Chapter 3

Estimation of F with Right Censored Data

3.1 Models and Basic Quantities

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F$$

Observations : $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ where $T_i = min\{C_i, X_i\}$ and $\delta_i = I(X_i \leq C_i)$. For simplicity, we assume $C_1, \ldots, C_n \sim G \perp \{X_i\}$. Here, we call X_is survival times, C_is censoring times, T_is observed times and δ_is censoring indicators.

Survival function : $S(t) = Pr(X > t) = 1 - F(t)$

Hazard function :

$$h(t) = \frac{f(t)}{S(t)}$$

provided that f, the probability density function (p.d.f.) of F, exists.

Theorem 1 *When f exists,*

$$S(t) = \exp\left(-\int_0^t h(s)ds\right).$$

proof:)

$$\begin{aligned}\int_0^t \frac{f(s)}{1-F(s)}ds &= \int_1^{1-F(t)} -\frac{1}{w}dw \\ &= [-\log w]_1^{1-F(t)} \\ &= -\log(1-F(t)) = -\log S(t).\end{aligned}$$

□

Meaning of Hazard function :

$$\frac{\Pr\left(\text{die at } (t, t+h] | \text{ survive until } t\right)}{h} = \frac{\int_t^{t+h} f(s)ds}{S(t)\cdot h} \to h(t) \quad \text{as} \quad h \to 0.$$

That is, the probability of a subject dying at Δt conditional on that it survives until time t is $h(t)$.

3.2 Common Parametric Family and Likelihood Based Inference

- Exponential (λ)
  - density : $f(t) = \lambda \exp(-\lambda t)$
  - survival function : $S(t) \exp(-\lambda t)$
  - hazard function : $h(t) = \lambda$ (**constant hazard**)

- Weibull (a, λ)
  - survival function : $S(t) = \exp(-\lambda t^a)$
  - hazard function : $h(t) = a\lambda t^{a-1}$
  - The hazard function in increasing when $a > 1$, decreasing when $a < 1$.
- Gamma (α, β)
  - density: $f(t) \propto t^{\alpha-1} \exp(-\beta t)$
  - α is the shape parameter and β is the scale parameter.
  - $\mathrm{E}(T) = \alpha/\beta$ and $\mathrm{Var}(T) = \alpha/\beta^2$.
- Log-Normal (μ, σ^2)
  - $\log X \sim N(\mu, \sigma^2)$.
  - heavy-tailed distribution.
- Pareto (θ)
  - $S(x)\lambda^\theta/x^\theta \qquad x \geq \lambda$.
  - Heavy-tailed (i.e. the tail is decreasing polynomially).

Likelihood inference:

- $f \in \{f_\theta : \theta \in \Theta\}$
- $\Pr(T = t, \delta = 1|\theta) = f_\theta(t) \cdot (1 - G(t))$
- $\Pr(T = t, \delta = 0|\theta) = (\int_t^\infty f_\theta(s)ds) \cdot g(t)$
- Hence, the likelihood is given as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \Big[f_\theta(t_i)(1 - G(t_i))\Big]^{\delta_i} \cdot \Big[(1 - F_\theta(t_i)) \cdot g(t_i)\Big]^{1-\delta_i} \\ &\propto \prod_{i=1}^{n} f_\theta(t_i)^{\delta_i} \cdot (1 - F_\theta(t_i))^{1-\delta i} \end{aligned} \tag{3.1}$$

provided G does not depend on θ (noninformative censoring).

Remark. If G depends on θ, we can still use (3.1) as a likelihood function (i.e. MLE and Fisher information). This likelihood is called a "partial likelihood".

[**HW**]

- Prove the asymptotic normality of the MLE with right censored data.
- Derive the MLE of the exponential distribution with mean λ and its asymptotic variance when right censored data are given.
- Find procedures or functions for fitting parametric survival functions in SAS and R. Estimate the mean survival time of the given dataset of survival times of prostate cancer patients ("prostate cancer data") in the course web page, assuming that the underlying distribution is Exponential. Do it by SAS and R both. Y

3.3 Nonparametric estimation of F

3.3.1 Empirical distribution function when no censored observations exists

Suppose that all observations are uncensored. A very natural estimator of F is the empirical distribution given as

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \le t).$$

Some properties of the empirical distribution are

- It is discrete putting masses on the observations $T_1, \ldots, T_n$.
- For fixed t, it is unbiased. That is, $\mathrm{E}(F_n(t)) = F(t)$.
- The plug-in estimator gives many interesting statistics.

  - Example: mean
    * Parameter: $\mu = \int t dF(t)$.
    * Plug-in estimator : $\hat{\mu} = \int t dF_n(t) = \bar{T}$.

- We will see that F_n is a maximum likelihood estimator in some sense.

3.3.2 Kaplan-Meier estimator

For given $(T_1, \delta_1), \ldots, (T_n, \delta_n)$, let q_n be the number of distinct uncensored observations and let $0 = t_0 < t_1 < \ldots < t_{q_n}$ be the ordered distinct times of uncensored observations. For given $t \in (t_{k-1}, t_k]$,

$$\begin{aligned} S(t) &= Pr(X > t) = \Pr(X > t | X > t_{k-1}) \cdot \Pr(X > t_{k-1}) \\ &= \Pr(X > t | X > t_{k-1}) \cdot \Pr(X > t_{k-1} | X > t_{k-2}) \cdot \Pr(X > t_{k-2}) \\ &= \Pr(X > t | X > t_{k-1}) \cdots \Pr(X > t_1 | X > t_0) \Pr(X > t_0) \\ &= P_t P_{k-1} \cdots P_0 \end{aligned}$$

where $P_k = Pr(X > t_k | X > t_{k-1})$. One obvious estimator of P_j is

$$\hat{P}_j = \frac{\text{number of patients alive at time } t_j}{\text{number of patients alive at time } t_{j-1}}.$$

Let

$$\begin{aligned} n_j &= \sum_{i=1}^{n} I(T_i \geq t_j), \\ d_j &= \sum_{i=1}^{n} I(T_i = t_j, \delta_i = 1). \end{aligned}$$

Then

$$\hat{P}_j = \frac{n_j - d_j}{n_j}.$$

Finally, we have

$$\hat{S}(t) = \prod_{t_k \leq t} \hat{P}_k,$$

which is the Kaplan-Meier estimator [Kaplan and Meier, 1958]. The K-M estimator is also called the *product limit estimator*.

Remark. The K-M estimator is discrete, given mass only to the uncensored observation.

Remark. If no censored observation exists, the K-M estimator becomes the empirical distribution function. That is,

$$1 - \hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \leq t) (= F_n).$$

Remark. If the largest observation is censored then,

$$\lim_{t \to \infty} \hat{S}(t) > 0.$$

That is, the K-M estimator is defective.

3.3.3 Nonparametric Maximum Likelihood Estimator

A question is how we can understand the K-M estimator. Is it an ad-hoc estimator? Or is there a general theory (such as MLE, MME etc) behind it? The aim of this section is to explain the K-M estimator in the framework of the likelihood principle. This explanation gives a way of extending the K-M estimator for more complicated censored/truncated data.

We start with the case of no censored observation. Let $X_1, \ldots, X_n$ be a random sample with the distribution function F. The objective is to estimate F based on the sample.

Suppose $F \in \mathcal{F}$ for some set of distribution functions. If there exists a $\sigma-$finite measure μ which dominates all F in $\mathcal{F}$, then we can define the density f by the Randon-Nykodym derivative of F with respect to μ :

$$f = \frac{dF}{d\mu}.$$

Then, the likelihood is given as $\prod_{i=1}^{m} f(X_i)$ and we can estimate f and hence F by maximizing the likelihood.

However, if $\mathcal{F}$ is large, there may not exist a dominating $\sigma-$finite measure. For example, consider $\mathcal{F} = \{$ all probability measure on $R\}$. In this case, there is no σ-finite dominating measure (why?), and hence the standard likelihood principle is not applicable.

A remedy for this is to use the **pairwise comparison**. For any two distribution functions F_1 and F_2, look at $F_1 + F_2$. Then we have $F_1 << F_1 + F_2$ and $F_2 << F_1 + F_2$. Hence, we can think of the density functions

$$f_1 = \frac{dF_1}{d(F_1 + F_2)}$$

and

$$f_2 = \frac{dF_2}{d(F_1 + F_2)}$$

By applying the likelihood principle, we prefer f_1 over f_2 if

$$\prod_{i=1}^{n} f_1(X_i) > \prod_{i=1}^{n} f_2(X_i).$$

In this way we can find the final winner, which is called the *Nonparametric maximum likelihood estimator* (NPMLE).

Definition. $\hat{F}$ is an NPMLE if for any $F \in \mathcal{F}$

$$\prod_{i=1}^{n} \frac{d\hat{F}}{d(\hat{F} + F)}(X_i) \geq \prod_{i=1}^{n} \frac{dF}{d(\hat{F} + F)}(X_i)$$

for all $F \in \mathcal{F}$.

The following theorem proves that the empirical distribution function is an NPMLE.

Theorem 2 *Let $F_n(t) = \sum_{i=1}^{n} I(X_i \leq t)/n$ be the empirical distribution. Then, F_n is an NPMLE.*

Lemma 1 *For given two probability measures μ_1 and μ_2, if $\mu_1\{x\} > 0$, then*

$$\frac{d\mu_1}{d(\mu_1+\mu_2)}(x) = \frac{\mu_1\{x\}}{\mu_1\{x\}+\mu_2\{x\}}$$
$$\frac{d\mu_2}{d(\mu_1+\mu_2)}(x) = \frac{\mu_2\{x\}}{\mu_1\{x\}+\mu_2\{x\}}.$$

Proof.

$$\begin{aligned}\mu_1\{x\} &= \int_{\{x\}} \frac{d\mu_1}{d(\mu_1+\mu_2)}(y)d(\mu_1+\mu_2) \\ &= \frac{d\mu_1}{d(\mu_1+\mu_2)}(x)\Big(\mu_1\{x\}+\mu_2\{x\}\Big).\end{aligned}$$

□

Proof of Theorem 2. Let $\mu_1 = F_n$ and μ_2 be any probability measure. Note that $\mu_1\{X_i\} > 0$ for all i=1, ..., n. Lemma 1 implies that if $\mu_2\{X_i\} = 0$ for any i, then

$$\frac{d\mu_2}{d(\mu_1+\mu_2)}(X_i) = 0$$

and hence

$$\prod_{i=1}^{n} \frac{d\mu_2}{d(\mu_1+\mu_2)}(X_i) = 0.$$

Hence without loss of generality, we assume $\mu_2\{X_i\} > 0$ for $i = 1, \ldots, n$. Moreover, we can assume without loss of generality $\sum_{i=1}^{n} \mu_2\{X_i\} = 1$. (why?). Hence, the NPMLE should be one among probability measures having positive masses at X_is. For such a probability measure μ,

$$\prod_{i=1}^{n} \mu\{x_i\} = \prod_{i=1}^{q_n} \mu\{t_j\}^{n_j}$$

where q_n is the number of distinct observations and $t_1 < \ldots < t_{q_n}$ are the ordered distinct

observations. Letting $\mu\{t_j\} = P_j$, the problem becomes to find $(P_1, \ldots, P_{q_n})$ maximizing $\prod_{i=1}^{q_n} P_i^{n_i}$ where $n_i = \sum_{j=1}^{n} I(X_j = t_i)$. The solution is

$$\hat{P}_i = \frac{n_i}{n},$$

which is F_n. $\square$

Remark. The NPMLE is a discrete measure having masses at data points.

Remark. Let $t_1 < \ldots < t_{q_n}$ be distinct ordered values of Xs and let $n_k = \sum_{i=1}^{n} I(X_i = t_k)$ for $k = 1, \ldots, q_n$. Let $P_k = \Delta F(t_i)$. If we assume that $F \in \{\sum_{k=1}^{q_n} P_k \cdot I(t_k \leq x)\}$, then the parameter is $\mathbf{P} = (P_1, \ldots, P_{q_n})$ and the likelihood is

$$L(\mathbf{P}) = \prod_{k=1}^{q_n} P_k^{n_k}. \tag{3.2}$$

The MLE of $\hat{\mathbf{P}}$ is $\hat{P}_k = n_k/n$, which yields the empirical distribution function. The likelihood (3.2) is called the *empirical likelihood* since it is a data-dependent likelihood. The empirical likelihood can be used for inference as well as estimation. For details, see Owen [2001].

Let's turn to the case of censored observations. The proof of showing that the K-M is an NPMLE consists of the following three steps.

1. An NPMLE has jumps only at uncensored observations (why?). That is, an NPMLE belongs to
$$\left\{ \sum_{k=1}^{q_n} P_k I(t_k \leq x) \right\}$$
where $t_1 \leq \ldots \leq t_{q_n}$ are ordered distinct uncensored observations.

2. The parameter is $\mathbf{P} = (P_1, \ldots, P_{q_n})$ and the empirical likelihood becomes
$$L(\mathbf{P}) = \prod_{i=1}^{n} (\Delta F(T_i))^{\delta_i} \left(1 - \sum_{t_j \leq T_i} \Delta F(t_j) \right)^{1-\delta_i} \tag{3.3}$$
where
$$\begin{aligned} \Delta F(T_i) &= P_j \qquad \text{if} \qquad T_i = t_j \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

3. Let $\hat{\mathbf{P}}$ be the MLE of $\mathbf{P}$ which maximizes $L(\mathbf{P})$. Then, the corresponding $\hat{F}$ is the K-M estimator.

Remark. Finding $\hat{\mathbf{P}}$ is not easy. However, after we learn the estimation of the hazard function, we will see how to get it easily.

Remark. Is an NPMLE good? For parametric problems, we know that under regularity conditions, the MLE is consistent and asymptotically optimal. What happen in nonparametric cases? The answers are not simple. An NPMLE may not be consistent and if it is consistent, it may not be optimal. In 70s, 80s and 90s, many smart statisticians have studied this issue. Now, most, but not all, mysteries are solved. For references, I recommend Bickel et al. [1993] abd vad der Vaart [1998]. Asymptotic optimality of the K-M estimator was first established by Wellner [1982].

3.3.4 E-M algorithm for the K-M estimator

We first review the general procedure of the E-M algorithm and explain how to apply the E-M algorithm to censored data problems.

The E-M algorithm is an optimization algorithm developed by Dempster et al. [1977] specialized for missing data. Let $X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} f(x|\theta)$. Instead of observing Xs, we observe $Y_i = f_i(X_i)$ for some known function f_i. Examples are right censoring where

$$f_i(x) = [x \cdot I(x \leq C_i) + C_i \cdot I(x > C_i), \quad I(x \leq C_i)]$$

and missing data problems (i.e. $X_i = (X_{i1}, X_{i2})$ and $f_i(X_i) = X_{i1}$). The E-M algorithm is an algorithm to find the MLE of θ based on Y_is.

Let

$$L(\theta|\mathbf{X}) = \prod_{i=1}^{n} f(X_i|\theta)$$

and

$$l(\theta|\mathbf{X}) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

be the complete likelihood and complete log-likelihood respectively where $\mathbf{X} = (X_1, \ldots, X_n)$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$. The E-M algorithm is given as follows.

1. Initialize θ_0 and let $m = 1$.
2. Repeat until converge.
   (a) Calculate the $l(\theta|\mathbf{Y}, \theta_{m-1}) = E_{\theta_{m-1}}\Big(l(\theta|\mathbf{X})|\mathbf{Y}\Big)$ which is the conditional expectation of the complete log-likelihood given data and parameters.
   (b) Maximize $l(\theta|\mathbf{Y})$ to update θ_m.
   (c) $m = m + 1$.

Now, we will explain how to apply the E-M algorithm for maximizing the empirical likelihood (3.3). Let $\theta = (\Delta F(t_1), \ldots, \Delta F(t_{q_n}))$. Then with $\mathbf{X}$(survival times), the empirical likelihood becomes

$$L(\theta|\mathbf{X}) = \prod_{i=1}^{q_n} (\Delta F(t_i))^{n_i},$$

where $n_i = \sum_{j=1}^{n} I(X_j = t_i)$. Let $Y_j = (T_j, \delta_j)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

1. **E step**

$$E_{\theta_{m-1}}\Big(l(\theta|\mathbf{X})|\mathbf{Y}\Big) = \sum_{\mathrm{i=1}}^{\mathrm{q_n}} \underbrace{E_{\theta_{m-1}}\Big(n_i \cdot \log \Delta F(t_i)|\mathbf{Y}\Big)}$$

$$\begin{aligned} \rightarrow E_{\theta_{m-1}}\Big(n_i \cdot \log \Delta F(t_i)|\mathbf{Y}\Big) &= \log \Delta F(t_i) \cdot \Big[\sum_{j=1}^{n} E_{\theta_{m-1}}(I(X_j = t_i)|\mathbf{Y})\Big] \\ &= \log \Delta F(t_i) \cdot \Big[\sum_{j=1}^{n} \underbrace{E_{\theta_{m-1}}(I(X_j = t_i)|Y_j)}\Big] \end{aligned}$$

$$\rightarrow E_{\theta_{m-1}}(I(X_j = t_i)|(T_j, \delta_j)) = \begin{cases} I(T_j = t_i) & \delta_j = 1 \\ \frac{\Delta F_{m-1}(t_i)}{1 - F_{m-1}(T_j)} I(t_i > T_j) & \delta_j = 0 \end{cases}$$

2. M step

$$l(\theta|\mathbf{Y}, \theta_{m-1}) = \sum_{i=1}^{q_n} \underbrace{E_{\theta_{m-1}}(n_i|\mathbf{Y})} \cdot \log \Delta F(t_i)$$

$$\rightarrow \sum_{i=1}^{q_n} E_{\theta_{m-1}}(n_i|\mathbf{Y}) = E_{\theta_{m-1}}(\sum_{i=1}^{q_n} n_i|\mathbf{Y})$$

$$\theta_m = \Delta\hat{F}(t_i) = \frac{E_{\theta_{m-1}}(n_i|\mathbf{Y})}{\sum_{i=1}^{q_n} E_{\theta_{m-1}}(n_i|\mathbf{Y})}$$

Example. For left censored data, the observations are $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ where $T_i = \max(C_i, X_i)$ and $\delta_i = I(X_I \geq C_i)$. And, the empirical likelihood becomes

$$\prod_{i=1}^{n} (\Delta F(T_i))^{\delta_i} \cdot \left(\sum_{t_j \leq T_i} \Delta F(t_j) \right)^{1-\delta_i} \tag{3.4}$$

where ΔF are positive only at $t_1, \ldots, t_{q_n}$. The E-M algorithm can be applied to find the maximizer of (3.4) similarly to the right censoring case.

3.3.5 Redistribution to the Right Algorithm [Efron, 1967]

Another explanation of the K-M estimator is given by Efron [1967] as follows. Assume there is no tie among $T_1, \ldots, T_n$. Let $T_{(1)}, \ldots, T_{(n)}$ be the order statistics and $\delta_{(1)}, \ldots, \delta_{(n)}$ be the corresponding censoring indicators.

1. Initialization: $w_i = 1/n$ for $i = 1, \ldots, n$.
2. For $i = 1$ to n :
   (a) If $\delta_{(i)} = 0$ then
      i. For $k = i + 1$ to n
      $$w_k = w_k + w_i/(n - i).$$
3. $\hat{F}(t) = \sum_{i=1}^{n} w_i I(T_i \leq t, \delta_i = 1)$.

The main idea of the above algorithm redistributes the weights of censored observation to observations on the right hand side equally. It can be shown that the estimator obtained by the above algorithm is the same as the K-M estimator (Try to prove!).

[HW]

- Show that the empirical distribution made by only uncensored observations is asymptotically biased.
- Prove that an NPMLE has jumps only at uncensored observations when right censored data exist.
- Write the E-M algorithm for left censored data.
- Write the redistribution to left algorithm for left censored data.
- Let $(1,1),(2,0),(3,1),(4,0),(5,1)$. Find the K-M estimator and the estimator by the redistribution to right algorithm to confirm that they are the same.
- Find the K-M estimator of the prostate cancer data using SAS or R.

3.4 Asymptotic distribution of the K-M estimator and estimation of the variance

The K-M estimator is an estimator of the distribution function, and so when we discuss the asymptotic distribution of the K-M estimator, we need to think about the convergence of the distribution as a function. First, we review various versions of the central limit theorem (CLT).

3.4.1 CLT for a random sample

The first and simplest form is the CLT for a random sample. Let $X_1, \ldots, X_n$ be a random sample from a distribution function F with $\mathrm{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Then, the CLT

implies that

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1). \tag{3.5}$$

3.4.2 Lindberg-Feller theorem

The extension of the above CLT to non indentical but still independent random variables is the Lindberg-Feller theorem. Consider the triangular array of random variables.

$$\begin{array}{cccc} X_{11}, & X_{12}, & \dots & X_{1k_1} \\ X_{21}, & X_{22}, & \dots & X_{2k_2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1}, & X_{n2}, & \dots, & X_{nk_n} \end{array}$$

Assume that $k_n \to \infty$

Let $S_n = \Sigma_{j=1}^{k_n} X_{nj}$. Assume $\mathrm{E}(X_{ij}) = 0$. Let $\sigma_{ij}^2 = \mathrm{Var}(X_{ij}) < \infty$ and let $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$. If

$$\sum_{j=1}^{k_n} \int_{|x|>\epsilon} x^2 dF_{nj}(x) \to 0$$

for all $\epsilon > 0$ where F_{nj} is the distribution function of X_{nj}, then

$$S_n \xrightarrow{d} N(0,1).$$

3.4.3 Functional CLT

Let $X_1, \dots, X_n$ be a random sample from a distribution function F with mean 0 and variance 1. Define $S_n(t)$ on $t \in [0,1]$ by

$$S_n(t) = \sum_{i=1}^{[nt]} X_i.$$

Then, for fixed t, we can show easily that

$$\frac{S_n(t)}{\sqrt{n}} \xrightarrow{d} N(0,t).$$

The question is what we can say about the probability measure of $S_n(\cdot)$. For this, we need a probability measure for a function which is analogue to the normal distribution for a random variable. And this is the Brownian motion.

Definition. A random function $W(\cdot)$ on $[0,1]$ is called a Brownian motion if

1. $W(0) = 0$
2. (Independent increment) For any sequence $0 = t_0 < t_1 < \cdots < t_k \leq 1$, $W(t_i) - W(t_{i-1}), i = 1, \ldots, k$ are mutually independent.
3. For given t, $W(t) \sim N(0, t)$.

Remark. The Brownian motion can be extended on $[0, \infty)$ easily.

Remark. Brownian motion has many interesting properties. For example, (i) it has a version which has continuous sample paths, (i) a sample path is nowhere differentiable with probability 1, (iii) a process with independent increment and having continuous sample paths is a Brownian motion. However, these properties are not used in this course.

The first functional CLT is that

$$\frac{S_n(\cdot)}{\sqrt{n}} \xrightarrow{d} W(\cdot). \tag{3.6}$$

Unfortunately, to understand the statement of (3.6), many advanced theories of probability are indispensable. Above all, we should have a definition of "in-distribution" convergence for random functions. For interesting readers, see Billingsley [1968], Pollard [1984], van der Vaart and Wellner [1996].

3.4.4 Convergence of the empirical distribution

A first application of functional CLT is the convergence of the empirical distribution to the Brownian bridge.

Let $X_1, \ldots, X_n$ be a random sample with distribution function F on $[0,1]$. Let F_n be the empirical distribution function defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t).$$

For fixed t, we can easily show that

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{d} N(0, F(t)(1 - F(t)). \tag{3.7}$$

For the functional version of (3.7), we need a random function called "Brwonian bridge."

Definition. A random function $B(\cdot)$ on $[0, 1]$ is called the Brownian bridge if

1. $B(0) = 0$ and $B(1) = 0$.
2. $B(\cdot)$ is continuous with probability 1.
3. For any sequence $0 \leq t_1 < t_2 < \ldots < t_k \leq 1$,

$$(B(t_1), B(t_2), \ldots, B(t_k)) \sim N(0, \Sigma)$$

where $\Sigma_{ij} = \min(t_i, t_j) - t_i t_j$.

Remark. We have

$$B(\cdot) \stackrel{d}{\equiv} W(\cdot) | \{W(1) = 0\}.$$

The following theorem is a functional version of the convergence of F_n.

Theorem 3

$$\sqrt{n}(F_n(\cdot) - F(\cdot)) \xrightarrow{d} B(F(\cdot)).$$

There are various applications of Theorem 3. Suppose we want to derive the asymptotic distribution of the sample median m_n. We can think m_n as a mapping from F_n to R. That is, $m_n = \theta(F_n)$ for some functional θ. If we can define the continuity and differentiability of θ, we can derive the asymptotic distribution of m_n by Theorem 3 and the (functional) delta method. For interesting readers, see Gill [1989]. This approach is very useful with the K-M estimator since the sample median with right censored data has a very complicate form.

3.4.5 Functional CLT for the K-M estimator

Under regularity conditions, we have

$$\sqrt{n}(S_n(\cdot) - S(\cdot)) \xrightarrow{d} S(\cdot)W(\sigma^2(\cdot))$$

where

$$\sigma^2(t) = \int_0^t \frac{dF(S)}{\overline{G}(S-)\overline{F}(S)\overline{F}(S-)}.$$

Here $\bar{F} = 1 - F$. We will prove this after we learn counting processes and martingale CLT.

3.4.6 Estimation of Variance: Greenwood's formula

The functional CLT of the K-M estimator gives a way of estimating the variance of S_n. Since $\mathrm{Var}(S(t)) \approx S(t)\sigma^2(t)/n$, a straightforward estimator of $\mathrm{Var}(S_n(t))$ is

$$S_n(t)\sigma_n^2(t)/n$$

where $\sigma_n^2(t)$ is obtained by replacing F and G by their estimators F_n and G_n in the formula of $\sigma^2(t)$.

But, in practice, people use another estimator called the " Greenwood's formula," which can be derived loosely as follows. Recall

$$\hat{S}(t) = \prod_{t_j \le t} \hat{P}_j$$

where

$$\hat{P}_j = 1 - \frac{d_j}{n_j}.$$

Again, recall the delta method: If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ and f is differentiable, then

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) \xrightarrow{d} N(0, (f^{'}(\theta))^2\sigma^2.).$$

First consider

$$\mathrm{Var}(\log \hat{S}(t)) \;=\; \mathrm{Var}(\sum_{t_j \le t} \log \hat{P}_j)$$

$$
\begin{aligned}
&= \sum_{t_j \le t} \mathrm{Var}(\log \hat{P}_j) \quad \text{assuming } \hat{P}_j \text{ are indep.} \\
&= \sum_{t_j \le t} \frac{1}{\hat{P}_j^2} \frac{\hat{P}_j(1-\hat{P}_j)}{n_j} \quad \text{assuming } n_j \hat{P}_j \sim Bin(n_j, P_j) \\
&= \sum_{t_j \le t} \frac{d_j}{(n_j - d_j) n_j}.
\end{aligned}
$$

Now we want to go back to $\hat{S}(t)$ from $\log \hat{S}(t)$ by the delta method to have

$$
\mathrm{Var}(\hat{S}(t)) = \underbrace{(\hat{S}(t))^2 \sum_{t_j \le t} \frac{d_j}{(n_j - d_j) n_j}}_{\text{Greenwood's formula}}.
$$

Some simulation studies support the Greenwood's formula and this is why people prefer it. After studying counting processes and stochastic integration, we will derive the Greenwood's formula as a very natural estimator.

[**HW**]

- Find the variance estimation of the K-M estimator of the prostate cancer data using SAS or R.

3.5 Bootstrap for right censored data

Suppose we want to estimate the median with right censored data. A very natural estimator is the median of the K-M estimator. Now, the question is how to estimate the variance of the median of the K-M estimator. As we mentioned, we can derive the theoretical asymptotic variance of it using the functional CLT and functional delta method. However, (i) the form of the asymptotic variance would be very messy and so it may not be easy to estimate the asymptotic variance and (ii) the functional CLT and functional delta method are too difficult for applied statisticians. One very powerful method for overcoming this issue is a bootstrap method.

3.5.1 General Idea of Bootstrap

Consider the symmetric location problem. Let $X_1, \ldots, X_n$ be a random sample from a distribution function F, which is known to be symmetric around θ. The objective is to estimate θ. Possible estimators are

1. $\bar{X}$
2. Sample Median
3. T_α : α-*th* trimmed mean
4. L-estimators (convex combination of order statistics)

For 1 and 2 we have

$$\begin{aligned} \sqrt{n}(\overline{X} - \theta) &\quad \to \quad N(0, \sigma^2) \\ \sqrt{n}(M_n - \theta) &\quad \to \quad N\Big(0, \frac{1}{4f'(\theta)^2}\Big). \end{aligned}$$

But for 3 and 4, no simple results are available.

Approach 1

We can get an exact expression for $\text{Var}(\hat{\theta})$. Suppose F is discrete with k atoms. The vector $(X_1, \ldots, X_n)$ has k^n possible values. So $\hat{\theta}$ has at most k^n possible values. We can estimate all k^n possible values with the corresponding probabilities, so get the distribution of $\hat{\theta}$ and so its variance.

Approach 2 (Monte Carlo Simulation)

Generate $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} F$ and calculate $\hat{\theta}(Y_1, \ldots, Y_n)$. Repeat it to get $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$. We estimate the variance of $\hat{\theta}$ by the sample variance of $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$. That is,

$$\hat{\text{Var}}(\hat{\theta}) = \frac{\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}^{(\cdot)})^2}{B},$$

which converges to $\text{Var}(\hat{\theta})$ as $B \to \infty$. Here

$$\hat{\theta}^{(\cdot)} = \frac{\sum_{b=1}^{B} \hat{\theta}^{(b)}}{B}.$$

What if we don't know F?

Approach 3 (Bootstrap)

Estimate F by F_n and use Approach 2. This consists of the following steps.

1. Sample with replacement n Xs from the urn. containing $X_1, \ldots, X_n$.
2. Call these artificial points $X_1^*, \ldots, X_n^*$.
3. Calculate $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$.
4. Repeat B times, getting $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$
5. Estimate $Var(\hat{\theta})$ by
$$\frac{\sum_{b=1}^{B} (\hat{\theta}^{*(b)} - \hat{\theta}^{*(\cdot)})^2}{B}.$$

Boostrap for more general situation

Instead of estimating the variance, we are interested in getting the C.I. of θ directly. So far, we have been dealing with statistics (i.e. a function of the data). We now want to to deal with random variables which may depend on things other than the data.

Let $R(\hat{\theta}, F) = (\hat{\theta} - \theta)/S_n$ where S_n is an estimation of the variance of $\hat{\theta}$. Assume that we know $R_{0.025}$ and $R_{0.975}$, the 0.025 and 0.975 quantiles of the distribution of R. Then

$$\Pr\left\{ R_{0.025} \leq \frac{\hat{\theta} - \theta}{S_n} \leq R_{0.975} \right\} = 0.95.$$

So, the interval $[\hat{\theta} - R_{0.975}S_n, \hat{\theta} - R_{0.025}S_n]$ is an exact 95% C.I. for θ.

We will use the bootstrap to estimate $R_{0.025}$ and $R_{0.975}$. Let $q_{0.025}(F)$ be the quantile of $R(\hat{\theta}, F)$. We will estimate $q_{0.025}(F)$ by $q_{0.025}(F_n)$ and construct the C.I. as follows, which is called the "bootstrap t-interval".

Algorithm

1. Generate $X_1^*, \ldots, X_n^* \sim F_n$.
2. Form
$$R^* = \frac{\hat{\theta}^* - \hat{\theta}}{S_n^*}.$$
3. Repeat 1 and 2 B times to get $R_1^*, \ldots, R_B^*$.
4. We form the bootstrap C.I. by
$$\left[\hat{\theta} - S_n R^*_{([0.975B])}, \hat{\theta} - S_n R^*_{([0.025B])} \right].$$

We cannot claim that the coverage probability is 95%. To say that the coverage probability is close to 95%, we need

$$\mathcal{L}(R^* | \text{ data }) \approx \mathcal{L}(R)$$

with probability 1. The following lemma is the simplest form of the bootstrap.

Lemma 2 *If $R(\hat{\theta}, F)$ converges weakly to a distribution G and*

$$\mathcal{L}(R^* | \textit{ data }) \xrightarrow{d} G$$

with probability 1. If G is continuous, then

$$\Pr\left\{\hat{\theta} - S_n R^*_{([0.975B])} \leq \theta \leq \hat{\theta} - S_n R^*_{([0.025B])}\right\} \to 0.95.$$

Remark. For studying bootstrap, we need a concept of random measures, which are mathematically involved.

Remark. Instead of using R, we can use $\sqrt{n}(\hat{\theta} - \theta)$. The C.I. obtained by this is called the "semi-pivotal interval".

Remark. We can also do the followings.

1. Generate $X_1^*, \ldots, X_n^* \sim F_n$.
2. Form $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$.
3. Repeat B times to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
4. We form the bootstrap C.I. by

$$\left[\hat{\theta}^*_{([0.25B])}, \hat{\theta}^*_{([0.975B])}\right].$$

This C.I. is called the "percentile interval", which has the weakest theoretical basis. In fact, the percentile interval is very similar to the Bayesian probability interval.

General idea of the bootstrap

- We have data $X \sim P$.
- Have a random variable $R(X, P)$ (ex. $\sqrt{n}(F_n - F)$)
- Let $D(P) = \mathcal{L}(R(X, P))$.
- We want to estimate $D(P)$, the sampling distribution of $R(X, P)$.

- The bootstrap has the following diagrams.

$$\underbrace{\text{True probability measure (unknown)} \quad \longrightarrow \quad \text{observed data}}_{R(X,P)}$$

$$\longrightarrow \underbrace{\text{Estimate} \quad \hat{P} \quad \longrightarrow \quad \text{bootstrap data}}_{R(X^*,P)}$$

- The bootstrap principle can be stated by "*As long as* $\hat{P}$ *is close to* P*, we can (may) use* $D(\hat{P})$ *to estimate* $D(P)$."

3.5.2 Bootstrap for censored data

Let $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ be data. Suppose we want to estimate $\text{Var}(\hat{F})$. There are two bootstrap procedures.

Method 1

Sample pairs $(T_i, \delta_i)^*$ with replacement from the bag containing $(T_1, \delta_1), \ldots, (T_n, \delta_n)$. Then, construct $\hat{F}^*$ by the K-M estimator.

Method 2

First, estimate $\hat{F}$ and $\hat{G}$ by the K-M estimators. For estimating G, we can simple replace δ_i by $1 - \delta_i$. Then, generate $X_i^* \sim \hat{F}$ and $C_i^* \sim \hat{G}$. Finally, form $T_i^* = \min\{X_i^*, C_i^*\}$ and $\delta_i^* = I(X_i^* \leq C_i^*)$.

Theorem 4 *The two bootstrap methods are identical.*

[HW]

- Prove Theorem 4.
- With the prostate cancer data
    1. Assuming that the true distribution of survival times is an exponential distribution, find the median estimator and derive the variance estimator. Construct 95% C.I. for the median.

2. Calculate the sample median using the K-M estimator.
3. Find the 95% bootstrap C.I.s of the median using semi-pivotal and percentile methods.
4. Compare the parametric C.I. and two bootstrap C.I.s.

* You should make the report format including introduction, explanation of methodologies, the flow chart of the algorithm, results and conclusion remarks. The program code should have comment lines which I can follow, and should be included in the appendix. Latex is recommended but HWP is also acceptable.

Bibliography

P.J. Bickel, C.A.J. Klassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

P. Billingsley. *convergence of Probability Measures*. Wiley, 1968.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

B. Efron. The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 831–853, 1967.

R.D. Gill. Non- and semi-parametirc maximum likelihood estimators and the von-mises method (part i). *Scandinavian Journal of Statistics*, 16:97–128, 1989.

E.L. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistics Association*, 53:457–481, 562–563, 1958.

A.B. Owen. *Empirical Likelihood*. Chapman and Hall, 2001.

D. Pollard. *Convergence of Stochastic Processes*. Springer, 1984.

A.W. vad der Vaart. *Asymptotic Statistics*. Cambridge, 1998.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

J.A. Wellner. Asymptotic optimality of the product limit estimator. *Annals of Statistics*, 10:595–602, 1982.