東海管理評論
民國九十年七月，第三卷，第一期，173-186

Jackknife Methods for Truncated Data

Pao-Sheng Shen[*] *Jacle Lin*[**]

Abstract

Let X and Y be two independent positive random variables with survival functions $\overline{F}$ and $\overline{G}$, respectively. Under random **truncation**, X and are both observable only when X is large than Y . The nonparametric MLE of $\overline{F}(x)$, $\overline{F}_n(x) = \Pi_{z \le x}[1 - d\Lambda_n(z)]$, was derived by Lynden-Bell（1971）, where $\Lambda_n(z)$ is the estimated cumulative hazard function . In this note, we derive an explicit formula for the delete-d **jackknife** estimate of $\Lambda_n(z)$. From this it is demonstrated that jackknifing may lead to a reduction of the bias. Besides, it is shown that the delete-1 jackknife variance estimator of $\overline{F}_n(x)$ consistently estimates the limit variance.

Keywords：Truncation, Jackknife.

1. Introduction

Let X and Y be two independent positive random variables with survival functions $\overline{F}$ and $\overline{G}$, respectively. Under random truncation, both X and Y are observable only when $X \ge Y$. Truncated data occur in astronomy,（e.g., Lynden-Bell, 1971）, epidemiology, biometry （see Wang, Jewell and Tsai, 1986）and possibly in other field such as economics.

Let (U_1, V_1), … , (U_n, V_n) denote the truncated sample. Let $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ be the ordered values of U_k and $V_{(k)}$, the concomitant of $U_{(k)}$ for $k = 1, \ldots, n$. The nonparametric MLE of $\overline{F}(x)$, $\overline{F}_n(x) = 1 - F_n(x) = 1 - \Pi_{z \le x}[1 - d\Lambda_n(z)]$, was derived by Lynden-Bell （1971）, where $\Lambda_n(z)$

---

* Professor of Department of Statistics, Tunghai University
** Unisys Taiwan Limited.

$=\sum_{U_{(k)}\le z}(1/n_k)$, $n_k=\sum_{j=k}^{n} I_{[V_{(j)}<U_{(k)}]}$, and $I_{[A]}$ is the indicator function of the event A. In Section 2, we derive an explicit formula for the delete-d jackknife estimator of $\Lambda_n(z)$. From this it is demonstrated that jackknifing may lead to a reduction of the bias. In Section 3, it is shown that the delete-1 jackknife variance estimator of $\overline{F}_n$ consistently estimates the limit variance. Simulation studies are conducted to compare confidence limits for the survival probability $\overline{F}(x)$ obtained via the delete-d jackknife with the Greenwood's formula (Tsai, Jewell and Wang, 1987).

2. Bias Reduction

Let a_f and a_g denote the lower boundaries of X and Y. Woodroofe (1985) showed that when $a_g \le a_f$, $\Lambda_n(x)$ underestimates $\Lambda(x)$ and the bias is $\int_0^x[1-C(z)]^n d\Lambda(z)$, where $C(z)=G(z)\{[1-F(z)]/P\}$ $(X \ge Y)$. As pointed out by Woodroofe(1985), although the bias of $\Lambda_n(x)$ converges to zero, but may do so arbitrarily slow. To reduce the bias of the estimate $\Lambda_n(x)$, we consider the delete-d jackknife estimator of $\Lambda_n(x)$. Let $D_{n,d}$ be the collection of subjects of $\{1,2,...,n\}$ which have size $n-d$, and $d>0$ is an integer less than n. For any $g=\{j_1,...,j_{n_d}\}\in D_{n,d}$, define $\Lambda_{n,g}(z)=\sum_{j\in g,U_{(j)}} 1/n_j$, for $0\le z\le\infty$.

Define $\overline{\Lambda}_{J(d)}(x)=\left[1/\binom{n}{d}\right]\sum_g \Lambda_{n,g}(x)$, where Σ_g denotes the summation over all the subsets in $D_{n,d}$. The delete-d jackknife estimator of $\Lambda_{J(d)}(x)$ (see Efron, 1982, p.7) is

$$\Lambda_{J(d)}(x)=\frac{n}{d}\Lambda_n(x)-(\frac{n}{d}-1)\overline{\Lambda}_{J(d)}(x).$$

Given $n_k>1$ and $d<n_k$, Lemma 2.1 derives the explicit form of $\overline{\Lambda}_{J(d)}(U_{(k)})$ $(k=1,...,n-1)$, where $U_{(k)}$ $(k=1,...,n-1)$ is computed from the

original sample and

Lemma 2.1.

Let $\overline{\Lambda}_{J(d)}(U_{(0)})=0$. For $k=1,\ldots,n-1$, given $n_k>1$ and $d<n_k$,

$$\overline{\Lambda}_{J(d)}(U_{(k)})=\overline{\Lambda}_{J(d)}(U_{(k-1)})+\frac{1}{n_k}.$$

Proof :

For $k=1,\ldots,n-1$, given $n_k>1$ and $d<n_k$, we have

$$\begin{aligned}
&\overline{\Lambda}_{J(d)}(U_{(k)})\\
&=\overline{\Lambda}_{J(d)}(U_{(k-1)})+\binom{n}{d}^{-1}\sum_{s=\max(0,d-(n-n_k))}^{\min(d,n_k-1)}\binom{n_k-1}{s}\binom{n-n_k}{d-s}\frac{1}{n_k-s}\\
&=\binom{n}{d}^{-1}\sum_{s=\max(0,d-(n-n_k))}^{d}\binom{n_k}{s}\binom{n-n_k}{d-s}\frac{1}{n_k}\\
&=\frac{1}{n_k}
\end{aligned}$$

This concludes the proof of Lemma **2.1.**.

Given $n_k>1$ and $d<n_k$, Lemma **2.2.** derives the explicit form of $\overline{\Lambda}_{J(d)}(U_{(k)})$ $(k=1,\ldots,n-1)$.

Lemma 2.2.

For $k=1,\ldots,n-1$, given $n_k>1$ and $d\geq n_k$,

$$\overline{\Lambda}_{J(d)}(U_{(k)})=\overline{\Lambda}_{J(d)}(U_{(k-1)})+\frac{1}{n_k}-\binom{n}{d}^{-1}\binom{n-n_k}{d-n_k}\frac{1}{n_k}.$$

Proof：

For $k=1,\dots,n-1$, given $n_k>1$ and $d\geq n_k$, we have

$$\bar{\Lambda}_{J(d)}(U_{(k)})$$

$$=\bar{\Lambda}_{J(d)}(U_{(k-1)})+\binom{n}{d}^{-1}\sum_{s=\max(0,d-(n-n_k))}^{\min(d,n_k-1)}\binom{n_k-1}{s}\binom{n-n_k}{d-s}\frac{1}{n_k-s}$$

$$=\bar{\Lambda}_{J(d)}(U_{(k-1)})+\binom{n}{d}^{-1}\sum_{s=0}^{n_k-1}\binom{n_k}{s}\binom{n-n_k}{d-s}\frac{1}{n_k}$$

$$=\bar{\Lambda}_{J(d)}(U_{(k-1)})+\frac{1}{n_k}-\binom{n}{d}^{-1}\binom{n-n_k}{d-n_k}\frac{1}{n_k}$$

This concludes the proof of Lemma 2.2..

Next, Lemma 2.3. derives the explicit form of $\bar{\Lambda}_{J(d)}(U_{(n)})$.

Lemma 2.3.

For $k=n$, we have

$$\bar{\Lambda}_{J(d)}(U_{(n)})$$

$$=\bar{\Lambda}_{J(d)}(U_{(n-1)})+\binom{n}{d}^{-1}\binom{n-1}{d}$$

$$=\bar{\Lambda}_{J(d)}(U_{(n-1)})+\frac{n-d}{n}$$

According to Lemma 2.1., 2.2. and 2.3., the following theorem derives the explicit form of the delete-d jackknife estimator, $\Lambda_{J(d)}(U_{(k)})$.

Theorem 2.1.

Given $n_k > 1$ $(k=1,\dots,n-1)$, the delete-d jackknife estimator, $\Lambda_{J(d)}(U_{(k)})$, for $k=1,\dots,n$ is given by

$$\Lambda_{J(d)}(U_{(k)}) = \Lambda_n(U_{(k)}) + \sum_{j=1}^{k} B_j ,$$

where $U_{(k)}$ $(k=1,\dots,n)$ is computed from the original sample and

$$B_j = \begin{cases} 0, & \text{for } d < n_j ; \\ \dfrac{n-d}{d}\dbinom{n}{d}^{-1}\dbinom{n-n_j}{d-n_j}\dfrac{1}{n_j}, & \text{for } d \ge n_j . \end{cases}$$

Proof :

Theorem **2.1.** follows from Lemma **2.1.** and Lemma **2.2.** upon

$$\overline{\Lambda}_{J(d)}(U_{(k)}) = \frac{n}{d}\overline{\Lambda}_{Jn}(U_{(k)}) - (\frac{n}{d}-1)\overline{\Lambda}_{J(d)}(U_{(k)}) \quad (k=1,\dots,n)$$

Next, we report on a small simulation study which is likely to demonstrate the impact of jackknifing procedure. The distributions for $X_i's$ are exponential: $X_i \sim \exp(1)$, The distribution for $Y_i's$ are Weibull: $Y_i \sim W(\beta,\delta)$, that is, $G(y) = 1 - e^{-(y/\beta)^{\delta}}$ for $y > 0$, with varing parameters $\beta = 0.25, 1.0, 4.0$, and $\delta = 1.0, 4.0$. We consider the estimation of survival function $\overline{F}(2) = e^{-2}$ $= 0.135$. The sample size is chosen as 25and the replication is 3,000 times. The delete-d jackknife estimator of $\overline{F}_n(x)$ is

$$\overline{F}_{J(d)}(x) = \frac{n}{d}\overline{F}_n(x) - (\frac{n}{d}-1)\frac{1}{\binom{n}{d}}\sum_{g}\overline{F}_{n,g}(x) .$$

where $\bar{F}_{n,g}(x)=1-\Pi_{j\in g,U_{(j)}\le x}[1-(1/n_j)]$. The d (denoted by $\hat{d}$) is chosen such that B_1 (see Theorem 2.1.) is maximized. 《Tab.1》shows the value of β_g, δ_g, $\hat{d}$, biases and mean-squared errors of $\bar{F}_n(2)$ and $\bar{F}_{J(\hat{d})}(2)$. Simulation results demonstrate that $\bar{F}_n(2)$ overestimates $\bar{F}(2)$. Jackknifing leads to a reduction of bias and the reduction is substantial for the case $\beta=4$ and $\delta=4$.

《Tab.1》 Simulation results of $\bar{F}_n(2)$ and $\bar{F}_{J(\hat{d})}(2)$ for $Y_i \sim W(\beta,\delta)$

n	β	δ	$\hat{d}$	*bias*		*mse*	
				$\bar{F}_n(2)$	$\bar{F}_{J(\hat{d})}(2)$	$\bar{F}_n(2)$	$\bar{F}_{J(\hat{d})}(2)$
25	0.25	1.0	22	0.002	-0.005	0.005	0.005
25	0.25	4.0	20	0.015	0.007	0.006	0.006
25	1.0	1.0	20	0.006	0.001	0.005	0.005
25	1.0	4.0	18	0.057	0.032	0.016	0.014
25	4.0	1.0	19	0.008	0.000	0.006	0.006
25	4.0	4.0	13	0.176	0.108	0.102	0.084

3. Estimation of variance

The delete-d jackknife variance estimator of $\bar{F}_n(x)$ is

$$V_d(\bar{F}_n(x))=\frac{n-d}{d\binom{n}{d}}\sum_g\left[\bar{F}_{n,g}(x)-\frac{1}{\binom{n}{d}}\sum_g\bar{F}_{n,g}(x)\right]^2.$$

In this section, it will be shown that the delete-1 jackknife variance estimator $V_1(\bar{F}_n(x))$ converges almost surely to the limit of the variance of $\bar{F}_n(x)$.

Since the estimator $\bar{F}_n(x)$ is closely related to the sample cumulative hazard function $\Lambda_n(x)$ it is convenient to start with the delete-1 jackknife variance estimator of $\Lambda_n(U_{(k)})$ $(k=1,\ldots,n)$

$$V_d\left(\Lambda_n(U_{(x)})\right)=\frac{1}{n}\sum_{m=1}^{n}\left[\Lambda_{n,m}(U_{(x)})-\overline{\Lambda}_{J(1)}(U_{(x)})\right]^2 .$$

where $\overline{\Lambda}_{J(1)}(U_{(k)})=(1/n)\sum_{m=1}^{n}\Lambda_{n,m}(U_{(k)})$ and $\Lambda_{n,m}(U_{(k)})$ denotes the delete-1 estimator of $\Lambda_n(U_{(k)})$ when $U_{(m)}$ $(m=1,\ldots,n)$ is deleted from the sample.

From Theorem **2.1.**, given $n_k>1$ $(k=1,\ldots,n-1)$, $\Lambda_{n,m}(U_{(k)})$ $(i=1,\ldots,n)$, is given by

$$\overline{\Lambda}_{J(1)}(U_{(k)})=\begin{cases}\Lambda_n(U_{(k)}), & \textbf{for } k=1,\ldots,n-1;\\ \Lambda_n(U_{(k)})-\dfrac{1}{n}, & \textbf{for } k=n.\end{cases}$$

Now, we shall show that the delete-1 jackknife variance estimator of $\sqrt{n}\Lambda_n(x)$ converges almost surely to the correct variance.

For $k=1,\ldots,n-1$, the delete-1 jackknife variance estimate of $\sqrt{n}\Lambda_n(U_{(k)})$ is given by

$$\begin{aligned}
&nV_1\left(\Lambda_{(n)}(U_{(k)})\right)\\
&=(n-1)\sum_{m=1}^{k}\left[\sum_{i=1}^{m-1}\frac{1}{n_i-\delta_{im}}+\sum_{i=m+1}^{k}\frac{1}{n_i}-\sum_{i=1}^{k}\frac{1}{n_i}\right]^2\\
&\quad+(n-1)\sum_{m=k+1}^{n}\left[\sum_{i=1}^{k}\frac{1}{n_i-\delta_{im}}-\sum_{i=1}^{k}\frac{1}{n_i}\right]^2\\
&=(n-1)\sum_{m=1}^{k}\left[\sum_{i=1}^{m-1}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}-\frac{1}{n_m}\right]^2\\
&\quad+(n-1)\sum_{m=k+1}^{n}\left[\sum_{i=1}^{k}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}\right]^2\\
&=(n-1)\sum_{m=1}^{k}\frac{1}{n_m^2}+(n-1)\sum_{m=1}^{k}\left[\sum_{i=1}^{m-1}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}\right]^2
\end{aligned}$$

$$+(n-1)\sum_{m=k+1}^{n}\left[\sum_{i=1}^{k}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}\right]^2-2(n-1)\sum_{i=1}^{k}\frac{1}{n_m}\sum_{i=1}^{m-1}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}$$

$$=(n-1)\sum_{m=1}^{k}\frac{1}{n_m(n_m-1)}$$

$$+2(n-1)\sum_{m=1}^{k}\sum_{i=1}^{m-1}\sum_{j=m+1}^{n}\frac{\delta_{mj}\delta_{ij}}{n_m n_i(n_m-\delta_{mj})(n_i-\delta_{ij})}$$

$$-2(n-1)\sum_{i=1}^{k}\frac{1}{n_m}\sum_{i=1}^{m-1}\frac{\delta_{im}}{n_i(n_i-\delta_{im})}$$

$$=(n-1)\sum_{m=1}^{k}\frac{1}{n_m(n_m-1)}+2(n-1)\sum_{m=1}^{k}\sum_{i=1}^{m-1}\frac{\sum_{j=m+1}^{n}(\delta_{ij}-\delta_{mj}\delta_{im})}{n_m(n_m-1)n_i(n_i-1)} \quad \cdots\cdots \quad (1)$$

The first term of (1), $(n-1)\sum_{m=1}^{k}\{1/[n_m(n_m-1)]\}$,is the analogue of Greenwood's formula (Tsai, Jewell and Wang, 1987) and converges almost surely to the asymptotic variance of $\sqrt{n}\Lambda_n(x)$ (see Wang, Jewell and Tsai, 1986), namely, $\int_0^x[dH(z)/C^2(z)]$, where $H(z)=P(X_i\le z\mid X_i\ge Y_i)$. The second term of (1) can be written as

$$2\sum_{i=1}^{k-1}\frac{1}{n_i(n_i-1)}\left[(n-1)\sum_{m=i+1}^{k}\frac{\frac{q_{im}}{(n_m-1)}-\delta_{im}}{n_m}\right], \quad \cdots\cdots\cdots\cdots\cdots \quad (2)$$

where $q_{im}=\sum_{j=m+1}^{n}\delta_{ij}$ $(i=1,\ldots,m-1,\ m=1,\ldots,k)$.

Since

$$E\left[\frac{q_{im}}{(n_m-1)}-\delta_{im}\middle|\, n_m\right]=\frac{\sum_{j=m+1}^{n}P(U_{(i)}>V_{(i)}\mid n_m)}{n_m-1}-P(U_{(i)}>V_{(i)}\mid n_m),$$

as $n \to \infty$ and $k/(n-p)$ $(0<p<1)$, for $i=1,\ldots,m-1$, $m=1,\ldots,k$, we have

$$\sum_{m=i+1}^{k} \frac{\frac{q_{im}}{n_m - 1} - \delta_m}{n_m} = O_p(n^{-1}).$$

Hence, (2) converges almost surely to zero and the delete-1 jackknife variance estimator, $nV_i(\Lambda_n(U_{(k)}))$ $(k=1,\ldots,n-1)$, converges almost surely to the limit variance of $\sqrt{n}\Lambda_n(U_{(k)})$

In order to study the estimate $\bar{F}_n(x)$, expand the logarithm

$$\ln \bar{F}_n(x) = -\Lambda_n(x) + \frac{1}{2}\sum_{U(i)\le x} \frac{1}{n_i^2} - \ldots, \qquad \text{(3)}$$

Now, jackknife and observe that the result of jackknifing the second and higher terms of (3) lead to expressions which are $o_p(1/n)$. Hence, the jackknife version of $\ln \bar{F}_n(x)$ has the same asymptotic (normal) distribution as $-\Lambda_n(x)$. Since $\exp[\ln \bar{F}_n(x)] = \bar{F}_n(x)$ and the exponential function is smooth, the difference between $V_1(\bar{F}_n(x))$ and $[\bar{F}_n(x)]^2 V_1(\Lambda_n(x))$ will tend to zero as n tends to infinity. Hence, the delete-1 jackknife estimate of variance of $\ln \bar{F}_n(x)$, $V_1(\bar{F}_n(x))$ converges almost surely to the correct variance.

We report on the results of some simulation investigations, comparing confidence limits for the survival probability $\bar{F}(x)$ obtained via the delete-d jackknife with the Greenwood's formula.

Using jackknife method an approximate $1-2\alpha$ confidence interval for $\bar{F}(x)$ is given by

$$\overline{F}_{J(d)}(x) \pm t_{\alpha,n-1}\sqrt{V_d\left(\overline{F}_n(x)\right)},$$

where $t_{\alpha,n-1}$ is the α upper percentile point of a t distribution with $n-1$ degrees of freedom.

Similarly, using Greenwood's formula an approximate $1-2\alpha$ confidence interval for $\overline{F}(x)$ can be constructed as

$$\overline{F}_n(x) \pm z_\alpha\sqrt{V_G\left(\overline{F}_n(x)\right)}.$$

where $V_G\left(\overline{F}_n(x)\right)=\left[\overline{F}_n(x)\right]^2\sum_{U(i)\le x}\left[1/n_i(n_i-1)\right]$ and z_α is the α upper percentile point of the standard normal distribution.

The $X_i's$ and $Y_i's$ distributions are the same as those used in Section **2**. The values of n, x and d are chosen as 25, 2 and 13, respectively. The significance level, α, is set at 0.025 and the replication is 3,000 times. 《Tab.2》 shows the results of the empirical coverages (E.C.) of confidence intervals based on the three estimators $V_G\left(\overline{F}_n(2)\right)$, $V_1\left(\overline{F}_n(2)\right)$ and $V_{\hat{d}}\left(\overline{F}_n(2)\right)$, which are denoted by C_G, C_1 and $C_{\hat{d}}$, respectively. 《Tab.2》 also shows the relative bias $\hat{B}_i/\hat{\Sigma}_i$, where $\hat{\Sigma}_i$ denotes the observed empirical variance of $\overline{F}_n(x)$, and $\hat{B}_i$ is the empirical bias of the estimator.

《Tab.2》 Empirical coverages (confidence level = 0.95) and relative biases for $n=25$, $x=2$, $Y_i \sim W(\beta_g, \delta_g)$

β	δ	$\hat{d}$	$\hat{B}_i/\hat{\Sigma}_i$			E.C.		
			$V_G(\overline{F}_n(2))$	$V_1(\overline{F}_n(2))$	$V_{\hat{d}}(\overline{F}_n(2))$	C_G	C_1	$C_{\hat{d}}$
0.25	1.0	13	-0.001	0.153	0.102	0.915	0.932	0.930
0.25	4.0	13	-0.112	0.075	0.052	0.907	0.918	0.921
1.0	1.0	13	-0.117	0.174	0.168	0.884	0.917	0.915
1.0	4.0	13	-0.419	0.192	0.177	0.784	0.881	0.899
4.0	1.0	13	-0.124	0.254	0.216	0.870	0.904	0.906
4.0	4.0	13	-0.538	0.532	0.375	0.520	0.696	0.753

The results of 《Tabl.2》 can be summarized as follows. $V_G(\overline{F}_n(2))$ underestimates, $V_1(\overline{F}_n(2))$ and $V_{\hat{d}}(\overline{F}_n(2))$ overestimate the variance of $\overline{F}_n(2)$. Compared to $V_1(\overline{F}_n(2))$, $V_{\hat{d}}(\overline{F}_n(2))$ have the advantages of smaller bias. The C_G is worse than C_1 and $C_{\hat{d}}$. The C_1 and $C_{\hat{d}}$ are very close except for $\beta_g = \delta_g = 4$, which is the case when the bias of $\overline{F}_{J(\hat{d})}(2)$ is smaller than that of $\overline{F}_{J(1)}(2)$.

References

Efron, B. (1982), "The jackknife, the bootstrap and other resampling plans", *Society for industrial and applied mathematics*. Philadelphia, Pennsylvania 19103.

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

Tsai, W.Y., N.P. Jewell, and M.C. Wang (1987), "A note on the product-limit estimator under right censoring and left truncation", *Biometrika*, 74 : 883∼

886.

Tukey, J. (1958), "Bias and confidence in not quite large samples", *Ann. Math. Statist.*, 29：614.

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

Woodroofe, M. (1985), "Estimating a distribution function with truncated data", *Ann. Statist.*, 13：163～177.

截取資料下之摺刀法

沈葆聖* 林明杰**

摘要

令 X 和 Y 分別表示兩獨立之連續變數具存活函數 $\overline{F}$ 和 $\overline{G}$。在**截取資料**下，僅當 $X \geq Y$ 時，方可同時觀察到 X 和 Y。Lynden-Bell（1971）提出 $\overline{F}(x)$ 的非參數最大概似估計值（nonparametric MLE），$\overline{F}_n(x) = \Pi_{z \leq x}\left[1 - d\Lambda_n(z)\right]$，其中 $\Lambda_n(z)$ 為累積危險函數估計值。本文中，我們推導去除 d 個的 $\Lambda_n(z)$ 摺刀估計值。依此，減低 $\overline{F}_n(x)$ 的估計偏差。此外，證明 $\overline{F}_n(x)$ 去除一個的摺刀雙方估計收斂至真正變方。

關鍵詞：截取資料，摺刀法。

* 東海大學統計系教授

** 台灣優利系統股份有限公司