Statistics 582, Problem Set 3

Wellner; 1/21/2009

Reading: Chapter 4, section 6.

Start reading Chapter 5 (to be handed out on Friday, 23 January).

Due: Wednesday, January 28, 2009.

Reminder: The make up lectures are scheduled for Friday, 30 January and Monday, 2 February, from 11:30 - 12:20.

1. Suppose, as in Example 4.3.10, that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\text{Mult}_k(1, \underline{p})$ so that $\underline{N}_n = \sum_{i=1}^n \underline{X}_i \sim \text{Mult}_k(n, \underline{p})$.
(a) Use Jensen's inequality to show that the log-likelihood
$$l_n(\underline{p}|\underline{X}) = \sum_{j=1}^k N_j \log p_j + \sum_{i=1}^n \log\left(\frac{1!}{X_{i1}! \cdots X_{ik}!}\right)$$
is maximized by $\underline{\hat{p}} = \underline{N}_n/n$. [Hint: write the first term of $l_n(\underline{p}|\underline{X})$ as $n\sum_{j=1}^k \hat{p}_j \log p_j$.]
(b) Relate $l_n(\underline{p})$ to $K(\underline{\hat{p}}, \underline{p})$ and hence show again that the maximizing value of $\underline{p}$ is $\hat{p}$.

2. (Right censored data). Suppose that $X, X_1, \ldots, X_n$ are i.i.d. survival times with unknown distribution function F, that $Y, Y_1, \ldots, Y_n$ are i.i.d. censoring times with unknown distribution function G, assumed to be independent of the X_i's, and that we can observe only the iid pairs $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ where $Z_i \equiv X_i \wedge Y_i$ and $\delta_i \equiv 1_{[X_i \le Y_i]}$; also let $Z \equiv X \wedge Y$ and $\delta = 1_{[X \le Y]}$.
A. Show that the joint distribution of (Z, δ) is given by
$$H^{(uc)}(z) = P(Z \le z, \delta = 1) = \int_{(0,z]} (1 - G(x-))dF(x)$$
where $G(x-) \equiv \lim_{y \uparrow x} G(y)$, and
$$H^{(c)}(z) = P(Z \le z, \delta = 0) = \int_{(0,z]} (1 - F(y))dG(y).$$
Furthermore, show that the survival function $1 - H(z) = P(Z > z)$ is given by $1 - H(z) = (1 - F(z))(1 - G(z))$ and also $H(z) = H^{(uc)}(z) + H^{(c)}(z)$.
B. Suppose that the cumulative hazard function corresponding to F is defined by
$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - F(y-)} dF(y).$$
Show that this can be expressed in terms of H and H_{uc} as
$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - H(y-)} dH^{(uc)}(y).$$
C. If $\mathbb{H}_n^{(uc)}(z) = n^{-1} \sum_{i=1}^n \delta_i 1\{Z_i \le z\}$ and $\mathbb{H}_n(z) = n^{-1} \sum_{i=1}^n 1\{Z_i \le z\}$, suggest an estimator of Λ_F based on the observed (Z_i, δ_i)'s.

3. 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

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

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

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

Here

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

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

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

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

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

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 $F(190)$ based for the data given in A.

4. 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 $T_1 < \ldots < T_k$.Let

$m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ 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 $p_j \equiv \Delta F(T_j) = F(T_j) - F(T_j-)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k)$. Show that the likelihood (for F) is

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

B. By defining $\lambda_i = p_i / \sum_{j=i}^{k+1} p_j$ for $i = 1, \ldots, k$ 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

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

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 ($\delta = 0$).

5. (Interval censored or current status data). Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables (survival times) with distribution function F as in Example 4.6.5. Suppose that $Y_1, \ldots, Y_n$ 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 $(Y_i, 1_{[X_i \le Y_i]}) \equiv (Y_i, \delta_i)$, $i = 1, \ldots, n$.
A. Consider the empirical functions

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

Show that for each fixed t we have

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

B. Plot the cumulative sum diagram $\{(n\mathbb{G}_n(T_{(i)}), n\mathbb{V}_n(T_{(i)})) : \; i = 1, \ldots, n\}$ 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 $\hat{F}_n(2)$.

6. **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 $t < \tau_H$.