東海大學應用統計研究所

碩士論文

指導教授：沈葆聖 博士

Estimation of the Truncation Probability
In the Random Truncation Model

NATIONAL CENTRAL LIBRARY ROC

研究生：張瑞芊

中華民國 九十二 年 七 月

Estimation of the Truncation Probability In the Random Truncation Model

Jui-Ping Chang

Dept. of Statistics Tunghai University

Taichung, 40704 Taiwan, R. O. C.

July 1,2003

Contents

ESTIMATION OF THE TRUNCATION PROBABILITY IN THE RANDOM TRUNCATION MODEL

ABSTRACT

Under random truncation, a pair of independent random variables U^* and V^* is observable only if U^* is larger than V^*. The resulting model is the conditional probability $H(x,y) = P(U^* \leq x, V^* \leq y | U^* \geq V^*)$. For the truncation probability $\alpha = P(U^* \geq V^*)$, a proper estimate 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. He and Yang (1998) showed that α_n is equivalent to a simpler representation $\hat{\alpha}_n$. In this article, using coupled inverse-probability-of-truncation weighted estimators, we propose an alternative proof of the equivalence. Similarly, for left-truncated and right-censored data, two estimators (denoted by $\tilde{\alpha}_n$ and $\hat{\alpha}_e$) are considered. It is shown that the equivalence of $\tilde{\alpha}_n$ and $\hat{\alpha}_e$ does not hold. Simulation results shows that the mean-squared error of $\tilde{\alpha}_n$ is smaller than that of $\hat{\alpha}_e$.

Key Words: Product-limit estimator; Truncation probability.

1 INTRODUCTION

Let U^* and V^* be the target and truncation variables with distribution functions F and G respectively. Assume that U^* and V^* are independent. For left-truncated data, both U^* and V^* are observable only when $U^* \geq V^*$. Truncated data occur in astronomy (e.g., Lynden-Bell (1), Woodroofe (2)), epidemiology, biometry (e.g., Wang, Jewell and Tsai (3), Tsai, Jewell and Wang (4)) and possibly in other field such as economics. For any distribution function K denote the left and right endpoints of its support by $a_K = inf\{t : K(t) > 0\}$ and $b_K = inf\{t : K(t) = 1\}$, respectively. Woodroofe (2) pointed out that if $a_G \leq a_F$ and $b_G \leq b_F$, then both F and G are identifiable. Let $(U_1, V_1), \ldots, (U_n, V_n)$ denote the truncated sample. Hence, $H(u, v) = P(U_i \leq u, V_i \leq v) = P(U^* \leq u, V^* \leq v | U^* \geq V^*)$. Let $I_{[A]}$ be the indicator function of the event A. Let $N_F(u) = \sum_{i=1}^n I_{[U_i \leq u]}$, $N_G(v) = \sum_{i=1}^n I_{[V_i \leq v]}$, and $R_n(u) = N_G(u) - N_F(u-) = \sum_{i=1}^n I_{[V_i \leq u \leq U_i]}$.

Let $U_{(1)} < U_{(2)} < \cdots < U_{(r)}$ denote the distinct ordered statistics of the sample $U_i's$. Let $d_i = N_F(U_{(i)}) - N_F(U_{(i)}-)$ denote the number of failure times at $U_{(i)}$ for $i = 1, \ldots, r$. Similarly, let $V_{(1)} < V_{(2)} < \cdots < V_{(q)}$ be the distinct order statistics of sample $V_1, V_2, \ldots, V_n$, and $e_j = N_G(V_{(j)}) - N_G(V_{(j)}-)$ denote the number of truncation times at $V_{(j)}$. A necessary and sufficient condition for the existence of the nonparametric maximum likelihood estimate (NPMLE) of $F(x)$ is $R_n(U_{(i)}) > d_i$ for $i = 1, \ldots, r$, for the existence of the NPMLE of $G(x)$ is $R_n(V_{(j)}) > e_j$ for $j = 1, \ldots, q-1$ (see Wang (5)). Under these regularity conditions, the NPMLEs of $F(x)$ and $G(x)$ are uniquely determined and given by

$$F_n(x) = 1 - \prod_{u \leq x} \Big[1 - \frac{dN_F(u)}{R_n(u)}\Big],$$

and

$$G_n(x) = \prod_{v > x} \Big[1 - \frac{dN_G(v)}{R_n(v)}\Big],$$

where $dN_F(u) = N_F(u) - N_F(u-)$ and $dN_G(v) = N_G(v) - N_G(v-)$.

For the truncation probability $\alpha = P(U^* \geq V^*)$, a proper estimate is $\alpha_n = \int G_n(s) dF_n(s)$. Since F_n and G_n have complicated product-limit forms, it is generally not easy to study the properties of α_n.

Since $R(x) = nP(V^* \leq x \leq U^* | U^* \geq V^*) = n\alpha^{-1}G(x)[1-F(x-)]$, replacing G, F and R by G_n, F_n and R_n yields another estimate $\hat{\alpha}_n(x) = nG_n(x)[1 - F_n(x-)]/R_n(x)$ for all x such that $R_n(x) > 0$. He and Yang (6) showed that α_n is equivalent to a $\hat{\alpha}_n(x)$ for all x such that $R_n(x) > 0$. In Section 2, using coupled inverse-probability-of-truncation weighted estimators, we give an alternative proof of equivalence. In Section 3, two estimators (denoted by $\tilde{\alpha}_n$ and $\hat{\alpha}_e$) are considered for left-truncated and right-censored data. It is shown that the equivalence of $\tilde{\alpha}_n$ and $\hat{\alpha}_e$ does not hold. In Section 4, a simulation study is conducted to examine the performance of $\tilde{\alpha}_n$ and $\hat{\alpha}_e$

2 EQUIVALENCE OF α_n and $\hat{\alpha}_n(x)$

First, we consider an inverse-probability-of-truncation weighted estimator of $F(x)$ and $G(x)$ (see Robins and Rotnitzky (7); Satten and Datta (8); Shen (9)). We simultaneously estimate $F(x)$ and $G(x)$ using coupled inverse-probability-of-truncation weighted estimators. Let $\hat{F}_c(x)$ and $\hat{G}_c(x)$ be given by

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

and

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

Shen (9) shows that $\hat{F}_c$ and $\hat{G}_c$ are equivalent to F_n and G_n, respectively. Based on $\hat{F}_c$ and $\hat{G}_c$, the following theorem shows that α_n and $\hat{\alpha}_n(x)$ are equivalent.

Lemma 2.1.

Suppose that $R_n(U_{(i)}) > 0$ and $R_n(V_{(k)}) > 0$ for $i = 1 \ldots, r$ and $k = 1, \ldots, q$. Then $\alpha_n = \hat{\alpha}_n(x)$ for all $x \in [V_{(1)}, U_{(r)}]$.

Proof:

First,

$$\alpha_n = \int G_n(x) dF_n(x) = \sum_{j=1}^{r} G_n(U_{(j)})[F_n(U_{(j)}) - F_n(U_{(j-1)})]$$

$$= \sum_{j=1}^{r} \hat{G}_c(U_{(j)})[\hat{F}_c(U_{(j)}) - \hat{F}_c(U_{(j-1)})] = \sum_{j=1}^{r} \frac{\hat{G}_c(U_{(j)}) d_j}{\hat{G}_c(U_{(j)})[\sum_{j=1}^{r} d_j/\hat{G}_c(U_{(j)})]}$$

$$= \frac{n}{\sum_{j=1}^{r} d_j/\hat{G}_c(U_{(j)})}. \tag{2.1}$$

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

$$\frac{d_i}{\hat{G}_c(U_{(i)})[\sum_{j=1}^{r} d_j/\hat{G}_c(U_{(j)})]} = \frac{d_i[1-F_n(U_{(i-1)})]}{R_n(U_{(i)})}.$$

Hence,

$$\alpha_n = \frac{n}{\sum_{j=1}^{r} d_j/\hat{G}_c(U_{(j)})} = nG_n(U_{(i)})[1-F_n(U_{(i-1)})]/R_n(U_{(i)})$$

$$= nG_n(U_{(i)})[1 - F_n(U_{(i)}-)]/R_n(U_{(i)}) = \hat{\alpha}_n(U_{(i)}).$$

Similarly,

$$\alpha_n = \int [1 - F_n(x-)]dG_n(x) = \sum_{j=1}^{q} [1 - F_n(V_{(j)}-)][G_n(V_{(j)}) - G_n(V_{(j-1)})]$$

$$= \sum_{j=1}^{q} [1 - \hat{F}_c(V_{(j)}-)][\hat{G}_c(V_{(j)}) - \hat{G}_c(V_{(j-1)})]$$

$$= \sum_{j=1}^{q} \frac{[1 - \hat{F}_c(V_{(j)}-)]e_j}{\Big[1 - \hat{F}_c(V_{(j)}-)\Big]\Big[\sum_{j=1}^{q} e_j/[1 - \hat{F}_c(V_{(j)}-)]\Big]} = \frac{n}{\sum_{j=1}^{q} e_j/[1 - \hat{F}_c(V_{(j)}-)]}.$$

Since $\hat{G}_c(V_{(k)}) - \hat{G}_c(V_{(k-1)}) = G_n(V_{(k)}) - G_n(V_{(k-1)})$, we have

$$\frac{e_k}{\Big[1 - \hat{F}_c(V_{(k)}-)\Big][\sum_{j=1}^{q} e_j/[1 - \hat{F}_c(V_{(j)}-)]} = \frac{e_k G_n(V_{(k)})}{R_n(V_{(k)})}.$$

Hence,

$$\alpha_n = \frac{n}{\sum_{j=1}^{q} e_j/[1 - \hat{F}_c(V_{(j)}-)} = nG_n(V_{(k)})[1 - F_n(V_{(k)}-)]/R_n(V_{(k)}) = \hat{\alpha}_n(V_{(k)}). \tag{2.2}$$

Note that the jumps of $\hat{\alpha}_n(x)$ occur at the distinct order statistics $U_{(i)}$'s and $V_{(k)}$'s. Since $\hat{\alpha}_n(U_{(i)}) = \hat{\alpha}_n(V_{(k)})$ for $i = 1 \ldots, r$ and $k = 1, \ldots, s$, it follows that $\alpha_n = \hat{\alpha}_n(x)$ for all $x \in [V_{(1)}, U_{(r)}]$.

Note that by (2.1) and (2.2) in Lemma 2.1 it follows that

$$\left[\sum_{i=1}^{n} \frac{1}{\hat{G}_c(U_i)}\right]^{-1} = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_c(V_i-)}\right]^{-1}.$$

3 LEFT-TRUNCATED AND RIGHT-CENSORED DATA

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 $U_i^* \geq V_i^*$. Woodroofe (2) 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 and individual follow-up study (see Wang (10)).

Notation

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 $l = 1, \ldots, h(t)$.

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

The likelihood function L can be decomposed into three factors (see Wang (10), Gross and Lai

(11)), 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 $\tilde{R}_n(u) = \sum_{i=1}^{n} I_{[V_i\le u\le X_i]}$ and $\tilde{N}_F(u) = \sum_{i=1}^{n} I_{[X_i\le u,\delta_i=1]}$. A necessary and sufficient condition for the existence of the NPMLE of L_1 is $\tilde{R}_n(U_{(s)}) > d_s = \tilde{N}_F(U_{(s)}) - \tilde{N}_F(U_{(s)}-)$ for $s = 1,\ldots,r$ (see Wang (5)). Under this regularity condition, the NPMLE of $F(x)$ from L_1 is uniquely determined and given by

$$\tilde{F}_n(x) = 1 - \prod_{u\le x}\Big[1-\frac{d\tilde{N}_F(u)}{\tilde{R}_n(u)}\Big],$$

where $d\tilde{N}_F(u) = \tilde{N}_F(u) - \tilde{N}_F(u-)$.

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

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

Next, let $\tilde{R}_n^t(u) = \sum_{i=1}^{n} I_{[V_i\le u\le X_i, V_i=V_{(t)}]}$ and $\tilde{N}_Q^t(u) = \sum_{i=1}^{n} I_{[X_i\le u,\delta_i=0,V_i=V_{(t)}]}$. For each $V_{(t)}$, a necessary and sufficient condition for the existence of the NPMLE of $Q(x|V_{(t)})$ is $\tilde{R}_n^t(C_{(l(t))}) > c_{l(t)} = \tilde{N}_Q^t(C_{l(t)}) - \tilde{N}_Q^t(C_{l(t)}-)$ 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

$$\tilde{Q}_n(x|V_{(t)}) = 1 - \prod_{u\le x}\Big[1-\frac{d\tilde{N}_Q^t(u)}{\tilde{R}_n^t(u)}\Big]$$

where $d\tilde{N}_Q^t(u) = N_Q^t(u) - N_Q^t(u-)$.

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

$$\tilde{Q}_n(x) = \sum_{t=1}^{q}\tilde{Q}_n(x|V_{(t)})[\tilde{G}_n(V_{(t)}) - \tilde{G}_n(V_{(t-1)})].$$

Note that when the bivariate distribution of (C_i, V_i^*) is continuous, the NPMLE of $Q(x|V_{(t)})$ does not exist.

Shen (9) 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

$$\hat{F}_e(x) = \left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{\delta_i I_{[X_i \le x]}}{\hat{G}_e(X_i) - \hat{Q}_e(X_i-)}, \tag{3.1}$$

$$\hat{G}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{I_{[V_i \le x]}}{1 - \hat{F}_e(V_i-)}, \tag{3.2}$$

and

$$\hat{Q}_e(x) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_e(V_i-)}\right]^{-1} \sum_{i=1}^{n} \frac{(1-\delta_i) I_{[X_i \le x]}}{1 - \hat{F}_e(X_i-)}. \tag{3.3}$$

Shen (9) showed the equivalence of $\tilde{F}_n$ and $\hat{F}_e$, and hence, the equivalence of $\tilde{G}_n$ and $\hat{G}_e$. However, the equivalence of $\tilde{Q}_n$ and $\hat{Q}_e$ does not hold. The justification of using $\hat{Q}_e$ is given in Shen (9).

For the truncation probability $\alpha = P(U^* \ge V^*)$, a proper estimate is

$\tilde{\alpha}_n = \int \tilde{G}_n(s) d\tilde{F}_n(s)$. Instead, under the assumption (C_i, V_i^*) is independent of U_i^* and $P(V_i^* < C_i) = 1$, we have

$$\frac{1}{n}\tilde{R}(x) = P(V_i \le x \le X_i) = P(V_i^* \le x \le \min\{U_i^*, C_i\} | V_i^* \le U_i^*)$$

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

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

For all x such that $\tilde{R}_n(x) > 0$, we can obtain an alternative estimator for α as

$$\hat{\alpha}_e(x) = n[\hat{G}_e(x) - \hat{Q}_e(x-)][1 - \hat{F}_e(x-)]/\tilde{R}_n(x).$$

To derive the explicit relationship between $\tilde{\alpha}_n$ and $\hat{\alpha}_e(x)$, 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 $\tilde{\alpha}_d = \int [\tilde{G}_n(x) - \hat{Q}_e(x-)] d\tilde{F}_n(x)$. For $\tilde{R}_n(x) > 0$, let

$$\hat{\alpha}_d(x) = n_d[\hat{G}_e(x) - Q_e(x-)][1 - \hat{F}_e(x-)]/\tilde{R}_n(x),$$

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

Lemma 3.1.

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

Proof:

By (3.1), we have

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

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

$$= n_d \left[\sum_{i=1}^{r} \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)}\right]^{-1}. \tag{3.4}$$

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

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

Hence,

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

The proof is completed.

Lemma 3.2.

Suppose that $\tilde{R}_n(U_{(i)}) > 0$ for $i = 1 \ldots, r$. Then $\hat{\alpha}_e(U_{(i)}) = \hat{\alpha}_e(U_{(1)})$ for $i = 2, \ldots, r$. **Proof:**

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

$$\hat{\alpha}_e(U_{(i)}) = \frac{n}{n_d} \hat{\alpha}_d(U_{(i)}) = \frac{n}{n_d} \tilde{\alpha}_d.$$

The proof is completed.

Lemma 3.3.

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

Proof: By (3.2) and (3.3), we have

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

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

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

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

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

By (3.4),

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

The proof is completed.

Lemma 3.4.

Suppose that $\tilde{R}_n(U_{(i)}) > 0$ and $R_n(V_{(j)}) > 0$ for $i = 1, \ldots, r$ and $j = 1, \ldots, t$. Then $\hat{\alpha}_e(U_{(i)}) = \hat{\alpha}_e(V_{(j)})$ 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 $B(x) = [1-\hat{F}_e(x-)]/\tilde{R}_n(x)$.

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

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

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

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 (3.2) and (3.3), we have

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

Note that

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

Hence,

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

Next,

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

Note that

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

Hence,

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

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

Proof:

Note that the jumps of $\hat{\alpha}_e(x)$ 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, $\hat{\alpha}_e(U_{(i)}) = \hat{\alpha}_e(V_{(t)}) = \alpha_e(C_{(l)})$ for $i = 1, \ldots, r$, $t = 1, \ldots, q$ and all $C_{(l)} \leq U_{(r)}$, it follows that $\hat{\alpha}_e(x)$ is constant for any $x \in [V_{(1)}, U_{(r)}]$.

Lemma 3.7.

Suppose that $\tilde{R}_n(U_{(i)}) > 0$ for $i = 1 \ldots, r$. Then for all $i = 1 \ldots, r$.

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

Proof:

$$\begin{aligned}
\tilde{\alpha}_n = \int \hat{G}_n(x) d\hat{F}_n(x) = \int \hat{G}_e(x) d\hat{F}_e(x) &= \int [\hat{G}_e(x) - \hat{Q}_e(x-)] d\hat{F}_e(x) + \int \hat{Q}_e(x-) d\hat{F}_e(x) \\
&= \hat{\alpha}_d(U_{(i)}) + \left[\sum_{i=1}^{r} \frac{d_i}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)} \right]^{-1} \sum_{i=1}^{r} \frac{d_i \hat{Q}_e(U_{(i)}-)}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)} \\
&= \frac{n_d}{n} \hat{\alpha}_e(U_{(i)}) + \frac{\hat{\alpha}_e(U_{(i)})}{n} \left[\sum_{i=1}^{r} \frac{d_i \hat{G}_e(U_{(i)})}{\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)} - n_d \right].
\end{aligned}$$

The proof is completed.

Note that

$$\sum_{i=1}^{r} d_i \hat{G}_e(U_{(i)})/[\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)] = \sum_{i=1}^{n} \delta_i \hat{G}_e(X_i)/[\hat{G}_e(X_i) - \hat{Q}_e(X_i-)].$$

Since $E[\delta_i G(X_i)|X_i] = P(X_i \leq C_i)P(V_i^* \leq X_i|X_i \leq C_i) = P(V_i^* \leq X_i \leq C_i) = G(X_i) - Q(X_i-)$, hence, $n_d/n + 1/n(\sum_{i=1}^{r} d_i \hat{G}_e(U_{(i)})/[\hat{G}_e(U_{(i)}) - \hat{Q}_e(U_{(i)}-)] - n_d)$ actually estimates 1.

4 SIMULATION STUDY

For left-truncated and right-censored data, a simulation study is conducted to examine the performance of the $\tilde{\alpha}_n$ and $\hat{\alpha}_e$. The U_i^*'s are exponentially distributed: $F(x) = 1 - e^{-x}$ for $x > 0$. The V_i^*'s are Weibull distribution: $G(x) = 1 - e^{-x^\beta}$ for $x > 0$, with varying parameters $\beta = 0.5, 1.0$, and 2.0. The C_i's are defined by $C_i = D_i^* + V_i^*$, where D_i^*'s are independent of V_i^* and exponentially distributed: $P(D_i^* \leq x) = 1 - e^{-x}$ for $x > 0$. Hence, $P(V_i^* < C_i) = 1$ The sample size is chosen as 100 and 200 and the replication is 5000 times. The estimator $\hat{F}_e(x) = \tilde{F}_n(x)$ is obtained based on the product-limit form. The estimators $\hat{G}_e(x)$ and $\hat{Q}_e(x)$ are obtained based on (3.2) and (3.3). To demonstrate the performances of $\hat{F}_e(x)$, $\hat{G}_e(x)$ and $\hat{Q}_e(x)$, Table 1 shows the biases, standard deviation (std) and squared root of mean squared error ($\sqrt{mse}$) of the three estimators at $x = 1.0$ and $x = 2.0$. Table 1 also shows the proportion of truncation (α) and the proportion of death ($p = n_d/n$). Table 2 shows the biases, standard deviation (std) and squared root of mean squared error ($\sqrt{mse}$) of the the two estimators $\hat{\alpha}_e(x)$ and $\tilde{\alpha}_n$. To obtain $\hat{\alpha}_e(x)$, the values of x are set at $U_{(50)}$ and $U_{(100)}$ for $n = 100$ and $n = 200$, respectively.

Table 1 shows that the three estimators $\hat{F}_e(x)$, $\hat{G}_e(x)$ and $\hat{Q}_e(x)$ work satisfactorily for moderate sample size. Table 2 shows that the $\tilde{\alpha}_n$ is less biased than $\hat{\alpha}_e$. The mean-squared error of $\tilde{\alpha}_n$ is smaller than that of $\hat{\alpha}_e$ for all the cases considered.

Table 1. Simulation results for biases, std and $\sqrt{mse}$
of the estimators $\hat{F}_e(x)$, $\hat{G}_e(x)$ and $\hat{Q}_e(x)$

				$\hat{F}_e(1.0)$			$\hat{G}_e(1.0)$			$\hat{Q}_e(1.0)$		
β	n	α	p	bias	std	mse	bias	std	mse	bias	std	mse
0.5	100	0.55	0.72	-0.021	0.062	0.065	0.121	0.120	0.171	0.116	0.079	0.140
0.5	200	0.55	0.72	-0.020	0.041	0.045	0.106	0.099	0.145	0.107	0.063	0.125
1.0	100	0.50	0.75	-0.032	0.073	0.080	0.022	0.089	0.092	0.007	0.062	0.062
1.0	200	0.50	0.75	-0.030	0.055	0.063	0.013	0.067	0.068	0.001	0.042	0.042
2.0	100	0.44	0.77	-0.046	0.126	0.134	0.017	0.083	0.085	-0.074	0.045	0.087
2.0	200	0.44	0.77	-0.027	0.115	0.118	0.005	0.082	0.082	-0.071	0.044	0.085
				$\hat{F}_e(2.0)$			$\hat{G}_e(2.0)$			$\hat{Q}_e(2.0)$		
β	n	α	p	bias	std	mse	bias	std	mse	bias	std	mse
0.5	100	0.55	0.72	-0.012	0.059	0.060	0.023	0.126	0.128	0.068	0.117	0.135
0.5	200	0.55	0.72	-0.013	0.043	0.045	0.002	0.108	0.108	0.051	0.088	0.102
1.0	100	0.50	0.75	-0.016	0.052	0.054	0.020	0.089	0.092	0.008	0.087	0.088
1.0	200	0.50	0.75	-0.015	0.035	0.038	0.006	0.077	0.078	-0.005	0.063	0.063
2.0	100	0.44	0.77	-0.018	0.059	0.062	0.115	0.033	0.117	0.021	0.135	0.135
2.0	200	0.44	0.77	-0.008	0.046	0.047	0.112	0.018	0.115	0.010	0.120	0.120

Table 2. Simulation results for biases, std and $\sqrt{mse}$
of the estimators $\hat{\alpha}_e$ and $\tilde{\alpha}_n$

				bias		std		$\sqrt{mse}$	
β	n	α	p	$\hat{\alpha}_e$	$\tilde{\alpha}_n$	$\hat{\alpha}_e$	$\tilde{\alpha}_n$	$\hat{\alpha}_e$	$\tilde{\alpha}_n$
0.5	100	0.55	0.72	0.1026	0.0904	0.1129	0.1202	0.1526	0.1505
0.5	200	0.55	0.72	0.0948	0.0818	0.0856	0.0895	0.1277	0.1212
1.0	100	0.50	0.75	0.0389	0.0192	0.0924	0.0963	0.1003	0.0982
1.0	200	0.50	0.75	0.0228	0.0087	0.0688	0.0691	0.0726	0.0697
2.0	100	0.44	0.77	0.0378	0.0143	0.1400	0.1328	0.1450	0.1336
2.0	200	0.44	0.77	0.0244	0.0069	0.1112	0.1074	0.1138	0.1077

References

[1] 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.

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

[3] 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.

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

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

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

[7] 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.

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

[9] Shen, P.-S. The product-limit estimates as an inverse-probability-weighted average. Communi. in Statist., Part A- Theory and Methods, **2003**, *32*, to appear.

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

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

SAS SIMULATION PROGRAM