GS01 1023: Survival Analysis

Lecture 1: Introduction to Failure Time Data

Lecturer: Ryan Sun *Scribes: Shu-Hsien Cho*

The goals of this unit are to introduce notation, discuss ways of probabilistically describing the distribution of a 'survival time' random variable, apply these to several common parametric families, and discuss how observations of survival times can be right-censored.

1.1 Failure Time Distribution

Suppose T is a non-negative random variable representing the time until some event of interest. For example, T might denote:

- the elapsed time from diagnosis of a disease until death
- the elapsed time between administration of a vaccine and development of an infection
- the elapsed time from the start of treatment of a symptomatic disease and the suppression of symptoms

1.1.1 T Continuous

We shall assume that T is continuous unless we specify otherwise. The probability density function (p.d.f.) and cumulative distribution function (c.d.f.) are most commonly used to characterize the distribution of any random variable, and we shall denote these by $f(\cdot)$ and $F(\cdot)$, respectively:

$$F(0) = P(T = 0) \qquad \begin{cases} p.d.f. : f(t) \\ c.d.f. : F(t) = P(T \leq t) \end{cases}$$

Because T is non-negative and usually denotes the elapsed time until an event, it is commonly characterized in other ways as well:

Survivor Function

$$S(t) \equiv 1 - F(t) = P(T > t) \quad \text{for} \quad t > 0$$

The survivor function simply indicates the probability that the event of interest has not yet occurred by time t; thus, if T denotes time until death, $S(t)$ denotes probability of surviving beyond time t. The right tail of the distribution is the important component for the incorporation of right censoring, so it is more convenient to concentrate on the survivor function in dealing with failure time distributions.

$F(\cdot)$ and $S(\cdot)$ are right continuous in t. For continuous survival time T, both functions are continuous in t. However, even when $F(\cdot)$ and $S(\cdot)$ are continuous, the nonparametric estimators $\hat{F}(\cdot)$ and $\hat{S}(\cdot)$ are discrete distributions. For example, $\hat{F}(\cdot)$ might be the cdf corresponding to the discrete distribution that weighted $w_1, w_2, \cdots, w_k$ at certain times $t_1, t_2, \cdots, t_k$. Thus, even though $F(\cdot)$ is continuous, its estimator $\hat{F}(\cdot)$ is (only) right continuous. (See Kaplan-Meier)

Hazard Function

$$\lambda(t) \equiv \lim_{h\to 0^+} P\left(\frac{t \le T < t+h|T \ge t}{h}\right) = \lim_{h\to 0^+} \frac{P\left(t \le T < t+h, T \ge t\right)/h}{P(T \ge t)} = \frac{f(t)}{S(t-)}$$

with $S(t^-) = \lim_{u\to t-} S(u)$. The hazard function is a conditional density, given that the event in question has not yet occurred prior to time t. Note that for continuous T, $\lambda(t) = -\frac{d}{dt}\log[1-F(t)] = -\frac{d}{dt}\log S(t)$

Cumulative Hazard Function

$$\Lambda(t) \equiv \int_0^t \lambda(u)du \quad\quad t > 0 = -\log[1-F(t)] = -\log S(t)$$

Note that

$$\begin{aligned} S(t) &= e^{-\Lambda(t)} \\ f(t) &= \lambda(t)e^{-\Lambda(t)}. \end{aligned}$$

Other representations of the failure time distribution are occasionally useful. An example is the expected residual life at time t which uniquely determines a continuous survival distribution with finite mean:

$$r(t) = E(T-t|T \ge t) = \frac{\int_t^\infty (v-t)f(v)dv}{S(t)} = \frac{\int_t^\infty S(v)dv}{S(t)} \quad\quad \text{(integral by parts)}$$

With $E(T) < \infty$ implying that $\lim_{t\to\infty} tS(t) = 0$,

$$E(T) = r(0) = \int_0^\infty S(v)dv$$

$$\implies \frac{1}{r(0)} = \frac{1}{\int_0^\infty S(v)dv} \implies \frac{1}{r(t)} = -\frac{d}{dt}\log\int_0^\infty S(v)dv \implies \int_0^t \frac{dv}{r(v)} = -\log\int_0^\infty S(v)dv + \log r(0)$$

$$\implies F(t) = \frac{r(0)}{r(t)}\exp\left[-\int_0^t \frac{du}{r(u)}\right]$$

1.1.2 T Discrete

If T is a discrete random variable taking values $a_1 < a_2 < \cdots$ with pmf $f(a_i) = P(T = a_i)$, the survivor function is $F(t) = \sum_{j|a_j>t} f(a_j)$. The hazard at a_i is defined as the conditional probability of failure at ai given that the individual has survived to a_i,

$$\lambda_i = P(T = a_i|T \ge a_i) = \frac{f(a_i)}{S(a_i^-)} \quad\quad \text{where } S(a^-) = \lim_{t\to a^-} S(t)$$

Note that fot $t \in [a_i, a_{i+1})$

$$\begin{aligned} S(t) &\equiv P(T > t) = P(T > a_i) \\ &= P(T > a_i|T > a_{i-1})P(T > a_{i-1}) \\ &= P(T > a_i|T > a_{i-1})P(T > a_{i-1}|T > a_{i-2})P(T > a_{i-2})\cdots \\ &= (1-\lambda_i)(1-\lambda_{i-1})\cdots(1-\lambda_1) = \prod_{i=1}^{j}(1-\lambda_i) \end{aligned}$$

Similarly, define $f(a_1) = \lambda_1$,

$$f(a_i) = \lambda_i \prod_{k=1}^{j-1} (1 - \lambda_k)$$

The results are quite easily deduced by considering the failure time process unfolding over time and a sequence of trials, each of which may or may not result in a failure.

1.1.3 T has Discrete and Continuous Component

$$S(t) = \exp\left[-\int_0^t \lambda_c(u)du\right] \sum_{j|a_j \le t} (1 - \lambda_j)$$

$$\Lambda(t) = \int_0^t \lambda_c(u)du + \sum_{j|a_j \le t} \lambda_j \text{ is a right-continuous nondecreasing function.}$$

$$d\Lambda(t) = \Lambda(t^- + dt) - \Lambda(t^-) = P[T \in [t, t+dt)|T \ge t] = \begin{cases} \lambda_i, & t = a_i, i = 1, 2, \cdots \\ \lambda_c(t)dt & \text{otherwise.} \end{cases}$$

The survivor function can be written as

$F(t) = \mathcal{P}_0^t[1-d\Lambda(u)]$, where the *product integral* is defined by $\mathcal{P}_0^t[1-d\Lambda(u)] = \lim \prod_{k=1}^{r} 1 - [\Lambda(u_k) - \Lambda(u_{k-1})]$

It is possible to examine survival experience by looking at the survival experience over each interval conditional upon the experience to that point. It also underlies failure time analysis by counting processes and martingales (Chapter 5), the construction of the likelihood under independent censoring (Section 6.2), the construction of partial likelihood in the Cox model (Section 4.3), and the analysis of multivariate failure times and life-history processes (Chapter 9).

With covariates x measured at the time origin of the study, we can then think of models for the corresponding hazard function

$$\lambda(t; x) = \lim_{h \to 0} \frac{P\{T \in [t, t+h)|T \ge t, x\}}{h}$$

Note 1: Note that $\lambda(t)dt = f(t)\,dt/S(t) \approx P(\text{fail in } [t, t+dt) \mid \text{survive until } t)$. Thus, the hazard function might be of more intrinsic interest than the p.d.f. to a patient who had survived a certain time period and wanted to know something about their prognosis.
Note 2: There are several reasons why it is useful to introduce the quantities $\lambda(t)$ and $\Lambda(t)$:

- Interpretability: Suppose T denotes time from surgery for breast cancer until recurrence. Then when a patient who had received surgery visits her physician, she would be more interested in conditional probabilities such as "Given that I haven't had a recurrence yet, what are my chances of having one in the next year" than in unconditional probabilities (as described by the p.d.f.).
- Analytic Simplifications: When the data are subject to right censoring, hazard function representations often lead to easier analyses. For example, imagine assembling a cohort of N men who just have turned 50 years of age and then following them for 1 year. Then if d of the men die during the year of follow-up, the ratio d/N estimates the (discrete) hazard function of T = age at death. We will see that $\Lambda(\cdot)$ has nice analytical properties.

- Modeling Simplifications: For many biomedical phenomena, T is such that $\lambda(t)$ varies rather slowly in t. Thus, $\lambda(\cdot)$ is well-suited for modeling.

Note 3: It is useful to think about real phenomena and how their hazard functions might be shaped. For example, if T denotes the age of a car when it first has a serious engine problem, then one might expect the corresponding hazard function $\lambda(t)$ to be increasing in t; that is, the conditional probability of a serious engine problem in the next month, given no problem so far, will increase with the life of the car. In contrast, if one were studying infant mortality in a region of the world where there was poor nutrition, one might expect $\lambda(t)$ to be decreasing during the first year of life. This is known to be due to selection during the first year of life. Finally, in some applications (such as when T is the lifetime of a light bulb), the hazard function will be approximately constant in t. This means that the chances of failure in the next short time interval, given that failure hasn't yet occurred, does not change with t; e.g., a 1-month old bulb has the same probability of burning out in the next week as does a 5-year old bulb. As we will see below, this "lack of aging" or "memoryless" property uniquely defines the exponential distribution, which plays a central role in survival analysis.

1.2 Time origin, (Right) Censoring, Truncation

It is important to have a clear and unambiguous definition of the time origin from which survival is measured. In some instances, time may represent age, with the time origin the birth of the individual. In other instances, the natural time origin may be the occurrence of some event, such as randomization or entry into a study or diagnosis of a particular disease. One would need to define carefully the clinical medical conditions that correspond to failure (and eligibility for the study)

A common feature of survival data is the presence of right censored observations. A right-censoring mechanism is said to be independent if the failure rates that apply to individuals on trial at each time $t > 0$ are the same as those that would have applied had there been no censoring. We briefly review settings in which right-censored data can arise and introduce notation to distinguish the underlying T from what is actually observed.

Type I Censoring Suppose that we "plug in" n bulbs at time 0, and then observe them for C time units, noting the times until burn out for those that burn out by time C. For the i^{th} bulb, let Ti = true lifetime. Note that we observe T_i if and only if $T_i < C$; otherwise, we know only that T_i exceeds c (right censored), i.e., $T_1, \cdots, T_n \sim i.i.d.F$. But in general we observe only (T_i^*, δ_i) for $i = 1, \cdots, n$ where $T_i^* = \min(T_i, C)$ and $\delta_i = I(T_i \leq C)$.

Type II Censoring Suppose instead that we "plug in" n bulbs, and then observe things until r (some predetermined #) bulbs fail. Here we end up with r uncensored lifetimes and $n - r$ censored (at time C) lifetimes, but unlike Type I censoring, here r is a constant.

Random Censoring (commonly arising in biostat) Define $C_1, \cdots, C_n$ constants and $T_1, \cdots, T_n \sim i.i.d.F$. Then suppose we observe $(T_1^*, \delta_1), \cdots, (T_n^*, \delta_n)$

Key Assumption: Censoring is noninformative

The large majority of statistical methods for failure time data assume that censoring acts 'noninformatively' of failure time. Loosely speaking, this means that being censored at time C tells us only that $T > C$. In terms of the potential censoring times C_j , noninformative censoring is achieved if C_j is independent of T_j , $j = 1, \cdots, n$; that is, if $C_j \perp T_j$. Mathematically weaker conditions have also been proposed and investigated (Kalbfleisch & Prentice, p. 212-214).

1.3 Estimation of the Survivor Function

1.3.1 Kaplan-Meier or Product Limit Estimator

We consider the nonparametric estimation of a survivor function $S(\cdot)$ based on n i.i.d. survival times that can be noninformatively right censored. The resulting estimator − commonly known as the Kaplan-Meier Estimator or the Product-Limit Estimator − is probably one of the most commonly-used estimators in medical/public health studies involving failure time data.

Suppose that $T_1, \cdots, T_n$ are i.i.d. survival times with survivor function $S(\cdot)$, with $C_1, \cdots, C_n$ the censoring times, i.i.d. and independent of the T_i, and suppose that our observations are denoted (T_i^*, δ_i). To begin, let us suppose that $F(\cdot)$ is discrete with mass points at $0 \leq v_1 < v_2 < \cdots$, and define the discrete hazard functions $\lambda_1 = P(T = v_1)$ and $\lambda_j = P(T = v_j | T > v_{j-1})$

Note that fot $t \in [a_i, a_{i+1})$

$$\begin{aligned} S(t) \equiv P(T > t) &= P(T > v_i) \\ &= P(T > v_i | T > v_{i-1}) P(T > v_{i-1}) \\ &= P(T > v_i | T > v_{i-1}) P(T > v_{i-1} | T > v_{i-2}) P(T > v_{i-2}) \cdots \\ &= (1-\lambda_i)(1-\lambda_{i-1}) \cdots (1-\lambda_1) = \prod_{i=1}^{j} (1-\lambda_i) \end{aligned}$$

Similarly, define $f(v_1) = \lambda_1$,

$$f(v_i) = \lambda_i \prod_{k=1}^{j-1} (1-\lambda_k)$$

Now consider making an inference about F based on the likelihood function corresponding to (T_i^*, δ_i) for $i = 1, \cdots, n$. This is just

$$\begin{aligned} L(F) &= \prod_{T_i^* : \delta_i = 1} f(T_i^*) \cdot \prod_{T_i^* : \delta_i = 0} [1 - F(T_i^*)] \\ &= \prod_j \lambda_j^{d_j} (1-\lambda_j)^{Y(v_j) - d_j} \end{aligned}$$

where $0 \leq h_j \leq 1$ and

$$d_j = \sum_{i=1}^{n} \delta_i \cdot I(T_i^* = v_j) = \# \text{ who fail at } v_j$$

and

$$Y(v_j) = \sum_{i=1}^{n} I(T_i^* \geq v_j) = \# \text{ ``at risk'' at } v_j$$

The maximizing solution is seen to be $\hat{h}_j = \dfrac{d_j}{Y(v_j)}$ (for $Y(v_j) > 0$) so that

$$\hat{S}(t) = \begin{cases} 1 & t < v_1 \\ \prod_{i=1}^{j} (1 - \hat{\lambda}_i) & v_j \leq t < v_{j+1} \end{cases}$$

Notice that the expression for $\hat{\lambda}_j$ makes sense: the probability of dying at v_j given you are alive before is estimated by $\dfrac{d_j}{Y(v_j)}$. Also the expression for $\hat{S}(t)$ makes sense: the probability of staying alive at v_j if alive before v_j is estimated by $1 - \dfrac{d_j}{Y(v_j)}$.

What if we didn't know, in advance, the times at which F had mass and did not necessarily want to assume that it had to be discrete? The likelihood function is just as before; i.e.,

$$L = L(F) = \prod_{i=1}^{n} \left\{ f(t_i^*)^{\delta_i} \left[1 - F(t_i^*)\right]^{1-\delta_i} \right\}$$

However, we now need to find the maximizing solution for $F \in \mathcal{F} = \{\text{all cdf}\}$

Kaplan and Meier argue that the maximizing solution must be a discrete distribution with mass on the observed times T_i^* only. The same algebra as above leads to the same form of solution as above. Notice that this means that the Kaplan-Meier estimator actually puts mass only on the observed failure times. That is, the Kaplan-Meier (or Product-Limit) estimator of $F(\cdot)$ is

$$\hat{S}(t) = \begin{cases} 1 & t < v_1 \\ \prod_{i=1}^{j} (1 - \dfrac{d_i}{Y(v_i)}) & v_j \le t < v_{j+1} \end{cases}$$

where $v_1 < v_2 < \cdots$ are distinct failure (uncensored) times. Thus, we can view $\hat{S}(\cdot)$ as a nonparametric MLE of $F(\cdot)$; this is sometimes denoted NPMLE. One equivalent representation of $\hat{S}(t)$ is given by:

$$\hat{S}(t) = \prod_{j: v_j \le t} \left(\frac{Y(v_j) - d_j}{Y(v_j)} \right), \text{ for } t \le \max(v_i)$$

It is instructive to think about how the Kaplan-Meier estimator places mass at the observed failure times. One way of gaining insight into this is by a construction of $\hat{S}(t)$ due to Efron (1967). This is known as the 'Redistribution of Mass' algorithm (for another algorithm, see Dinse 1985).

Step 1 Arrange data in increasing order, with censored observations to the right of uncensored observations in the case of ties.

Step 2 Put mass $\dfrac{1}{n}$ at each observation.

Step 3 Start from the smallest observation and move 'right'. Each time a censored observation is reached, redistribute its mass evenly to all observations to the right.

Step 4 Repeat Step 3 until all censored observations (except largest observations) have no mass. If largest (v_g) is censored, regard this mass as $> v_g$.

Since $\hat{S}(\cdot)$ is a nonparametric estimator of $S(\cdot)$, it follows that a nonparametric estimator of $H(\cdot) = -\log[\hat{S}(\cdot)]$ is given by $\hat{H}(t) = -\log[\hat{S}(t)] = -\sum_{i=1}^{j} \log(1 - \hat{h}_i)$ for $v_j \le < v_{j+1}$. However, for small x, $\log(1-x) \sim -x$ and thus this sum is approximately $\sum_{i=1}^{j} \hat{h}_i$. This suggests the alternative estimator $\hat{H}(t) = \sum_{i=1}^{j} \hat{h}_i$ (for $j \ge 1$). This estimator is called the Nelson-Aalen estimator of $H(\cdot)$.

Next consider how we might approximate the distribution of $\hat{S}(t)$. One approach is to use the large-sample properties of maximum likelihood estimators, assuming that such results apply in this setting (the usual regularity conditions do not hold here since the space we are maximizing over is not a finite-dimensional parameter space).

$$L = L(\lambda_1, \lambda_2, \cdots) = \prod_j \lambda_j^{d_j}(1-\lambda_j)^{Y(v_j)-d_j}$$

and hence

$$\begin{aligned} -\frac{\partial^2 \log L}{\partial \lambda_j \partial \lambda_k} &= 0 \qquad j \neq k \\ -\frac{\partial^2 \log L}{\partial \lambda_j^2} &= \frac{Y(v_j)}{\hat{\lambda}_j(1-\hat{\lambda}_j)} \end{aligned}$$

Because $\hat{\lambda}_1, \hat{\lambda}_2, \cdots$ are approximately uncorrelated, with approximate means $\lambda_1, \lambda_2, \cdots$ and

$$Var(\hat{\lambda}_j) \approx \frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{Y(v_j)} = \frac{d_j(Y(v_j)-d_j)}{Y(v_j)^3}$$

Since $\hat{S}(t)$ is a product of terms of the form $\prod_{i=1}^{j}(1-\hat{\lambda}_i)$, it follows that $\hat{S}(t)$ is approximately unbiased in the discrete time setting.

The variance of Kaplan-Meier estimator: For $v_j \leq t < v_{j+1}$

$$\begin{aligned} Var[\log \hat{S}(t)] &\approx \sum_{i=1}^{j} Var\{\log(1-\hat{\lambda}_i)\} \\ &\overset{\text{delta method}}{=} \sum_{i=1}^{j} Var(\hat{\lambda}_i) \cdot \frac{1}{(1-\hat{\lambda}_i)^2} \\ &= \sum_{i=1}^{j} \frac{d_i}{Y(v_j)(Y(v_j)-d_j)} \end{aligned}$$

Greenwood's Formula: For $v_j \leq t < v_{j+1}$

$$\begin{aligned} Var[\hat{S}(t)] &\approx Var[\log \hat{S}(t)] \left[\exp(\log \hat{S}(t))\right]^2 \\ &= \hat{S}(t)^2 Var[\log \hat{S}(t)] \\ &\approx \hat{S}(t)^2 \sum_{i=1}^{j} \frac{d_i}{Y(v_j)(Y(v_j)-d_j)} \end{aligned}$$

One use of Greenwood's formula is to get an approximate confidence interval (e.g., a 95% CI) for $S(t)$. One obvious choice is $\hat{S}(t) \pm 1.96\sqrt{Var[\hat{S}(t)]}$. However, this could give limits that are greater than 1 or less than 0. One alternative is to note that $\log[-\log \hat{S}(t)]$ can take values in $(-\infty, \infty)$. Thus, using the delta method, we can approximate the variance of $\log[-\log \hat{S}(t)]$ from $Var[\hat{S}(t)]$, resulting in

$$Var\left\{\log\left[-\log \hat{S}(t)\right]\right\} \approx \frac{\sum_{i=1}^{j} \frac{d_i}{Y(v_j)(Y(v_j)-d_j)}}{\left[\log \hat{S}(t)\right]^2}$$

Given an approximate 95% CI for $\log\left[-\log\hat{S}(t)\right]$ we can re-express to get the corresponding CI for $S(t)$.

Breslow and Crowley (1974) show that as $n\to\infty$,

$$\sqrt{n}\left[\hat{S}(\cdot)-S(\cdot)\right]\overset{?}{\to}\text{zero mean Gaussian process}$$

1.3.2 Life-Table and Related Estimates

A life table is a summary of the survival data grouped into convenient intervals. In some applications (e.g., actuarial), the data are often collected in such a grouped form. In other cases, the data might be grouped to get a simpler and more easily understood presentation.

The data are grouped into intervals $I_1,\cdots,I_k$ such that $I_j=(b_0+\cdots+b_{j-1},b_0+\cdots+b_j)$ is of width b_j with $b_0=0$. The life table then presents the number of failures and censored survival times falling in each interval.

Suppose that m_j censored times and d_j failure times fall in the interval I_j, and let $n_j=\sum\limits_{l\geq j}(d_l+m_l)$ be the number of individuals at risk at the start of the j^{th} interval. The standard life-table estimator of the conditional probability of failure in I_j given survival to enter I_j, is $\hat{q}_j=1$ if $n_j=0$ and $\hat{q}_j=\dfrac{d_j}{n_j-\frac{m_j}{2}}$ otherwise. The $\dfrac{m_j}{2}$ term in the denominator is used in an attempt to adjust for the fact that not all the n_j individuals are at risk for the whole of I_j. The corresponding life-table estimator of the survivor function at the end I_j is

$$\tilde{F}(b_1+\cdots+b_j)=\prod_{l=1}^{j}(1-\hat{q}_l)$$

Using Greenwood's formula, with n_j replaced by $n_j-\dfrac{m_j}{2}$ provides an estimator of the variance of $\tilde{F}$.

(needed to be updated)

1.4 Comparison of Survival Curves: Log-rank test

The log-rank test is the most commonly-used statistical test for comparing the survival distributions of two or more groups (such as different treatment groups in a clinical trial). Assume that we have 2 groups of individuals, say group 0 and group 1. In group j, there are N_j i.i.d. underlying survival times with common cdf denoted $F_j(\cdot)$, for $j=0,1$. The corresponding hazard and survival functions for group j are denoted $\lambda_j(\cdot)$ and $S_j(\cdot)$, respectively. As usual, we assume that the observations are subject to noninformative right censoring.

We want a nonparametric test of $H_0: F_0(\cdot)=F_1(\cdot)$; or equivalently, of $H_0: S_0(\cdot)=S_1(\cdot)$ or $H_0:\lambda_0(\cdot)=\lambda_1(\cdot)$. If we knew F_0 and F_1 were in the same parametric family, then H_0 is expressible as a point/region in a Euclidean parameter space. However, we instead want a nonparametric test; that is, a test whose validity does not depend on any parametric assumptions. It is intuitively clear that a UMP (Uniformly Most Powerful) test cannot exist. Two options in this case are to select a directional test or an omnibus test.

	observed to fail at τ_j		at risk at τ_j
group 0	d_{0j}	$Y_0(\tau_j) - d_{0j}$	$Y_0(\tau_j)$
group 1	d_{1j}	$Y_1(\tau_j) - d_{1j}$	$Y_1(\tau_j)$
	d_j	$Y(\tau_j) - d_j$	$Y(\tau_j)$

1.4.1 Log-rank test

Early work (1960s) in this area fell along 2 lines: (a) Modify rank tests to allow censoring (Gehan, 1965). (b) Adapt methods used for analyzing 2 × 2 contingency tables to accommodate censoring (Mantel, 1966). We introduce the log-rank test from the latter perspective as it easily includes tests developed from the former and provides good insight into the properties of the log-rank test.

Log-rank test construction: Denote the distinct times of observed failures as $\tau_1 < \cdots < \tau_k$ and define

$$\begin{aligned}
Y_i(\tau_j) &= \#\text{ persons in group } i \text{ who are at risk at } \tau_j \ (i = 0, 1 j = 1, \cdots, k) \\
Y(\tau_j) &= Y_0(\tau_j) + Y_1(\tau_j) = \#\text{ at risk at } \tau_j \text{ (both groups)} \\
d_{ij} &= \#\text{ in group } i \text{ who fail (uncensored) at } \tau_j \ (i = 0, 1 j = 1, \cdots, k) \\
d_j &= d_{0j} + d_{1j} = \text{total } \#\text{ failures at } \tau_j
\end{aligned}$$

The information at time τ_j can be summarized in the following 2 × 2 table:

Note: $\dfrac{d_{0j}}{Y_0(\tau_j)}$ can be viewed as an estimator of $\lambda_0(\tau_j)$. Suppose $H_0 : F_0(\cdot) = F_1(\cdot)$ holds. Conditional on the 4 marginal totals, a single element (say d_{1j}) defines the table. Furthermore, with this conditioning and assuming H_0, d_{1j} has the hypergeometric distribution; that is:

$$P[d_{ij} = d] = \frac{\binom{d_j}{d}\binom{Y(\tau_j)-d_j}{Y_1(\tau_j)-d}}{\binom{Y(\tau_j)}{Y_1(\tau_j)}} \text{ for } d = \max(0, d_j - Y_0(\tau_j)), \cdots, \min(d_j, Y_1(\tau_j)).$$

The mean and variance of d_{1j} under H_0 are thus

$$\begin{aligned}
E_j &= \left(\frac{Y_1(\tau_j)}{Y(\tau_j)}\right) d_j \\
V_j &= \frac{Y(\tau_j) - Y_1(\tau_j)}{Y(\tau_j) - 1} \cdot Y_1(\tau_j) \left(\frac{d_j}{Y(\tau_j)}\right)\left(1 - \frac{d_j}{Y(\tau_j)}\right) = \frac{Y_0(\tau_j) Y_1(\tau_j) d_j (Y(\tau_j) - d_j)}{Y(\tau_j)^2 (Y(\tau_j) - 1)}
\end{aligned}$$

Define $O_j = d_{1j}$. Fisher's test would tell us to consider extreme values of d_{1j} as evidence against H_0. Thus, define

$$O = \sum_{j=1}^{k} O_j = \text{total } \#\text{ failures in group } 1E \qquad = \sum_{j=1}^{k} E_j V = \sum_{j=1}^{k} V_j$$

and let

$$Z = \frac{O - E}{\sqrt{V}} = \frac{\sum_j (O_j - E_j)}{\sqrt{\sum_j V_j}}$$

Then under H_0, it is argued that $Z \overset{approx}{\sim} N(0,1)$ or $Z^2 \overset{approx}{\sim} \chi_1^2$

Question

- While E_j may be a conditional expectation for each j, it is not clear that E has such an interpretation. Also, the creation of Z and its approximation as a $N(0,1)$ r.v. suggests that the contributions from each τ_j are independent. Is this accurate? Then, is $Z \overset{\mathcal{L}}{\to} N(0,1)$ under H_0?
- Note the similarity of the log-rank test to techniques for combining 2×2 tables across strata (e.g., cities).
- Note that the sequences $Y_0(\tau_1)$; $Y_0(\tau_2)$; $Y_0(\tau_3),\cdots$ and $Y_1(\tau_1)$; $Y_1(\tau_2)$; $Y_1(\tau_3),\cdots$ are nonincreasing, and as soon as one reaches 0 [e.g., $Y_0(\tau_5) = 0$ at $\tau_5 = 18.7$], it must follow that $O_j = E_j$ and $V_j = 0$ at and beyond this time. Thus, we would get the same answer (i.e., Z) if the construction stopped at the last time when both $Y_0(\tau_j)$ and $Y_1(\tau_j)$ are > 0.
- Although it is not obvious from the construction, the log-rank test is a directional test oriented towards alternatives where $S_1(t) = [S_0(t)]^\theta$, or equivalently, when $\dfrac{\lambda_1(t)}{\lambda_0(t)} = \theta$. We will see later that the log-rank statistic arises as a score test from a partial likelihood function for Cox's proportional hazards model.
- While the heuristic arguments leading to the approximation of the null distribution of the log-rank test seem reasonable, is the result correct? In addition, how does the test behave as a function of the amount of censoring or the hazard functions in the two treatment groups? We return to these important practical questions in later units.

1.4.2 Stratified Log-rank test

Stratified Log-rank test: Suppose that we have two groups (say, 2 treatments), as before, but that we want to control (adjust) for a categorical covariate (e.g., gender). Then there are $4 = 2 \times 2$ types of individuals. For example, their respective survivor functions might be as shown below. If we still want to compare treatment groups, but also 'adjust' for gender, a stratified log-rank test could be used. Suppose $S_j^{(l)}(\cdot)$ denotes the survival function for group j in stratum l, and consider $H_0 : S_0^{(l)}(\cdot) = S_1^{(l)}(\cdot)$, $l = 1, \cdots, L$

Construction

1. Separate data into L groups, where $: =$ # levels of the categorical covariates on which you want to stratify (e.g., $L = 2$ when stratifying by gender)
2. Compute O, E, V (say, $O^{(l)}, E^{(l)}, V^{(l)}$) within each group, just as with the ordinary logrank

References

[1] J.D. Kalbfleisch and R.L. Prentice, "The Statistical Analysis of Failure Time Data. Second Edition." 2002, Chapter1.

[2] Harvard BIO244 Lecture Note, Unit 1, 2, 5.