#287

Variance Estimation from Censored Data

Liangjun Tang and Yuly Koshevnik
Department of Statistical Science
Southern Methodist University
Dallas, TX 75275-0332

Abstract [1]

The inference on the variance of the cumulative hazard function estimate and its asymptotic variance are discussed. Three kinds of estimates are analyzed. In the finite sample case, the mean and variance for the variance estimate under the Koziol-Green model are derived.

1. Introduction

Consider the study described by means of the following right random censoring model. Let T and C be the survival time and censoring time, respectively. Random variables $T_1, T_2, \ldots, T_n$ and $C_1, C_2, \ldots, C_n$ represent n independent copies (independent and identially distributed, within each group, or i.i.d.) of T and C, respectively. They can be thought of as samples drawn from the populations of T–values and C–values, with the cumulative distribution functions (CDF's), F and G, respectively. The observable data are not pairs (T, C) themselves, but their transforms, $(Y_1, D_1), (Y_2, D_2), \ldots, (Y_n, D_n)$, where $Y_i = \min(T_i, C_i)$ and $D_i = \mathbf{1}(T_i \leq C_i)$. The Kaplan-Meier estimate (KME) of a survival function, $S(t) = 1 - F(t)$, is

$$\hat{S}_{KM}(t) = \prod_{i:\ Y_{(i)} \leq t} \left(1 - \frac{1}{n - i + 1}\right)^{\delta_i} \tag{0.1}$$

where δ_i is the D-value associated with $Y_{(i)}$, that is, $\delta_i = D_j$, when $Y_{(i)} = Y_j$. See Kaplan and Meier (1958). Similarly, the KME of the cumulative hazard function (CHF), $\Lambda(t) = -\log S(t)$ is

$$\hat{\Lambda}_1(t) = \sum_{i:\ Y_{(i)} \leq t} -\delta_i \log\left(1 - \frac{1}{n - i + 1}\right). \tag{0.2}$$

By the Greenwood's formula, see Miller (1981), the variance estimate of (1) is

$$\hat{S}^2_{KM}(t) \cdot \sum_{i:\ Y_{(i)} \leq t} \frac{\delta_i}{(n - i + 1)(n - i)}.$$

Then, the variance estimate of (2) is

$$\hat{V}_1(t) = \sum_{i:\ Y_{(i)} \leq t} \frac{\delta_i}{(n - i + 1)(n - i)} \tag{0.3}$$

Here, we call it the KM variance estimate of estimated CHF (KMVE).

This paper is focused on the variance estimate for the estimated CHF. In Section 2 we discuss an alternative variance estimate, similar to the Nelson-Aalen estimate for a survival function. In Section 3 we derive the asymptotic variance by the influence curve method. In Section 4 a Bayes variance estimate is discussed, and also the minimax type of estimate is obtained. At last, in Section 5 the exact mean and variance are found for the special case of a Koziol-Green model and those estimates are compared.

[1] The second author's research was supported by NSF Grant DMS–931147.

2. Kaplan-Meier and Nelson-Aalen Variance Estimates

First, look at the KME. At the time $Y_{(i)}$, the estimated probability of individual $Y_{(i)}$, if not censored, facing death is

$$\hat{p}_i = \frac{1}{n-i+1}.$$

The instant hazard of $Y_{(i)}$ facing is

$$\hat{\lambda}_i = -\log(1-\hat{p}_i).$$

By using the δ-method and binomial distribution,

$$\begin{aligned} \mathrm{Var}(\hat{\lambda}_i) &\approx \frac{1}{(1-\hat{p}_i)^2}\cdot \mathrm{Var}(\hat{p}_i) \approx \frac{\hat{p}_i}{(n-i+1)(1-\hat{p}_i)} \\ &\approx \frac{1}{(n-i)(n-i+1)}. \end{aligned}$$

If $Y_{(i)}$ is censored, then $\hat{p}_i = 0$. So, we assume its instant hazard is zero and then variance is also zero. Hence, the variance of CHF estimate is

$$\mathrm{Var}(\hat{\Lambda}_1(t)) \approx \sum_{i:\ Y_{(i)}\le t} \frac{\delta_i}{(n-i+1)(n-i)}$$

that is (3).

Nelson and Aalen used the approximation,

$$\log(1-\frac{1}{n-i+1}) \approx -\frac{1}{n-i+1},$$

under which (2) becomes

$$\hat{\Lambda}_2(t) = \sum_{i:\ Y_{(i)}\le t} \frac{\delta_i}{n-i+1}.$$

So, the Nelson-Aalen estimate (NAE) of a survival function S(t) is

$$\hat{S}(t) = \prod_{i:\ Y_{(i)}\le t} \left(\exp(-\frac{1}{n-i+1})\right)^{\delta_i}.$$

At the time $Y_{(i)}$, the probability of individual $Y_{(i)}$ facing death is

$$\hat{p}_i = 1-\exp(-\frac{1}{n-i+1})$$

and its instant hazard is

$$\hat{\lambda}_i = \frac{1}{n-i+1}.$$

Similarly,

$$\begin{aligned}\mathrm{Var}(\hat{\lambda}_i) &\approx \frac{1}{n-i+1}\left(\exp(\frac{1}{n-i+1})-1\right)\\ &\approx \frac{1}{(n-i+1)^2}.\end{aligned}$$

So, we define the variance estimate

$$\hat{V}_2(t) = \sum_{i:\ Y_{(i)}\leq t} \frac{\delta_i}{(n-i+1)^2}. \tag{0.4}$$

We call it the NA variance estimate of estimated CHF (NAVE). which is a little smaller than the Kaplan-Meier type.

3. Asymptotic Variance

As we know, the asymptotic variance of KME for CHF is

$$\mathrm{V}(t) = \int_0^t \frac{dF(x)}{(1-F(x))^2(1-G(x))} \tag{0.5}$$

see Breslow and Crowley (1974). Our interest is how to estimate it. When F and G are not continuous, integration above should be specified. Like re-expressing the KME, see Peterson (1977), here, we express V(t) as follows.

$$\begin{aligned}\Psi(S^{(1)},S^{(0)},t) &= -\int_0^t \frac{dS^{(1)}(x)}{(1-S(x))^2}\\ &+ \sum_{s\leq t\ :\ \text{jumps of } S^{(1)}(s)} \left(\frac{1}{S^{(1)}(s^+)+S^{(0)}(s^+)} - \frac{1}{S^{(1)}(s^-)+S^{(0)}(s^-)}\right)\end{aligned} \tag{0.6}$$

if $S^{(1)}$ and $S^{(0)}$ have no common jump points, where $S^{(1)}(t)$ and $S^{(0)}(t)$ are sub-survival functions,

$$\begin{aligned}S^{(1)}(t) &= p(Y_i > t, D_i = 1)\\ S^{(0)}(t) &= p(Y_i > t, D_i = 0)\end{aligned}$$

and first term is the integral over the intervals on which $S^{(1)}(x)$ is continuous. Hence, in the finite sample case, the variance estimate should be

$$\Psi(\hat{S}_n^{(1)},\hat{S}_n^{(0)},t)$$

where $\hat{S}_n^{(1)}$ and $\hat{S}_n^{(0)}$ are empirical sub-survival function of $S^{(1)}$ and $S^{(0)}$, those are

$$\begin{aligned}\hat{S}_n^{(1)}(t) &= \frac{1}{n}\sum_{i=1}^{n} 1(Y_i > t, D_i = 1)\\ \hat{S}_n^{(0)}(t) &= \frac{1}{n}\sum_{i=1}^{n} 1(Y_i > t, D_i = 0)\end{aligned}$$

From the calculation, we obtain

$$\begin{aligned}\Psi(\hat{S}_n^{(1)}, \hat{S}_n^{(0)}, t) &= \sum_{i:Y_{(i)}\le t,\ \delta_i=1} \left(\frac{1}{n-i} - \frac{1}{n-i+1}\right) \\ &= \sum_{i:Y_{(i)}\le t} \frac{\delta_i}{(n-i)(n-i+1)},\end{aligned}$$

that is the KMVE as (3). Hence, it is a consistent estimate of asymptotic variance, i.e.

$$n\Psi(\hat{S}_n^{(1)}, \hat{S}_n^{(0)}, t) \rightarrow V(t)$$

as n goes to infinity.

Next, assume that F and G are continuous, hence $S^{(1)}$ and $S^{(0)}$ are continuous, and then derive the influence curves for the above asymptotic variance, which is an analog of deriving the asymptotic variance V(t) as (5), see Reid (1981). Set

$$\begin{aligned}S^{(1)}(t,\epsilon,x) &= S^{(1)}(t) + \epsilon\cdot(1 - D_x(t)), \\ S^{(0)}(t,\delta,y) &= S^{(0)}(t) + \delta\cdot(1 - D_y(t)),\end{aligned}$$

where $0 < \epsilon,\ \delta < 1$, $x \neq y$ and $D_x(t) = \mathbf{1}(x \le t)$. The influence curves can be defined as the derivatives with respect to ϵ and δ at $\epsilon = 0$ and $\delta = 0$, respectively. i.e.

$$\begin{aligned}\mathrm{IC}_1(S^{(1)}, S^{(0)}, t, x) &= \lim_{\epsilon\to 0} \frac{\Psi(S^{(1)}(t,\epsilon,x), S^{(0)}(t), t) - \Psi(S^{(1)}, S^{(0)}, t)}{\epsilon} \\ \mathrm{IC}_2(S^{(1)}, S^{(0)}, t, y) &= \lim_{\delta\to 0} \frac{\Psi(S^{(1)}(t), S^{(0)}(t,\delta,y), t) - \Psi(S^{(1)}, S^{(0)}, t)}{\delta}.\end{aligned}$$

Briefly they are denoted as $\mathrm{IC}_1(x)$ and $\mathrm{IC}_2(y)$. So,

$$\begin{aligned}\Psi(S^{(1)}(t,\epsilon,x), S^{(0)}(t), t) &= -\int_0^t \frac{dS^{(1)}(s)}{(S^{(1)}(s) + \epsilon\cdot(1 - D_x(t)) + S^{(0)}(s))^2} \\ &\quad + D_x(t)\left(\frac{1}{S^{(1)}(x) + S^{(0)}(x)} - \frac{1}{S^{(1)}(x) + \epsilon + S^{(0)}(x)}\right)\end{aligned}$$

Hence,

$$\mathrm{IC}_1(x) = 2\int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s) + S^{(0)}(s))^3} + \frac{D_x(t)}{(S^{(1)}(x) + S^{(0)}(x))^2}.$$

Similarly,

$$\Psi(S^{(1)}(t), S^{(0)}(t,\delta,y), t) = -\int_0^t \frac{dS^{(1)}(s)}{(S^{(1)}(s) + \delta\cdot(1 - D_y(t)) + S^{(0)}(s))^2}.$$

Since $S^{(1)}(t)$ has no jumps and the second term in (6) vanishes, we obtain:

$$\mathrm{IC}_2(y) = 2\int_0^{y\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}.$$

Due to the identity,

$$\mathrm{P}\,(Y_i \le s_1, D_i = 1, Y_i \le s_2, D_i = 0) = 0,$$

we have that $\mathrm{E}(\mathrm{IC}_1(x)\mathrm{IC}(y)) = 0$. Ignoring the higher order of the infinitesimal terms, the asymptotic variance can be found,

$$\begin{aligned}
\mathrm{As.Var}(\hat{V}(t)) &= \lim_{n\to\infty} n\cdot \mathrm{E}(\Psi(\hat{S}_n^{(1)}, \hat{S}_n^{(0)}, t) - \Psi(S^{(1)}, S^{(0)}, t))^2 \\
&= \mathrm{E}\,(\mathrm{IC}_1(x) - \mu_1 + \mathrm{IC}_2(y) - \mu_2)^2 \\
&= \mathrm{E}(\mathrm{IC}_1(x))^2 + \mathrm{E}(\mathrm{IC}_2(y))^2 - (\mu_1+\mu_2)^2
\end{aligned}$$

where $\mu_1 = \mathrm{E}(\mathrm{IC}_1(x))$ and $\mu_2 = \mathrm{E}(\mathrm{IC}(y))$. Integration by parts yields,

$$\begin{aligned}
\mu_1+\mu_2 &= -2\int_0^\infty \int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}\cdot d(S^{(1)}+S^{(0)})(s) \\
&\quad -\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^2} dS^{(1)}(x) \\
&= 2\int_0^\infty (S^{(1)}(x)+S^{(0)}(x)) \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^3} dS^{(1)}(x) \\
&\quad -\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^2} dS^{(1)}(x) \\
&= -\mathrm{V}(t).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}(\mathrm{IC}_1(x))^2 + \mathrm{E}(\mathrm{IC}_2(y))^2 = &-\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^4} dS^{(1)}(x) \\
&-4\int_0^\infty \left(\int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}\right)^2 \cdot d(S^{(1)}+S^{(0)})(x) \\
-4\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^2} &\int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}\cdot dS^{(1)}(s).
\end{aligned}$$

Among those terms, again by the integration by parts,

$$\text{Second Term} = 8\cdot \int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^2} \left(\int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}\right)\cdot dS^{(1)}(x).$$

So,

$$
\begin{aligned}
& \mathrm{E}(\mathrm{IC}_1(x))^2 + \mathrm{E}(\mathrm{IC}_2(y))^2 \\
= \; & -\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^4} dS^{(1)}(x) \\
& +4\int_0^\infty \frac{D_x(t)}{(S^{(1)}(x)+S^{(0)}(x))^2} \int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3} \cdot dS^{(1)}(s) \\
= \; & \int_0^t \frac{1}{(S^{(1)}(x)+S^{(0)}(x))^2} dV(x) \\
& -4\int_0^\infty \left(\int_0^{x\wedge t} \frac{dS^{(1)}(s)}{(S^{(1)}(s)+S^{(0)}(s))^3}\right) d\left(\int_x^\infty \frac{D_s(t)}{(S^{(1)}(s)+S^{(0)}(s))^2} \cdot dS^{(1)}(s)\right) \\
= \; & \int_0^t \frac{1}{(S^{(1)}(x)+S^{(0)}(x))^2} dV(x) + 4\int_0^t \left(\frac{V(t)-V(x)}{S^{(1)}(x)+S^{(0)}(x)}\right) \cdot dV(x).
\end{aligned}
$$

Finally, the asymptotic variance of $\hat{V}_1(t)$ can be written as

$$
\begin{aligned}
\mathrm{As.Var}(\hat{V}(t)) \;=\; & \int_0^t \frac{1}{(S^{(1)}(x)+S^{(0)}(x))^2} dV(x) \\
& +4\int_0^t \left(\frac{V(t)-V(x)}{S^{(1)}(x)+S^{(0)}(x)}\right) \cdot dV(x) - \mathrm{V}^2(t).
\end{aligned}
\tag{0.7}
$$

Since

$$
\begin{aligned}
\int_0^t \left(\frac{V(t)-V(x)}{S^{(1)}(x)+S^{(0)}(x)}\right) \cdot dV(x) \;\geq\; & \int_0^t (V(t)-V(x)) \cdot dV(x) \\
\geq\; & \frac{V^2(t)}{2},
\end{aligned}
$$

so,

$$
\mathrm{As.Var}(\hat{V}(t)) \geq \int_0^t \frac{1}{(S^{(1)}(x)+S^{(0)}(x))^2} dV(x) + \mathrm{V}^2(t).
$$

This shows the asymptotic variance of the variance estimate increases much faster than V(t). Since

$$
\lim_{n\to\infty} \sqrt{n}(\hat{V}_1(t) - \hat{V}_2(t)) = 0,
$$

this implies that the KMVE and NAVE have the similar asymptotic behavior.

4. Bayes Approach

Now, consider the Bayesian procedure. We choose the Dirichlet distribution as the prior with parameter α and the quadratic loss function, i.e.

$$L(F,\hat{F}) = \int_0^{+\infty} \left(F(t) - \hat{F}(t)\right)^2 dW(t)$$

where W is a nonnegative nondecreasing weight function on $\mathcal{R} = (0, +\infty)$. Then, the Bayes estimate is the posterior mean. According to Ferguson (1973), the Bayes CDF estimate and survival function estimate are

$$\begin{aligned} \hat{F}_n(t,\alpha) &= \frac{\alpha(t) + \sum_{i=1}^n D_{X_i}(t)}{\alpha(+\infty) + n}, \\ \hat{S}_n(t,\alpha) &= \frac{\alpha(+\infty) - \alpha(t) + \sum_{i=1}^n (1 - D_{X_i}(t))}{\alpha(+\infty) + n} \end{aligned}$$

respectively. Here $\alpha(t)$ is a nonnegative increasing function on $\mathcal{R}$ with the finite total mass, $\alpha(\mathcal{R}) = \alpha(+\infty)$ and $\alpha(0) = 0$. As $\alpha(\mathcal{R})$ goes to zero, the Bayes estimates transfer into empirical estimates, for a CDF or survival function. Compared to the non-Bayes CDF estimate, $\hat{F}_n(t,\alpha)$ is not unbiased, but it has a smaller variance. However, they have the same limiting distribution. For the censoring case, consider the KME of survival function S(t) given by (1). Set $N_Y(t) = \#(Y_i > t)$ and $N_{Y^-}(t) = \#(Y_i \geq t)$. Then the KME can be written as

$$\hat{S}_{KM}(t) = \frac{N_Y(t)}{n} \prod_{i:Y_{(i)} \leq t} \left(\frac{N_{Y^-}(Y_{(i)})}{N_Y(Y_{(i)})} \right)^{1-\delta_i}.$$

The Bayes estimate is

$$\begin{aligned} \hat{S}_{KM}(t,\alpha) &= \frac{\alpha(\mathcal{R}) - \alpha(t) + N_Y(t)}{\alpha(\mathcal{R}) + n} \prod_{i:Y_{(i)} \leq t} \left(\frac{\alpha(\mathcal{R}) - \alpha(Y_{(i)}) + N_{Y^-}(Y_{(i)})}{\alpha(\mathcal{R}) - \alpha(Y_{(i)}) + N_Y(Y_{(i)})} \right)^{1-\delta_i} \\ &= \frac{\alpha(\mathcal{R}) - \alpha(t) + N_Y(t)}{\alpha(\mathcal{R}) + n} \prod_{i:Y_{(i)} \leq t} \left(\frac{\alpha(\mathcal{R}) - \alpha(Y_{(i)}) + n - i + 1}{\alpha(\mathcal{R}) - \alpha(t) + n - i} \right)^{1-\delta_i} \end{aligned}$$

where we assume $\alpha(t)$ is continuous. See Susarla and van Ryzin (1976). Similarly, for the variance estimate,

$$\begin{aligned} \hat{V}(t) &= \sum_{i:Y_{(i)} \leq t} \frac{\delta_i}{N_{Y^-}(Y_{(i)}) N_Y(Y_{(i)})} \\ &= \frac{1}{N_Y(t)} - \frac{1}{n} - \sum_{i:Y_{(i)} \leq t} \frac{1 - \delta_i}{N_{Y^-}(Y_{(i)}) N_Y(Y_{(i)})} \end{aligned}$$

The Bayes estimate is,

$$\begin{aligned}\hat{V}_\alpha(t) &= \frac{\alpha(\mathcal{R})+n}{n(\alpha(\mathcal{R})-\alpha(t)+N_Y(t))}-\frac{1}{n}\\ &\quad -\sum_{i:Y_{(i)}\le t}\frac{(\alpha(\mathcal{R})+n)^2}{n^2}\frac{1-\delta_i}{(\alpha(\mathcal{R})-\alpha(Y_{(i)})+N_{Y^-}(Y_{(i)}))(\alpha(\mathcal{R})-\alpha(Y_{(i)})+N_Y(Y_{(i)}))}\\ &= \frac{\alpha(t)+(n-N_Y(t))}{n(\alpha(\mathcal{R})-\alpha(t)+N_Y(t))}\\ &\quad -\frac{(\alpha(\mathcal{R})+n)^2}{n^2}\sum_{i:Y_{(i)}\le t}\frac{1-\delta_i}{(\alpha(\mathcal{R})-\alpha(Y_{(i)})+n-i+1)(\alpha(\mathcal{R})-\alpha(Y_{(i)})+n-i)}\end{aligned}$$

Last, consider the minimax approach. For a quadratic loss function, if the Bayesian estimate has a constant risk, then it is minimax. Recall the Bayes empirical CDF estimate is

$$\hat{F}_n(t,\alpha)=\frac{\alpha(t)+\sum_{i=1}^n D_{X_i}(t)}{\alpha(+\infty)+n}$$

Let Φ be a class of estimate of F(t),

$$\Phi=\{\phi:\phi(t)=a+\sum_{i=1}^n b_i\cdot D_{X_i}(t)\}$$

We try to find $\phi_0(t)\in\Phi$ such that the risk function, which is defined as the mean of loss, is a constant. From Prakasa Rao (1983) or Phadia (1973), it holds if and only if

$$a=\frac{1}{2(\sqrt{n}+1)},\quad b=\frac{1}{n+\sqrt{n}}.$$

So,

$$\phi_0(t)=\frac{1}{2(\sqrt{n}+1)}+\frac{\sqrt{n}}{\sqrt{n}+1}\cdot\hat{F}_n(t).$$

Equivalently,

$$\alpha(t)=\frac{\sqrt{n}}{2},\quad \alpha(\mathcal{R})=\sqrt{n},$$

that is a classical result of Bayes theory. This is the minimax CDF estimate. Similarly, the minimax survival function estimate is

$$\hat{S}^*(t)=\frac{1}{2(\sqrt{n}+1)}+\frac{\sqrt{n}}{\sqrt{n}+1}\cdot\hat{S}_n(t).$$

The minimax modification of the KM estimate is

$$\begin{aligned}\hat{S}^*_{KM}(t) &= \frac{\sqrt{n}/2+N_Y(t)}{\sqrt{n}+n}\prod_{i:Y_{(i)}\le t}\left(\frac{\sqrt{n}/2+N_{Y^-}(Y_{(i)})}{\sqrt{n}/2+N_Y(Y_{(i)})}\right)^{1-\delta_i}\\ &= \frac{1+2\sqrt{n}}{2(1+\sqrt{n})}\prod_{i:Y_{(i)}\le t}\left(\frac{\sqrt{n}/2+n-i}{\sqrt{n}/2+n-i+1}\right)^{\delta_i}\end{aligned}$$

The corresponding version of the variance estimate is

$$\begin{aligned}\hat{V}_3(t) &= \frac{\sqrt{n}/2+n-N_Y(t)}{n(\sqrt{n}/2+N_Y(t))} \\ &\quad -\frac{(\sqrt{n}+1)^2}{n}\cdot\sum_{i:Y_{(i)}\le t}\frac{1-\delta_i}{(\sqrt{n}/2+N_{Y-}(Y_{(i)}))(\sqrt{n}/2+N_Y(Y_{(i)}))} \\ &= \frac{\sqrt{n}/2+n-N_Y(t)}{n(\sqrt{n}/2+N_Y(t))}-\frac{(\sqrt{n}+1)^2}{n}\frac{n-N_Y(t)}{(\sqrt{n}/2+n)(\sqrt{n}/2+N_Y(t))} \\ &\quad +\frac{(\sqrt{n}+1)^2}{n}\sum_{i:Y_{(i)}\le t}\frac{\delta_i}{(\sqrt{n}/2+n-i+1)(\sqrt{n}/2+n-i)}\end{aligned}$$

Approximately, it can be written,

$$\hat{V}_3(t) = \frac{(\sqrt{n}+1)^2}{n}\sum_{i:Y_{(i)}\le t}\frac{\delta_i}{(\sqrt{n}/2+n-i+1)(\sqrt{n}/2+n-i)} \tag{0.8}$$

Here, we call it the minimax variance estimate of estimated CHF (MMVE). As the same augument of Section 3, $\hat{V}_3(t)$ has the same asymptotic behavior as $\hat{V}_1(t)$ and $\hat{V}_2(t)$.

5. Mean and Variance for the Koziol-Green Model

Generally speaking, it is very complicated to find the mean and variance of those variance estimates above. But, for the Koziol-Green model, these mean and variance depend on the values of CDF at observed time t and rate of censoring only. So, they are much easier to compute and tabulate.

Koziol and Green (1976) introduced a model, by assuming that the observed Y_i and indicator of censoring D_i are independent, or equivalently two CHFs are proportional. Let H(t) be the CDF of Y_i, and p be the probability of the uncensored rate. Then $n - N_Y(t)$ follows the binomial distribution with the success rate H(t), and D_i is Bernoulli distributed with the parameter p. Let E_1 and E_2 be the expectations with respect to above two distributions respectively. In this case, the survival function $S(t) = (1 - H(t))^p$, and the asymptotic variance is

$$V(t) = p\frac{H(t)}{1-H(t)} = p\frac{1-S^p(t)}{S^p(t)}. \tag{0.9}$$

Since Y_i and D_i are independent, the double integral can be replaced by the repeated integral. See Chen, Hollander and Langberg (1982). For any $\gamma > 0$, the moment of order γ for $\hat{\Lambda}_1(t)$ can be calculated as follows:

$$\mathrm{E}(\hat{\Lambda}_1^\gamma(t)) = \mathrm{E}\left(\sum_{i:Y_{(i)}\le t}-\delta_i\log\left(1-\frac{1}{n+1-i}\right)\right)^\gamma$$

$$\begin{aligned} &= \mathrm{E}_2\left(\mathrm{E}_1\left(\sum_{i=1}^{n-N_Y(t)} -\delta_i \log\left(1-\frac{1}{n+1-i}\right)\right)^{\gamma}\right) \\ &= \mathrm{E}_2\left(\sum_{q=1}^{n-1}\left(\sum_{i=1}^{q} -\delta_i \log\left(1-\frac{1}{n-q+1}\right)\right)^{\gamma} \cdot \binom{n}{q} H^q(t)\,(1-H(t))^{n-q}\right). \end{aligned}$$

In particular, the first and second moments are:

$$\begin{aligned} \mathrm{E}(\hat{\Lambda}_1(t)) &= \sum_{q=1}^{n-1} \log\left(\frac{n}{n-q}\right) p \cdot \binom{n}{q} H^q(t)\,(1-H(t))^{n-q}, \\ \mathrm{E}(\hat{\Lambda}_1^2(t)) &= \sum_{q=1}^{n-1}\left(p(1-p)\sum_{i=1}^{q} \log^2\left(\frac{n-i}{n+1-i}\right) + p^2\left(\log\frac{n}{n-q}\right)^2\right) \\ &\quad \cdot \binom{n}{q} H^q(t)\,(1-H(t))^{n-q}, \end{aligned}$$

respectively. To consider the mean and variance of the variance estimates above, we have to multiply those by the sample size n and n^2 respectively. Set the exact variance of CHF estimate above as:

$$V_0(t) = \left(\mathrm{E}(\hat{\Lambda}_1^2(t)) - \left(\mathrm{E}(\hat{\Lambda}_1(t))\right)^2\right) \cdot n \qquad (0.10)$$

Similarly, for the KMVE $\hat{V}_1(t)$, the first and second moments are:

$$\begin{aligned} \mathrm{E}(\hat{V}_1(t)) &= \sum_{q=1}^{n-1} \frac{q}{n(n-q)}\, p \cdot \binom{n}{q} H^q(t)\,(1-H(t))^{n-q} \\ \mathrm{E}(\hat{V}_1^2(t)) &= \sum_{q=1}^{n-1}\left(\sum_{i=1}^{q} \frac{p(1-p)}{(n+1-i)^2(n-i)^2} + p^2\left(\frac{q}{n(n-q)}\right)^2\right) \\ &\quad \cdot \binom{n}{q} H^q(t)\,(1-H(t))^{n-q} \end{aligned}$$

Hence, it makes sense to introduce the rescaled (asymptotic) mean and variance as

$$\mathrm{AM}(\hat{V}_1(t)) = E(\hat{V}_1(t)) \cdot n \qquad (0.11)$$

$$\mathrm{AV}(\hat{V}_1(t)) = \left(E(\hat{V}_1(t))^2 - (E(\hat{V}_1(t)))^2\right) \cdot n^2. \qquad (0.12)$$

Also, for the NAVE,

$$\mathrm{E}(\hat{V}_2(t)) = \sum_{q=1}^{n-1}\left(\sum_{i=1}^{q} \frac{1}{(n-i+1)^2}\right) p \cdot \binom{n}{q} H^q(t)(1-H(t))^{n-q},$$

$$\begin{aligned} \mathrm{E}(\hat{V}_2^2(t)) &= \sum_{q=1}^{n-1} \left(\sum_{i=1}^{q} \frac{p(1-p)}{(n-i+1)^4} + p^2 \left(\sum_{i=1}^{q} \frac{1}{(n-i+1)^2} \right)^2 \right) \times \\ &\quad \binom{n}{q} H^q(t)(1-H(t))^{n-q}, \end{aligned}$$

and for the MMVE $\hat{V}_3(t)$,

$$\begin{aligned} \mathrm{E}(\hat{V}_3(t)) &= \sum_{q=1}^{n-1} \frac{q}{(\sqrt{n}/2+n-q)(\sqrt{n}/2+n)} \times \\ &\quad p\binom{n}{q} H^q(t)(1-H(t))^{n-q} \cdot \frac{(\sqrt{n}+1)^2}{n}, \\ \mathrm{E}(\hat{V}_3^2(t)) &= \sum_{q=1}^{n-1} \{ \sum_{i=1}^{q} \frac{p(1-p)}{(\sqrt{n}/2+n-i+1)^2(\sqrt{n}/2+n-i)^2} \\ &\quad + p^2 \left(\frac{q}{(\sqrt{n}/2+n-q)(\sqrt{n}/2+n-q+1)} \right)^2 \} \times \\ &\quad \binom{n}{q} H^q(t)(1-H(t))^{n-q} \cdot \frac{(\sqrt{n}+1)^4}{n^2} \end{aligned}$$

Similarly, we define rescaled means and variances for $\hat{V}_2(t)$ and $\hat{V}_3(t)$ as in (11) and (12). Then calculate the exact variance (10), asymptotic variance (9) and the means and variances of those three estimates, which are tabulated below.

TABLE I.
Variance and Asymptotic Variance of CHF Estimate (2) for Koziol-Green Model.

	CDF	0.1	0.3	0.5	0.7
p=0.25	n=20, V	0.1427	1.0364	3.6281	1.8883
	n=100, V	0.1332	0.8305	4.7584	37.5694
	n=500, V	0.1315	0.7986	3.8938	51.2521
	n=∞, AV	0.1310	0.7912	3.7500	30.6143
p=0.5	n=20, V	0.1255	0.5939	1.9811	7.1222
	n=100, V	0.1188	0.5328	1.5775	5.9962
	n=500, V	0.1176	0.5228	1.5145	5.2004
	n=∞, AV	0.1173	0.5204	1.5000	5.0556
p=0.75	n=20, V	0.1204	0.5049	1.3761	4.0981
	n=100, V	0.1145	0.4653	1.1768	3.2013
	n=500, V	0.1134	0.4584	1.1470	3.0240
	n=∞, AV	0.1131	0.4567	1.1399	2.9845

TABLE II.
Mean and Variance of KMVE for Koziol-Green Model.

	CDF	0.1	0.3	0.5	0.7
p=0.25	n=20, mean	0.1422	1.0274	2.4274	0.6843
	n=100, mean	0.1331	0.8269	4.6249	11.1228
	n=500, mean	0.1314	0.7979	3.8782	39.6736
p=0.5	n=20, mean	0.1251	0.5844	1.9341	4.7732
	n=100, mean	0.1188	0.5314	1.5647	5.8007
	n=500, mean	0.1176	0.5225	1.5122	5.1730
p=0.75	n=20, mean	0.1201	0.4983	1.3269	3.9453
	n=100, mean	0.1144	0.4642	1.1698	3.1486
	n=500, mean	0.1134	0.4582	1.1457	3.0148
p=0.25	n=20, variance	0.0154	1.7006	12.5355	5.7689
	n=100, variance	0.0025	0.0988	22.0388	331.6009
	n=500, variance	0.0005	0.01643	1.2121	2165.8523
p=0.5	n=20, variance	0.0097	0.1461	3.4166	21.7371
	n=100, variance	0.0015	0.0185	0.2144	11.7610
	n=500, variance	0.0003	0.0035	0.3590	0.9732
p=0.75	n=20, variance	0.0084	0.0753	0.7464	11.1144
	n=100, variance	0.0015	0.0115	0.0715	0.8504
	n=500, variance	0.0003	0.0022	0.0130	0.1331

TABLE III.
Mean and Variance of NAVE for Koziol-Green Model.

	CDF	0.1	0.3	0.5	0.7
p=0.25	n=20, mean	0.1329	0.8441	1.6314	0.4330
	n=100, mean	0.1314	0.8045	4.0555	7.4706
	n=500, mean	0.1311	0.7938	3.8088	32.1314
p=0.5	n=20, mean	0.1180	0.5359	1.6026	3.3145
	n=100, mean	0.1174	0.5232	1.5237	5.3557
	n=500, mean	0.1173	0.5210	1.5045	5.1082
p=0.75	n=20, mean	0.1136	0.4644	1.1911	3.1198
	n=100, mean	0.1132	0.4581	1.1488	3.0481
	n=500, mean	0.1131	0.4570	1.1416	2.9965
p=0.25	n=20, variance	0.0131	0.7279	4.1114	1.9517
	n=100, variance	0.0024	0.0909	10.5353	112.9980
	n=500, variance	0.0005	0.0162	1.1375	822.4976
p=0.5	n=20, variance	0.0085	0.1111	1.4947	7.4112
	n=100, variance	0.0016	0.0178	0.1970	7.4747
	n=500, variance	0.0003	0.0034	0.0354	0.9304
p=0.75	n=20, variance	0.0075	0.0625	0.4787	4.3072
	n=100, variance	0.0015	0.0111	0.0678	0.7570
	n=500, variance	0.0003	0.0022	0.0129	0.1305

TABLE IV.
Mean and Variance of MMVE for Koziol-Green Model.

	CDF	0.1	0.3	0.5	0.7
p=0.25	n=20, mean	0.1609	0.8066	1.1860	0.2904
	n=100, mean	0.1423	0.7784	2.5240	2.4754
	n=500, mean	0.1357	0.7783	3.0024	8.6391
p=0.5	n=20, mean	0.1466	0.6186	1.5505	2.5177
	n=100, mean	0.1288	0.5540	1.4845	3.9377
	n=500, mean	0.1221	0.5334	1.4801	4.3877
p=0.75	n=20, mean	0.1421	0.5578	1.3138	2.8251
	n=100, mean	0.1246	0.4943	1.1917	2.8512
	n=500, mean	0.1180	0.4721	1.1574	2.8909
p=0.25	n=20, variance	0.0182	0.3746	1.4397	0.7134
	n=100, variance	0.0028	0.0696	1.2597	6.9892
	n=500, variance	0.0005	0.0143	0.4540	8.1137
p=0.5	n=20, variance	0.0128	0.1211	0.7603	2.7477
	n=100, variance	0.0020	0.0187	0.1482	1.5594
	n=500, variance	0.0004	0.0035	0.0309	0.4820
p=0.75	n=20, variance	0.0115	0.0802	0.3995	1.7764
	n=100, variance	0.0018	0.0123	0.0644	0.4623
	n=500, variance	0.0003	0.0023	0.0125	0.1042

6. Conclusions

From the analysis above, three kinds of the variance estimate are found. They have the same asymptotic behavior. For a finite sample, the special case of the Koziol-Green model is studied. By the rough comparisons, the KMVE is closer to the exact variance (10), but the MMVE has a smaller variance and NAVE is closer to the asymptotic variance (9). For the high values of CDF, since the variance and asymptotic variance increase very fast, they are less meaningful and then are omitted here.

Bibliography

Breslow, N. and Crowley, J.(1974): A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship, Annals of Statistics 2 437-453.

Chen, Y. Y., Hollander, M., and Langberg, N. A.(1982): Small-Sample Results for the Kaplan-Meier Estimator, JASA 77, 141-144.

Ferguson, T. S. (1973): A Bayesian Analysis of Some Nonparametric Problems, Annals of Statistics, 1, 209-230.

Kaplan, E. L. and Meier, P.(1958): Nonparametric Estimation from Incomplete Observations, JASA 53, 457-81.

Koziol, J.A. and Green, S.B.(1976): A Cramer-Von Mises Statistic for Randomly Censored Data. Biometrika 63, 465-74.

Miller, R. G.(1981): *Survival Analysis,* John Wiley and Sons.

Peterson, A. V.(1977): Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions, JASA 72, 854-858.

Phadia, E. G. (1973): Minimax Estimation of a Cumulative Distribution Function, Annals of Statistics, 1, 1149-1157.

Prakasa Rao, B.L.S. (1983): *Nonparametric Functional Estimation,* Academic Press.

Reid, N.(1981): Influence Functions for Censored Data, Annals of Statistics 9, 78-92.

Susarla, V. and van Ryzin, J. (1976): Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations, JASA 71, 897-902.