CONSISTENCY OF THE KAPLAN-MEIER ESTIMATOR WHEN DATA ARE INDEPENDENT NON-IDENTICALLY DISTRIBUTED

Mai Zhou

Department of Statistics, University of Kentucky, Lexington, KY 40506-0027

The Kaplan-Meier estimator computed from non identically distributed observations (for both survival and censoring times) is not estimating the simple average of the survival functions. The difference of the two can be substantial in the tails. The Greenwood formula is less sensitive. We then examine applications of this related to the interim power analysis in the double-blind two sample tests.

Key Words and Phrases: Censored data, the Nelson-Aalen estimator, Greenwood formula, Interim analysis, Double blind.

1. Introduction, Notation and Preliminary

The Kaplan-Meier (1958) estimator is a popular estimator in survival analysis. It estimates the distribution (survival) function when observations are subject to right censoring. But most studies of the large sample property of the Kaplan-Meier estimator assume that the observations are iid (Breslow and Crowley, 1974; Gill 1980) or at least one set of the survival times/censoring times are iid (Zhou 1991). Yang (1997) introduces some weighted Kaplan-Meier estimators. He had some limit result (TH2) restricted the setting to be a regression set up.

In practice, however, it often happens that neither the survival times nor the censoring times are iid. This could be due to neglected covariates that are different for different patients or gradual change of environment and lifestyle over time, etc. that are not accounted for. For instance, Andersen et. al. (1993) calculated Kaplan-Meier estimator (p.268) and Nelson-Aalen estimator (p.213) for 79 male patients with Malignant Melanoma *assuming* the survival times are iid. Further analysis later reveal that thickness of the tumor among the 79 patients do influence the survival times. Thus, the group of 79 patients are not really iid since they have different tumor thickness. Similarly, other covariate effects on the competing risk can make the censoring times non iid.

Another instance that this problem could arise is in a double blind two sample clinical trial. In the trial often the need of early/interim evaluation arise *before* un-blinding is authorized. In the absence of the identifying information for control/treatment, i.e. still blinded, if we treat them as one combined sample then we are dealing with a situation of non-identical distribution in both the life time and censoring, at least when the null hypothesis do not hold. See Shih (1992) and Govindarajuru (1998).

What is the limit, if exist, of the Kaplan-Meier estimator computed from those data? Is it the average of the survival functions? Another way of posing the same question is: how robust is the Kaplan-Meier estimator against the assumption of identical distribution? Our Theorem 2.1 show that the limit of the Kaplan-Meier estimator do exist but is not the average of the survival functions. Simulation in section 4 show how different this limit could be from the the average of the survival functions.

How is the Greenwood's formula affected? In particular, if a two sample test is to be based on the difference of 5 year survival probabilities of treatment and control, can we get some estimate of the power of the eventual test without un-blinding? Shih (1992) and Govindarajulu (1998) considered a similar testing problem for normal populations and without censoring.

The rest of this section is to establish notation and a lemma. Suppose we have two sets of non-negative random variables: $Y_1, Y_2, \cdots, Y_n$ which are censoring times, independent but non-identically distributed with continuous distribution functions $G_1(t)$, $G_2(t)$, $\cdots$, $G_n(t)$; and $X_1, X_2, \cdots, X_n$ which are survival times, also independent but non-identically distributed with continuous distribution functions $F_1(t), F_2(t), \cdots, F_n(t)$. We also assume the Y_i's are independent of the X_i's. To make the minimal digress to technical details we also assume all the distributions are continuous.

The actual data that we can observe are

$$Z_i = \min(Y_i, X_i), \ \ \delta_i = I_{[Z_i = X_i]} \quad i = 1, 2, \cdots n. \tag{1.1}$$

A popular estimator based on (1.1) is the Kaplan-Meier (1958) estimator

$$1 - \hat{F}_K(t) = \prod_{s \le t} \left(1 - \frac{\Delta N(s)}{R^+(s)} \right) \ , \quad \text{for} \quad t \le T^n \ , \tag{1.2}$$

here $R^+(t) = \sum I_{[Z_i \ge t]}$; $N(t) = \sum I_{[Z_i \le t, \, \delta_i = 1]}$; $\Delta N(s) = N(s) - N(s-)$ and $T^n = \max_{i \le n}\{Z_i\}$. Another popular estimator is the Nelson-Aalen estimator of cumulative hazard function

$$\hat{\Lambda}(t) = \sum_{s \le t} \frac{\Delta N(s)}{R^+(s)} = \int_0^t \frac{dN(s)}{R^+(s)} \, , \quad \text{for } t \le T^n \ . \tag{1.3}$$

The Altshuler's estimator of survival function, which is related to the Nelson-Aalen estimator, is given by

$$1 - \hat{F}_A(t) = e^{-\hat{\Lambda}(t)} \, . \tag{1.4}$$

We further denote $1 - H_i(t) = P(Z_i \ge t) = [1 - G_i(t)] \, [1 - F_i(t)]$ and $\Lambda_i(t) = \int_0^t (1 - F_i(s))^{-1} dF_i(s)$.

The following lemma provides a link between Nelson-Aalen estimator and Kaplan-Meier estimator through Altshuler's estimator. We can therefore deduce the property for the Kaplan-Meier estimator by those of Nelson-Aalen estimator. Notice this lemma is purely algebraic in nature and has nothing to do with random variable or distributions. Therefore it is valid for non iid data too.

Lemma 1.1 *Let $\hat{F}_K(t)$, $\hat{F}_A(t)$ be the Kaplan-Meier and the Altshuler's estimator defined in (1.2) and (1.4) above. We have, $\forall t$, if $1-\hat{F}_K(t)>0$, then*

$$\left|\hat{F}_K(t)-\hat{F}_A(t)\right| < \left[1-\hat{F}_K(t)\right]\frac{4}{R^+(t)}\,.$$

PROOF: See Cuzick (1985) for a similar inequality and proof. This exact inequality was first stated without proof in Gu (1987), Lemma 2.2. The proof is purely algebraic. $\diamondsuit$

2. Limits of Estimators

The following theorem identifies the limit of the Kaplan-Meier and Nelson-Aalen estimator computed from non iid observations. Notice the limit of the Kaplan-Meier estimator in Theorem 2.1 not only depends on the $F_i(t)$ but also involves the censoring distributions $G_i(t)$. In contrast, when at least one set of survival times/censoring times are iid, the almost sure limit of the Kaplan-Meier estimator is free from $G_i(t)$ (cf. Zhou 1991).

Theorem 2.1 *The Kaplan-Meier estimator $\hat{F}_K(t)$ and the Nelson-Aalen estimator $\hat{\Lambda}(t)$ converge almost surely as $n\to\infty$. In fact we have, if τ is such that $\frac{1}{n}E[R^+(\tau)]=\frac{1}{n}\sum 1-H_i(\tau)\ge\delta>0$ for all large n, then*

$$\lim_n \sup_{0<t<\tau}\left|\hat{F}_K(t)-\exp[-\int_0^t \frac{\sum[1-G_i(s)]dF_i(s)}{\sum[1-G_i(s)][1-F_i(s)]}]\right| = 0 \quad a.s. \tag{2.1}$$

Furthermore, for $\epsilon\le 2$ and $n>4/(\epsilon^2\delta^4)$,

$$P\left(\sup_{t<\tau}\left|\hat{\Lambda}(t)-\int_0^t \frac{\sum[1-G_i(s)]dF_i(s)}{\sum[1-G_i(s)][1-F_i(s)]}\right|>\epsilon\right) < 16(n+2)\exp[-\frac{n\delta^4\epsilon^2}{288}]. \tag{2.2}$$

For later reference, we define two sequences of (nonrandom) functions $F_n^*(t)$ and $\Lambda_n^*(t)$ to be the limit in (2.1) and (2.2). i.e. $\lim|\hat{F}_K(t)-F_n^*(t)|\to 0$.

We prelude the proof with three lemmas. The first lemma can be proved following Pollard (1984) pp.14-16. The remaining two are more or less consequences of the first lemma.

Lemma 2.1 *Let Z_i, $i = 1, \ldots, n$ and $P(Z_i \geq t) = 1 - H_i(t)$ be as defined in section one. We have $\forall \epsilon > 0$, for those n such that $\frac{1}{n} \leq \frac{\epsilon^2}{2}$*

$$P\left(\sup_t \left|\sum \frac{1}{n}\left[I_{[Z_i \geq t]} - 1 + H_i(t)\right]\right| > \epsilon\right) \leq 8(n+1)\exp\left[-\frac{n\epsilon^2}{8}\right]\ , \text{ and} \tag{2.3}$$

$$P\left(\sup_t \left|\sum \frac{1}{n}\left[I_{[Z_i \leq t, \delta_i = 1]} - EI_{[Z_i \leq t, \delta_i = 1]}\right]\right| > \epsilon\right) \leq 8(n+1)\exp\left[-\frac{n\epsilon^2}{8}\right]\ .$$

Lemma 2.2 *For any $\epsilon > 0$, if a real $\tau > 0$ and an integer n is such that $\frac{1}{n}E[R^+(\tau)] \geq \delta > 0$ and $\frac{1}{n} < \min(\frac{\epsilon^2}{2}, \frac{\epsilon^2\delta^4}{4})$, we have*

$$P\left(\sup_{t \leq \tau}\left|\int_0^t (\frac{1}{R^+(s)} - \frac{1}{E[R^+(s)]})dN(s)\right| > \epsilon\right) \leq 8(n+2)\exp[-\frac{n\min(\epsilon^2\delta^4, \delta^2)}{32}]\,.$$

Proof: Rewrite the integrals as

$$\sup_{t \leq \tau}\left|\int_0^t \frac{\frac{E[R^+(s)]}{n} - \frac{R^+(s)}{n}}{\frac{R^+(s)}{n}\frac{ER^+(s)}{n}} d\frac{1}{n}N(s)\right|.$$

Since $N(s)$ is increasing and $N(\tau) - N(0) \leq n$ the above is bounded by

$$\sup_{t \leq \tau}\frac{|\frac{E[R^+(t)]}{n} - \frac{R^+(t)}{n}|}{\frac{E[R^+(t)]}{n}\frac{R^+(t)}{n}}\int_0^\tau d\frac{1}{n}N(s) \leq \sup_{t \leq \tau}\frac{|\frac{E[R^+(t)]}{n} - \frac{R^+(t)}{n}|}{\frac{ER^+(t)}{n}\frac{R^+(t)}{n}}\ .$$

Notice that $\{R^+(t) \cdot ER^+(t)\}^{-1}$ is increasing in t, therefore the above is again bounded by

$$\sup_{t \leq \tau}\frac{1}{n}|ER^+(t) - R^+(t)|\frac{1}{\frac{ER^+(\tau)}{n}}\frac{1}{\frac{R^+(\tau)}{n}} \leq \frac{1}{\delta}\frac{1}{\frac{R^+(\tau)}{n}}\sup_{t \leq \tau}\frac{1}{n}|ER^+(t) - R^+(t)|\ ;$$

in view of the assumption of this lemma. Consequently, we have

$$P\left(\sup_{t \leq \tau}\left|\int_0^t (\frac{1}{R^+(s)} - \frac{1}{ER^+(s)})dN(s)\right| > \epsilon\right) \leq P\left(\frac{1}{\delta}\frac{1}{\frac{R^+(\tau)}{n}}\sup_{t \leq \tau}\frac{1}{n}|ER^+(t) - R^+(t)| > \epsilon\right). \tag{2.4}$$

By a similar argument to Lemma 2.1, we have

$$P\left(\left|\frac{R^+(\tau)}{n} - \frac{ER^+(\tau)}{n}\right| > \frac{\delta}{2}\right) < 8\exp[-\frac{n\delta^2}{32}]\,.$$

Therefore aside from a set with probability at most $8\exp[-\frac{n\delta^2}{32}]$, the term $\{\frac{R^+(\tau)}{n}\}^{-1}$ is bounded by $\frac{2}{\delta}$ in view of the assumption of this lemma. We thus have

$$\text{right hand side of (2.4)} \leq 8\exp[-\frac{n\delta^2}{32}] + P\left(\frac{1}{\delta}\frac{2}{\delta}\sup_{t \leq \tau}\frac{1}{n}|ER^+(t) - R^+(t)| > \epsilon\right)\,.$$

Now we can use the Lemma 2.1 to bound the second term and finish the proof.

Lemma 2.3 *For any* $\epsilon > 0$, *if* $\frac{1}{n}E[R^+(\tau)] \geq \delta > 0$, *we have*

$$P\left(\sup_{t\leq\tau}\left|\int_0^t \frac{d[N(s)-EN(s)]}{E[R^+(s)]}\right| > \epsilon\right) \leq 8(n+1)\exp\{-\frac{n\epsilon^2\delta^2}{72}\}\,.$$

Proof: Integration by parts, we obtain (for any $t \leq \tau$),

$$\left|\int_0^t \frac{1}{ER^+(s)} d[N(s)-EN(s)]\right| =$$

$$\left|\frac{N(s)-EN(s)}{ER^+(s)}|_0^t - \int_0^t [N(s)-EN(s)]d[\frac{1}{ER^+(s)}]\right|$$

$$\leq 2\sup_{0\leq t\leq\tau}\left|\frac{N(t)-EN(t)}{ER^+(t)}\right| + \sup_{0\leq t\leq\tau}|N(t)-EN(t)|\int_0^\tau d[\frac{1}{ER^+(s)}]$$

$$\leq 2\sup_{0\leq t\leq\tau}\frac{1}{n}\left|N(t)-EN(t)\right|\frac{1}{\frac{ER^+(\tau)}{n}} + \sup_{0\leq t\leq\tau}\frac{1}{n}\left|N(t)-EN(t)\right|\frac{1}{\frac{ER^+(\tau)}{n}}\,.$$

By the assumption of the lemma, $ER^+(\tau)/n \geq \delta$, we thus can continue

$$\leq \frac{3}{\delta}\sup_{0\leq t\leq\tau}\frac{1}{n}\left|N(t)-EN(t)\right|\,.$$

Therefore we have

$$P\left(\sup_{t\leq\tau}\left|\int_0^t \frac{d[N(s)-EN(s)]}{E[R^+(s)]}\right| > \epsilon\right) \leq P\left(\frac{3}{\delta}\sup_{t\leq\tau}\frac{1}{n}|N(t)-EN(t)| > \epsilon\right)\,.$$

Now use Lemma 2.1, we obtain the desired result.

Proof of Theorem 2.1: We first establish the probability inequality concerning the Nelson-Aalen estimator. The almost sure limit of the Kaplan-Meier estimator follows easily from this inequality, Lemma 1.1 and Borel-Cantelli Lemma.

By standard argument we have,

$$P\left(\sup_{t\leq\tau}\left|\hat{\Lambda}(t) - \int_0^t \frac{dEN(s)}{E[R^+(s)]}\right| > \epsilon\right)$$

$$\leq P\left(\sup_{t\leq\tau}\left|\int_0^t \left[\frac{1}{R^+(s)} - \frac{1}{E[R^+(s)]}\right] dN(s)\right| > \frac{\epsilon}{2}\right) + P\left(\sup_{t\leq\tau}\left|\int_0^t \frac{d[N(s)-EN(s)]}{E[R^+(s)]}\right| > \frac{\epsilon}{2}\right)\,.$$

Now use lemma 2.2 and 2.3 to bound the two probability terms on the right to finish proof. ♢

Appendix

Here we specialize the above consistency results to the iid case, where both the lifetimes and the censoring times are assumed to be iid. The results will be simpler and easier to remember. And the general case can be described as 'similar' results hold for non-identically distributed observations.

Theorem 2.1* *The Kaplan-Meier estimator $\hat{F}_K(t)$ and the Nelson-Aalen estimator $\hat{\Lambda}(t)$ converge almost surely, uniformly on a finite interval, as $n \to \infty$.*

In fact we have, if τ is a real number such that $\frac{1}{n}E[R^+(\tau)] = 1-H(\tau) = [1-F(\tau)][1-G(\tau)] \geq \delta > 0$, then for $\epsilon \leq 2$ and $n > 4/(\epsilon^2\delta^4)$,

$$P\left(\sup_{t\leq\tau}\left|\hat{\Lambda}(t) - \Lambda(t)\right| > \epsilon\right) < 16(n+2)\exp\left[-\frac{n\delta^4\epsilon^2}{288}\right]. \tag{1}$$

In particular, using the above inequality, we have for any $0 < \eta < 1/2$,

$$\sup_{t\leq\tau}\left|\hat{\Lambda}(t) - \Lambda(t)\right| = o(n^{-\eta}) \quad \textit{in probability} \tag{2}$$

and

$$\sup_{t\leq\tau}\left|\hat{\Lambda}(t) - \Lambda(t)\right| = o(n^{-\eta}) \quad a.s. \tag{3}$$

Similar inequality for the Kaplan-Meier estimator:

$$\sup_{0<t\leq\tau}\left|\hat{F}_K(t) - F(t)\right| = o(n^{-\eta}) \quad a.s. \tag{4}$$

The last result can be obtained by using the inequality $|1 - e^{-\hat{\Lambda}(t)} - \hat{F}_k(t)| < \frac{4[1-\hat{F}_k(t)]}{R^+(t)}$. One note: easy to show $4/R^+(\tau) = o(n^{-2\eta}) \quad a.s.$.

Remark: Since our basic inequality hold for EVERY finite n (not asymptotic), the inequality (1) above can also be explored to obtain results like when $\delta = \delta_n \to 0$ but $n\delta^4 \approx n^{0.00001}$.

We shall not pursue this here. (this will give uniform consistancy on expanding intervals, i.e. $\tau = \tau_n \to \infty$).

Remark: We can also obtain simpler results when the censoring is independent but non-identically distributed. Again, we leave this to reader.

Four useful facts about counting process martingales:

(1) $M_i(t) = I[T_i \leq t, \delta_i = 1] - \int_0^t I[T_i \geq s] d\Lambda(s)$ is an $\mathcal{F}_t$-martingale, $t \in [0, \infty)$.

We recall the filtration definition:

$$\mathcal{F}_t = \sigma\{[T_i \leq s]; s \leq t\}.$$

(2) $\langle M_i(t) \rangle = \int_0^t I[T_i \geq s] d\Lambda(s)$.

(3) $M(t) = \int_0^t g(s) dM_i(s)$ is also an $\mathcal{F}_t$-martingale, provided $g(\cdot)$ is a predictable function.

(4) $\langle M(t) \rangle = \int_0^t g^2(s) d\langle M_i(t) \rangle = \int_0^t g^2(s) I[T_i \geq s] d\Lambda(s)$.

3. Variance Estimates and Two Sample Tests

In this section we specialize the non-iid model (1.1) to the two sample case, i.e. $F_i(t) =$ either $F_1(t)$ or $F_2(t)$. We also examine the limit of Greenwood's formula for estimating variance under this setting.

The Greenwood formula states that the variance of the Kaplan-Meier estimator can be estimated by

$$\sigma^2_{GW} = (1 - \hat{F}_K(t))^2 \int_0^t \frac{dN(s)}{R^+(s)[R^+(s-)]} \ . \tag{3.1}$$

But this is developed for a sample with iid survival times. Will this still be a good estimator of the variance of the Kaplan-Meier estimators calculated on non-iid data?

Similar analysis to those in section two show that when sample size n is large, (3.1) approaches to

$$\frac{(1 - F^*_n(t))^2}{n} \int_0^t \frac{\sum[1 - G_i(s)]dF_i(s)}{(\sum[1 - G_i(s)][1 - F_i(s)])^2}$$

with $F_i =$ either F_1 or F_2, and $F^*_n(t)$ defined as in section two. The simulations in section four (for two or more samples) show that (3.1) is not too far from the (sample) variance of $\hat{F}_K(t)$.

Another, may be more relevant question is: how will (3.1) relate to the two variances of the two Kaplan-Meier estimators *had we been able* to separate the data into two iid samples? This question is seen to arise from the following situation.

In a two-arm double-blind clinical trial, the power analysis may require something like: have at least 80% power to detect a difference of more than 10% in 5-year survival rates. Or require to have at least 80% power when the ratio of the 5-year survival rates is 1.1 or above etc. Since logarithm of the survival is the (cumulative) hazard, the latter is equivalent of requiring the difference of cumulative hazards be larger than $\log 1.1$.

To design a trial to achieve such power needs information (or assumptions) on the variance of the Kaplan-Meier (or Nelson-Aalen) estimator which among other things depends on the unknown censoring patterns. In the interim analysis of the double-blind trial *without* unblinding (see Shih 1992), it is tempting to re-assess the variance assumption made in the design. Without unblinding we are dealing with a non-iid sample. We therefore ask if we can estimate the variances *without* unblinding.

Due to independence of the two arms, we have, at $t = 5$, say

$$Var(\hat{F}_1(5) - \hat{F}_2(5)) = Var(\hat{F}_1(5)) + Var(\hat{F}_2(5)). \tag{3.2}$$

If we had the un-blinding information we would use Greenwood formula to estimate the two variances on the right side of (3.2) separately, yielding the estimator (3.4) below.

Assume the two samples have approximate equal sample sizes (which many designed trials do), we suggest an estimator of the above variance *without* un-blinding be

$$4 \times [1 - \hat{F}(5)]^2 \int_0^5 \frac{dN(s)}{R^+(s)[R^+(s-)]} \ . \tag{3.3}$$

In the next section we compared this estimator with the "correct variance estimator"

$$[1 - \hat{F}_{1k}(5)]^2 \int_0^5 \frac{dN_1(s)}{R_1(s)[R_1(s-)]} + [1 - \hat{F}_{2k}(5)]^2 \int_0^5 \frac{dN_2(s)}{R_2(s)[R_2(s-)]} \ . \tag{3.4}$$

When the target power is specified in the ratio of survival rates or in the difference of the cumulative hazards, the estimator of $\log[1 - F(5)]$ is usually the Nelson-Aalen estimator rather then $\log[1 - \hat{F}_k(5)]$. The variance estimator of the Nelson-Aalen estimator in the iid case is (see, eg. Andersen et. al. 1993)

$$\hat{Var}(\hat{\Lambda}(t)) = \int_0^t \frac{dN(s)}{[R^+(s)]^2} \ .$$

Without un-blinding and assume approximate equal sample sizes we suggest to use

$$4 \times \int_0^t \frac{dN(s)}{[R^+(s)]^2} \tag{3.5}$$

to estimate the variance

$$Var(\hat{\Lambda}_1(t) - \hat{\Lambda}_2(t)) = Var(\hat{\Lambda}_1(t)) + Var(\hat{\Lambda}_2(t)) \ .$$

The "correct variance estimator" of the above is (when separation information are available)

$$\int_0^t \frac{dN_1(s)}{R_1^2(s)} + \int_0^t \frac{dN_2(s)}{R_2^2(s)} \ . \tag{3.6}$$

The two variance estimators, (3.5) and (3.6), will be compared in the simulation section below.

4. Simulations

There are numerous possibilities of how the observations can deviate from iid setting. We shall only look at two kinds of those non iid situations here: proportional hazards type and location shift type.

4.1 Example one: proportional hazards model

In our first simulation we took sample size 100 for each run and each entry in the table is based on 10,000 runs. Let $r_i = 1.01, 1.03, 1.05, \cdots, 2.99$ be fixed and the random observations be generated through (1.1) by

$$X_i \sim \exp(r_i); \qquad Y_i \sim \exp(r_i) \quad i = 1, 2, \cdots, 100. \tag{4.1}$$

The entry $\hat{S}_1(t)$ below is the Kaplan-Meier estimator for this (non iid) data setup. The sampleSd entry is the sample standard deviation of the 10,000 $\hat{S}_1(t)$. Greenwood entry is the average of 10,000 Greenwood estimator (of standard deviation) based on the non iid sample.

For comparison we also computed a Kaplan-Meier estimator from iid observations. The iid observations are generated from (4.1) except r_i are now identical to 2, the average of the $r_i^{\prime}s$. In other words, we used one distribution that has the average hazard of the non iid sample. This is reported as $\hat{S}_3(t)$.

t	$\hat{S}_1(t)$	$\hat{S}_3(t)$	$\hat{S}_{3.5}(t)$	Sample Sd	Greenwood
1.8	0.0759422	0.0597146	0.0611599	0.07084787	0.06555177*
1.1	0.1589313	0.1143878	0.1243290	0.07235635	0.06737697*
0.7	0.2874665	0.2467853	0.2627724	0.06788468	0.06708834
0.5	0.3990381	0.3681328	0.3847122	0.06419775	0.06443758
0.3	0.5657511	0.5485494	0.5634977	0.05713725	0.05799505
0.2	0.6790735	0.6708442	0.6825978	0.05048886	0.05160385
0.15	0.7462364	0.7408596	0.7506964	0.04600569	0.04675237
0.08	0.8535374	0.8519782	0.8579444	0.03639471	0.03645285

* the average is calculated after the NA's are removed.

Table 1. iid ($\hat{S}_3$, $\hat{S}_{3.5}$) versus proportional hazard ($\hat{S}_1$) data, Kaplan-Meier estimator

Our comparison in table 1 is also valid for the general proportional hazards situation as the following argument shows. If a strict increasing function $g(\cdot)$ were to apply to every X_i and Y_i we generated above, it will not change the censor/non-censor status of observations and it will not change the relative order of the observations. Thus the Kaplan-Meier estimators of the transformed data will be the same as those computed before transformation, except that the time t may be different. Since g(exponential random variable with rate r_i) is a random variable with survival function $e^{-r_i A(t)}$ with $A(t) = g^{-1}(t)$, we obtain a general proportional hazards model.

In view of proportional hazards model, we rewrite the parameters r_i as e^{β_i} as is customary when formulating a proportional hazards model. We therefore also include as comparison a third Kaplan-Meier estimator computed from iid observations, where the iid observations are generated by (4.1) except the parameters r_i are now all equal to $e^{\bar{\beta}}$, $\bar{\beta}$ been the average of the β_i. The results is reported as $\hat{S}_{3.5}(t)$ in table 1.

What is reported above represents a small portion of the simulations we did. For example, we computed Kaplan-Meier estimates when the censoring times and survival times are from a sample of exponential distribution with rates r_i generated independently and separately from a uniform distribution. The same pattern of difference still exists although less severe.

From the simulation we can see the following pattern:

(1). When the underlying data are not iid but follow a proportional hazards model, the Kaplan-Meier estimator computed from those non iid data is different from the Kaplan-Meier estimator computed from iid data that follows the average hazard.

(2). The largest difference occurs at the tail when $P(X > t)$ is around 10%. The relative difference can be large (20% – 30%).

(3). The difference is smaller when we replace iid data that follows the average hazard by iid data with hazard equal to the average of the β_i (where $r_i = e^{\beta_i}$) i.e. $\hat{S}_1$ is closer to $\hat{S}_{3.5}$ than is to $\hat{S}_3$.

(4). Greenwood's formula gives an estimate that is not too far from the sample standard deviation. See also the last two rows of table 3 and table 4 for a two sample simulation for this.

4.2 Example two: location shift model

Our second simulation will be the location shift non-iid case: let r_i be as before and

$$X_i \sim r_i + N(\mu = 100, sd = 10); \quad Y_i \sim r_i + N(\mu = 100, sd = 10). \quad i = 1, 2, \cdots, 100. \qquad (4.2)$$

The entry $\hat{S}_1(t)$ below is the average of 10,000 Kaplan-Meier estimates, each computed from non iid observations generated through (1.1) by the distributions (4.2) above. The entry $\hat{S}_3(t)$ below is the average of 10,000 Kaplan-Meier estimates, each computed from iid observations generated through (1.1) by (4.2) except with r_i all equal to 2, the average of r_i.

t	$\hat{S}_1(t)$	$\hat{S}_3(t)$	Sample Sd	Greenwood
94	0.7887621	0.7890140	0.04375003	0.04311677
100	0.5797238	0.5795154	0.05828902	0.05728733
102	0.5010833	0.5002693	0.06186100	0.06083789
106	0.3466497	0.3451841	0.06688993	0.06582047
112	0.1592873	0.1584064	0.07434973	0.06713526*
117	0.0796410	0.07936349	0.07269148	0.06561400*

* average computed after removing the NA's

Table 2. iid ($\hat{S}_3$) versus location shift ($\hat{S}_1$) data, Kaplan-Meier estimator

(5) The Kaplan-Meier estimator is less sensitive to the non-iid of shift alternative compared to the proportional hazards alternative.

4.3 Example three: two sample problem

In this example we take equal sample size in the two samples and $n = 100$ or $n = 50$. We generate

$$X_{11} \cdots X_{1n} \sim F_1(\cdot) \qquad X_{21}, \cdots, X_{2n} \sim F_2(\cdot) \tag{4.3}$$

and censoring distributions

$$C_{11} \cdots C_{1n} \sim G_1(\cdot) \qquad C_{21}, \cdots, C_{2n} \sim G_2(\cdot) \tag{4.4}$$

These were used to form the two samples of censored observations. In computing the blind version of the variance estimator we merge the two samples of censored observations into one.

We compute and compare (3.5) with (3.6) and (3.3) with (3.4) at time $t = 0.5$.

If the power analysis were to be performed based on the difference of the cumulative hazard functions at time t, or the (log of the) ratio of the survival functions at time t, we can proceed as follows:

Estimate the variance of $\hat{\Lambda}_1(t) - \hat{\Lambda}_2(t)$ by (3.5)$= Var$, then if the difference of the cumulative hazard at time t is η, the power of the test can be approximated by

$$P(|N(\eta, Var)| > 1.96\sqrt{Var}).$$

From the table 3 and 4 we see that

(6) the differences between the blind and un-blind variance estimators are small, often less then 3%. Considering the random fluctuations exit even when two samples are indeed iid

(under hull hypothesis, simu0, where both variance estimates are unbiased), these blind versions of variance estimators are a good substitute for the "correct" variance estimator.

	simu1	simu2	simu3	simu4	simu0
$F_1(\cdot)$	exp(1)	exp(1)	exp(1)	exp(1)	exp(1)
$G_1(\cdot)$	exp(1)	exp(1)	exp(1)	exp(1)	exp(1)
$F_2(\cdot)$	exp(0.9)	exp(1.15	exp(1.25)	exp(0.8)	exp(1)
$G_2(\cdot)$	exp(0.85)	exp(1.3)	exp(1.5)	exp(0.9)	exp(1)
sample size	100/100	100/100	100/100	100/100	50/50
(3.4)	.006030772	.006676352	.006948632	.005939583	.01239244
(3.3)	.006057681	.006689299	.006925038	.005984055	.01252191
(3.6)	.0159109	.02013357	.02238024	.01500604	.03509194
(3.5)	.0157728	.01972328	.02132062	.01478237	.03476628
Sample Var	.00147403	.001620016	.001686649	.001428136	.003142266
Greenwood	.00151442	.001672325	.00173126	.001496014	.003130477

Table 3. Comparison of blind and un-blind variance estimators

	simu5	simu6	simu7	simu8	simu9
$F_1(\cdot)$	exp(0.6)	exp(0.6)	exp(1.6)	exp(2)	exp(2.2)
$G_1(\cdot)$	exp(1)	U[1,9]	U[2,7]	U[0,5]	U[0,5]
$F_2(\cdot)$	exp(0.8)	exp(0.8)	exp(1.3)	exp(1.5)	exp(1.5)
$G_2(\cdot)$	exp(0.9)	U[2,10]	U[2,9]	U[1,7]	U[1,7]
sample size	100/100	100/100	100/100	100/100	100/100
(3.4)	.0053072	.004088382	.00492142	.004915672	.004808153
(3.3)	.0053737	.004133301	.00497129	.004990339	.004928135
(3.6)	.01097273	.008427386	.02148737	.02963218	.03271447
(3.5)	.01089167	.008348447	.02120381	.02828058	.03031182
Sample Var	.00131513	.001021852	.001240745	.001211992	.001180526
Greenwood	.00134342	.001033325	.001242823	.001247585	.001232034

Table 4. Comparison of blind and un-blind variance estimators

References

Andersen, P.K., Borgan, O., Gill, R. and Keiding, N. (1993), *Statistical Models Based on Counting Processes.* Springer, New York.

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

Cuzick, J. (1985), Asymptotic properties of censored linear rank tests *Ann. Statist.* 13, 133-141.

Gill, R. (1980), *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Mathematisch Centrum, Amsterdam.

Govindarajuru, Z. (1998). Robustness of a sample size re-estimation procedure in clinical trials. Tech. Report Univ. Kentucky Dept. of Statist.

Gu, M. G. (1987), *Sequential analysis of survival data in staggered-entry clinical trials.* Ph. D. thesis. Columbia University.

Kaplan, E. and Meier, P. (1958), "Non-parametric estimator from incomplete observations," *J. Amer. Statist. Assoc.* 53, 457–481.

Pollard, D. (1984), *Convergence of Stochastic Processes.* Springer, New York.

Shih, W.J. (1992). Sample size re-estimation in clinical trials. *Biopharmaceutical Sequential Statistical Applications.* 285-301. (Ed. K.E. Peace). Marcel Dekker, New York.

Zhou, M. (1991), Some properties of the Kaplan-Meier estimator for independent non-identically distributed random variables. *Ann. Statist.* 19, 2266-2274.

Yang, S. (1997). A generalization of the product-limit estimator with an application to censored regression. *Ann. Statist.* **25**, 1088-1108.