

Statistics 582, Problem Set 3

Wellner; 1/24/2007

Reading: Chapter 4, section 6. Start reading Chapter 5.

Due: Wednesday, January 31, 2007.

1. We showed in class that the nonparametric maximum likelihood estimator of F in the (right) censored data problem, possibly with ties, is the Kaplan-Meier (product limit) estimator $\widehat{\mathbb{F}}_n(t)$ given by

$$1 - \widehat{\mathbb{F}}_n(t) = \prod_{s \leq t} (1 - \Delta \widehat{\Lambda}(s))$$

where $\widehat{\Lambda}_n(t)$ is the *Nelson-Aalen* estimator of

$$\Lambda(t) \equiv \Lambda_F(t) \equiv \int_0^t \frac{1}{1 - F_-} dF,$$

given by

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s) \quad 0 \leq s \leq t.$$

Here

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n} \sum_{i=1}^n \delta_i 1_{[Z_i \leq t]}, \quad \mathbb{H}_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{[Z_i \leq t]}$$

are the sub-empirical distribution function of the uncensored observations and the marginal empirical distribution of all the Z 's uncensored or censored.

A. Compute $1 - \widehat{\mathbb{F}}_n$ for the following data (time in days until vaginal cancer in rats, group 1; from Kalbfleisch and Prentice, 1980, page 2):

143, 164, 188, 188, 190, 192, 206, 209, 213, 216,
220, 227, 230, 234, 246, 265, 304, 216+, 244+

here + indicates censoring ($\delta = 0$).

B. In class I gave a heuristic derivation of

$$\sqrt{n}(\widehat{\mathbb{F}}_n(t) - F(t)) \Rightarrow (1 - F(t))B(C(t))$$

as a process uniformly in $t \in [0, \tau]$ for any $\tau < \tau_H$ (i.e. for any τ with $1 - H(\tau) = (1 - F(\tau))(1 - G(\tau)) > 0$, where B is a standard Brownian motion process and where

$$C(t) \equiv \int_0^t \frac{1}{(1 - H_-(s))^2} dH^{uc}(s), \quad 0 \leq s \leq t$$

Thus we have, for each fixed $t < \tau$,

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \rightarrow_d N(0, (1 - F(t))^2 C(t))$$

Suggest an estimator of $C(t)$ and hence an estimator of $(1 - F(t))^2 C(t)$.

C. Show that your estimator of $(1 - F(t))^2 C(t)$ is consistent.

D. Use the estimator you suggest in B to obtain an approximate 90% confidence interval for $F(200)$ based for the data given in A.

2. Consider nonparametric maximum likelihood estimation of F in the right-censored data problem considered in class, but extend the argument to include ties as follows:

A. When there are ties, let the distinct Z 's be denoted by $T_1 < \dots < T_k$. Let m_1, \dots, m_k and n_1, \dots, n_k be defined by $m_j \equiv \#$ of $Z_i \delta_i = T_j$, $n_j \equiv \#$ of $Z_i(1 - \delta_i) = T_j$, and let $p_j \equiv \Delta F(T_j) = F(T_j) - F(T_{j-})$, $j = 1, \dots, k$, $p_{k+1} = 1 - F(T_k)$. Show that the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\delta}) = \prod_{i=1}^k p_i^{m_i} \left(\sum_{j=i+1}^{k+1} p_j \right)^{n_i}.$$

B. By defining $\lambda_i = p_i / \sum_{j=i}^{k+1} p_j$ for $i = 1, \dots, k$ and $\lambda_{k+1} = 1$, and rewriting the likelihood in terms of the λ_i 's, show that the likelihood is maximized by

$$\hat{\lambda}_i = m_i / \sum_{j=i}^k (m_j + n_j) = \frac{n \Delta \mathbb{H}_n^{uc}(T_i)}{n(1 - \mathbb{H}_n(T_i -))}.$$

and hence that the nonparametric MLE of F is (again) the Kaplan - Meier estimator

$$1 - \widehat{F}_n(t) = \prod_{s \leq t} (1 - \Delta \widehat{\Lambda}_n(s)).$$

C. Compute $1 - \widehat{F}_n$ for the following data (length of time until complete remission in weeks for the “maintained group”) from a study of the efficacy of chemotherapy for acute Myelogenous leukemia (AML):

9, 13, 13+, 18, 23, 28+, 31, 31, 34, 45+, 48, 161+;

here “+” indicates censoring ($\delta = 0$).

3. 4. (Interval censored or current status data). Suppose that X_1, \dots, X_n are i.i.d. random variables (survival times) with distribution function F as in Example 4.6.5. Suppose that Y_1, \dots, Y_n are i.i.d. random variables (“observation times”)

with a distribution function G which are independent of the X_i 's. Unfortunately, we cannot observe the X_i 's directly but can only observe $(Y_i, 1_{[X_i \leq Y_i]}) \equiv (Y_i, \delta_i)$, $i = 1, \dots, n$.

A. Consider the empirical functions

$$\mathbb{G}_n(t) = n^{-1} \sum_{i=1}^n 1\{Y_i \leq t\} = \mathbb{P}_n 1\{Y \leq t\},$$

$$\mathbb{V}_n(t) = n^{-1} \sum_{i=1}^n \delta_i 1\{Y_i \leq t\} = \mathbb{P}_n \delta 1\{Y \leq t\}.$$

Show that for each fixed t we have

$$\mathbb{G}_n(t) \rightarrow_{a.s.} G(t), \quad \text{and} \quad \mathbb{V}_n(t) \rightarrow_{a.s.} \int_0^t F dG \equiv V(t).$$

B. Plot the cumulative sum diagram $\{(n\mathbb{G}_n(T_{(i)}), n\mathbb{V}_n(T_{(i)})) : i = 1, \dots, n\}$ and the MLE \hat{F}_n of F as described in example 4.6.5, page 38 of the notes, for the following data: $(3.5, 0)$, $(1.2, 1)$, $(5.7, 1)$, $(6.1, 0)$, $(4.2, 1)$.

C. What would the MLE of F be (at $t = 2$) if we assumed that F is exponential θ distribution (with $1 - F_\theta(x) = \exp(-\theta x)$ for $x > 0$)? Compare with the value of the MLE $\hat{F}_n(2)$.

4. **Optional bonus problem:** (a) Show that if \mathbb{U} is a standard Brownian bridge process on $[0, 1]$ and \mathbb{B} is a standard Brownian motion process on $[0, \infty)$, then $(1 + t)\mathbb{U}(t/(1 + t)) \stackrel{d}{=} \mathbb{B}(t)$ as processes on $[0, \infty)$.
 (b) Use the result of (a) to show that the limit process for the Kaplan-Meier estimator $(1 - F(t))\mathbb{B}(C(t))$ satisfies

$$(1 - F(t))\mathbb{B}(C(t)) \stackrel{d}{=} \left(\frac{1 - F(t)}{1 - K(t)} \right) \mathbb{U}(K(t))$$

as processes on $[0, \tau]$ for any $\tau < \tau_H$ where $C(t) = \int_0^t (1 - H(s))^{-2} dH^{(uc)}(s)$ and $K(t) \equiv C(t)/(1 + C(t))$.

(c) Show that when there is no censoring (so $G \equiv 0$), $K(t) = F(t)$ for $t < \tau_H$.