東海大學統計學系碩士班

碩士論文

指導教授:沈葆聖 教授

ESTIMATION OF THE TRUNCATION PROBABILITY WITH LEFT-TRUNCATED AND RIGHT-CENSORED DATA

研究生:許躍耀

中華民國九十三年七月

ESTIMATION OF THE TRUNCATION PROBABILITY WITH LEFT-TRUNCATED AND RIGHT-CENSORED DATA

Yueh-Yao Hsu
Dept. of Statistics
Tunghai University
Taichung, 40704
Taiwan, R. O. C.

July 6, 2004

Contents

ABSTRACT

Let (U_i^*, C_i, V_i^*) be i.i.d. random vectors such that (C_i, V_i^*) is independent of U_i^*. Let F, Q and G denote the common distribution function of U_i^*, C_i and V_i^*, respectively. For left-truncated and right-censored data, one can observe nothing if $U_i^* < V_i^*$ and observe (X_i^*, δ_i^*), with $X_i^* = \min(U_i^*, C_i)$ and $\delta_i^* = I_{[U_i^* \leq C_i]}$. In this note, we consider the estimation of the truncation probability $\alpha = P(U^* \geq V^*)$. A proper estimate of α is $\alpha_n = \int G_n(s) dF_n(s)$, where F_n and G_n are nonparametric maximum likelihood estimate (NPMLE) of the distributions F and G, respectively. When the largest observation is not censored, we obtain an alternative representation $\hat{\alpha}_n$ for α_n. For the special case of $C_i = \infty$, the results are reduced to those obtained by He and Yang (1998).

Key Words: Left truncation, right censoring, truncation probability.

1. INTRODUCTION

Let (U_i^*, C_i, V_i^*) be i.i.d. random vectors such that (C_i, V_i^*) is independent of U_i^*. It will be assumed throughout this section that $C_i \geq V_i^*$. Let F, Q and G denote the common distribution function of U_i^*, C_i and V_i^*, respectively. For left-truncated and right-censored data, one can observe nothing if $U_i^* < V_i^*$ and observe (X_i^*, δ_i^*), with $X_i^* = \min(U_i^*, C_i)$ and $\delta_i^* = I_{[U_i^* \leq C_i]}$, if $U_i^* \geq V_i^*$. For any distribution function H denote the left and right endpoints of its support by $a_H = inf\{t : H(t) > 0\}$ and $b_H = inf\{t : H(t) = 1\}$, respectively. Woodroofe (1985) pointed out that if $a_G \leq \min(a_F, a_Q)$ and $b_G \leq \min(b_F, b_Q)$, then F, Q and G are all identifiable. Data of this kind often arise in epidemiology, individual follow-up study (see Wang (1991), Wang, Jewell and Tsai (1987), Tsai, Jewell and Wang (1987)) and possibly in other fields. Consider the following application.

Example:

In hemophilia AIDS-data sets the time of infection T_s can be quite accurately determined. A database will cover patients from, say 1978, till 1995, and hence a patient with a longer survival time will have a larger probability of being part of the sample than a patient with a short survival time. Let U_i^* be the time between T_s and death and let $V_i^* = 1978 - T_s$ if $T_s < 1978$ and $V_i^* = 0$ if $T_s \geq 1978$. Then a patient will only be part of the sample if $U_i^* \geq V_i^*$. Let $C_i = 1995 - T_s$ denote the the time from T_s to the end of study. Hence, $P(C_i > V_i^*) = 1$ and U_i^* is subject to censoring due to termination of study.

In this note, under the assumption that $P(C_i^* > V_i^*) = 1$, we consider the estimation of the truncation probability $\alpha = P(U_i^* \geq V_i^*)$.

2. The α_n and $\hat{\alpha}_n$ Estimator

2.1. Notations

Let $(X_1, \delta_1, V_1), \ldots, (X_n, \delta_n, V_n)$ denote the left-truncated and right-censored sample.

Let $U_{(1)} < U_{(2)} < \cdots < U_{(r)}$ be the distinct ordered failure times and d_s be the number of failure times at $U_{(s)}$ for $s = 1, \ldots, r$.

Similarly, let $V_{(1)} < V_{(2)} < \cdots < V_{(q)}$ be the distinct ordered truncation times and e_t be the number of truncation times at $V_{(t)}$ for $t = 1, \ldots, q$.

Let $C_{(1)} < C_{(2)} < \cdots < C_{(h)}$ be the distinct ordered censoring times and c_l be the number of censoring times at $C_{(l)}$ for $l = 1, \ldots, h$.

For each $V_{(t)}$ $(t = 1, \ldots, q)$, let $C_{(1(t))} < C_{(2(t))} < \cdots < C_{(h(t))}$ be the distinct ordered censoring times and $c_{l(t)}$ be the number of censoring times at $C_{(l(t))}$ for $l = 1, \ldots, h(t)$.

2.2. The NPMLE of F, G and Q

Let $Q(x|v) = P(C_i \le x | V_i^* = v)$ denote the conditional distribution function of C given $V^* = v$. Let $dF(x) = F(x) - F(x-)$, $dG(x) = G(x) - G(x-)$, and $dQ(x|v) = Q(x|v) - Q(x - |v)$.

The likelihood function L can be decomposed into three factors (see Wang (1991), Gross and Lai (1996)), yielding

$$L = \prod_{i=1}^{n} \Big\{ dF(X_i) dG(V_i) [1 - Q(X_i - |V_i)]/\alpha \Big\}^{\delta_i} \times \prod_{i=1}^{n} \Big\{ dQ(X_i|V_i) dG(V_i) [1 - F(X_i)]/\alpha \Big\}^{1-\delta_i}$$

$$= \left\{ \prod_{i=1}^{n} \frac{[F(X_i)]^{\delta_i} [1 - F(X_i)]^{1-\delta_i}}{1 - F(V_i -)} \right\} \times \left\{ \prod_{t=1}^{q} \left[\frac{dG(V_{(t)}) [1 - F(V_{(t)} -)]}{\alpha} \right]^{e_t} \right\}$$

$$\times\left\{\prod_{t=1}^{q}\left[\prod_{V_i=V_{(t)}}[1-Q(X_i-|V_{(t)})]^{\delta_i}[dQ(X_i|V_{(t)})]^{1-\delta_i}\right]\right\}=L_1L_2L_3,$$

where L_1, L_2, and L_3 represent the likelihoods in the first, second, and third brace, respectively.

Let $R_n(u)=n^{-1}\sum_{i=1}^{n}I_{[V_i\le u\le X_i]}$ and $N_F(u)=\sum_{i=1}^{n}I_{[X_i\le u,\delta_i=1]}$. A necessary and sufficient condition for the existence of the nonparametric maximum likelihood estimate (NPMLE) of L_1 is $nR_n(U_{(s)})>d_s=[N_F(U_{(s)})-N_F(U_{(s)}-)]$ for $s=1,\ldots,r$ (see Wang (1987)). Under this regularity condition, the NPMLE of $F(x)$ from L_1 is uniquely determined and given by

$$F_n(x)=1-\prod_{u\le x}\Big[1-\frac{dN_F(u)}{nR_n(u)}\Big]=1-\prod_{U_{(s)}\le x}\Big[1-\frac{d_s}{nR_n(U_{(s)})}\Big],$$

where $dN_F(u)=N_F(u)-N_F(u-)$.

Based on L_2, the NPMLE of $G(y)$ is uniquely determined and given by

$$G_n(y)=\left[\sum_{t=1}^{q}\frac{e_t}{1-F_n(V_{(t)}-)}\right]^{-1}\sum_{t=1}^{q}\frac{e_tI_{[V_{(t)}\le y]}}{1-F_n(V_{(t)}-)}.$$

Based on F_n and G_n, a proper estimator of α is $\alpha_n=\int G_n(s)dF_n(s)$.

Next, let $R_n^t(u)=n^{-1}\sum_{i=1}^{n}I_{[V_i\le u\le X_i,V_i=V_{(t)}]}$ and $N_Q^t(u)=\sum_{i=1}^{n}I_{[X_i\le u,\delta_i=0,V_i=V_{(t)}]}$. For each $V_{(t)}$, a necessary and sufficient condition for the existence of the NPMLE of $Q(x|V_{(t)})$ is $R_n^t(C_{(l(t))})>c_{l(t)}=N_Q^t(C_{l(t)})-N_Q^t(C_{l(t)}-)$ for $l=1,\ldots,h(t)$. Under these regularity conditions, the NPMLE of $Q(x|V_{(t)})$ from L_3 is uniquely determined and given by

$$Q_n(x|V_{(t)})=1-\prod_{u\le x}\Big[1-\frac{dN_Q^t(u)}{nR_n^t(u)}\Big]=1-\prod_{C_{l(t)}\le x}\Big[1-\frac{c_{l(t)}}{nR_n^t(C_{l(t)})}\Big],$$

where $dN_Q^t(u)=N_Q^t(u)-N_Q^t(u-)$.

When $Q_n(x|V_{(t)})$ exists for all $V_{(t)}$'s, the NPMLE of Q (denoted by Q_n) can be written as

$$Q_n(x) = \sum_{t=1}^{q} Q_n(x|V_{(t)})[G_n(V_{(t)}) - G_n(V_{(t-1)})].$$

Note that when the bivariate distribution of (C_i, V_i^*) is continuous, we have $nR_n^t(C_{l(t)}) = c_{l(t)} = 1$, and the NPMLE of $Q(x|V_{(t)})$ does not exist. To circumvent this difficulty, Shen (2003) considered the inverse-probability-weighted estimators by simultaneously estimating F, G and Q. Let $\hat{F}_e(x)$, $\hat{G}_e(x)$ and $\hat{Q}_e(x)$ be given by

$$\hat{F}_e(x) = \left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \le x]}}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}$$

$$= \left[\sum_{s=1}^{r} \frac{d_s}{\hat{G}_e(U_{(s)}) - \hat{Q}_e(U_{(s)}-)}\right]^{-1} \sum_{s=1}^{r} \frac{d_s I_{[U_{(s)} \le x]}}{\hat{G}_e(U_{(s)}) - \hat{Q}_e(U_{(s)}-)}, \tag{2.1}$$

$$\hat{G}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{1 - \hat{F}_e(V_i-)}$$

$$= \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{t=1}^{q} \frac{e_t I_{[V_{(t)} \le x]}}{1 - \hat{F}_e(V_{(t)}-)}, \tag{2.2}$$

and

$$\hat{Q}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{(1-\delta_i) I_{[X_i \le x]}}{1 - \hat{F}_e(X_i-)}$$

$$= \left[\sum_{t=1}^{q} \frac{e_t}{1 - \hat{F}_e(V_{(t)}-)}\right]^{-1} \sum_{l=1}^{h} \frac{(c_l) I_{[C_{(l)} \le x]}}{1 - \hat{F}_e(C_{(l)}-)}. \tag{2.3}$$

The justification of using $\hat{F}_e$, $\hat{G}_e$, and $\hat{Q}_e$ is given as follows. We consider the subdistribution function

$W_F(x) = P(X_i \le x, \delta_i = 1) = P(U_i^* \le x, U_i^* \le C_i | U_i^* \ge V_i^*)$

$= \alpha^{-1} P(U_i^* \le x, V_i^* \le U_i^* \le C_i) = \alpha^{-1} \int_{a_F}^{x} P(V_i^* \le u \le C_i) dF(u)$

$= \alpha^{-1} \int_{a_F}^{x} [G(u) - Q(u-)] dF(u)$. Thus, we have $dF(x) = \alpha \frac{dW_F(x)}{G(x)-Q(x-)}$. When $G(x)$, $Q(x-)$ and α are known, $F(x)$ can be estimated by

$n^{-1}\alpha \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \le x]}}{G(X_i)-Q(X_i-)}$. Let $x = \infty$. It follows that α can be estimated by

$n\Big[\sum_{i=1}^{n} \frac{\delta_i}{G(X_i)-Q(X_i-)}\Big]^{-1}$. This justifies the use of the estimator $\hat{F}_e(x)$.

The justification of using $\hat{G}_e(x)$ can be obtained by considering the subdistribution function $W_G(x) = P(V_i \le x)$. When $1 - F(x)$ and α are known, $G(x)$ can be estimated by $n^{-1}\alpha \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{1-F(V_i-)}$. Let $x = \infty$. It follows that α can be estimated by $n\Big[\sum_{i=1}^{n} \frac{1}{1-F(V_i-)}\Big]^{-1}$. This justifies the use of the estimator $\hat{G}_e(x)$.

Similarly, the justification of using $\hat{Q}_e(x)$ can be obtained by considering the subdistribution function $W_Q(x) = P(X_i \le x, \delta_i = 0) = P(C_i^* \le x, C_i^* \le U_i^* | U_i^* \ge V_i^*)$ $= \alpha^{-1} \int_0^x [1 - F(u-)] dQ(u)$. When $1 - F(u-)$ and α are known, $Q(x)$ can be estimated by $n^{-1}\alpha \sum_{i=1}^{n} \frac{(1-\delta_i) I_{[X_i \le x]}}{1-F(X_i-)}$.

Shen (2003) showed the equivalence of F_n and $\hat{F}_e$, and hence, the equivalence of G_n and $\hat{G}_e$. However, the equivalence of Q_n and $\hat{Q}_e$ does not hold.

Based on the arguments above, two alternative estimators of α are

$n\Big[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i)-\hat{Q}_e(X_i-)}\Big]^{-1}$ and $n\Big[\sum_{i=1}^{n} \frac{1}{1-\hat{F}_e(V_i-)}\Big]^{-1}$.

Instead, under the assumption (C_i, V_i^*) is independent of U_i^* and $P(C_i > V_i^*) = 1$, we have

$$R(x) = P(V_i \le x \le X_i) = P(V_i^* \le x \le \min\{U_i^*, C_i\} | V_i^* \le U_i^*)$$

$$= P(V_i^* \le x, C_i \ge x) P(U_i^* \ge x)/\alpha = [P(V_i^* \le x) - P(C_i < x)] P(U_i^* \ge x)/\alpha$$

$$= [G(x) - Q(x-)][1 - F(x-)]/\alpha.$$

For all x such that $nR_n(x) > 0$, we can obtain another estimator for α as $\hat{\alpha}_n(x) = [G_n(x) - \hat{Q}_e(x-)][1 - F_n(x-)]/R_n(x)$. In the following section, we will establish the equivalence of all the estimators suggested above.

3. THE EQUIVALENCE OF α_n AND $\hat{\alpha}_n$

To derive the explicit relationship between α_n and $\hat{\alpha}_n(x)$, we consider the estimation of $\alpha_d = P(V_i^* \le U_i^* \le C_i)$. Note that $\alpha = \alpha_d + \alpha_c$, where $\alpha_c = P(C_i < U_i^*)$. Let $\tilde{\alpha}_d = \int [G_n(x) - \hat{Q}_e(x-)]dF_n(x)$. For $R_n(x) > 0$, let

$$\hat{\alpha}_d(x) = \frac{n_d}{n}\hat{\alpha}_n(x) = \frac{n_d}{n}[G_n(x) - \hat{Q}_e(x-)][1 - F_n(x-)]/R_n(x),$$

where $n_d = \sum_{i=1}^r d_i$ denotes the number of death.

Lemma 3.1.

Suppose that $nR_n(U_{(i)}) > 0$ for $i = 1 \ldots, r$. Then $\tilde{\alpha}_d = \hat{\alpha}_d(U_{(i)})$ for all $i = 1, \ldots, r$.

Proof:

By (2.1), we have

$$\tilde{\alpha}_d = \int [G_n(x) - \hat{Q}_e(x-)]dF_n(x) = \sum_{i=1}^r [\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)})][\hat{F}_e(U_{(i)}) - \hat{F}_e(U_{(i-1)})]$$

$$= \left[\sum_{i=1}^r \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)}\right]^{-1} \sum_{i=1}^r [\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)] \frac{d_i}{[\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)]}$$

$$= n_d \left[\sum_{i=1}^r \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)}\right]^{-1}. \tag{3.1}$$

Since $\hat{F}_e(U_{(i)}) - \hat{F}_e(U_{(i-1)}) = F_n(U_{(i)}) - F_n(U_{(i-1)})$, we have

$$\left[\sum_{i=1}^r \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)}\right]^{-1} \frac{d_i}{[\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)]} = \frac{d_i[1 - F_n(U_{(i-1)})]}{nR_n(U_{(i)})}.$$

Hence,

$$\tilde{\alpha}_d = n_d \frac{[\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)][1 - \hat{F}_e(U_{(i-1)})]}{nR_n(U_{(i)})} = \hat{\alpha}_d(U_{(i)}).$$

The proof is completed.

Lemma 3.2.

Suppose that $R_n(U_{(i)}) > 0$ for $i = 1\ldots, r$.

Then $\hat{\alpha}_n(U_{(i)}) = \hat{\alpha}_n(U_{(1)}) = n\left[\sum_{i=1}^{r} \frac{d_i}{\hat{G}_e(U_{(i)})-\hat{Q}_e(U_{(i)}-)}\right]^{-1}$ for $i = 2, \ldots, r$.

Proof:

From Lemma 3.1, for $i = 1, \ldots, r$, we have

$$\hat{\alpha}_n(U_{(i)}) = \frac{n}{n_d}\hat{\alpha}_d(U_{(i)}) = \frac{n}{n_d}\tilde{\alpha}_d = n\left[\sum_{i=1}^{r} \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)}\right]^{-1}.$$

The proof is completed.

Lemma 3.3.

When the last observation is not censored, we have

$$\alpha_n = n\left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} = n\left[\sum_{i=1}^{n} \frac{1}{1-\hat{F}_e(V_i-)}\right]^{-1}.$$

Proof:

First, it is easily shown that when the largest observation is not censored, $\int G_n(x)dF_n(x) = \int(1 - F_n(x-))dG_n(x)$ and $\int \hat{Q}_e(x)dF_n(x) = \int(1 - F_n(x-))d\hat{Q}_e(x)$. Hence, we have

$$\tilde{\alpha}_d = \int [G_n(x) - \hat{Q}_e(x-)]dF_n(x) = \int (1 - F_n(x-))d[G_n(x) - \hat{Q}_e(x-)]$$

$$= \int [1 - \hat{F}_e(x-)]d[\hat{G}_e(x) - \hat{Q}_e(x)] = \int [1 - \hat{F}_e(x-)]d\hat{G}_e(x) - \int [1 - \hat{F}_e(x-)]d\hat{Q}_e(x)$$

$$= \left[\sum_{i=1}^{n} \frac{1}{1-\hat{F}_e(V_i-)}\right]^{-1} \left\{\sum_{t=1}^{q}[1 - \hat{F}_e(V_{(t-1)})]\frac{e_t}{1-\hat{F}_e(V_{(t-1)})} - \right.$$

$$\left. \sum_{l=1}^{h}[1 - \hat{F}_e(C_{(l-1)})]\frac{c_l}{1-\hat{F}_e(C_{(l-1)})}\right\} = \left[\sum_{i=1}^{n} \frac{1}{1-\hat{F}_e(V_i-)}\right]^{-1} [\sum_{t=1}^{q} e_t - \sum_{l=1}^{h} c_l]$$

$$= (n - n_c)\left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} = n_d\left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1}.$$

By (3.1), it follows that

$$\tilde{\alpha}_d = n_d\left[\sum_{i=1}^{r} \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i-1)})}\right]^{-1} = n_d\left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1}.$$

Note that

$$\alpha_n = \int G_n(x)dF_n(x) = \int (1 - F_n(x-))dG_n(x)$$

$$= \int (1 - F_e(x-))dG_e(x) = n\left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1}.$$

This completes the proof.

Lemma 3.4.

Suppose that the largest observation is not censored; $R_n(U_{(i)}) > 0$ and $R_n(V_{(j)}) > 0$ for $i = 1, \ldots, r$ and $j = 1, \ldots, t$. Then $\hat{\alpha}_n(U_{(i)}) = \hat{\alpha}_n(V_{(j)})$ for $i = 1, \ldots, r$ and $j = 1, \ldots, t$.

Proof:

Let us denote by $V^*_{(1)} < V^*_{(2)} < \cdots < V^*_{(h)}$ the distinct ordered values of V_j in $[U_{(i-1)}, U_{(i)}]$, i.e.,

$$U_{(i-1)} < V^*_{(1)} < V^*_{(2)} < \cdots < V^*_{(m)} < U_{(i)}.$$

Let $A(x) = \hat{G}_e(x) - \hat{Q}_e(x-)$ and $B(x) = [1 - \hat{F}_e(x-)]/R_n(x)$.

For any $V^*_{(j)}$ in $[U_{(i-1)}, U_{(i)}]$, we have

$$\hat{\alpha}_n(U_{(i)}) - \hat{\alpha}_n(V^*_{(j)}) = A(U_{(i)})B(U_{(i)}) - A(V^*_{(j)})B(V^*_{(j)})$$

$$= [A(U_{(i)}) - A(V^*_{(j)})]B(V^*_{(j)}) + A(U_{(i)})[B(U_{(i)}) - B(V^*_{(j)})].$$

Note that for any V_k in $[V^*_{(j)}, U_{(i)}]$, $1 - \hat{F}_e(V_k-) = 1 - \hat{F}_e(U_{(i-1)})$. Similarly, for any X_k in $[V^*_{(j)}, U_{(i)}]$, $1 - \hat{F}_e(X_k-) = 1 - \hat{F}_e(U_{(i-1)})$.

Hence, by (2.2) and (2.3), we have

$$[A(U_{(i)}) - A(V^*_{(j)})]B(V^*_{(j)}) = \left[\sum_{i=1}^n \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \frac{\sum_{k=1}^n (I_{[V^*_{(j)} < V_k \le U_{(i)}]} - I_{[V^*_{(j)} \le X_k < U_{(i)}]})}{nR_n(V^*_{(j)})}.$$

Note that

$$\begin{aligned}
&\sum_{k=1}^n \left(I_{[V^*_{(j)} < V_k \le U_{(i)}]} - I_{[V^*_{(j)} \le X_k < U_{(i)}]}\right) \\
&= \sum_{k=1}^n \left(I_{[V_k \le U_{(i)}]} - I_{[X_k < U_{(i)}]}\right) - \sum_{k=1}^n \left(I_{[V_k \le V^*_{(j)}]} - I_{[X_k < V^*_{(j)}]}\right) \\
&= \sum_{k=1}^n I_{[V_k \le U_{(i)} \le X_k]} - \sum_{k=1}^n I_{[V_k \le V^*_{(j)} \le U_k]} = nR_n(U_{(i)}) - nR_n(V^*_{(j)}).
\end{aligned}$$

Hence,

$$[A(U_{(i)}) - A(V^*_{(j)})]B(V^*_{(j)}) = \left[\sum_{i=1}^n \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} [R_n(U_{(i)}) - R_n(V^*_{(j)})]/R_n(V^*_{(j)}).$$

Next,

$$A(U_{(i)})[B(U_{(i)}) - B(V^*_{(j)})] = [\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)][1 - \hat{F}_e(U_{(i-1)})] \frac{R_n(V^*_{(j)}) - R_n(U_{(i)})}{nR_n(V^*_{(j)})R_n(U_{(i)})}.$$

Note that

$$\begin{aligned}
&[1 - \hat{F}_e(U_{(i-1)})]/nR_n(U_{(i)}) = [1 - F_n(U_{(i-1)})]/nR_n(U_{(i)}) = [F_n(U_{(i)}) - F_n(U_{(i-1)})]/d_i \\
&= [\hat{F}_e(U_{(i)}) - \hat{F}_e(U_{(i-1)})]/d_i = \left[\sum_{i=1}^n \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} \frac{1}{\hat{G}_e(U_i) - \hat{Q}_e(U_i-)}.
\end{aligned}$$

Hence,

$$A(U_{(i)})[B(U_{(i)}) - B(V^*_{(j)})] = \left[\sum_{i=1}^n \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} [R_n(V^*_{(j)}) - R_n(U_{(i)})]/R_n(V^*_{(j)}).$$

By Lemma 3.3, it follows that

$$[A(U_{(i)}) - A(V^*_{(j)})]B(V^*_{(j)}) + A(U_{(i)})[B(U_{(i)}) - B(V^*_{(j)})] = 0.$$

The proof is completed.

Lemma 3.5.

Suppose that the largest observation is not censored, $nR_n(U_{(i)}) > 0$ and $nR_n(C_{(l)}) > 0$ for $i = 1, \ldots, r$

and $l = 1, \ldots, h$. Then $\hat{\alpha}_n(U_{(i)}) = \hat{\alpha}_n(C_{(l)})$ for $i = 1, \ldots, r$ and $l = 1, \ldots, h$.

Proof:

The proof is similar to that of Lemma 3.4 and is omitted.

Lemma 3.6.

Suppose that the largest observation is not censored, $nR_n(U_{(i)}) > 0$, $nR_n(V_{(t)}) > 0$

and $R_n(C_{(l)}) > 0$ for $i = 1 \ldots, r$, and $t = 1, \ldots, q$ and $l = 1, \ldots, h$. Then $\hat{\alpha}_n(x)$ is constant for all $x \in [V_{(1)}, U_{(r)}]$, and

$$\hat{\alpha}_n(x) = \alpha_n = n\left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} = n\left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1}.$$

Proof:

Note that the jumps of $\hat{\alpha}_n(x)$ occur at the distinct order statistics $U_{(i)}$'s, $V_{(t)}$'s and $C_{(l)}$'s. By Lemma 3.2, 3.4 and 3.5, $\hat{\alpha}_n(U_{(i)}) = \hat{\alpha}_n(V_{(t)}) = \alpha_n(C_{(l)})$ for $i = 1, \ldots, r$, $t = 1, \ldots, q$ and all $C_{(l)} \leq U_{(r)}$, it follows that $\hat{\alpha}_n(x)$ is constant for any $x \in [V_{(1)}, U_{(r)}]$. By (3.1) and Lemma 3.3, whence the result.

Under the condition $P(C_i > V_i^*) = 1$, Wang (1991) show that $\sqrt{n}\{n[\sum_{i=1}^{n} 1/(1 -$

$\hat{F}_e(V_i-))]^{-1} - \alpha\} = \sqrt{n}(\alpha_n - \alpha)$ converges weakly to $N(0, \sigma^2_{\alpha_n})$, where

$$\sigma^2_{\alpha_n} = \alpha^3 \int_{a_G}^{b_G-} \frac{1}{S(s-)} dG(s) + \alpha^2 \int_{a_G}^{b_G-} \frac{(1-G(s))^2 dF(s)}{R(s)S(s-)} - \alpha^2, \tag{3.2}$$

where $S(s) = 1 - F(s)$.

When $C_i^* = \infty$, U_i^* is only subject to left-trucation, i.e., left-truncated data (see Lynden-Bell (1971), Woodroofe (1985)). In that case, He and Yang (1998), showed the equivalence of α_n and $\hat{\alpha}_n$. Their approaches are different from those presented in this note. Besides, they showed that $\sqrt{n}(\hat{\alpha}_n(x) - \alpha)$ converges weakly to $N(0, \sigma^2_{\hat{\alpha}_n(x)})$, where

$$\sigma^2_{\hat{\alpha}_n(x)} = \alpha^2 \int_{a_G}^{x} \frac{dW_F(s)}{R^2(s)} + \alpha^2 \int_{x}^{b_G-} \frac{dW_G(s)}{R^2(s)} - \alpha^2 \frac{1}{R(x)} + 2\alpha^3 - \alpha^2 \tag{3.3}$$

for $x \in (a_G, b_G)$, is a constant, where $W_F(s) = P(X_i \le s, \delta_i = 1)$ and $W_G(s) = P(V_i \le s)$. The following Lemma shows the equivalence of the two expressions.

Lemma 3.7.

When $C_i = \infty$, we have $\sigma^2_{\alpha_n} = \sigma^2_{\hat{\alpha}_n(x)}$ for all $x \in (a_G, b_G)$.

Proof:

It suffices to show that

$$\underbrace{\int_{a_G}^{b_G-} \frac{(1-G(s))^2}{R(s)S(s-)} dF(s)}_{(3.2.1)} + \underbrace{\alpha \int_{a_G}^{b_G-} \frac{1}{S(s-)} dG(s)}_{(3.2.2)} = \underbrace{\int_{a_G}^{x} \frac{dW_F(s)}{R^2(s)}}_{(3.3.1)} + \underbrace{\int_{x}^{b_G-} \frac{dW_G(s)}{R^2(s)}}_{(3.3.2)} - \frac{1}{R(x)} + 2\alpha.$$

First,

$$(3.2.1) = \underbrace{\int_{a_G}^{b_G-} \frac{1}{R(s)S(s-)} dF(s)}_{(3.2.1.1)} + \underbrace{\int_{a_G}^{b_G-} \frac{G^2(s)}{R(s)S(s-)} dF(s)}_{(3.2.1.2)} - \underbrace{\int_{a_G}^{b_G-} \frac{2G(s)}{R(s)S(s-)} dF(s)}_{(3.2.1.3)}.$$

$$(3.2.1.1) = \underbrace{\int_{a_G}^{x} \frac{1}{R(s)S(s-)} dF(s)}_{(3.2.1.1.1)} + \underbrace{\int_{x}^{b_G-} \frac{1}{R(s)S(s-)} dF(s)}_{(3.2.1.1.2)}.$$

Since $dF(s) = \alpha \frac{1}{G(s)} dW_F(s)$ and $R(s) = \alpha^{-1} G(s) S(s-)$, we have

$(3.2.1.1.1) = \int_{a_G}^{x} \frac{\alpha}{R(s)G(s)S(s-)} dW_F(s) = \int_{a_G}^{x} \frac{1}{R^2(s)} dW_F(s) = (3.3.1)$.

Next, $(3.2.1.2) = \int_{a_G}^{b_G-} \frac{\alpha G(s)}{S^2(s-)} dF(s) = \alpha \int_{a_G}^{b_G-} G(s) d\big[\frac{1}{S(s)}\big]$,

$(3.2.1.3) = -2\alpha \int_{a_G}^{b_G-} 1 d\big[\frac{1}{S(s-)}\big] = 2\alpha - 2\alpha \frac{1}{S(b_G-)}$, and $(3.2.2) = \alpha \frac{1}{S(b_G-)} - (3.2.1.2)$.

It follows that $(3.2.1) + (3.2.2) = (3.3.1) + 2\alpha - \alpha \frac{1}{S(b_G-)} + (3.2.1.1.2)$.

Next, since $dW_G(s) = \alpha^{-1} S(s-) dG(s)$, we have

$$(3.3.2) = \alpha^{-1} \int_x^{b_G-} \frac{S(s-)}{R^2(s)} dG(s) = \int_x^{b_G-} \frac{1}{R(s)G(s)} dG(s) =$$

$$-\alpha \int_x^{b_G-} \frac{1}{S(s-)} d\Big[\frac{1}{G(s)}\Big] = -\alpha \frac{1}{S(b_G-)} + \frac{1}{R(x)} + \alpha \int_x^{b_G-} \frac{1}{G(s)} d\Big[\frac{1}{S(s)}\Big].$$

Since $\alpha \int_x^{b_G-} \frac{1}{G(s)} d\big[\frac{1}{S(s)}\big] = \int_x^{b_G-} \frac{1}{R(s)S(s-)} dF(s) = (3.2.1.1.2)$, we have $(3.3.2) - \frac{1}{R(x)} + 2\alpha = (3.2.1) + (3.2.2)$. The proof is completed.

4. DISCUSSION

For the case where no assumption is made on the distribution of V_i^* and C_i, the truncation probability is defined as $\alpha = P(\min(U_i^*, C_i) \geq V_i^*)$ and

$$R(x) = P(V_i \leq x \leq X_i) = P(V_i^* \leq x \leq \min\{U_i^*, C_i\} | V_i^* \leq \min(U_i^*, C_i))$$

$$= P(V_i^* \leq x, C_i \geq x) P(U_i^* \geq x)/\alpha = K(x)[1 - F(x-)]/\alpha,$$

where $K(x) = P(V_i^* \leq x \leq C_i)$. Note that for this general case, when $a_G \leq \min(a_F, a_Q)$ and $b_G \leq \min(b_F, b_G)$, the product limit estimator F_n is still a consistent estimator of F (see Tsai, Jewell and Wang (1987)). Hence, given $K(x)$, for all x such that $R_n(x) > 0$, we can obtain an estimator for α as $\hat{\alpha}_n(x) = K(x)[1 - F_n(x-)]/R_n(x)$. However, $K(x)$ cannot be estimated from the data since there is no distributional assumption on V_i^* and C_i (see He and Yang (2000)).

BIBLIOGRAPHY

Gross, S. T. and Lai, T. L. Boostrap methods for truncated data and censored data. Statist. Sinica, **1996**, *6*, 509-530.

He, S. and Yang, G. L. Estimation of the truncation probability in the random truncation model. Ann. Statist., **1998**, *26*, 1011-1027.

He, S. and Yang, G. L. On the strong convergence of the product-limit estimator and its integrals under censoring and random truncation. Statis. & Probab. Lett. **2000**, *49*, 235-244.

Lynden-Bell, D. A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon. Not. R. Astr. Soc. **1971**, *155*, 95-118.

Robins, J. M. and Rotnitzky, A. Rocovery of information and adjustment for dependent censoring using surrogate markers, in AIDS Epidemiology-Methodological Issues, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhauser, **1992**, pp. 297-331.

Satten, G. A. and Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. Amer. Statist. Ass., **2001**, *55*, 207-210.

Shen, P.-S. The product-limit estimates as an inverse-probability-weighted average. Communi. in Statist., Part A- Theory and Methods, **2003**, *32*, 1119-1133.

Tsai, W.-Y., Jewell, N. P. and Wang, M.-C. A note on the product-limit estimate under right censoring and left truncation. Biometrika, **1987**, *74*, 883-886.

Wang, M.-C.; Jewell, N. P.; Tsai, W.-Y. Asymptotic properties of the product-limit estimate under random truncation. Ann. Statist., **1986**, *14* 1597-1605.

Wang, M.-C. Product-limit estimates: a generalized maximum likelihood study. Communi. in Statist., Part A- Theory and Methods, **1987**, *6*, 3117-3132.

Wang, M.-C. Nonparametric estimation from cross-sectional survival data. J. Amer. Statist. Ass., **1991**, **86**, 130-143.

Woodroofe, M. Estimating a distribution function with truncated data. Ann. Statist., **1985**, *13*, 163-167.