Statistics for Biology and Health
Chapter 3 Nonparametric Estimation of Basic Quantities for Right-Censored and Left-Truncated Data

Qi Guo

July 20,2019

THE UNIVERSITY OF TEXAS AT DALLAS
School of Natural Sciences and Mathematics

Contents

Introduction

Introduction

- We assume the potential censoring time is unrelated to the potential event time.
- Suppose that the events occur at D distinct times $t_1 < t_2 < \cdots < t_D$, and that at time t_i there are d_i events, Y_i be the number of individuals who are at risks at time t_i.
- Y_i is the number of individuals who are at risk at time t_i, and count of the number of individuals with a time on study of t_i or more.
- d_i / Y_i provides an estimate of the conditional probability that an individual who survives to just prior to time t_i experiences the event at time t_i.
- And now we need to find some basic estimator, like $S(t)$, $H(t)$, s.d, etc.

Estimators of the Survival and Cumulative Hazard Functions for Right-Censored Data

KM and NA estimator

- The standard estimator of the survival function, proposed by Kaplan and Meier, is called the Product-Limit estimator, and we introduce it in the notes before, it provides an efficient means of estimating the survival function for right-censored data.
- An alternate estimator of the cumulative hazard rate, which has better small-sample-size performance than the estimator based on the KM estimator, is Nelson-Aalen estimator.

The Nelson-Aalen estimator

- The Nelson-Aalen estimator has two primary uses in analyzing data.
  1. The first is in selecting between parametric models for the time to event. If a given parametric model fits the data, the resulting graph will be approximately linear. For example, a plot of $\tilde{H}(t)$ versus t will be approximately linear if the exponential distribution, with hazard rate λ, fits the data.
  2. A second is in providing crude estimates of the hazard rate $h(t)$. These estimates are the slope of the Nelson-Aalen estimator, and better estimate of hazard rate is about kernel smoothing and we introduce it before.

Pointwise Confidence Intervals for the Survival Function

Confidence intervals for the survival function at a fixed time t_0

- The intervals are constructed to assure, with a given confidence level $1 - \alpha$ that the true value of the survival function, at a predetermined time t_0, falls in the interval we shall construct.
- And we need some additional notation except: $\hat{S}(t)$, like the KM estimator, $\hat{V}[\hat{S}(t)]$, like the variance of the KM estimator. Let

$$\delta_s{}^2(t) = \hat{V}[\hat{S}(t)]/\hat{S}^2(t)$$

Confidence Interval

- The most commonly used $100 \times (1 - \alpha)\%$ confidence interval for the survival function at time t_0, termed the linear confidence interval, is defined by

$$[\hat{S}(t_0) - Z_{1-\alpha/2}\delta_s(t_0)\hat{S}(t_0), \hat{S}(t_0) + Z_{1-\alpha/2}\delta_s(t_0)\hat{S}(t_0)]$$

  This is the confidence interval routinely constructed by most statistical packages.
- Better confidence intervals can be constructed by first transforming $\hat{S}(t_0)$.

Confidence Interval

- 1. A log transformation of the cumulative hazard rate, The $100 \times (1-\alpha)\%$ log-transformed confidence interval for the survival function at t_0 is given by:

$$[\hat{S}(t_0)^{1/\theta}, \hat{S}(t_0)^{\theta}] \textit{where } \theta = \exp\left\{\frac{Z_{1-\alpha/2}\delta_s(t_0)}{\ln[\hat{S}(t_0)]}\right\}$$

2. The second transformation is an arcsine-square root transformation of the survival function which yields the following $100 \times (1-\alpha)\%$ confidence interval for the survival function:

$$\sin^2\left\{max\left[0, arcsin(\hat{S}(t_0)^{1/2}) - 0.5Z_{1-\alpha/2}\delta_s(t_0)\left(\frac{\hat{S}(t_0)}{1-\hat{S}(t_0)}\right)^{1/2}\right]\right\}$$

$$\leq S(t_0) \leq$$

$$\sin^2\left\{min\left[\frac{\pi}{2}, arcsin(\hat{S}(t_0)^{1/2}) + 0.5Z_{1-\alpha/2}\delta_s(t_0)\left(\frac{\hat{S}(t_0)}{1-\hat{S}(t_0)}\right)^{1/2}\right]\right\}$$

Confidence Bands for the Survival function

Confidence Bands

- Sometimes it is of interest to find upper and lower confidence bands which guarantee, with a given confidence level, that the survival function falls within the band for all t in some interval
- And we need to find two random functions $L(t)$ and $U(t)$, so that $1 - \alpha = Pr[L(t) \leq S(t) \leq U(t), \textit{ for all } t_L \leq t \leq t_U]$, and we call such a $[L(t), U(t)]$ a $(1 - \alpha) \times 100\%$ confidence band for $S(t)$.
- There are two main approaches to constructing confidence bands for $S(t)$.

Constructing Confidence Bands

- The first approach provides confidence bounds which are proportional to the pointwise confidence intervals, and these bands are called the equal probability or EP bands.
- To implement these bands we pick $t_L < t_U$ so that t_L is greater than or equal to the smallest observed event time and t_U is less than or equal to the largest observed event time.
- To construct confidence bands for $S(t)$, based on a sample of size n, define

$$a_L = \frac{n\delta_s^2(t_L)}{1 + n\delta_s^2(t_L)}$$

and

$$a_U = \frac{n\delta_s^2(t_U)}{1 + n\delta_s^2(t_U)}$$

The construction of the EP confidence bands requires that $0 < a_L < a_U < 1$

Constructing Confidence Bands

- To construct a $100 \times (1 - \alpha)\%$ confidence band for $S(t)$ over the range $[t_L, t_U]$,first find a confidence coefficient, $c_\alpha(a_L, a_U)$ from Table C.3 in Appendix C.
- As in the case of $100 \times (1 - \alpha)\%$ pointwise confidence intervals at a fixed time, there are three possible forms for the confidence bands.
  1. Linear:

$$[\hat{S}(t) - c_\alpha(a_L, a_U)\delta_s(t)\hat{S}(t), \hat{S}(t) + c_\alpha(a_L, a_U)\delta_s(t)\hat{S}(t)]$$

  2. Log-Transformed
  3. Arcsine-Square Root Transformed

Constructing Confidence Bands

- The second approach to construct bands are not proportional to the pointwise confidence bounds, it is find the appropriate confidence coefficient $k_\alpha(a_L, a_U)$, from Table C.4 of Appendix C.
- And there are still have three possible forms for the confidence bands.
  1. Linear:

$$[\hat{S}(t) - \frac{k_\alpha(a_L, a_U)[1 + n\delta_s^2(t)]}{n^{1/2}}\hat{S}(t), \hat{S}(t) + \frac{k_\alpha(a_L, a_U)[1 + n\delta_s^2(t)]}{n^{1/2}}\hat{S}(t)]$$

  2. Log-Transformed
  3. Arcsine-Square Root Transformed

Point and Interval Estimates of the Mean Survival Time

Estimator of the Mean Survival Time

- Before we know the mean time to the event μ is given by $\mu = \int_0^\infty S(t)dt$, and a natural estimator of μ is obtained by substituting $\hat{S}(t)$ for $S(t)$ in this expression.
- But this estimator is appropriate only when the largest observation corresponds to a death because in other cases, the KM estimator is not defined beyond the largest observation.
- Two solutions:
  1. Use Efron's tail correction to the KM estimator which changes the largest observed time to a death if it was a censored observation, and the mean restricted to the interval $[0, t_{max}]$ is made.
  2. Estimate the mean restricted to some preassigned interval $[0, \tau]$, where τ is chosen by the investigator to be the longest possible time to which anyone could survive.

Estimator of the Mean Survival Time

- For either case, the estimated mean restricted to the interval $[0, \tau]$ with τ either the longest observed time or preassigned by the investigator, is given by:

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt$$

- The variance of this estimator is:

$$\hat{V}[\hat{\mu}_\tau] = \sum_{i=1}^{D} \Big[\int_{t_i}^{\tau} \hat{S}(t)dt \Big]^2 \frac{d_i}{Y_i(Y_i - d_i)}$$

- And the $100 \times (1 - \alpha)\%$ CI:

$$\hat{\mu}_\tau \pm Z_{1-\alpha/2} \sqrt{\hat{V}[\mu_\tau]}$$

Summary Curves for Competing Risks

Summary Curves for Competing Risks

- The summary survival curves presented before are based on the assumption that the event and censoring times are independent, but here in the case of competing risks data, it's suspect, and we have three techniques for summarizing competing risks data(we mentioned in Chapter 1).
    1. The first estimator which is commonly used is the complement of the KM estimator. Here occurrences of the other event are treated as censored observations
    2. The second estimator is the cumulative incidence function. This estimator is constructed as follows.

Summary Curves for Competing Risks

- Let $t_1 < t_2 < \cdots < t_k$ be the distinct times where one of the competing risk occurs. At time t_i let Y_i be the number of subjects at risk, r_i be the number of subjects with an occurrence of the event of interest at this time, and d_i be the number of subjects with an occurrence of any of the other events of interest at this time, so $(d_i + r_i)$ is the number of subjects with an occurrence of any one of the competing risks at this time.
- The cumulative incidence function is defined by:

$$\begin{cases} 0 & \text{if } t \leq t_1 \\ \sum_{t_i \leq t} \left\{ \prod_{j=1}^{i-1} \frac{1-[d_j+r_j]}{Y_j} \right\} \frac{r_i}{Y_i} & \text{if } t_1 \leq t \end{cases}$$

Summary Curves for Competing Risks

- The variance of the cumulative incidence is estimated by:

$$V[CI(t)] = \sum_{t_i \leq t} \hat{S}(t_i)^2 \left\{ [CI(t) - CI(t_i)]^2 \frac{r_i + d_i}{Y_i^2} + [1 - 2(CI(t) - CI(t_i))] \frac{r_i}{Y_i^2} \right\}$$

- Confidence pointwise $(1 - \alpha)100\%$ confidence intervals for the cumuative incidence are given by $CI(t) \pm Z_{1-\alpha/2} V[CI(t)]^{1/2}$
- The third probability used to summarize competing risks data is the conditional probability function for the competing risk.
  For a particular risk, K. Let $CI_K(t)$ and $CI_{Kc}(t)$ be the cumulative incidence functions for risk K and for all other risks lumped together, respectively.

Summary Curves for Competing Risks

- The conditional probability function:

$$CP_K(t) = \frac{CI_K(t)}{1 - CI_{Kc}(t)}$$

- The variance of this statistic is estimated by:

$$V[CP_K(t)] = \frac{\hat{S}(t^-)^2}{[1 - CI_{Kc}(t)]^4} \sum_{t_i \le t} \frac{[1 - CI_{Kc}(t_i)]^2 r_i + CI_K(t_i)^2 d_i}{Y_i^2}$$

- The conditional probability is an estimate of the conditional probability of event K's occurring by t given that none of the other causes have occurred by t.