Statistics 582, Problem Set 4

Wellner; 1/27/2010

Reading: Chapter 4, sections 6-7.

Start reading Chapter 5 (handed out on Wednesday, 27 January).

Due: Wednesday, February 3, 2010.

Reminder: Make up lecture 2, 11:30 - 12:20, Friday, 29 January (in Balmer 411).

1. 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):

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

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

2. 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 (times of remission (in weeks) of leukemia patients (Gehan (1965), 6-MP group; from Cox and Oakes (1984), page 8):

$$6*,\ 6,\ 6,\ 6,\ 7,\ 9*,\ 10*,\ 10,\ 11*,\ 13,\ 16,\ 17*,\ 19*,$$
$$20*,\ 22,\ 23,\ 25*,\ 32*,\ 32*,\ 34*,\ 35*\,.$$

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

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

This derivation proceeded under the assumption that F is continuous. 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 $(1-F(t))^2C(t)$.
C. Show that your estimator of $(1-F(t))^2C(t)$ is consistent.
D. Use the estimator you suggest in B to obtain an approximate 90% confidence interval for $F(18)$ based for the data given in A.
E. Compare the variance estimator in D to the estimator of variance based on Greenwood's formula:

$$\begin{aligned}\widehat{Var(\hat{F}_n(t)} &= n^{-1}(1-\hat{F}_n(t))^2 \int_{[0,t]} \frac{1}{(1-\mathbb{H}_n(s-)-\Delta\mathbb{H}_n^{uc}(s))(1-\mathbb{H}_n(s-))} d\mathbb{H}_n^{uc}(s) \\ &= (1-\hat{F}_n(t))^2 \sum_{j:T_j<t} \frac{d_j}{r_j(r_j-d_j)}\end{aligned}$$

where $r_j \equiv n(1-\mathbb{H}_n(T_j-))$, $d_j \equiv n\Delta\mathbb{H}_n^{uc}(T_j)$, and $T_1 < \cdots < T_m$ are the distinct values of $Z_{n:1} \le \cdots \le Z_{n:n}$.

3. (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 $\{(n\mathbb{G}_n(Y_{(i)}), n\mathbb{V}_n(Y_{(i)})) : i = 1, \ldots, n\}$ and the MLE $\hat{F}_n$ of F as described in example 4.6.5, page 38 of the notes, for the following data: $(3.3, 0)$, $(2.1, 1)$, $(4.7, 1)$, $(7.3, 0)$, $(5.1, 1)$, $(8.4, 1)$.
C. What would the MLE of F be (at $t = 4$) 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 $\hat{F}_n(4)$.

4. (a) Use Jensen's inequality to extend the treatment of Example 4.6.1 given in class to the case where ties are possible. That is, suppose that $Y_1, \ldots, Y_k$ are the distinct values appearing in the sample $X_1, \ldots, X_n$ and let $m_j \equiv \#\{i \le n : X_i = Y_j\}$, $q_j \equiv Q(\{Y_j\})$ for $j = 1, \ldots, k$ so that $\sum_{j=1}^k m_j = n$, and $\sum_{j=1}^k q_j \le 1$. Then show that

$$\prod_{j=1}^{k} q_j^{m_j} \le \prod_{j=1}^{k} (\frac{m_j}{n})^{m_j},$$

and that the resulting maximizer yields the empirical measure $\mathbb{P}_n$. (b) Does the argument you gave in (a) have any connection to the Kullback-Leibler divergence $K(\underline{\hat{p}}, \underline{q})$?

5. **Optional bonus problem 1:** (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 $(1 - F(t))\mathbb{B}(C(t))$ (assuming F continuous) 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 $t < \tau_H$.

6. **Optional bonus problem 2:** (An example where the nonparametric MLE fails.) Suppose that $\mathcal{F}$ is the collection of all cumulative distribution functions on $[0, 1]$ such that $x \mapsto F(x)/x$ is nondecreasing. (This is the class of *star-shaped* distributions.)
(a) Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}$. Show that there is a maximizer $\hat{F}_n$ of the empirical likelihood $F \mapsto \prod_{i=1}^n F\{X_i\}$ over $\mathcal{F}$.
(b) Show that $\hat{F}_n$ satisfies $\hat{F}_n(x) \to_{a.s.} xF(x)$ for every x.