Statistics 582, Problem Set 5, Solutions

Wellner; 2/12/98

1. In class on 1/28 we showed 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 (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 ($\delta = 0$).

B. In class on 1/30 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 $F(30)$ based for the data given in A.

Solution: A. For the given AMP data, the distinct times T_i are
9, 13, 18, 23, 28, 31, 34, 45, 48, 161.
If we let $r_i \equiv n(1 - \mathbb{H}_n(T_i-))$ and $d_i = n\Delta\mathbb{H}_n^{uc}(T_i)$, then we obtain the following table and calculated values of the estimator:

T_i	r_i	d_i	$1 - \frac{d_i}{r_i}$	$\prod_{j \le i}(1 - \frac{d_j}{r_j})$
9	12	1	11/12	.917
13	11	1	10/11	.833
18	9	1	8/9	0.741
23	8	1	7/8	0.648
28	7	0	1	0.648
31	6	2	2/3	0.432
34	4	1	3/4	0.324
45	3	0	1	0.324
48	2	1	1/2	0.162
161	1	0	1	0.162

B. A natural estimator for

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

is

$$\widehat{C}_n(t) \equiv \int_0^t \frac{1}{(1 - \mathbb{H}_n(s-))^2} d\mathbb{H}_n^{uc}(s)$$

and hence a natural estimator of $D(t) \equiv (1-F(t))^2 C(t)$ is
$\widehat{D}_n(t) \equiv (1 - \widehat{F}_n(t))^2 \widehat{C}_n(t)$.
C. By the Glivenko-Cantelli theorem it follows that

$$||\mathbb{H}_n - H||_\infty \equiv \sup_{0 \le t < \infty} |\mathbb{H}_n(t) - H(t)| \to_{a.s.} 0\,,$$

and

$$||\mathbb{H}_n^{uc} - H^{uc}||_\infty \equiv \sup_{0 \le t < \infty} |\mathbb{H}_n^{uc}(t) - H^{uc}(t)| \to_{a.s.} 0\,.$$

Thus we can write

$$\begin{aligned}
|\widehat{C}_n(t) - C(t)| &= \Big| \int_0^t \frac{1}{(1-\mathbb{H}_n(s-))^2} d\mathbb{H}_n^{uc}(s) - \int_0^t \frac{1}{(1-H_-(s))^2} dH^{uc}(s) \Big| \\
&\le \Big| \int_0^t \left\{ \frac{1}{(1-\mathbb{H}_n(s-))^2} - \frac{1}{(1-H_-(s))^2} \right\} d\mathbb{H}_n^{uc}(s) \Big| \\
&\quad + \Big| \int_0^t \frac{1}{(1-H_-(s))^2} d(\mathbb{H}_n^{uc}(s) - H^{uc}(s)) \Big| \\
&\le \int_0^t \Big| \left\{ \frac{1}{(1-\mathbb{H}_n(s-))^2} - \frac{1}{(1-H_-(s))^2} \right\} \Big| d\mathbb{H}_n^{uc}(s) \\
&\quad + \Big| \frac{1}{n} \sum_{i=1}^n \left(\frac{\delta_i}{1-H(Z_i-)^2} 1_{[Z_i \le t]} - E(\frac{\delta}{(1-H(Z-))^2} 1_{[Z \le t]}) \right) \Big| \\
&\le \frac{2||\mathbb{H}_n - H||_\infty}{(1-\mathbb{H}_n(\tau))^2 (1-H(\tau))^2} \mathbb{H}_n^{uc}(\tau) + |\overline{R}_n(t) - E(R(t))| \\
&\to_{a.s.} 0
\end{aligned}$$

since $\mathbb{H}_n(\tau) \to_{a.s.} H(\tau) < 1$, $\mathbb{H}_n^{uc}(\tau) \le 1$, $||\mathbb{H}_n - H||_0^\tau \le ||\mathbb{H}_n - H||_\infty \to_{a.s.} 0$, and

$$R_i \equiv R_i(t) \equiv \frac{\delta_i}{(1-H(Z_i-))^2} 1_{[Z_i \le t]} \,,$$

$i = 1, \ldots, n$ are i.i.d. with finite mean $ER(t) < \infty$ for any $t \le \tau < \tau_H$, so the second term converges to 0 by the SLLN.

D. Now

$$\widehat{C}_n(t) = \int_0^t \frac{1}{(1-\mathbb{H}_n(s-))^2} d\mathbb{H}_n^{uc}(s) = n \sum_{i:T_i \le t} \frac{d_i}{r_i^2} \,,$$

so we find that

$$\widehat{C}_n(30) = 12 \left\{ \frac{1}{12^2} + \frac{1}{11^2} + \frac{1}{9^2} + \frac{1}{8^2} \right\} = .5182 \,,$$

and hence $\widehat{\sigma} = \sqrt{(1-\widehat{\mathbb{F}}_n(30))^2 \widehat{C}_n(30)/n} = \sqrt{(.648)^2(.5182)/12} = .1346\ldots$. Thus an approximate 90% confidence interval for $F(30)$ is given by

$$\widehat{\mathbb{F}}_n(30) \pm 1.645\widehat{\sigma} = 1 - .648 \pm .2214 = .352 \pm .2214 = (.1306, .5734) \,.$$

2. Let $\Theta = \{1, 2\} = \mathcal{A}$ where 1 = a patient has tuberculosis, 2 = a patient does not have tuberculosis. Let X be the number of positive reactions to two different tuberculosis tests, so that $\mathcal{X} = \{0, 1, 2\}$, and suppose that X has the following distributions

x	0	1	2
$p_1(x)$	.04	.12	.84
$p_2(x)$	.64	.28	.08

If the losses are given by $L(1,1) = L(2,2) = 0$ $L(1,2) = 100$, $L(2,1) = 10$, and the prior $\lambda = (\lambda_1, \lambda_2) = (.3, .7)$, find the Bayes rule d_B and the minimax rule d_M. Plot the risk set and label the non-randomized decision rules.

Solution: Let $d = (d_0, d_1, d_2)$ with d_i =prob of action 2 when $x = i$ is observed, $i = 0, 1, 2$. Then the risks are

$$R(1,d) = 100\{d_0(.04) + d_1(.12) + d_2(.84)\}$$

$$R(2,d) = 10\{(1-d_0)(.64) + (1-d_1)(.28) + (1-d_2)(.08)\}\,,$$

and, for $\underline{\lambda} = (.3, .7)$, the Bayes risk of d is

$$\begin{aligned} \mathcal{R}(\lambda, d) &= (.3)R(1,d) + (.7)R(2,d) \\ &= 7 + (.01)\{-328d_0 + 164d_1 + 2464d_2\} \end{aligned}$$

which is minimized by $d = (1,0,0) \equiv d_B \equiv d_4$ (in the list of nonrandomized rules below).
To find a minimax rule, equate $R(1,d) = R(2,d)$: this yields

$$\{4d_0 + 12d_1 + 84d_2\} = 10 - 6.4d_0 - 2.8d_1 - .8d_2\,.$$

Solving for d_0 yields

$$d_0 = (100 - 148d_1 - 848d_2)/104\,,$$

and plugging this back into $R(1,d)$ yields

$$R(1,d) = R(2,d) = \frac{(4)(100)}{104}) + (12 - \frac{(148)(4)}{104})d_1 + (84 - \frac{(4)(848)}{104})d_2$$

which is minimized by $d_1 = 0$, $d_2 = 0$; then $d_1 = 100/104 = 25/26 \doteq .962\ldots$. Hence the minimax rule is $d_M = (25/26, 0, 0)$, and the corresponding common risk is $R(1,d_M) = R(2,d_M) = 100/26 = 50/13 \doteq 3.846\ldots$. Note that for the Bayes rule we have $R(1,d_B) = 4$, $R(2,d_B) = 3.6$.
The nonrandomized rules and their risks are:

x	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
0	0	0	0	1	1	1	0	1
1	0	0	1	0	1	0	1	1
2	0	1	0	0	0	1	1	1
$R(1,d)$	0	84	12	4	16	88	96	100
$R(2,d)$	10	9.2	7.2	3.6	0.8	2.8	6.4	0