Estimation of Survival Time Distribution

Two general approaches:

- Parametric estimation: specify a parametric distribution for T and estimate the parameter
  - Maximum likelihood
  - Special care for censored observations
  - Convenient for inferences and efficient on computation
  - Lack robustness
- Nonparametric estimation: develop an empirical estimate of survival function
  - More flexible
  - Widely used in practice
  - Life table estimator and Kaplan-Meier estimator

Example

Acute Leukemia

- 42 patients with acute leukemia were randomized to receive 6-mercaptopurine (6-MP) or placebo.
- Interested in evaluating the treatment effect on maintaining remission.
- T is the duration of remission
  - 6-MP 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, $n_2 = 21$
    $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$ (months)

Nonparametric Estimation with Complete Data

Recall $S(t) = \mathbb{P}(T > t)$ is the population fraction surviving beyond t.

- Assume we observe independent samples of T, denoted by $t_1, \cdots, t_n$.
- An empirical estimator of $S(t)$ is the sample fraction surviving beyond t:

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

- The asymptotic CI of $\hat{S}(t)$ can be derived using CLT.

Example

In the acute leukemia example, let us focus on the placebo group:

$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$ (months)

Value of t	$\hat{S}(t)$
$0 \leq t \leq 1$	$21/21 = 1.000$
$1 < t \leq 2$	$19/21 = 0.905$
$2 < t \leq 3$	$17/21 = 0.809$
$3 < t \leq 4$	$16/21 = 0.762$
$\vdots$	$\vdots$
$22 < t \leq 23$	$1/21 = 0.048$
$23 < t < \infty$	0

For Incomplete Data: Kaplan-Meier Estimator

In a 1958 paper in the Journal of the American Statistical Association, Kaplan and Meier proposed a way to estimate S(t) nonparametrically, even in the presence of censoring. The method is based on the ideas of conditional probability

In most applications, right censoring is inevitable. We only observe $(y_i, \delta_i)_{i=1,\cdots,n}$, where δ_i is an event indicator (if $\delta_i = 1$, $y_i = t_i$ is a complete observation). The empirical estimator is inadequate.

For example, if the placebo group has

$$1^+, 1, 2^+, 2, 3, \cdots$$

The survival function $S(t) = 21/21 = 1$ for $0 < t < 1$; but for $1 \leq t < 2$, should we use $20/21$ or $19/21$?

Kaplan-Meier Estimator

Kaplan-Meier Estimator of $S(t)$:

- Based on the idea of conditional probability.
- $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$ are ordered observed times

$$\begin{aligned}
S(t_k) &= \mathbb{P}(T > t_k) = \mathbb{P}(T > t_k | T \geq t_k)\mathbb{P}(T \geq t_k) \\
&= \mathbb{P}(T > t_k | T \geq t_k)\mathbb{P}(T \geq t_k | T \geq t_{k-1})\mathbb{P}(T \geq t_{k-1}) \\
&= \mathbb{P}(T > t_k | T \geq t_k)\mathbb{P}(T > t_{k-1} | T \geq t_{k-1})\mathbb{P}(T \geq t_{k-1}) \\
&= \prod_{i=0}^{k} \mathbb{P}(T > t_i | T \geq t_i) \\
&= \prod_{i=0}^{k} [1 - \mathbb{P}(T = t_i | T \geq t_i)]
\end{aligned}$$

The conditional probability $\mathbb{P}(T = t_i | T \geq t_i)$ can be estimated by $\hat{\lambda}_i = d_i/n_i$, where n_i is the number at risk at t_i^- and d_i is the number of deaths at t_i^+.

- $n_i - d_i$ is the number of patients who survive beyond t_i.
- $n_i - n_{i+1} - d_i$ is the number of censored observation at t_i.
- λ_i is the conditional probability of death at t_i given that the individual is still alive at t_i.

The Kaplan-Meier estimator of survival function is defined as

$$\hat{S}(t) = \prod_{i=1}^{k}(1 - \hat{\lambda}_i) = \prod_{i=1}^{k}(1 - d_i/n_i), \ \ t_k \leq t < t_{k+1}$$

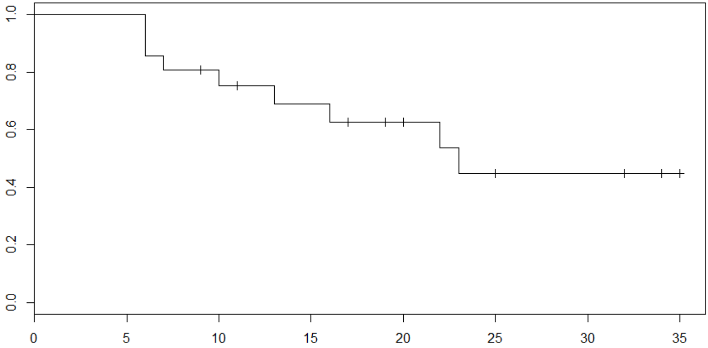
- Order observed y_i as $y_{(1)} \leq \cdots \leq y_{(n)}$
- Obtain m discrete time stamps
- At each discrete time point, calculate $\hat{\lambda}_i$
- Calculate $\hat{S}(t)$
- For complete data, KM estimator is equivalent to the empirical estimator.

Example

Survival function for 6-MP group:

$6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+,$
$20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+$

t_i	n_i	d_i	c_i	$\hat{\lambda}_i$	$\hat{S}(t)$
6	21	3	1	3/21	1*(1-3/21)=0.857
7	17	1	0	1/17	0.857*(1-1/17)=0.807
9	16	0	1	0/16	0.807*(1-0/16)=0.807
...					


0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35

Confidence Intervals for KM

- Greenwood formula
- Based on delta method on $\hat{S}(t) = \prod_{i=1}^{k}(1 - \hat{\lambda}_i)$
- Asymptotically, we have

$$\text{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{i=1}^{k} \frac{d_i}{(n_i - d_i)n_i}$$

- 95% CI is $\hat{S}(t) \pm 1.96 \times se(\hat{S}(t))$ (should be truncated at 0 and 1)
- Appropriate for $0 \ll S(t) \ll 1$ and moderate to large sample size (>20 uncensored observations).
- In practice, a better approach is to get 95% CI for $\log S(t)$ or $\log(-\log S(t))$ and transform it back.

Estimation of Cumulative Hazard Function

Recall $H(t) = -\log S(t)$. A KM estimator is

$$\hat{H}(t) = -\log \hat{S}(t) = -\sum_{i=1}^{k} \log(1 - d_i/n_i)$$

Alternatively, a better one is the Nelson-Aalen estimator:

$$\tilde{H}(t) = \begin{cases} 0, & 0 \le t < t_1 \\ \displaystyle\sum_{t_i \le t} d_i/n_i, & t \ge t_1 \end{cases}$$

- $\mathrm{var}(\tilde{H}(t)) = \sum_{i=1}^{k} d_i/n_i^2$
- $\exp(-\tilde{H}(t))$ is the Fleming-Harrington estimator of $S(t)$.