AISM (2008) 60:427–440
DOI 10.1007/s10463-006-0087-z

Small-sample studies on right censored data with discrete failure times

Jiantian Wang

Received: 26 September 2005 / Revised: 6 July 2006 / Published online: 16 December 2006

Abstract It has been already noticed that the classical Greenwood formula may not be an unbiased estimator for the variance of the Kaplan–Meier Product Limit estimator (PLE). However, a rigorous proof for such a suggestion has not been available. In this paper, we investigate some small-sample properties of the PLE and show that the Greenwood formula strictly underestimates the variance of the PLE. Besides, some existing estimators for the variance of the PLE are also discussed.

Keywords Kaplan–Meier estimator · Greenwood formula · Discrete data · Survival analysis

1 Introduction

Let T be a random variable representing failure time of the subject under study. The inference of T can be characterized either by its survival function or its cumulative hazard function. In practice, the failure time T may not be always observable due to the presence of censoring. Under the assumption of independent censoring, the survival function and the cumulative hazard function of T can be well estimated by the celebrated Kaplan–Meier (1958) Product Limit estimator (PLE) and the Nelson (1969) estimator (NE), respectively.

The estimation of the variances of these estimators is critical in the analysis of the inference of these estimators and in the estimation of median survival time (see, Brookmeyer and Crowley 1982; Slud et al. 1984; Jennison and Turnbull

J. Wang (✉)
Department of Mathematics and Computer Sciences,
Kean University, Union, 07083, NJ, USA
e-mail: jwang@kean.edu

1985). The classical Greenwood formula is commonly employed for estimating the variance of the PLE.

The Greenwood formula is asymptotically a consistent estimator for the variance of the PLE (see, Fleming and Harrington 1991). However, as suggested by Jennison and Turnbull (1985), it may underestimate the true variance of the PLE when sample is small. Some alternative estimators have been advocated, such as the Peto estimator (PE) proposed by Peto et al. (1977) and a homogenetic estimator (HE) proposed by Zhao (1996). Slud et al. (1984) pointed out that the PE can seriously overestimate the variance of the PLE. Shen (2002) conducted a simulation showing that the HE can significantly overestimate the variance of the PLE. Theoretical proofs for these conclusions, however, have not been available.

Since most clinical studies are expensive and sample size can not be large enough as desired, therefore, the study for small-sample properties of the PLE and the NE has some value in theory as well as in practice. In the case that failure time is continuous, Chen et al. (1982) derived the exact moments of the PLE and the NE under the assumption of the proportional hazard model. Phadia and Shao (1999) dropped such assumption and gave out the formulae for the kth moment of the PLE and the NE in the form of integration. They also provided an approximation of the variance of the PLE (see, Phadia and Shao 1999, Corollary 2). However, those formulae are not so readily amenable to the theoretical study of Greenwood formula and other variance estimators of the PLE.

In this paper, we investigate the case of small sample censored data with discrete failure times. We study the biases and variances of the PLE and the NE without any assumption on the underlying distribution model. Specially, we prove that the Greenwood formula strictly underestimates the variance of the PLE, while the PE and HE can either seriously overestimate or underestimate the variance of the PLE.

The paper is organized as follows. Section 2 sketches the derivation of the PLE and the NE. Some results on moments and biases of the PLE and the NE are presented in Sect. 3. Section 4 devotes to the study of the Greenwood formula and other variance estimators of the PLE. The article closes with a short discussion in Sect. 5.

2 The PLE and the NE for discrete failure time

Conventionally, the failure time T is assumed to be continuous. However, in some situations, T may be genuinely discrete, such as when T is a counting variable. Also, in practice, continuous data are frequently modeled as discrete data (see, e.g. Willett and Singer 1993). Therefore, it is necessary and also instructive to study discrete censored data.

As usual, for $0 \le t < \infty$, define $S(t) = P(T \ge t)$, $f(t) = -\mathrm{d}S(t)/\mathrm{d}t$, and $\lambda(t) = f(t)/S(t)$ as the survival function, the density function, and the hazard function of T, respectively. For any function g, denote $g(t-)$, $g(t+)$ as the left and the right limits of g at t. Let $H(t)$ be the Heaviside function defined as $H(t) = 1$ for $t \geq 0$ and 0 otherwise. The derivative of $H(t)$, denoted as $\delta(t)$, is the Dirichlet function which gives unit mass at point t. Assume T is a discrete random variable taking values at t_j with probability a_j for $j = 1, \ldots, m$. Then, $S(t) = 1 - \sum_{j=1}^{m} a_j H((t - t_j)-)$, $f(t) = \sum_{j=1}^{m} a_j \delta((t - t_j)-)$, $\lambda(t) = f(t)/S(t) = \sum_{j=1}^{m} a_j \delta((t - t_j)-)/S(t) = \sum_{j=1}^{m} \{a_j/S(t_j)\}\delta((t - t_j)-) = \sum_{j=1}^{m} \lambda_j \delta((t - t_j)-)$, here $\lambda_j = a_j/S(t_j) = P(T = t_j | T \geq t_j)$. The cumulative hazard function $\Lambda(t) = \int_0^t \lambda(t)\mathrm{d}t = \sum_{j=1}^{m} \lambda_j H((t - t_j)-)$. In terms of λ_j, $S(t)$ can also be written as

$$\begin{aligned}S(t) &= \prod_{t_j<t} \frac{S(t_j+)}{S(t_j)} = \prod_{t_j<t} \left\{1 - \frac{a_j}{S(t_j)}\right\} = \prod_{j=1}^{m} \left\{1 - \frac{a_j}{S(t_j)} H((t - t_j)-)\right\} \\ &= \prod_{j=1}^{m} \{1 - \lambda_j H((t - t_j)-)\}.\end{aligned}$$

Let C be the censoring time. Under the presence of censoring, the observable variables are $X = \min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the usual indicator function. Under the assumption that C is independent of T, we have

$$\lambda_j = P(T = t_j | T \geq t_j) = P(T = t_j | T \geq t_j, C \geq t_j) = \frac{P(T = t_j, C \geq t_j)}{P(X \geq t_j)}.$$

Let $(X_i, \Delta_i), i = 1, \ldots, n$, be the observed data. Then, a natural estimator for λ_j is $\hat{\lambda}_j = d_j/n_j$, where $d_j = \sum_{i=1}^{n} I(T_i = t_j, C_i \geq t_j)$ is the number of failures at t_j, and $n_j = \sum_{i=1}^{n} I(X_i \geq t_j)$ is the number at risk at t_j. Plugging $\hat{\lambda}_j$ into the expressions of $S(t)$ and $\Lambda(t)$ yields the classical PLE

$$\hat{S}(t) = \prod_{j=1}^{m} \left\{1 - \frac{d_j}{n_j} H((t - t_j)-)\right\}, \tag{1}$$

and the NE

$$\hat{\Lambda}(t) = \sum_{j=1}^{m} \frac{d_j}{n_j} H((t - t_j)-), \tag{2}$$

where 0/0 is defined as 0.

Remark 1 The PLE (1) and the NE (2) traditionally are derived by the maximum likelihood approach. However, as pointed out in Kalbfleisch and Prentice (1981, p. 12), some investigations are needed to assure the reasonability of such maximization since it involved many parameters. Comparatively, the derivation here is simple and of mathematical rigor.

3 Moments and biases of the PLE and the NE

So far, much study has been done on the asymptotic properties of the PLE and the NE when sample size is large and T is continuous. However, relatively only a few attempts have been made on the exact small-sample results for the PLE and the NE even when failure time is continuous. These attempts include Chen et al. (1982) who derived the exact moments of the PLE under the assumption of proportional hazard model, Gillespie et al. (1992) who investigated the bias of several modified version of the PLE, Guerts (1985, 1987) who discussed small-sample performance via computational simulations, and Phadia and Shao (1999) who derived the formula of exact moments of the PLE and the first two moments of NE also. To our best knowledge, no study has been conducted on the properties of the PLE and the NE when sample size is small and T is discrete.

Let $t_0 = 0$ and $t_{m+1} = \infty$, and assume $0 = t_0 < t_1 < t_2 < \cdots < t_m$. Denote $G(t) = P(C \geq t)$ and $\pi(t) = P(X \geq t)$. Assume C is continuous and thus $P(C = t_j) = 0$ for all $j = 1, \ldots, m$. We have

Theorem 1 *Denote* $k_j = P(X = t_j) = a_j G(t_j)$, $h_j = P(t_{j-1} < X < t_j) = S(t_j)\{G(t_{j-1}) - G(t_j)\}$. *Let* $\hat{S}(t)$ *and* $\hat{\Lambda}(t)$ *be defined as in* (1) *and* (2). *Then, the kth moments of* $\hat{S}(t)$ *and* $\hat{\Lambda}(t)$ *are given respectively by*

$$E[\hat{S}^k(t)] = \sum_{j=1}^{m+1} I(t_{j-1} < t \leq t_j) \sum_{\sum_j (l_j + d_j) = n} \binom{n}{l_1\, d_1 \ldots l_m\, d_m} \prod_{p<j} \left(1 - \frac{d_p}{n_p}\right)^k \prod_{j=1}^{m} h_j^{l_j} k_j^{d_j},$$

and

$$E[\hat{\Lambda}^k(t)] = \sum_{j=1}^{m+1} I(t_{j-1} < t \leq t_j) \sum_{\sum_j (l_j + d_j) = n} \binom{n}{l_1\, d_1 \ldots l_m\, d_m} \left(\sum_{p<j} \frac{d_p}{n_p}\right)^k \prod_{j=1}^{m} h_j^{l_j} k_j^{d_j},$$

where $\binom{n}{l_1\, d_1 \ldots l_m\, d_m}$ *is the combinatorial coefficient,* $n_p = n - \sum_{i \leq p} l_i - \sum_{j<p} d_j$, *and the interval* $I(t_m < t \leq \infty)$ *is understood as* $I(t > t_m)$.

Proof Let $\sum_{i=1}^{n} I(t_{j-1} < X_i < t_j) = l_j$, $\sum_{i=1}^{n} I(X_i = t_j) = d_j$. Then $E\{I(t_{j-1} < X_i < t_j)\} = h_j$ and $E\{I(X_i = t_j)\} = k_j$. Thus,

$$\begin{aligned} E[\hat{S}^k(t)] &= \sum_{j=1}^{m+1} I(t_{j-1} < t \leq t_j) E\{\hat{S}^k(t)\} \\ &= \sum_{j=1}^{m+1} I(t_{j-1} < t \leq t_j) \sum_{\sum_j (l_j + d_j) = n} \binom{n}{l_1\, d_1 \ldots l_m\, d_m} \hat{S}^k(t) \prod_{j=1}^{m} h_j^{l_j} k_j^{d_j} \end{aligned}$$

$$= \sum_{j=1}^{m+1} I(t_{j-1} < t \le t_j) \sum_{\sum_j (l_j + d_j) = n} \binom{n}{l_1\, d_1 \ldots l_m\, d_m} \prod_{p<j} \left(1 - \frac{d_p}{n_p}\right)^k \prod_{j=1}^{m} h_j^{l_j} k_j^{d_j}.$$

Similarly for $E[\hat{\Lambda}^k(t)]$. □

For continuous failure case, it is already proved by martingale theory that the nonnegative bias of the PLE and the nonpositive bias of the NE converge to zero exponentially as $n \to \infty$ (see Fleming and Harrington 1991). The exact biases for the PLE and the NE in small sample have been derived only for some typical data models, such as the so-called proportional hazard model in which $G(t) = S(t)^\alpha$ for some α (see Chen et al. 1982; Phadia and Shao 1999, for details).

Whereas, in discrete failure time case, we can derive the biases of these estimators without any assumption on the underlying survival distribution model.

For usage later on, we list some identities as a lemma whose proof is elementary.

Lemma 1 *Define* $0/0$ *as* 0*, then the following identities hold:*

$$\sum_{\substack{k_1,k_2,k_3 \ge 0 \\ k_1+k_2+k_3=n}} \binom{n}{k_1\, k_2\, k_3} \left(1 - \frac{k_1}{n-k_2}\right) x^{k_1} y^{k_2} z^{k_3} = \frac{z}{x+z}(x+y+z)^n + \frac{x}{x+z} y^n. \quad (3)$$

$$\sum_{\substack{k_1,k_2,k_3 \ge 0 \\ k_1+k_2+k_3=n}} \binom{n}{k_1\, k_2\, k_3} \frac{k_1}{n-k_2} x^{k_1} y^{k_2} z^{k_3} = \frac{x}{x+z}\{(x+y+z)^n - y^n\}. \quad (4)$$

$$\begin{aligned} &\sum_{\substack{k_1,k_2,k_3 \ge 0 \\ k_1+k_2+k_3=n}} \binom{n}{k_1\, k_2\, k_3} \left(1 - \frac{k_1}{n-k_2}\right)^2 x^{k_1} y^{k_2} z^{k_3} \\ &\quad = \frac{z^2}{(x+z)^2}(x+y+z)^n + \frac{(x+z)^2 - z^2}{(x+z)^2} y^n \\ &\qquad + \frac{xz}{(x+z)^2} \sum_{k_2=0}^{n-1} \binom{n}{k_2} \frac{1}{n-k_2} y^{k_2} (x+z)^{n-k_2}. \end{aligned} \quad (5)$$

Theorem 2 *For the biases of the PLE and the NE, we have*

$$E[\hat{S}(t)] - S(t) = \sum_{j=2}^{m+1} I(t_{j-1} < t \le t_j) \sum_{1 \le k \le j} \alpha_{jk} \tilde{\pi}_k^n,$$

and

$$E[\hat{\Lambda}(t)] - \Lambda(t) = -\sum_{j=2}^{m+1} I(t_{j-1} < t \le t_j) \lambda_j \tilde{\pi}_j^n,$$

where $\tilde{\pi}_k = P(X < t_k)$, and α_{jk} are some nonnegative constants depending on $\{h_p, k_p, p \le j\}$.

Proof As before, denote $n_k = n - \sum_{p\le k} l_p - \sum_{p<k} d_p$, $\tilde{n}_k = n - \sum_{p\le k} l_p - \sum_{p\le k} d_p$. Let $\alpha_j = 1 - \sum_{p\le j}(k_p + h_p)$. For $m = 1$, clearly if $l_1 < n$, $\hat{S}(t) = 1 - H((t-t_1)-)$, $\hat{\Lambda}(t) = H((t-t_1)-)$, and, if $l_1 = n$, $\hat{S}(t) = 1$, $\hat{\Lambda}(t) = 0$. Thus,

$$E[\hat{S}(t)] - S(t) = I(t > t_1)E\{I(l_1 = n)\} = I(t > t_1)\tilde{\pi}_1^n,$$
$$E[\hat{\Lambda}(t)] - \Lambda(t) = -I(t > t_1)E\{I(l_1 = n)\} = -I(t > t_1)\tilde{\pi}_1^n.$$

Now, consider $m \ge 2$. For $t_1 < t \le t_2$, by (3) in Lemma 1,

$$\begin{aligned} E[\hat{S}(t)] &= \sum_{l_1,d_1,\tilde{n}_1} \binom{n}{l_1, d_1, \tilde{n}_1}\left(1 - \frac{d_1}{n_1}\right) h_1^{l_1} k_1^{d_1} \alpha_1^{\tilde{n}_1} \\ &= \frac{\alpha_1}{k_1+\alpha_1} + \frac{k_1}{k_1+\alpha_1} h_1^n = S(t) + a_1\tilde{\pi}_1^n. \end{aligned}$$

While for $t_2 < t \le t_3$, again by (3) in Lemma 1,

$$\begin{aligned} E[\hat{S}(t)] &= \sum \binom{n}{l_1\, d_1\, l_2\, d_2\, \tilde{n}_2}\left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) h_1^{l_1} k_1^{d_1} h_2^{l_2} k_2^{d_2} \alpha_2^{\tilde{n}_2} \\ &= \sum_{l_1,d_1} \binom{n}{l_1\, d_1\, \tilde{n}_1}\left(1 - \frac{d_1}{n_1}\right) h_1^{l_1} k_1^{d_1} \sum_{l_2,d_2,\tilde{n}_2} \binom{\tilde{n}_1}{l_2\, d_2\, \tilde{n}_2}\left(1 - \frac{d_2}{n_2}\right) h_2^{l_2} k_2^{d_2} \alpha_2^{\tilde{n}_2} \\ &= \sum_{l_1,d_1} \binom{n}{l_1\, d_1\, \tilde{n}_1}\left(1 - \frac{d_1}{n_1}\right) h_1^{l_1} k_1^{d_1} \left[\frac{\alpha_2}{k_2+\alpha_2}(h_2+k_2+\alpha_2)^{\tilde{n}_1} + \frac{k_2}{k_2+\alpha_2} h_2^{\tilde{n}_1}\right] \\ &= \frac{\alpha_2}{k_2+\alpha_2}\left[\frac{\alpha_1}{k_1+\alpha_1} + \frac{k_1}{k_1+\alpha_1}\tilde{\pi}_1^n\right] + \frac{k_2}{k_2+\alpha_2}\left[\frac{h_2}{k_1+h_2}\tilde{\pi}_2^n + \frac{k_1}{k_1+h_2}\tilde{\pi}_1^n\right] \\ &= S(t) + \alpha_{21}\tilde{\pi}_1^n + \alpha_{22}\tilde{\pi}_2^n. \end{aligned}$$

Applying the same arguments to the interval $t_{j-1} < t \le t_j$, we have,

$$\begin{aligned} E[\hat{S}(t)] &= \sum \binom{n}{l_1\, d_1\, \cdots, l_j\, d_j\, \tilde{n}_j} \prod_{k\le j}\left(1 - \frac{d_k}{n_k}\right) \prod_{p\le j}\left(h_p^{l_p} k_p^{d_p}\right) \alpha_j^{\tilde{n}_j} \\ &= \sum \binom{n}{l_1\, d_1\, \cdots, \tilde{n}_{j-1}} \prod_{k<j}\left(1 - \frac{d_k}{n_k}\right) \prod_{p<j}\left(h_p^{l_p} k_p^{d_p}\right) \\ &\quad \times \sum_{l_j,d_j,\tilde{n}_j} \binom{\tilde{n}_{j-1}}{l_j\, d_j\, \tilde{n}_j}\left(1 - \frac{d_j}{n_j}\right) h_j^{l_j} k_j^{d_j} \alpha_j^{\tilde{n}_j} \\ &= \sum \binom{n}{l_1\, d_1\, \cdots, \tilde{n}_{j-1}} \prod_{k<j}\left(1 - \frac{d_k}{n_k}\right) \prod_{p<j}(h_p^{l_p} k_p^{d_p}) \end{aligned}$$

$$
\begin{aligned}
&\times\left[\frac{\alpha_j}{k_j+\alpha_j}(h_j+k_j+\alpha_j)^{n_j}+\frac{k_j}{k_j+\alpha_j}h_j^{n_j}\right]\\
&=\prod_{p\le j}\frac{\alpha_p}{k_p+\alpha_p}+\sum_{p\le j}\alpha_{jp}\tilde{\pi}_p^n\\
&=S(t)+\sum_{p\le j}\alpha_{jp}\tilde{\pi}_p^n.
\end{aligned}
$$

The last equality comes from the fact $\prod_{p\le j}\alpha_p/(k_p+\alpha_p)=S(t_j)$, which can be easily verified since $k_p+\alpha_p=P(X\ge t_p)=P(T\ge t_p)P(C\ge t_p)$, while $\alpha_p=P(X>t_p)=P(T>t_p)P(C>t_p)$. The continuity assumption of C yields $\alpha_p/(k_p+\alpha_p)=S(t_{p+1})/S(t_p)$.

Now we turn to $E[\hat{\Lambda}(t)]$. For $t_1<t\le t_2$, by (4) in Lemma 1, we have

$$E[\hat{\Lambda}(t)]=\sum_{l_1,d_1,\tilde{n}_1}\tbinom{n}{l_1,d_1,\tilde{n}_1}\tfrac{d_1}{n_1}h_1^{l_1}k_1^{d_1}\alpha_1^{\tilde{n}_1}=\tfrac{k_1}{k_1+\alpha_1}-\tfrac{k_1}{k_1+\alpha_1}h_1^n.$$

Similarly,

$$
\begin{aligned}
&\sum\binom{n}{l_1\,d_1\;\cdots,l_j\,d_j\,\tilde{n}_j}\frac{d_j}{n_j}\prod_{p\le j}h_p^{l_p}k_p^{d_p}\alpha_j^{\tilde{n}_j}\\
&\quad=\sum\binom{n}{l_1\,d_1\;\cdots,l_{j-1}\,d_{j-1}\,\tilde{n}_{j-1}}\prod_{p<j}h_p^{l_p}k_p^{d_p}\sum\binom{\tilde{n}_{j-1}}{l_j\,d_j\,\tilde{n}_j}\frac{d_j}{n_j}h_j^{l_j}k_j^{d_j}\alpha_j^{\tilde{n}_j}\\
&\quad=\sum\binom{n}{l_1\,d_1\;\cdots,l_{j-1}\,d_{j-1}\,\tilde{n}_{j-1}}\\
&\qquad\times\prod_{p<j}h_p^{l_p}k_p^{d_p}\left[\frac{k_j}{k_j+\alpha_j}\left\{(h_j+k_j+\alpha_j)^{\tilde{n}_{j-1}}-h_j^{\tilde{n}_{j-1}}\right\}\right]\\
&\quad=\frac{k_j}{k_j+\alpha_j}-\frac{k_j}{k_j+\alpha_j}\tilde{\pi}_j^n.
\end{aligned}
$$

Clearly, $k_p/(k_p+\alpha_p)=a_p/S(t_p)=\lambda_p$ and therefore, for $t_j<t\le t_{j+1}$, $E[\hat{\Lambda}(t)]-\Lambda(t)=-\sum_{p\le j}\frac{k_p}{k_p+\alpha_p}\tilde{\pi}_p^n=-\sum_{p\le j}\lambda_p\tilde{\pi}_p^n$. □

Corollary 1 *We have*

$$0\le E[\hat{S}(t)]-S(t)\le\{1-S(t)\}\{1-\pi(t)\}^n, \tag{6}$$

and

$$-\Lambda(t)\{1-\pi(t)\}^n\le E[\hat{\Lambda}(t)]-\Lambda(t)\le 0. \tag{7}$$

Proof For $t_{j-1} < t \le t_j$, $E[\hat{S}(t)] - S(t) = \sum_{k\le j} \alpha_{jk}\tilde{\pi}_k^n$. Letting all $\tilde{\pi}_k \equiv 1$ for $k \le j$, it implies that $P(X \ge t_k) = 0$, or equivalently, $P(C > t_1) = 0$. Therefore, all failure times are censored and hence $\hat{S}(t) \equiv 1$ and so, $E[\hat{S}(t)] - S(t) = 1 - S(t) = \sum_{k\le j} \alpha_{jk}$. Thus, $E[\hat{S}(t)] - S(t) = \sum_{k\le j} \alpha_{jk}\tilde{\pi}_k^n \le \{1 - S(t)\}\tilde{\pi}_j^n = \{1 - S(t)\}\{1 - \pi(t)\}^n$. That is (6). Again for $t_{j-1} < t \le t_j$, from Theorem 2, we have $\Lambda(t) - E[\hat{\Lambda}(t)] = \sum_{k\le j} \lambda_k \tilde{\pi}_k^n \le (\sum_{k\le j} \lambda_k)\tilde{\pi}_j^n = \Lambda(t)\{1 - \pi(t)\}^n$, (7) is thus proved. □

4 The variance of the PLE and its estimators

The formula for the variance of the PLE emerges as a consequence of Theorem 1 and Theorem 2. If we neglect the terms that diminish exponentially, we have

$$\begin{aligned}\operatorname{var}[\hat{S}(t)] &= E[\hat{S}^2(t)] - S^2(t)\\ &= \sum_{j=1}^{m+1} I(t_{j-1} < t \le t_j) \sum_{\sum_j (l_j + d_j) = n} \binom{n}{l_1\, d_1 \ldots l_m\, d_m}\\ &\quad \times \prod_{k<j} \left(1 - \frac{d_k}{n_k}\right)^2 \prod_{j=1}^{m} h_j^{l_j} k_j^{d_j} - S^2(t).\end{aligned}$$

The asymptotic variance of the PLE for continuous failure time given in Fleming and Harrington (1991) is

$$\operatorname{avar}[\hat{S}(t)] = S^2(t) \int_0^t \frac{-dS(s)}{nS^2(s)G(s)}, \tag{8}$$

which was first derived by Breslow and Crowley (1974).

The formula (8) does not readily lend it to the case where failure time is discrete. Since, even when there is no censoring, according to (8) for $t_{j-1} < t < t_j$,

$$\operatorname{avar}[\hat{S}(t)] = S^2(t) \int_0^t \frac{-dS(s)}{nS^2(s)G(s)} = \frac{1}{n} S^2(t_j) \sum_{k<j} \frac{a_k}{S^2(t_k)} \ne \frac{1}{n} S(t_j)\{1 - S(t_j)\}.$$

Incidentally, the discrete version of (8) is

$$\begin{aligned}\operatorname{avar}[\hat{S}(t)] &= S^2(t) \sum \frac{a_j}{nS^2(t_j)G(t_j)} H((t - t_j)-)\\ &= S^2(t) \sum \frac{S(t_j) - S(t_{j+1})}{nS(t_j)\pi(t_j)} H((t - t_j)-),\end{aligned}$$

and thus a corresponding estimator would be $\hat{\Sigma}_0(t) = \hat{S}^2(t) \sum_{j=1}^m (d_j/n_j^2) H((t - t_j)-)$.

A classical estimator for var$[\hat{S}(t)]$ is the Greenwood formula, defined as

$$\hat{\Sigma}_G(t) = \hat{S}^2(t) \sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)} H((t - t_j)-). \tag{9}$$

Compared with $\hat{\Sigma}_0(t)$, $\hat{\Sigma}_G(t)$ would be preferable, since in the case of no censoring, $\hat{\Sigma}_G(t) = n^{-1}\hat{S}(t)\{1 - \hat{S}(t)\}$, while $\Sigma_0(t) < n^{-1}\hat{S}(t)\{1 - \hat{S}(t)\}$.

As suggested by Jennison and Turnbull (1985), the Greenwood formula may underestimate the true variance of the PLE.

In the following theorem, we prove this conclusion mathematically when failure time is discrete. Since a continuous variable can be regarded as a limit of discrete ones, the conclusion should also hold for continuous failure time.

Theorem 3 *Let $\hat{S}(t)$ be the PLE defined as in* (1) *and $\hat{\Sigma}_G(t)$ be the Greenwood formula defined as in* (9). *Then, for any sample size n, we have, var$[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \geq 0$.*

Proof We first check the interval $t_1 < t \leq t_2$. By the identity (5) in Lemma 1, we have

$$\begin{aligned}
E[\hat{S}^2(t)] &= \sum_{l_1,d_1,\tilde{n}_1} \binom{n}{l_1, d_1, \tilde{n}_1} \left(1 - \frac{d_1}{n_1}\right)^2 h_1^{l_1} k_1^{d_1} \alpha_1^{\tilde{n}_1} \\
&= \frac{\alpha_1^2}{(k_1+\alpha_1)^2} + \frac{(k_1+\alpha_1)^2 - \alpha_1^2}{(k_1+\alpha_1)^2} h_1^n \\
&\quad + \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \sum_{k=1}^{n} \binom{n}{k} \frac{1}{k} (k_1+\alpha_1)^k h_1^{n-k} \\
&= S^2(t) + a_1(2 - a_1)h_1^n + a_1(1 - a_1) \sum_{k=1}^{n} \binom{n}{k} \frac{1}{k} (k_1+\alpha_1)^k h_1^{n-k}.
\end{aligned}$$

Denote $\Gamma_n^{(1)}(x, y) = \sum_{k=1}^n \binom{n}{k} k^{-1} x^k y^{n-k}$, and $\Gamma_n^{(2)}(x, y) = \sum_{k=1}^n \binom{n}{k} k^{-2} x^k y^{n-k}$. Neglecting the terms that converge to zero exponentially, we obtain

$$\mathrm{var}[\hat{S}(t)] = a_1(1 - a_1)\Gamma_n^{(1)}(k_1 + \alpha_1, h_1) = a_1(1 - a_1)\Gamma_n^{(1)}\{\pi(t_1), 1 - \pi(t_1)\}.$$

For the mean of the Greenwood formula on $t_1 < t \leq t_2$, we have

$$\begin{aligned}
E[\hat{\Sigma}_G(t)] &= \sum_{l_1,d_1,\tilde{n}_1} \binom{n}{l_1, d_1, \tilde{n}_1} \left(1 - \frac{d_1}{n_1}\right)^2 \frac{d_1}{n_1\tilde{n}_1} h_1^{l_1} k_1^{d_1} \alpha_1^{\tilde{n}_1} \\
&= \sum_{l_1,d_1,\tilde{n}_1} \binom{n}{l_1, d_1, \tilde{n}_1} \left\{\frac{\tilde{n}_1}{n_1^2} - \frac{\tilde{n}_1^2}{n_1^3}\right\} h_1^{l_1} k_1^{d_1} \alpha_1^{\tilde{n}_1}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l_1} \binom{n}{l_1} \frac{h_1^{l_1}}{n_1^2} \sum_{d_1,\tilde{n}_1} \binom{n_1}{d_1\ \ \tilde{n}_1} \tilde{n}_1 k_1^{d_1} \alpha_1^{\tilde{n}_1} \\
&\quad - \sum_{l_1} \binom{n}{l_1} \frac{h_1^{l_1}}{n_1^3} \sum_{d_1,\tilde{n}_1} \binom{n_1}{d_1\ \ \tilde{n}_1} \tilde{n}_1^2 k_1^{d_1} \alpha_1^{\tilde{n}_1} \\
&= \frac{\alpha_1}{(k_1+\alpha_1)} \Gamma_n^{(1)}(k_1+\alpha_1, h_1) - \frac{\alpha_1^2}{(k_1+\alpha_1)^2} \Gamma_n^{(1)}(k_1+\alpha_1, h_1) \\
&\quad - \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \Gamma_n^{(2)}(k_1+\alpha_1, h_1) \\
&= \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \Gamma_n^{(1)}(k_1+\alpha_1, h_1) - \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \Gamma_n^{(2)}(k_1+\alpha_1, h_1) \\
&= a_1(1-a_1)\{\Gamma_n^{(1)} - \Gamma_n^{(2)}\}(\pi(t_1), 1-\pi(t_1)).
\end{aligned}
$$

Thus, on $t_1 < t \le t_2$,

$$
\mathrm{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] = a_1(1-a_1)\Gamma_n^{(2)}(\pi(t_1), 1-\pi(t_1)) \ge 0.
$$

On the interval $t_2 < t \le t_3$,

$$
\begin{aligned}
E[\hat{S}^2(t)] &= \sum \binom{n}{l_1\ d_1\ l_2\ d_2\ \tilde{n}_2} \left(1-\frac{d_1}{n_1}\right)^2 \left(1-\frac{d_2}{n_2}\right)^2 h_1^{l_1} k_1^{d_1} h_2^{l_2} k_2^{d_2} \alpha_2^{\tilde{n}_2} \\
&= \sum \binom{n}{l_1\ d_1\ \tilde{n}_1} \left(1-\frac{d_1}{n_1}\right)^2 h_1^{l_1} k_1^{d_1} \sum \binom{\tilde{n}_1}{l_2\ d_2\ \tilde{n}_2} \left(1-\frac{d_2}{n_2}\right)^2 h_2^{l_2} k_2^{d_2} \alpha_2^{\tilde{n}_2} \\
&= \sum \binom{n}{l_1\ d_1\ \tilde{n}_1} \left(1-\frac{d_1}{n_1}\right)^2 h_1^{l_1} k_1^{d_1} \left\{ \frac{\alpha_2^2}{(k_2+\alpha_2)^2} \alpha_1^{\tilde{n}_1} \right. \\
&\quad \left. + \frac{(k_2+\alpha_2)^2 - \alpha_2^2}{(k_2+\alpha_2)^2} h_2^{\tilde{n}_1} + \frac{k_2\alpha_2}{(k_2+\alpha_2)^2} \sum \binom{\tilde{n}_1}{l_2} \frac{h_2^{l_2}(k_2+\alpha_2)^{n_2}}{n_2} \right\} \\
&= \frac{\alpha_2^2}{(k_2+\alpha_2)^2} \left\{ \frac{\alpha_1^2}{(k_1+\alpha_1)^2} + \frac{(k_1+\alpha_1)^2 - \alpha_1^2}{(k_1+\alpha_1)^2} h_1^n \right. \\
&\quad \left. + \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \Gamma_n^{(1)}(k_1+\alpha_1, h_1) \right\} \\
&\quad + \frac{(k_2+\alpha_2)^2 - \alpha_2^2}{(k_2+\alpha_2)^2} \left\{ \frac{h_2^2}{(k_1+h_2)^2} (h_1+k_1+h_2)^n + \frac{(k_1+h_2)^2 - h_2^2}{(k_1+h_2)^2} h_1^n \right. \\
&\quad \left. + \frac{k_1 h_2}{(k_1+h_2)^2} \Gamma_n^{(1)}(k_1+h_2, h_1) \right\} \\
&\quad + \frac{k_2\alpha_2}{(k_2+\alpha_2)^2} \sum \binom{n}{l_1\ d_1\ l_2\ \tilde{n}_2} \left(1-\frac{d_1}{n_1}\right)^2 \frac{1}{\tilde{n}_2} h_1^{l_1} k_1^{d_1} h_2^{l_2} (k_2+\alpha_2)^{\tilde{n}_2}.
\end{aligned}
$$

While for the mean of Greenwood formula on $t_2 < t \leq t_3$, we have

$$\begin{aligned}E[\hat{\Sigma}_G(t)] &= \sum \binom{n}{l_1\, d_1\, l_2\, d_2\, \tilde{n}_2}\left(1-\frac{d_1}{n_1}\right)^2\left(1-\frac{d_2}{n_2}\right)^2 \\ &\quad\times\left(\frac{d_1}{n_1(n_1-d_1)}+\frac{d_2}{n_2(n_2-d_2)}\right)\prod_{j=1}^{2} h_j^{l_j}k_j^{d_j}\alpha_2^{\tilde{n}_2} \\ &= I_1 + I_2.\end{aligned}$$

where

$$\begin{aligned}I_1 &= \sum \binom{n}{l_1\, d_1\, l_2\, d_2\, \tilde{n}_2}\left(1-\frac{d_1}{n_1}\right)^2\left(1-\frac{d_2}{n_2}\right)^2 \frac{d_1}{n_1(n_1-d_1)} h_1^{l_1}k_1^{d_1}h_2^{l_2}k_2^{d_2}\alpha_2^{\tilde{n}_2} \\ &= \sum \binom{n}{l_1\, d_1\, \tilde{n}_1}\left(1-\frac{d_1}{n_1}\right)^2 \frac{d_1}{n_1(n_1-d_1)} h_1^{l_1}k_1^{d_1} \\ &\quad\times \sum \binom{\tilde{n}_1}{l_2\, d_2\, \tilde{n}_2}\left(1-\frac{d_2}{n_2}\right)^2 h_2^{l_2}k_2^{d_2}\alpha_2^{\tilde{n}_2} \\ &= \sum \binom{n}{l_1\, d_1\, \tilde{n}_1}\left(1-\frac{d_1}{n_1}\right)^2 \frac{d_1}{n_1(n_1-d_1)} h_1^{l_1}k_1^{d_1} \\ &\quad\times\left\{\frac{\alpha_2^2}{(k_2+\alpha_2)^2}\alpha_1^{\tilde{n}_1}+\frac{(k_2+\alpha_2)^2-\alpha_2^2}{(k_2+\alpha_2)^2}h_2^{\tilde{n}_1}\right. \\ &\quad\left.+\frac{k_2\alpha_2}{(k_2+\alpha_2)^2}\sum\binom{\tilde{n}_1}{l_2}\frac{1}{n_2}h_2^{l_2}(k_2+\alpha_2)^{n_2}\right\} \\ &= I_{11}+I_{12}+I_{13}.\end{aligned}$$

With the same arguments as in the calculation of $E[\hat{\Sigma}_G(t)]$ on $t_1 < t \leq t_2$, we obtain,

$$\begin{aligned}I_{11} &= \frac{\alpha_2^2}{(k_2+\alpha_2)^2}\frac{k_1\alpha_1}{(k_1+\alpha_1)^2}\{\Gamma_n^{(1)}-\Gamma_n^{(2)}\}(k_1+\alpha_1,h_1) \\ I_{12} &= \frac{(k_2+\alpha_2)^2-\alpha_2^2}{(k_2+\alpha_2)^2}\frac{k_1h_2}{(k_1+h_2)^2}\{\Gamma_n^{(1)}-\Gamma_n^{(2)}\}(k_1+h_2,h_1) \\ I_{13} &= \frac{(k_2\alpha_2)}{(k_2+\alpha_2)^2}\sum\binom{n}{l_1\, d_1\, l_2\, n_2}\left(1-\frac{d_1}{n_1}\right)^2\frac{d_1}{n_1(n_1-d_1)}\frac{1}{n_2}h_1^{l_1}k_1^{d_1}h_2^{l_2}(k_2+\alpha_2)^{n_2}.\end{aligned}$$

For I_2, we have

$$I_2 = \sum \binom{n}{l_1\, d_1\, l_2\, d_2\, \tilde{n}_2}\left(1-\frac{d_1}{n_1}\right)^2\left(1-\frac{d_2}{n_2}\right)^2 \frac{d_2}{n_2(n_2-d_2)} h_1^{l_1}k_1^{d_1}h_2^{l_2}k_2^{d_2}\alpha_2^{\tilde{n}_2}$$

$$
\begin{aligned}
&= \sum \binom{n}{l_1\, d_1\, \tilde{n}_1} \left(1-\frac{d_1}{n_1}\right)^2 h_1^{l_1} k_1^{d_1} \sum \binom{\tilde{n}_1}{l_2\, d_2\, \tilde{n}_2} \left(1-\frac{d_2}{n_2}\right)^2 \frac{d_2}{n_2(n_2-d_2)} h_2^{l_2} k_2^{d_2} \alpha_2^{\tilde{n}_2} \\
&= \sum \binom{n}{l_1\, d_1\, \tilde{n}_1} \left(1-\frac{d_1}{n_1}\right)^2 h_1^{l_1} k_1^{d_1} \frac{k_2\alpha_2}{(k_2+\alpha_2)^2} \{\Gamma_{\tilde{n}_1}^{(1)} - \Gamma_{\tilde{n}_1}^{(2)}\}(k_2+\alpha_2, h_2) \\
&= \frac{k_2\alpha_2}{(k_2+\alpha_2)^2} \sum \binom{n}{l_1\, d_1\, l_2\, \tilde{n}_2} \left(1-\frac{d_1}{n_1}\right)^2 \left(\frac{1}{n_2}-\frac{1}{n_2^2}\right) h_1^{l_1} k_1^{d_1} h_2^{l_2} (k_2+\alpha_2)^{n_2}.
\end{aligned}
$$

Therefore, for $t_2 < t \le t_3$,

$$
\begin{aligned}
&\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \\
&\quad \ge \frac{\alpha_2^2}{(k_2+\alpha_2)^2} \frac{k_1\alpha_1}{(k_1+\alpha_1)^2} \Gamma_n^{(2)}(k_1+\alpha_1, h_1) \\
&\quad + \frac{(k_2+\alpha_2)^2 - \alpha_2^2}{(k_2+\alpha_2)^2} \frac{k_1 h_2}{(k_1+h_2)^2} \Gamma_n^{(2)}(k_1+h_2, h_1) \\
&\quad + \frac{k_2\alpha_2}{(k_2+\alpha_2)^2} \sum \binom{n}{l_1\, d_1\, l_2\, \tilde{n}_2} \left(1-\frac{d_1}{n_1}\right)^2 \Big\{A\Big\} h_1^{l_1} k_1^{d_1} h_2^{l_2} (k_2+\alpha_2)^{n_2}.
\end{aligned}
$$

Since

$$
\begin{aligned}
A &= \frac{1}{n_2} - \left\{ \frac{1}{n_2} - \frac{1}{n_2^2} + \frac{d_1}{n_1(n_1-d_1)} \frac{1}{n_2} \right\} = \frac{1}{n_2^2} - \frac{d_1}{n_1(n_1-d_1)} \frac{1}{n_2} \\
&= \frac{1}{n_2} \left(\frac{1}{n_2} - \frac{1}{n_1-d_1} + \frac{1}{n_1} \right) = \frac{1}{n_2} \left(\frac{1}{n_2} - \frac{1}{\tilde{n}_1} + \frac{1}{n_1} \right) > 0,
\end{aligned}
$$

so, $\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \ge 0$ holds on $t_2 < t \le t_3$.

Notice that if we fix (l_1, d_1) and replace n by $\tilde{n}_1 = n - l_1 - d_1$, then, the expression $\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)]$ on $(t_2, t_3]$ is similar to that on $(t_1, t_2]$. With this idea, we can complete the proof by mathematical induction. Assume that we have proved $\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \ge 0$ holds on $t_j < t \le t_{j+1}$. On the interval $t_{j+1} < t \le t_{j+2}$, for each pair of (l_1, d_1), freeze it temporarily, regard the sample size now as $n - l_1 - d_1$, and t_i as t'_{i-1}, $(i = 1, \ldots, j+2)$, then, by the same arguments as on interval $(t_j, t_{j+1}]$ and the mathematical induction assumption, we have the corresponding $\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \ge 0$. Defreezing (l_1, d_1) and taking expectation on all possible pairs of (l_1, d_1) leads to the inequality $\operatorname{var}[\hat{S}(t)] - E[\hat{\Sigma}_G(t)] \ge 0$ on $t_{j+1} < t \le t_{j+2}$. □

Now, we investigate the Peto estimator (PE) proposed by Peto et al. (1977) and the homogenetic estimator (HE) proposed by Zhao (1996). Specifically, on the interval $t_j < t \le t_{j+1}$, the PE can be written as $\hat{\Sigma}_P(t) = \hat{S}(t)^2\{1 - \hat{S}(t)\}/n_j$, while the HE as $\hat{\Sigma}_H(t) = \hat{S}^2(t)\{1 - \hat{S}(t)\}/(n_j - d_j)$. Slud et al. (1984) proved that $\lim_{n\to\infty} n\hat{\Sigma}_G(t) \le \lim_{n\to\infty} n\hat{\Sigma}_P(t)$, and pointed out that the PE can seriously overestimate the variance of the PLE. Shen (2002) conducted a simulation to show that the HE can overestimate the variance of the PLE too. It is not so clear whether these two estimators always overestimate the variance of the PLE. Here we show by simple arguments that both PE and HE can either overestimate or underestimate the variance of the PLE.

Just check with the interval $t_1 < t \leq t_2$. On $t_1 < t \leq t_2$, we have

$$\hat{\Sigma}_H(t) = \frac{\hat{S}^2(t)\{1 - \hat{S}(t)\}}{n_1 - d_1} = \frac{d_1 \tilde{n}_1}{n_1^3} = \hat{\Sigma}_G(t).$$

Hence, HE, as the Greenwood formula, can underestimate the variance of the PLE. Since HE is always larger than PE, PE thus can also underestimate the variance of the PLE. On the other hand, in the case of no censoring,

$$\hat{\Sigma}_H(t) = \frac{\hat{S}^2(t)\{1 - \hat{S}(t)\}}{n_j - d_j} = \frac{n_j}{n_{j+1}} \frac{\hat{S}(t)\{1 - \hat{S}(t)\}}{n}.$$

Since $n_j/n_{j+1} > 1$, so the HE will overestimate its estimand when there is no censoring. Conclusively, both of these estimators can either overestimate or underestimate the variance of the PLE.

5 Discussion

Whereas much is known about the asymptotic properties for the PLE and the NE when failure time is continuous and sample size is large, relatively little is explored for the corresponding results when failure time is discrete and sample size is small. As revealed in this paper, the estimation of the variances of the PLE and the NE with small sample size is not so straightforward. Although several estimators for the variance of the PLE have been proposed, none of them is consistent in general. Therefore, additional information may be needed in choosing a desirable variance estimator in applications.

Acknowledgement The author would like to thank the referees for their insightful suggestions and constructive comments which lead to significant improvement of this paper.

References

Breslow, N. E., Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics, 2*, 437–453.

Brookmeyer, R., Crowley, J. (1982). A confidence interval for the median survival time. *Biometrics, 38*, 29–41.

Chen, Y. Y., Hollander, M., Langberg, N. A. (1982). Small-sample results for the Kaplan–Meier estimator. *Journal of the American Statistical Association, 77*, 141–144.

Guerts, J. H. J. (1985). Some small-sample non-proportional hazard results for the Kaplan–Meier estimator. *Statistica Neerlandica, 39*, 1–13.

Guerts, J. H. J. (1987). On the small-sample performance of Efron's and Gill's version of the product limit estimator under nonproportional hazards. *Biometrics*, *43*, 683–692.

Fleming, T. R., Harrington, D. P. (1991). *Counting processes and survival analysis.* New York: Wiley.

Greenwood, M. (1926). The natural duration of cancer, reports on public health and medical subjects. *Her Majesty's Stationery Office, 35*, 1–39.
Jennison, C., Turnbull, B. W. (1985). Repeated confidence interval for the median survival time. *Biometrika, 72*, 619-625.
Kalbfleisch. J. D., Prentice, R. L. (1981). *The statistical analysis of failure time data.* New York: Wiley.
Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.
Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology, 1*, 27–52.
Peto, R., Pike, M. C., Armitage, P., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *II Analysis and Examples. British Journal of Cancer, 35*, 1–39.
Phadia, E. G., Shao, P. Y. S. (1999). Exact moments of the product limit estimator. *Statistics and Probability Letters, 41*, 277–286.
Shen, P. S. (2002). A note on the homogenetic estimate for the variance of the Kaplan–Meier estimator. *Communications in Statistics—Theory and Methods, 31*, 1595–1603.
Slud, E. V., Byar, D. P., Green, S. B. (1984). A comparison of Reflected vs. Test- based confidence intervals for the median survival time, based on censored data. *Biometrics, 40*, 587–600.
Willett, J. B., Singer, J. D. (1993). Investigating onset, cessation, relapse and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology, 61*, 952–965.
Zhao, G. L. (1996). The homogenetic estimate for the variance of survival rate. *Statistics in Medicine, 15*, 51–60.