Journal of the Korean Statistical Society (2003), **32**: *1, pp 25–32*

ON THE EMPIRICAL MEAN LIFE PROCESSES FOR RIGHT CENSORED DATA

HYO-IL PARK[1]

ABSTRACT

In this paper, we define the mean life process for the right censored data and show the asymptotic equivalence between two kinds of the mean life processes. We use the Kaplan-Meier and Susarla-Van Ryzin estimates as the estimates of survival function for the construction of the mean life processes. Also we show the asymptotic equivalence between two mean residual life processes as an application and finally discuss some difficulties caused by the censoring mechanism.

AMS 2000 subject classifications. Primary 62G20.

Keywords. Empirical mean life process, empirical mean residual life process, Kaplan-Meier estimate, right censored data, survival function, Susarla-Van Ryzin estimate.

1. INTRODUCTION

The estimation of the mean survival time for right censored data has been long considered by many authors. However the results have not been so satisfactory because of the possibility that the largest observation may be censored. Only Susarla and Van Ryzin (1980), and Gill (1983) obtained the most successful results using the different estimates for survival function. Susarla and Van Ryzin used a variant of the Bayesian estimates proposed by Susarla and Van Ryzin (1976) whereas Gill used the Kaplan-Meier estimate. We note that the two estimates are asymptotically equivalent in the sense that the results for the asymptotic normality coincide. For the asymptotic normality, Gill applied conditions in a natural way but Susarla and Van Ryzin introduced a set of assumptions in a complicated manner. Furthermore, since Gill used the Kaplan-Meier estimate, the martingale theory based on the point processes could be adopted for

Received June 2002; accepted October 2002.

[1]Department of Statistics, Chongju University, Chongju 360-764, Korea (e-mail : hipark@chongju.ac.kr)

the investigation of the large sample behavior. On the other hand, if we look more closely into the Susarla-Van Ryzin estimate, then we will find that the estimate is well defined besides the biasedness problem since the values of the Susarla-Van Ryzin estimate become 0 beyond the largest observation with disregard that it is censored or not. Therefore it would be worthwhile to work with the two estimates more deeply. In this paper, we define the mean life process and show the two mean life processes based on the two estimates of survival function are asymptotically equivalent and apply this asymptotic equivalence to show the equivalence of two estimates of the mean residual life processes based on the Kaplan-Meier and Susarla-Van Ryzin estimate.

2. MAIN RESULT

Let $X_1, \ldots, X_n$ be a random sample of non-negative survival times with a continuous survival function S and $Y_1, \ldots, Y_n$, an independent random sample of non-negative censoring times with an arbitrary distribution function G. Since the right censoring schemes are involved, we only observe $(T_1, \delta_1), \ldots, (T_n, \delta_n)$, where $T_i = \min\{X_i, Y_i\}$ and $\delta_i = I(X_i \leq Y_i)$ for each i. $I(\cdot)$ is an indicator function. We assume that the survival function S has a finite mean. Then it is well-known that

$$E(X) = \int_0^\infty x dF(x) = \int_0^\infty S(x)dx$$

since the life time random variable is non-negative, where $F(\cdot) = 1 - S(\cdot)$ is the distribution function of X. Let $\hat{S}_n$ and $\tilde{S}_n$ be the Kaplan-Meier and Susarla-Van Ryzin estimates of S, respectively. Then we define two mean life processes, $\hat{\mu}_n(t)$ and $\tilde{\mu}_n(t)$ based on $\hat{S}_n$ and $\tilde{S}_n$ as follows: let for each $t \in [0, \tau)$,

$$\hat{\mu}_n(t) = \sqrt{n} \int_0^t (\hat{S}_n(u) - S(u))du$$

and

$$\tilde{\mu}_n(t) = \sqrt{n} \int_0^t (\tilde{S}_n(u) - S(u))du,$$

where $\tau = \inf\{t : 1 - H(t) = 0\}$ with the notation that $1 - H(t) = S(t)(1 - G(t))$. Also let $\tau_S = \inf\{t : S(t) = 0\}$ and $\tau_G = \inf\{t : 1 - G(t) = 0\}$ for the later use. We note that H is continuous from the assumption that S is continuous. Then we will show that the two processes, $\hat{\mu}_n(t)$ and $\tilde{\mu}_n(t)$, are asymptotically equivalent in the following sense.

THEOREM 2.1. *Let $\tau = \tau_S < \tau_G$. Then we have that for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\sup_{0\le t<\tau} |\hat{\mu}_n(t) - \tilde{\mu}_n(t)| > \epsilon\right) = 0.$$

Before we prove Theorem 2.1, first of all, we obtain a relation between $\hat{S}_n$ and $\tilde{S}_n$. We note that from equation (7.0.11, p. 295) in Shorack and Wellner (1985) that

$$\begin{aligned}\tilde{S}_n(t) &= \frac{\sum_{j=1}^n I[T_{(j)} > t]}{n} \prod_{j:T_{(j)}\le t} \left\{1 + \frac{I[\delta_{(j)} = 0]}{n-j+1}\right\} \\ &= \hat{S}_n(t)(1 - \hat{G}_n(t)) \prod_{j:T_{(j)}\le t} \left\{1 + \frac{I[\delta_{(j)} = 0]}{n-j+1}\right\} \\ &= \hat{S}_n(t) \prod_{j:T_{(j)}\le t} \left\{1 - \frac{I[\delta_{(j)} = 0]}{(n-j+1)^2}\right\},\end{aligned}$$

where $\hat{G}$ is the Kaplan-Meier estimate of G obtained by switching the roles of life time and censoring random variables, $T_{(j)}$ is the j^{th} largest observation among $T_1, \ldots, T_n$ and $\delta_{(j)}$ is the concomitant of the j^{th} order statistic $T_{(j)}$, that is $\delta_{(j)} = \delta_i$ if $T_{(j)} = T_i$. Therefore we have that

$$\begin{aligned}\hat{S}_n(t) - \tilde{S}_n(t) &= \hat{S}_n(t)\left[1 - \prod_{j:T_{(j)}\le t}\left\{1 - \frac{I[\delta_{(j)} = 0]}{(n-j+1)^2}\right\}\right] \\ &= \hat{S}_n(t)K_n(t), \text{ say.}\end{aligned}$$

Then we have

$$\begin{aligned}\hat{\mu}_n(t) - \tilde{\mu}_n(t) &= \sqrt{n}\int_0^t (\hat{S}_n(u) - \tilde{S}_n(u))du \\ &= \sqrt{n}\int_0^t \hat{S}_n(u)K_n(u)du.\end{aligned}$$

We note that $K_n(t) \ge 0$ for all $t \in [0, \tau)$. In the following lemma, we obtain a useful bound for $K_n(t)$.

LEMMA 2.1. *For every $t \in [0, \tau)$, we have*

$$K_n(t) \le \frac{k_0}{n(n-k_0+1)} \le \frac{\hat{H}_n(T_{(k_0)})}{n(1-\hat{H}_n(T_{(k_0)}-))} \le \frac{1}{n(1-\hat{H}_n(T_{(k_0)}-))},$$

where $k_0 = \max\{j : T_{(j)} \leq t, \delta_{(j)} = 0\}$ and $\hat{H}_n$ is the empirical distribution function of H.

PROOF. First of all, we note that

$$\begin{aligned}
&1 - \prod_{j=1}^{k_0} \left\{1 - \frac{1}{(n-j+1)^2}\right\} \\
&= 1 - \left\{1 - \frac{1}{n^2}\right\}\left\{1 - \frac{1}{(n-1)^2}\right\} \cdots \left\{1 - \frac{1}{(n-k_0+1)^2}\right\} \\
&= 1 - \left(1 + \frac{1}{n}\right)\left(1 - \frac{1}{n-k_0+1}\right) \\
&= \frac{k_0}{n(n-k_0+1)}.
\end{aligned}$$

Also we note that since

$$1 > \prod_{j:T_{(j)} \leq t} \left\{1 - \frac{I[\delta_{(j)} = 0]}{(n-j+1)^2}\right\} \geq \prod_{j:T_{(j)} \leq t} \left\{1 - \frac{1}{(n-j+1)^2}\right\} \geq 0,$$

we have

$$1 - \prod_{j:T_{(j)} \leq t} \left\{1 - \frac{I[\delta_{(j)} = 0]}{(n-j+1)^2}\right\} \leq 1 - \prod_{j:T_{(j)} \leq t} \left\{1 - \frac{1}{(n-j+1)^2}\right\}.$$

Thus Lemma 2.1 follows. □

We note that Lemma 2.1 shows that when no censoring occurs, the two estimates exactly coincide since $k_0 = 0$ for all $t \in [0, \tau)$. Now we prove Theorem 2.1 as follows.

LEMMA 2.2. *Let $\tau = \tau_S < \tau_G$. Then we have that for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\sqrt{n} \sup_{0 \leq t < T_{(n)}} K_n(t) > \epsilon\right) = 0.$$

PROOF. From Lemma 2.1 and the fact that $1/(1-\hat{H}_n(t-))$ is a non-decreasing function in t, we note that

$$\sup_{0 \leq t < T_{(n)}} K_n(t) \leq \frac{1}{n(1 - \hat{H}_n(T_{(k_0^n)}-)}},$$

where $k_0^n = \max\{j : T_{(j)} \leq T_{(n)},\ \delta_{(j)} = 0\}$. Therefore it is enough to show that

$$P\left(\frac{1}{\sqrt{n}}\frac{1}{1-\hat{H}_n(T_{(k_0^n)}-)} > \epsilon\right) \to 0.$$

In order to show this, we note that $1-G(\tau) > 0$. Also we note that $T_{(n)}$ converges with probability one to τ and so $1-\hat{G}_n(T_{(n)})$ converges with probability one to $1-G(\tau) > 0$. However if $\delta_{(n)} = 0$, then $1-\hat{G}_n(T_{(n)}) = 0$. Therefore since $1-\hat{G}_n(T_{(n)}) = 1-\hat{G}_n(T_{(k_0^n)})$ and $1-\hat{G}_n(T_{(k_0^n)})$ converges with probability one to $1-G(\tau) > 0$, k_0^n should become proportional to n in the long run. Otherwise, $1-\hat{G}_n(T_{(k_0^n)})$ can not converge with probability one to $1-G(\tau) > 0$. This in turn, implies that $1-\hat{H}_n(T_{(k_0^n)}-) = 1-k_0^n/n$ converges with probability one to $1-u$, with $u \in (0,1)$. Therefore Lemma 2.2 follows by applying the weak law of large numbers. □

PROOF OF THEOREM 2.1. Since for each $u \in [0,\tau)$,

$$\begin{aligned} |\ \hat{\mu}_n(t) - \tilde{\mu}_n(t)| &\leq \sqrt{n}\int_0^t \hat{S}_n(u)K_n(u)\ du \\ &\leq \sqrt{n}K_n(t)\int_0^t \hat{S}_n(u)\ du, \end{aligned}$$

and $\int_0^t \hat{S}_n(u)\ du$ is non-decreasing in t, we have that

$$\sup_{0\leq t<\tau} |\ \hat{\mu}_n(t) - \tilde{\mu}_n(t)| \leq \sqrt{n} \sup_{0\leq t<\tau} K_n(t)\int_0^\tau \hat{S}_n(u)\ du.$$

For each n, we note that

$$\sqrt{n} \sup_{0\leq t<\tau} K_n(t) = \sqrt{n} \sup_{0\leq t<T_{(n)}} K_n(t).$$

Also since $\int_0^\tau \hat{S}_n(u)\ du$ converges with probability one to $\int_0^\tau \hat{S}(u)du \leq \int_0^\infty \hat{S}(u)du$, which is finite, from Lemma 2.2 and Slutsky's Theorem, we see that for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{0\leq t<\tau} |\ \hat{\mu}_n(t) - \tilde{\mu}_n(t)| > \epsilon\right) = 0.$$

□

3. APPLICATION TO THE EQUIVALENCE OF TWO MEAN RESIDUAL LIFE PROCESSES

The residual life time is an important biometric function to be estimated. The characterization and properties for the mean residual life time are well summarized in Hall and Wellner (1981). Yang (1978) proposed an estimation for complete data. For the right censored data, Kumazawa (1987) and Park *et al.* (1993) proposed estimates based on the Kaplan-Meier estimate and a version of the Bayesian estimate for the survival function, respectively. However, we note that their asymptotic results are equivalent. In this section, also we show the asymptotic equivalence between two forms of empirical mean residual life processes using the main result. In general the mean residual life time is defined as follows: for any survival random variable T with survival function S,

$$e(t) = E(T - t|T > t) = \frac{1}{S(t)} \int_t^\infty S(u)\, du$$

for all $t \in (0, \infty)$. We assume that e is bounded on $[0, \tau)$. The estimate of Kumazawa, $\hat{e}_n(t)$ and the estimate of Park *et al.*, $\tilde{e}_n(t)$ are as follows:

$$\hat{e}_n(t) = \frac{1}{\hat{S}_n(t)} \int_t^{T_{(n)}} \hat{S}_n(u)\, du \quad \text{and} \quad \tilde{e}_n(t) = \frac{1}{\tilde{S}_n(t)} \int_t^{T_{(n)}} \tilde{S}_n(u)\, du.$$

Then the two empirical mean residual life processes are defined as follows based on $\hat{e}_n(t)$ and $\tilde{e}_n(t)$:

$$\hat{L}_n(t) = \sqrt{n}\, (\hat{e}_n(t) - e(t)) \quad \text{and} \quad \tilde{L}_n(t) = \sqrt{n}\, (\tilde{e}_n(t) - e(t)).$$

We now show that $\hat{L}_n(t)$ and $\tilde{L}_n(t)$ are asymptotically equivalent in the following sense:

THEOREM 3.1. *Suppose that $e(\cdot)$ is bounded on $[0, \tau)$ and $\tau = \tau_S < \tau_G$. Then we have that for every $\epsilon > 0$*

$$\lim_{n \to \infty} P\left(\sup_{0 \le t < \tau} |\, \hat{L}_n(t) - \tilde{L}_n(t)| > \epsilon \right) = 0.$$

PROOF. From the definition,

$$\tilde{L}_n(t) - \hat{L}_n(t) = \frac{\sqrt{n}}{\tilde{S}_n(t)} \int_t^{T_{(n)}} \tilde{S}_n(u)\, du - \frac{\sqrt{n}}{\hat{S}_n(t)} \int_t^{T_{(n)}} \hat{S}_n(u)\, du$$

$$
\begin{aligned}
&= \sqrt{n}\left(\frac{1}{\tilde{S}_n(t)} - \frac{1}{\hat{S}_n(t)}\right)\int_t^{T_{(n)}} \tilde{S}_n(u)\, du \\
&\quad + \frac{\sqrt{n}}{\hat{S}_n(t)}\left(\int_t^{T_{(n)}} \tilde{S}_n(u)\, du - \int_t^{T_{(n)}} \hat{S}_n(u)\, du\right) \\
&= \sqrt{n}\,\frac{K_n(t)}{\hat{S}_n(t)}\int_t^{T_{(n)}} \tilde{S}_n(u)\, du + \frac{\sqrt{n}}{\hat{S}_n(t)}\int_t^{T_{(n)}} (\tilde{S}_n(u) - \hat{S}_n(u))\, du \\
&= \sqrt{n}\,\frac{K_n(t)}{\hat{S}_n(t)}\int_t^{T_{(n)}} \tilde{S}_n(u)\, du + \frac{\sqrt{n}}{\hat{S}_n(t)}\int_t^{T_{(n)}} K_n(u)\hat{S}_n(u)\, du \\
&\le \sqrt{n}\, K_n(t)\tilde{e}_n(t) + \sqrt{n} \sup_{t\le u\le T_{(n)}} K_n(u)\hat{e}_n(t).
\end{aligned}
$$

Therefore we have that

$$
\begin{aligned}
\sup_{0\le t\le T_{(n)}} |\, \hat{L}_n(t) - \tilde{L}_n(t)| \le \sqrt{n} \sup_{0\le t\le T_{(n)}} K_n(t) \sup_{0\le t\le T_{(n)}} \tilde{e}_n(t) \\
+\sqrt{n} \sup_{0\le t\le T_{(n)}} K_n(t) \sup_{0\le t\le T_{(n)}} \hat{e}_n(t).
\end{aligned}
$$

Therefore by Lemma 2.2, it is enough to show that $\hat{e}_n(t)$ and $\tilde{e}_n(t)$ are bounded to apply the Slutsky's Theorem. Since $e(\cdot)$ is bounded on $[0,\tau)$ and $\hat{e}_n(t)$ and $\tilde{e}_n(t)$ can take values at most n different values on $[0,\tau)$, it easy to show that $\hat{e}_n(t)$ and $\tilde{e}_n(t)$ are bounded on $[0,\tau)$. □

4. Discussion

In the previous two sections, we showed the asymptotic equivalences only for the case of $\tau = \tau_S < \tau_G$. In order to complete our discussion, we should have shown that the asymptotic equivalences for the case of $\tau = \tau_G \le \tau_S$. In this case, we may show the asymptotic equivalences by choosing a sequence (M_n) instead of using $T_{(n)}$ such as $M_n \to \tau$. Then the choices of the sequence completely depend on the censoring distributions. This can be seen from Lemma 2.2 since the conclusion of Lemma 2.2 completely relys on the censoring manners at the tail part of the whole observations. However since our main concern is for the survival function of the life time not for the distribution function of the censoring time, we do not handle this matter.

The estimations of the mean life time, mean residual life time and the mean difference in two sample problem based on right censored data, have been considered for a long time in survival analysis. However up to now, the results are not so quite satisfactory because of the possibility of the censoring of the largest

observation. Therefore still a lot of results are being introduced to overcome this difficulty. However for any case, their asymptotic properties coincide. Therefore it would be worthwhile to show the equivalence among the results.

ACKNOWLEDGEMENTS

The author wishes to express his appreciation to the referees for pointing out errors.

REFERENCES

GILL, R. D. (1983). "Large sample behaviour of the product-limit estimator on the whole line", *The Annals of Statistics*, **11**, 49–58.

HALL, W. J. AND WELLNER, J. A. (1981). "Mean residual life", In *Statistics and Related Topics* (M. Csorgo, D. A. Dawson, J. N. K. Rao and A. K. Md. E. Saleh, eds.), 169–184, North-Holland, Amsterdam.

KUMAZAWA, Y. (1987). "A note on an estimator of life expectancy with random censorship", *Biometrika*, **74**, 655–658.

PARK, B. G., SOHN, J. K. AND LEE, S. B. (1993). "Nonparametric estimation of mean residual life function under random censorship", *Journal of the Korean Statistical Society*, **22**, 147–157.

SHORACK, G. R. AND WELLNER, J. A. (1985). *Empirical Processes with Applications to Statistics*, Wiley & Sons, New York.

SUSARLA, V. AND VAN RYZIN, J. (1976). "Nonparametric Bayesian estimation of survival curves from incomplete observations", *Journal of the American Statistical Association*, **71**, 897–902.

SUSARLA, V. AND VAN RYZIN, J. (1980). "Large sample theory for an estimator of the mean survival time from censored samples", *The Annals of Statistics*, **8**, 1002–1016.

YANG, G. L. (1978). "Estimation of a biometric function", *The Annals of Statistics*, **6**, 112–116.