)}} n_{A_{(i)}}}{N_{(i)}} \right] , \end{aligned}$$

we may derive the variance of the Gehan statistic by the previous formula. To see the equivalence, note that

$$\begin{aligned} G &= \sum_{y_i \text{ censored}} \sum_{j \in R_i} U_{ij} \\ &\quad + \sum_{y_{(i)}} \sum_{j \in R_{(i)}} U_{ij} \end{aligned}$$

$$= \quad I + II$$

Clearly, $I = 0$. For II, if the failure at $y_{(i)}$ is from group "A", then the score is

$$-\left\{(N_{(i)} - n_{A_{(i)}}) - \underbrace{(m_{D_{(i)}} - d_{(i)})}_{\searrow\ \#\text{ of failure at } y_{(i)} \text{ from ``}B\text{''}}\right\}$$

and $n_{A_{(i)}} - d_{(i)}$ otherwise. Thus the total score evaluated $y_{(i)}$ is

$$-\left[d_{(i)}\left(N_{(i)} - n_{A_{(i)}} - m_{D_{(i)}} + \ d_{(i)}\right) \ - \ \left(m_{D_{(i)}} - d_{(i)}\right)\left(n_{A_{(i)}} - d_{(i)}\right)\right]$$

$$= \quad -\left[d_{(i)}N_{(i)} - m_{D_{(i)}}n_{A_{(i)}}\right].$$

Thus

$$\begin{aligned} G &= -\sum_{y_{(i)}}\left[d_{(i)}N_{(i)} - m_{D_{(i)}}n_{A_{(i)}}\right] \\ &= -\sum_{(i)} N_{(i)}\left[d_{(i)} - \frac{m_{D_{(i)}}n_{A_{(i)}}}{N_{(i)}}\right], \end{aligned}$$

and

$$\frac{G}{\sqrt{\sum_{(i)} \frac{N_{(i)}^2 n_{A_{(i)}} n_{B_{(i)}} m_{D_{(i)}} m_{\bar{D}_{(i)}}}{N_{(i)}^2 (N_{(i)}-1)}}} \quad \overset{\text{approx}}{\underset{n \text{ large}}{\sim}} \quad N(0,1)$$

5 Truncation Models

Statistical techniques for truncated data have been integrated into survival analysis in last two decades. Truncation is a sampling mechanism for observing incomplete data where a random variable is observable only if it falls in a certain region (untruncated region). When the random variable of interest falls outside the region, the information about the variable is lost and therefore excluded from the data set. Truncated survival data typically arise in observational studies.

5.1 Left-Truncation and Length-Biased Sampling

When studying the natural history of a disease, an incident cohort is defined as a group of subjects whose initial events are randomly sampled from a pre-determined calendar time interval. The subjects are followed for detecting the occurrence of the failure event until loss to follow-up or end-of-study. The data collected from an incident cohort are the typical right-censored data. The observed data include observations (y, δ)s, where $y = \min(t, c)$, $\delta = I(t \leq c)$, t and c are the failure and censoring times.

When the failure times are long, the incident cohort design is inefficient for natural history studies because it usually takes a long follow-up time to observe enough failure events. In contrast, a prevalent sampling design which draws samples from a disease prevalent population is more focused and thus more practical in real studies. The prevalent sample is formed by subjects whose initial events had occurred but have not experienced the failure event at the time of recruitment, τ. The prevalent sampling can be described by one of the following two models:

I. Define T as the time from the disease incidence to the failure event for subjects who became diseased in a calendar time interval $[a, b)$, where $a < 0$. The variable W is the time from the disease incidence to the (potential) recruitment time. The variable W is called left truncation time. Under the left truncation sampling, the probability density of the *observed* (w, t) is the population probability density of (w, t) given $T \geq W$:

$$p_s(w, t) = p(w, t | T \geq W) \ .$$

Without further complication of censoring, the observations include (w, t)s, where

Let g and f respectively be the marginal density function of W and T. Assume the time to failure, T, is independent of when the initiating event occurs, then it implies T and W are independent of each other, forming the *non-informative truncation model.*

II. Assume the initial events occur over the calendar time as a nonstationary Poisson process with intensity $\lambda(u)$, $u \in [0, \tau]$, and the distribution of T is independent of u, when the initial event occurs. Define the pdf $\lambda_0(u) = \lambda(u)/\int_0^\tau \lambda(v)dv$ as the normalized $\lambda(t)$ in $[0, \tau]$. Conditioning on the number of initial events occurring in $[0, \tau]$, the event times u's are order statistics of iid random variables with pdf g. Pick an event time U randomly from U's and define $W = \tau - U$, then the pdf of W is $g(w) = \lambda_0(\tau - w)$.

Example. Suppose a random sample of women with breast cancer (b.c.) are recruited for observation of survival. The failure time T is defined as the time from onset of b.c. to death and f is the probability density function of T. Suppose the time of recruitment, τ, is a fixed calendar time. Then, g can be interpreted as the the rate of occurrence of b.c. over time.

5.2 Left-Truncation and Length-Biased Sampling

The joint density of the observed (w, t) can then be expressed as

$$\begin{aligned} p_s(w,t) &= \frac{g(w)f(t)I(t \geq w)}{\mathrm{P}(T \geq W)} \\ &= \frac{g(w)f(t)I(t \geq w)}{\int S(u)g(u)du} \ . \end{aligned} \tag{1}$$

In the situation that g is uniformly distributed then the observed t follows the **length-biased distribution**. Length-biased sampling could arise in many epidemiological studies when survival data are collected from a disease population. In the breast cancer (b.c.) example, assume (i) the rate of occurrence of b.c. remains constant over time, and (ii) the density function of the time from b.c. to death, f, is independent of when b.c. occurred. Conditions (i) and (ii) together are referred to as the equilibrium condition. The equilibrium condition typically holds for so-called 'stable diseases'. When the equilibrium condition is satisfied, we observe length-biased failure time which has the following density function:

$$p_s(t) = \int p_s(w,t)dw = tf(t)/\mu \ , \tag{2}$$

where $\mu = \mathrm{E}[T]$ is the mean failure time. In general, treating length-biased data as the 'usual data' would lead to biased analytical results because of the bias of data. When length-biased data are encountered, we should use bias-adjusted methods for analysis; see Wang (1997, 'length-bias', Encyclop. of Biostat.) and references therein. Although statistical methods can be formulated for length-biased observations, Assumption (i) is required for validating the length-biased model as well as the corresponding methods (Vardi, 1982 Annal. Stat.; Wang, 1996, Biometrika).

Let $I(u)$ represent the disease incidence (occurrence) rate at the calendar time u and S_u the survival function of T for those patients whose disease was initiated at u. Then, the disease prevalence rate at the calendar time τ can be obtained as $P(\tau) = \int_{-\infty}^{\tau} I(u) S_u(\tau - u) du$. When the equilibrium condition is satisfied, the incidence rate is a constant ($I(u) = I_0$) and the survival function is independent of u ($S_u = S$), and

$$P(\tau) = I_0 \int_{-\infty}^{\tau} S(\tau - u) du = I_0 \int_0^{\infty} S(u) du = I_0 \times \mu$$

is independent of τ. Thus, let $P(\tau) = P_0$ and we derive

$$P_0 = I_0 \times \mu \quad \text{(Prevalence = Incidence} \times \text{duration))} \ .$$

Length-biased data can be viewed as a special case of left truncated data, since the conditional density of the observed t given w is

$$f(t) I(t \geq w) / S(w), \tag{3}$$

which corresponds to the density function of *left truncated* failure time. By viewing length-biased data as left truncated data, we next consider how to analyze left truncated data in a general setting. It is important to indicate that the validity of the truncated density in (3) depends only on Assumption (ii) and not on Assumption (i).

5.3 Left Truncated Data: Product-Limit Estimator

Suppose n individuals are recruited into a propective follow-up study by prevalent sampling. Suppose the observed data $(w_1, t_1), \ldots, (w_n, t_n)$ are independent and identically distributed observations. Let $t_{(1)} < \ldots < t_{(J)}$ be the distinct and ordered values of $t_1, \ldots, t_n$. Define

$R_{(j)} = \{i \ : \ w_i \leq t_{(j)} \leq t_i\}$

$d_{(j)} =$ Number of failures at $t_{(j)}$

$N_{(j)} =$ Number of individuals in $R_{(j)}$

$\lambda_{(j)} = f(t_{(j)}) / S(t_{(j)}^-)$

Product-limit estimator

For $t_{(i-1)} \leq t < t_{(i)}$, recall

$$S(t) \approx \frac{Pr(T \geq t_{(2)})}{Pr(T \geq t_{(1)})} \cdot \frac{Pr(T \geq t_{(3)})}{Pr(T \geq t_{(2)})} \cdots \frac{Pr(T \geq t_{(i)})}{Pr(T \geq t_{(i-1)})}.$$

Now estimate $\frac{Pr(T \geq t_{(j+1)})}{Pr(T \geq t_{(j)})}$ by $\frac{N_{(j)} - d_{(j)}}{N_{(j)}}$, $j = 1, 2, \ldots, i-1$. The product-limit estimator is thus

$$\hat{S}(t) = \left(1 - \frac{d_{(1)}}{N_{(1)}}\right) \left(1 - \frac{d_{(2)}}{N_{(2)}}\right) \cdots \left(1 - \frac{d_{(i-1)}}{N_{(i-1)}}\right)$$

$$= \boxed{\prod_{t_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{N_{(j)}}\right)}$$

Example Data: $(4,5), (0,4), (5,7), (1,2), (2,8), (1,5)$

failure times	2	4	5	7	8
$d_{(i)}$	1	1	2	1	1
$N_{(i)}$	4	4	4	2	1

$R_{(1)} = \{(0,4), (1,2), (2,8), (1,5)\}$

$R_{(2)} = \{(4,5), (0,4), (2,8), (1,5)\}$

$\cdots\cdots\cdots$

The truncation product-limit estimate is thus

$$\hat{S}(1) = 1$$

$$\hat{S}(2) = \left(1 - \frac{1}{4}\right) = \frac{3}{4}$$

$$\hat{S}(4) = \left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{4}\right) = \frac{3}{4} \cdot \frac{3}{4}$$

$$\hat{S}(5) = \left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{2}{4}\right) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{2}{4}$$

Note: Unlike right censored data, risk sets usually are NOT nested!

Example Data: $(4,5), (0,1^+), (5,7), (1,2), (2,4^+), (1,5)$

failure times	2	5	7
$d_{(i)}$	1	2	1
$N_{(i)}$	3	3	1

$R_{(1)} = \{(1,2), (2,4^+), (1,5)\}$

$R_{(2)} = \{(4,5), (5,7), (1,5)\}$

$R_{(3)} = \{(5,7)\}$

. . . .

The truncation product-limit estimate is thus

$$\begin{aligned}
\hat{S}(1) &= 1 \\
\hat{S}(2) &= \left(1 - \frac{1}{3}\right) = \frac{2}{3} \\
\hat{S}(5) &= \left(1 - \frac{1}{3}\right)\left(1 - \frac{2}{3}\right) = \frac{2}{3} \cdot \frac{1}{3} \\
\hat{S}(7) &= \left(1 - \frac{1}{3}\right)\left(1 - \frac{2}{3}\right)\left(1 - \frac{1}{1}\right) = 0
\end{aligned}$$

Note that the applicability of the product-limit estimator requires that the truncation time w_i be observable, and such a requirement might not be met in some applications.

Remarks: For left truncated and right censored data,

- **modified Greenwoods Formula** still holds for the estimation of the asymptotic variance of the product-limit estimator - just use the revised risk sets.
- **modified partial likelihood method** still holds for the estimation of β in the **proportional hazards model** - just use the revised risk sets.
- **modified log-rank tests** still hold for testing the difference between two groups - just use the revised risk sets.

Essentially, censoring and truncation share some significant similarities in statistical analysis - especially, the similarities in the 'risk set methods'. Nevertheless, regardless of the similarities, there still exist significant dissimilarities (i.e., different statistical properties) that are not emphasized in this course. References include Woodroofe (1985, Ann. Statist.), Wang et al. (1986, Ann. Statist.), Tsai et al. (1987, Biometrika), Keiding and Gill (1988, Ann. Statist.) and Wang (1989, 1991, JASA).

5.4 Right Truncation

Suppose that a certain disease can be characterized by an initial event and a failure event. An example is the study of the natural history of Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS), where the HIV-infection is the initial event and the AIDS diagnosis is the failure event. Let X denote the calendar time of the initial event and T the time from the initial event to the failure event. Then an observation (x, t) is observed only if $x + t \leq \tau$, where τ is the closing date of data collection. This is an example of right truncation: the failure time T is observed only when $T \leq \tau - X$. Let $W = \tau - X$. Then W is called the truncation time.

Product-Limit Estimator

Suppose the observed observations $\{(W_i, T_i) : T_i \leq W_i, i = 1, \ldots, n\}$ are independent and identically distributed. Let $t_{(1)} < \ldots < t_{(J)}$ be the distinct and ordered values of $t_1, \ldots, t_n$. A practical constraint in nonparametric estimation is that a nonparametric distribution estimator cannot estimate the distribution function beyond the largest observed $t_{(J)}$. Thus, what can be estimated is the conditional distribution function $F^*(t) = F(t)/F(t_{(J)})$ for $t \leq t_{(J)}$. Define

$R_{(j)} = \{i \ : \ t_i \leq t_{(j)} \leq w_i\}$

$d_{(j)} =$ Number of failures at $t_{(j)}$

$N_{(j)} =$ Number of individuals in $R_{(j)}$

$\lambda_{(j)} = f(t_{(j)})/F(t_{(j)})$

For $t \leq t_{(J)}$, the product-limit estimator is

$$\boxed{\hat{F}^*(t) = \prod_{t_{(j)} > t} \left(1 - \frac{d_{(j)}}{N_{(j)}}\right)}$$