Hindawi
Computational and Mathematical Methods in Medicine
Volume 2020, Article ID 7827434, 11 pages
https://doi.org/10.1155/2020/7827434

Research Article

Improved Kaplan-Meier Estimator in Survival Analysis Based on Partially Rank-Ordered Set Samples

Samane Nematolahi,[1] Sahar Nazari,[2] Zahra Shayan,[1] Seyyed Mohammad Taghi Ayatollahi ,[1] and Ali Amanati[3]

[1] *Department of Biostatistics, Medical School, Shiraz University of Medical Sciences, Shiraz, Iran*
[2] *Department of Medicine, University of Alberta, Edmonton, Canada*
[3] *Professor Alborzi Clinical Microbiology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran*

Correspondence should be addressed to Seyyed Mohammad Taghi Ayatollahi; ayatolahim@sums.ac.ir

Received 19 January 2020; Revised 3 April 2020; Accepted 25 April 2020

Academic Editor: John Mitchell

This study presents a novel methodology to investigate the nonparametric estimation of a survival probability under random censoring time using the ranked observations from a Partially Rank-Ordered Set (PROS) sampling design and employs it in a hematological disorder study. The PROS sampling design has numerous applications in medicine, social sciences and ecology where the exact measurement of the sampling units is costly; however, sampling units can be ordered by using judgment ranking or available concomitant information. The general estimation methods are not directly applicable to the case where samples are from rank-based sampling designs, because the sampling units do not meet the identically distributed assumption. We derive asymptotic distribution of a Kaplan-Meier (KM) estimator under PROS sampling design. Finally, we compare the performance of the suggested estimators via several simulation studies and apply the proposed methods to a real data set. The results show that the proposed estimator under rank-based sampling designs outperforms its counterpart in a simple random sample (SRS).

1. Introduction

The idea of ranked set sampling (RSS) was introduced by McIntyre [1] for the first time. It can provide a more structural method for collecting the sample units. A generalization of RSS is the PROS sampling design. Both sampling methods are similar with a clear difference; in the PROS sampling design that we use in this paper, the ranker divides the sampling units into ranked subsets of prespecified sizes based on their partial ranks [2]. These sampling designs are techniques to obtain more representative samples from the underlying population where measurement of the units is costly and/or time-consuming. In such sampling designs, sampling units are ordered fairly accurately by using available auxiliary information which may be costly to some extent (see [3]).

After the PROS sampling design was introduced by Ozturk [4], many statisticians became interested in this rank-based sampling method. For example, Ozturk [5] and Frey [6] have relaxed the assumption concerning the prespecification of the number of subsets in each set. Nazari et al. [7] have developed nonparametric kernel density estimators using PROS data. Hatefi et al. [3] have applied PROS sampling in mixture modeling to estimate the age structures of short-lived fish species. Ozturk [8] have used the properties of PROS samples under multiple auxiliary information in the estimation of the population mean and total infinite population settings. Nazari et al. [9] have estimated the distribution function using PROS samples. Hatefi et al. [10] have studied the information and uncertainty structures of PROS data.

Currently, survival study is one of the important statistical tools for analyzing the data extracted from medical studies and social sciences. Presence of censoring observations is the distinction between survival analysis and other statistical

Table 1: An example of a Partially Rank-Ordered Set sample.

Cycle	Set	Subset	Observation
1	S_1	$D_1 = \{\mathbf{d}_1, d_2, d_3\} = \{\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}, \{4, 5, 6\}, \{7, 8, 9\}\}$	$X_{[d1]1}$
	S_2	$D_2 = \{d_1, \mathbf{d}_2, d_3\} = \{\{1, 2, 3\}, \{\mathbf{4}, \mathbf{5}, \mathbf{6}\}, \{7, 8, 9\}\}$	$X_{[d2]1}$
	S_3	$D_3 = \{d_1, d_2, \mathbf{d}_3\} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{\mathbf{7}, \mathbf{8}, \mathbf{9}\}\}$	$X_{[d3]1}$
2	S_1	$D_1 = \{\mathbf{d}_1, d_2, d_3\} = \{\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}, \{4, 5, 6\}, \{7, 8, 9\}\}$	$X_{[d1]2}$
	S_2	$D_2 = \{d_1, \mathbf{d}_2, d_3\} = \{\{1, 2, 3\}, \{\mathbf{4}, \mathbf{5}, \mathbf{6}\}, \{7, 8, 9\}\}$	$X_{[d2]2}$
	S_3	$D_3 = \{d_1, d_2, \mathbf{d}_3\} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{\mathbf{7}, \mathbf{8}, \mathbf{9}\}\}$	$X_{[d3]2}$

analyses (see [11]). However, survival analyses are expensive due to the need of a large sample size and the potentially long follow-up duration [12]. For the sake of parsimony, we may consider the cost-effective sampling methods, in which only a small proportion of the available units is measured; however, they contain a portion of the information contributed by all of the units; for more information, see [13].

In this study, we develop the KM nonparametric estimator using the PROS sampling design. The KM estimator measures the probability that a person survives longer than a specific time, which is fundamental in survival analysis. We study the asymptotic properties of this new estimator and compare it with SRS and RSS counterparts. What distinguishes the present research from previous endeavors is that we employ the PROS sampling design for incomplete data containing censored observations, while all research on PROS sampling design has been concerned with the inference procedure for complete data. There are only a few results available when the researcher has incomplete data and the sampling design is based on RSS not PROS samples. For example, Yu and Tam [14] have considered maximum likelihood estimation of parameters of the log-normal distribution and have introduced a KM estimator for RSS. Zhang et al. [15] have used RSS for estimating the KM estimator of a reliability function with random right-censored data where the population distribution is unknown. Strzalkowska-Kominiak and Mahdizadeh [13] have proposed a KM estimator based on RSS when censored data are under random detection limit assumption. Mahdizadeh and Strzalkowska-Kominiak [16] have proposed a confidence interval for a distribution function when data are right-censored with random censoring time by applying RSS design.

In Section 2, we present some primary notes. In Section 3, we introduce the nonparametric KM estimator. In Section 4, we show the asymptotic normality of the KM estimator based on imperfect PROS sampling design. We compare the performance of the PROS KM estimator with respect to its SRS and RSS counterparts using simulation studies in Section 5. In addition, we illustrate our proposed method with a real example. We consider a dataset collected in Amir Medical Oncology Center, as our population in Section 6.

2. Necessary Background

2.1. Ranked Set Sampling. To obtain a RSS of size nL, with set size n and L cycles, from the underlying population, a set of n units is randomly selected from the population. The units are ranked via some mechanisms. Then, the unit that ranked as the smallest was selected for the final measurement. Another set of n units is drawn and ranked, and the unit ranked as the second smallest is selected for measurement. This process is continued until the unit ranked as the maximum is selected and measured. This is one cycle of the RSS procedure; the cycle can be repeated L times to generate RSS of size nL (see [17]).

2.2. Partially Rank-Ordered Set Sampling. In this section, we introduce the PROS sampling design and present the necessary notation. This sampling design is of the form G^{**} design in Ozturk (see [4]). In order to extract a PROS sample of size $N = nL$, we choose a set size $s = nm$ and a design parameter $D = \{d_1, \cdots, d_n\}$ that partitions the set $\{1, 2, \cdots, s\}$ into n mutually exclusive subsets $d_j = \{(j-1)m + 1, \cdots, jm\}, j = 1, \cdots, n$. Sampling units are then assigned to the subsets d_j, $j = 1, \cdots, n$, based on visual inspection, judgment ranking, or using a concomitant variable such that all units in the subset d_j are judged to have smaller ranks than all units in the subset $d_{j'}$, when $j < j'$. A unit is then randomly selected from the subset d_1 for full measurement and denoted by $X_{[d_1]1}$. Again, we randomly select a set containing s units and assign them to n subsets; after that, we randomly draw a member from subset d_2 and denote it by $X_{[d_2]1}$. These steps are continued until we randomly extract a unit from d_n, $X_{[d_n]1}$. These observations constitute one cycle of the PROS sampling design; after L repetitions of this process, we achieve a PROS sample of size nL, denoted by $X_{\text{PROS}} = \{X_{[d_j]i}, i = 1, \cdots, L, j = 1, \cdots, n\}$; for more details, see [9].

Table 1 presents a simple example of the construction of a PROS sample when $s = 9, n = 3$, and $m = 3$, the cycle size is $L = 2$, and the design parameter is $D = \{d_1, d_2, d_3\} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}$. Each set contains nine units assigned to three partially rank-ordered subsets. In this process, units in each subset have equal chance to take any place in the subset. One unit, in each set from the bold-faced subset, is randomly drawn and measured. The resulting PROS sample is denoted by $\{X_{[d_j]i}, i = 1, 2, j = 1, 2, 3\}$.

It should be noted that, if all members in the subset d_j have exactly smaller ranks than all members in $d_{j'}, j < j'$, the PROS sampling design is perfect. Otherwise, we have an imperfect PROS sampling design. Suppose that α

is a doubly stochastic matrix; we model the subsetting error probabilities in the imperfect PROS as follows (see [7] and [9]):

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{d_1,d_1} & \cdots & \alpha_{d_1,d_n} \\ \vdots & \ddots & \vdots \\ \alpha_{d_n,d_1} & \cdots & \alpha_{d_n,d_n} \end{bmatrix}, \tag{1}$$

where α_{d_j,d_h} is the probability of assigning a unit into the subset d_j when it belongs to the subset d_h with $\sum_{h=1}^{n} \alpha_{d_j,d_h} = \sum_{j=1}^{n} \alpha_{d_j,d_h} = 1$.

Throughout this paper, we use $\mathrm{PROS}_\alpha(n, L, s, D)$ as a symbol of an imperfect PROS sampling design with the design $D = \{d_j, j = 1, \cdots, n\}$, where α represents a subsetting error probability matrix, n shows the number of subsets, and L and s exhibit the number of cycles and the set size, respectively. It should be pointed out that $m = s/n$.

SRS and RSS designs are special cases of the PROS sampling design when $s = 1$ and $s = n$, respectively. For a perfect PROS design, since $\alpha_{d_j,d_h} = 0$ for $h \neq j$ and $\alpha_{d_j,d_j} = 1$ for $j = 1, \cdots, n$, the subsetting error matrix is an identity matrix and the notation $\mathrm{PROS}_I(n, L, s, D)$ can be used.

In this paper, the cumulative distribution function (CDF) of the studied variable in the population, CDF of $X_{[d_j]i}$ for $i = 1, \cdots, L$, and CDF of the rth-order statistic among a simple random sample of size s are denoted by F, $F_{[d_j]}$, and $F_{(r:s)}$, respectively. In addition, the corresponding probability density functions are represented by f, $f_{[d_j]}$, and $f_{(r:s)}$.

3. Kaplan-Meier Estimator Based on PROS Sampling Design

Definition 1. Let $X_1, \cdots, X_n \sim F$ and $C_1, \cdots, C_n \sim G$ be two independent random variables where we observe $Y_i = \min\{X_i, C_i\} \sim H$ and $\delta_i = 1\{X_i \leq C_i\}$ be the indicator variable which specifies the event/censored status. The KM estimator defined as

$$1 - \widehat{F}_{\mathrm{SRS}}(t) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right)^{1\left\{Y_{(i)} \leq t\right\}}, \tag{2}$$

where $Y_{(1)}, \cdots, Y_{(n)}$ are ordered values of the simple random sample (SRS) with related $\delta_{[1]}, \cdots, \delta_{[n]}$ values; see [18] for more information.

Based on the above Definition 1 and Definition 1 in [9], we estimate the KM estimator based on the imperfect PROS sampling design $\mathrm{PROS}_\alpha(n, L, s, D)$.

The KM estimator based on the $\mathrm{PROS}_\alpha(n, L, s, D)$ sample, $\mathrm{X}_{\mathrm{PROS}}$, defined as

$$1 - \widehat{F}_{\mathrm{PROS}}(t) = \frac{1}{n} \sum_{j=1}^{n} \left(1 - \widehat{F}_{[d_j]}(t)\right), \tag{3}$$

where $1 - \widehat{F}_{[d_j]}$ is the KM estimator based on the independent and identically distributed (SRS) $\{Y_{[d_j]1}, Y_{[d_j]2}, \cdots, Y_{[d_j]L}\}$, defined as

$$1 - \widehat{F}_{[d_j]}(t) = \prod_{k=1}^{L} \left(1 - \frac{\delta^*_{[d_j]k}}{L - k + 1}\right)^{1\left\{Y^*_{[d_j]k} \leq t\right\}}, \tag{4}$$

where $Y^*_{[d_j]1}, \cdots, Y^*_{[d_j]L}$ are ordered values of $Y_{[d_j]1}, \cdots, Y_{[d_j]L}$ and $\delta^*_{[d_j]k}$ values are related to $Y^*_{[d_j]k}$ values for $k = 1, \cdots, L$.

4. Asymptotic Properties

In this section, we study the behavior of the nonparametric KM estimator in large samples based on the imperfect PROS sampling design. The asymptotic properties of the KM estimator under the SRS were widely available in the literature survey [19–21].

We demonstrate that no stronger assumptions are needed while using the imperfect PROS-based KM estimator. At first, we introduce the following lemma, which is a straight result of Lemma 2.1 in Stute and Wang [18].

Lemma 1. *Suppose $X \sim F$ and $C \sim G$ are two independent random variables. In addition, let $X_{[d_j]i} \sim F_{[d_j]}$ be the PROS sample from subset d_j in the ith cycle and $C_{[d_j]i}$ be the corresponding censored time.*

Set $H_{[d_j]}(t) = P(\min(X_{[d_j]i}, C_{[d_j]i}) \leq t)$, then we have

$$1 - H_{[d_j]}(t) = \left(1 - F_{[d_j]}(t)\right)(1 - G(t)). \tag{5}$$

Proof. See Appendix A.

Due to the expressed lemma, we can define

$$\begin{aligned} \tilde{H}^0_{[d_j]}(z) &= \int_{-\infty}^{z} \left(1 - F_{[d_j]}(y)\right) G(dy), \\ \tilde{H}^1_{[d_j]}(z) &= \int_{-\infty}^{z} (1 - G(y-)) F_{[d_j]}(dy). \end{aligned} \tag{6}$$

We also set

$$\gamma_{0d_j}(X) = \exp\left\{\int_{-\infty}^{x} \frac{\tilde{H}^0_{[d_j]}(dz)}{1 - H_{[d_j]}(z)}\right\}. \tag{7}$$

Let $\varphi(w)$ be a score function

$$\gamma_{1d_j}(X) = \frac{1}{1 - H_{[d_j]}(x)} \int 1\{x < w\}\varphi(w)\gamma_{0d_j}(w)\tilde{H}^1_{[d_j]}(dw),$$
$$\gamma_{2d_j}(X) = \iint \frac{1\{v < x, v < w\}\varphi(w)\gamma_{0d_j}(w)}{\left(1 - H_{[d_j]}(v)\right)^2} \tilde{H}^0_{[d_j]}(dv)\tilde{H}^1_{[d_j]}(dw). \tag{8}$$

Now, we present Theorem 1.

Theorem 1. *Assume F and G are continuous and*

$$\int \varphi^2(x)\gamma_0^2(x)\tilde{H}^1(dx) < \infty, \tag{9}$$

$$\int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{(1 - H(z))(1 - G(z))} \right)^{1/2} F(dx) < \infty, \tag{10}$$

where

$$\tilde{H}^0(z) = P(X \le z, \delta = 0) = \int_{-\infty}^{z} (1 - F(y))G(dy), z \in \mathbb{R},$$
$$\tilde{H}^1(z) = P(X \le z, \delta = 1) = \int_{-\infty}^{z} (1 - G(y))F(dy), z \in \mathbb{R}. \tag{11}$$

Also, set

$$\gamma_0(x) = \exp\left\{ \int_{-\infty}^{x} \frac{\tilde{H}^0(dz)}{1 - H(z)} \right\},$$
$$\gamma_1(x) = \frac{1}{1 - H(x)} \int 1\{x < w\}\varphi(w)\gamma_0(w)\tilde{H}^1(dw),$$
$$\gamma_2(x) = \int \frac{\int 1\{v < x, v < w\}\varphi(w)\gamma_0(w)}{(1 - H(v))^2} \tilde{H}^0(dv)\tilde{H}^1(dw). \tag{12}$$

As $L \longrightarrow \infty$ and $N = nL$, we have

$$\sqrt{N} \int \varphi(x)\left(\widehat{F}_{PROS}(dx) - F(dx)\right) \sim N\left(0, \sigma_n^2\right), \tag{13}$$

where

$$\sigma_n^2 = \frac{1}{n}\sum_{j=1}^{n} \mathrm{var}\left[\varphi_{\left(Y_{[dj]}\right)}\, \gamma_{0dj}\left(Y_{[dj]}\right)\delta_{[dj]} + \gamma_{1dj}\left(Y_{[dj]}\right)\left(1 - \delta_{[dj]}\right) - \gamma_{2dj}\left(Y_{[dj]}\right)\right]. \tag{14}$$

Proof. In view of the equivalent theorem in SRS sampling design [21], it suffices to show that, for every $j = 1, \cdots, n$,

$$\int \varphi^2(x){\gamma^2}_{0dj}(x){\tilde{H}^1}_{[dj]}(dx) < \infty, \tag{15}$$

$$\int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{\left(1 - H_{[d_j]}(z)\right)(1 - G(z))} \right)^{1/2} F_{[d_j]}(dx) < \infty. \tag{16}$$

As to equation (15), under continuity of F and G and $\gamma_{0dj}(X) = (1 - G(x))^{-1}$, we also have

$$\tilde{H}^1_{[d_j]}(dx) = \frac{d\left(\tilde{H}^1_{[d_j]}(x)\right)}{dx} = \frac{d\left(\int_{-\infty}^{x} (1 - G(y))F_{[d_j]}(dy)\right)}{dx}$$
$$= (1 - G(x))F_{[d_j]}(dx). \tag{17}$$

Under the continuity of F, there exists a density f. We have $F(dx) = (1/n)\sum_{j=1}^{n} F_{[d_j]}(dx)$; hence, $nF(dx) = \sum_{j=1}^{n} F_{[d_j]}(dx)$.

By using the above relationship,

$$\begin{aligned} &\int \varphi^2(x){\gamma^2}_{0d_j}(x)\tilde{H}^1_{[d_j]}(dx) \\ &= \int \frac{\varphi^2(x)}{(1 - G(x))^2}(1 - G(x))F_{[d_j]}(dx) \\ &= \int \frac{\varphi^2(x)}{(1 - G(x))} F_{[d_j]}(dx) \\ &\le \sum_{j=1}^{n} \int \frac{\varphi^2(x)}{(1 - G(x))} F_{[d_j]}(dx) \\ &= n\int \frac{\varphi^2(x)}{(1 - G(x))} F(dx) < \infty. \end{aligned} \tag{18}$$

By equation (9), this phrase is finite, so we prove equation (15).

To prove that (16) holds, we have to determine a lower bound for $1 - F_{[d_j]}(z)$.

We know that

$$\sum_{i=s-u+1}^{s} \binom{s}{s-i} (1 - F(z))^{i-(s-u+1)}(F(z))^{s-i}$$
$$= \sum_{i=0}^{u-1} \binom{s}{u-1-i} (1 - F(z))^{i}(F(z))^{u-1-i}, \tag{19}$$

for $i \le u-1$, we have

$$\begin{aligned}
&\begin{pmatrix} s \\ u-1-i \end{pmatrix} \ge \begin{pmatrix} u-1 \\ i \end{pmatrix}, \\
&1 - F_{(u:s)}(z) = \sum_{i=s-u+1}^{s} \begin{pmatrix} s \\ i \end{pmatrix} (1-F(z))^{i} (F(z))^{s-i} \\
&\quad = (1-F(z))^{s-u+1} \sum_{i=s-u+1}^{s} \begin{pmatrix} s \\ s-i \end{pmatrix} \\
&\qquad \cdot (1-F(z))^{i-(s-u+1)} (F(z))^{s-i} \\
&\quad = (1-F(z))^{s-u+1} \sum_{i=0}^{u-1} \begin{pmatrix} s \\ u-1-i \end{pmatrix} \\
&\qquad \cdot (1-F(z))^{i} (F(z))^{u-1-i} \\
&\quad \ge (1-F(z))^{s-u+1} \sum_{i=0}^{u-1} \begin{pmatrix} u-1 \\ i \end{pmatrix} \\
&\qquad \cdot (1-F(z))^{i} (F(z))^{u-1-i} \\
&\quad = (1-F(z))^{s-u+1}, \Rightarrow 1 - F_{(u:s)}(z) \\
&\quad \ge (1-F(z))^{s-u+1}.
\end{aligned} \tag{20}$$

Therefore, we have

$$\begin{aligned}
1 - F_{[d_j]}(z) &= \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \left(1 - F_{(u:s)}(z)\right) \\
&\ge \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(z))^{s-u+1}.
\end{aligned} \tag{21}$$

We know

$$\begin{aligned}
f_{[d_j]}(x) &= \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} f_{(u:s)}(x) \\
&= \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \frac{s!}{(u-1)!(s-u)!} f(x) F(x)^{u-1} (1-F(x))^{s-u} \\
&\le \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \frac{s!}{(u-1)!(s-u)!} f(x) (1-F(x))^{s-u}.
\end{aligned} \tag{22}$$

Also,

$$z \le x \Rightarrow (1-F(z))^{s-u} \ge (1-F(x))^{s-u}, \tag{23}$$

so,

$$\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(z))^{s-u} \ge \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}. \tag{24}$$

Then,

$$\begin{aligned}
&\left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(z))^{s-u}\right)^{-(1/2)} \\
&\quad \le \left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}\right)^{-(1/2)}.
\end{aligned} \tag{25}$$

Based on Lemma 1 and the above equations

$$\begin{aligned}
&\int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{\left(1 - H_{[d_j]}(z)\right)(1-G(z))}\right)^{1/2} F_{[d_j]}(dx) = \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{\left(1 - F_{[d_j]}(z)\right)(1-G(z))^2}\right)^{1/2} f_{[d_j]}(x) dx \\
&\quad \le \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{\left((1/m)\sum_{h=1}^{n}\sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(z))^{s-u+1}\right)(1-G(z))^2}\right)^{1/2} \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \frac{s!}{(u-1)!(s-u)!} f(x) (1-F(x))^{s-u} dx \\
&\quad = \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{(1-F(z))(1-G(z))^2} \times \frac{1}{(1/m)\sum_{h=1}^{n}\sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(z))^{s-u}}\right)^{1/2} \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \frac{s!}{(u-1)!(s-u)!} (1-F(x))^{s-u} F(dx) \\
&\quad \le \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{(1-H(z))(1-G(z))}\right)^{1/2} \left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}\right)^{-(1/2)} \left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \frac{s!}{(u-1)!(s-u)!} (1-F(x))^{s-u}\right) F(dx) \\
&\quad \le \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{(1-H(z))(1-G(z))}\right)^{1/2} \left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}\right)^{-(1/2)} \left(\frac{C}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}\right) F(dx) \\
&\quad = C \int |\varphi(x)| \left(\int_{-\infty}^{x} \frac{G(dz)}{(1-H(z))(1-G(z))}\right)^{1/2} \left(\frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} (1-F(x))^{s-u}\right)^{1/2} F(dx).
\end{aligned} \tag{26}$$

We define constant C as

$$C = \max_{u=1,\cdots,s}\left\{\frac{s!}{(u-1)!(s-u)!}\right\}. \tag{27}$$

Because $s-u \geq 0$, we have

$$\frac{1}{m}\sum_{h=1}^{n}\sum_{u\in d_h}\alpha_{d_j,d_h}(1-F(x))^{s-u} \leq \frac{1}{m}\sum_{h=1}^{n}\sum_{u\in d_h}\alpha_{d_j,d_h} = 1, \tag{28}$$

so (26) is smaller than

$$C\int|\varphi(x)|\left(\int_{-\infty}^{x}\frac{G(dz)}{(1-H(z))(1-G(z))}\right)^{1/2}F(dx). \tag{29}$$

In view of equation (10), this equation is finite, and this completes the proof. □

It should be noted that Theorem 1 has been proven only for the imperfect model, which has already been described in Section 2.2, and this model is not completely general.

5. Simulation Study

In this section, we compare the performance of the KM estimator of survival function under the PROS sampling design relative to its SRS and RSS counterparts.

To do so, we considered two situations in which the original random variables were generated from an exponential distribution with mean 1 (model A) and standard log-normal distribution with mean 1.649 (model B). The censored variables in the two cases are supposed to have an exponential distribution; a common rate of exponential distribution was determined when the desired censoring level was prespecified. In all simulation scenarios $s = nm$ and the set size for the RSS sampling design is n. The algorithm of the simulation study is explained in Appendix B.

By using distribution theory, if D and E are independent and distributed exponentially with means θ_1 and θ_2, respectively, then $P(D \leq E) = \theta_2/(\theta_1 + \theta_2)$. On the other hand, $P(D \leq E) = 1 - p$. Setting the values of the censoring level (p) and $\theta_1 = 1$ in these equations, we can find the appropriate value of the exponential rate in model A. Given the fact that there is no such expression for model B, we found the exponential common rate for the censoring variable by trial and error, although one can easily solve this problem numerically by using software like R. The values of the exponential rate were equal to 0.013 and 0.190 and led to censoring levels of 0.1 and 0.6, respectively.

For each combination of sample sizes $N = 30$, 120, and 240 and the mentioned censoring levels 0.1 and 0.6, 5000 samples were generated under the SRS, RSS, and PROS sampling designs. For different values of n, m, and L and the misplacement probabilities $\alpha_{d_i,d_i} = \alpha_0$ and $\alpha_{d_i,d_j} = (1-\alpha_0)/(n-1)$ for $i \neq j$, the values of the mean squared error (MSE) were computed for the three estimators from each sample when $\alpha_0 = 0$, 0.5, 0.7, and 1.

5.1. Comparing the Kaplan-Meier Estimators.

5.1. Comparing the Kaplan-Meier Estimators. We compare the performance of the KM estimators of the survival function between the studied sampling designs. The efficiency of the PROS estimation with respect to its SRS and RSS counterparts, at the point t, is defined as

$$\mathrm{RP} = \frac{\mathrm{MSE}\left(1-\widehat{F}_{\mathrm{RSS}}(t)\right)}{\mathrm{MSE}\left(1-\widehat{F}_{\mathrm{PROS}}(t)\right)}, \mathrm{SP} = \frac{\mathrm{MSE}\left(1-\widehat{F}_{\mathrm{SRS}}(t)\right)}{\mathrm{MSE}\left(1-\widehat{F}_{\mathrm{PROS}}(t)\right)}, \tag{30}$$

where $1-\widehat{F}_{\mathrm{PROS}}(t)$, $1-\widehat{F}_{\mathrm{RSS}}(t)$, and $1-\widehat{F}_{\mathrm{SRS}}(t)$ are the KM estimators of the survival function at point t based on PROS, RSS, and SRS sampling designs, respectively.

Note that $\mathrm{MSE}(1-\widehat{F}_{\mathrm{PROS}}(t)) = E[F(t)-\widehat{F}_{\mathrm{PROS}}(t)]^2$. MSE $(1-\widehat{F}_{\mathrm{RSS}}(t))$ and $\mathrm{MSE}(1-\widehat{F}_{\mathrm{SRS}}(t))$ are similarly defined. Also, $t = F^{-1}(q)$ for a fixed percentile $q \in (0,1)$, and $F^{-1}(.)$ is the inverse of the underlying distribution function. The values of RP and SP calculate for $m = 3$ and $n = 3$ and 5 in both models when we consider $q = 0.10$, 0.25, 0.50, 0.75, and 0.90. Because of the large volume of output and similar results in both models, we only report the results for model A in this article.

In the literature, the sample sizes in the PROS and RSS designs were similar but they have used a much smaller set size for RSS sampling design than for PROS. However, simulation studies that are not presented here show that the RSS-based estimator may performs better than the one using the PROS sampling design under the same sample size and the same set size.

As shown in Figures 1 and 2, in model A, the KM estimator based on the PROS sampling design in most cases is more efficient than the KM estimator based on the RSS and SRS sampling designs with similar sample sizes. The best performance of the PROS design over the SRS and RSS designs happens when the ranking errors are small or zero, i.e., when $\alpha_0 = 0.7$ and 1. The efficiency of the KM estimator based on PROS relative to SRS is as good as or higher than the efficiency of the KM estimator based on the PROS relative to the RSS procedure, regardless of the censoring level and ranking error. Assuming a fixed sample size and censoring level, by increasing the n for large values of α_0, the efficiency of the KM estimator based on the PROS sampling design is enhanced. It should be noted that in an imperfect PROS sampling design $(\alpha_0 = 0)$, the efficiency reduced as n increased.

We can conclude that increasing the level of censorship in a smaller sample size leads to a reduction in efficiency in both models, but for a larger sample size, this rarely happens; in other words, the level of censored data in the smaller sample size has a greater impact on the performance of the PROS sampling design compared to the that in the larger sample size.

We conclude that, regardless of the censoring level and ranking error, increasing the sample size leads to increased efficiency. The perfect PROS KM estimator performs three times more efficiently than the SRS KM estimator in several simulation scenarios. It is worth noting that RP might decrease when one considers the same set size in the PROS

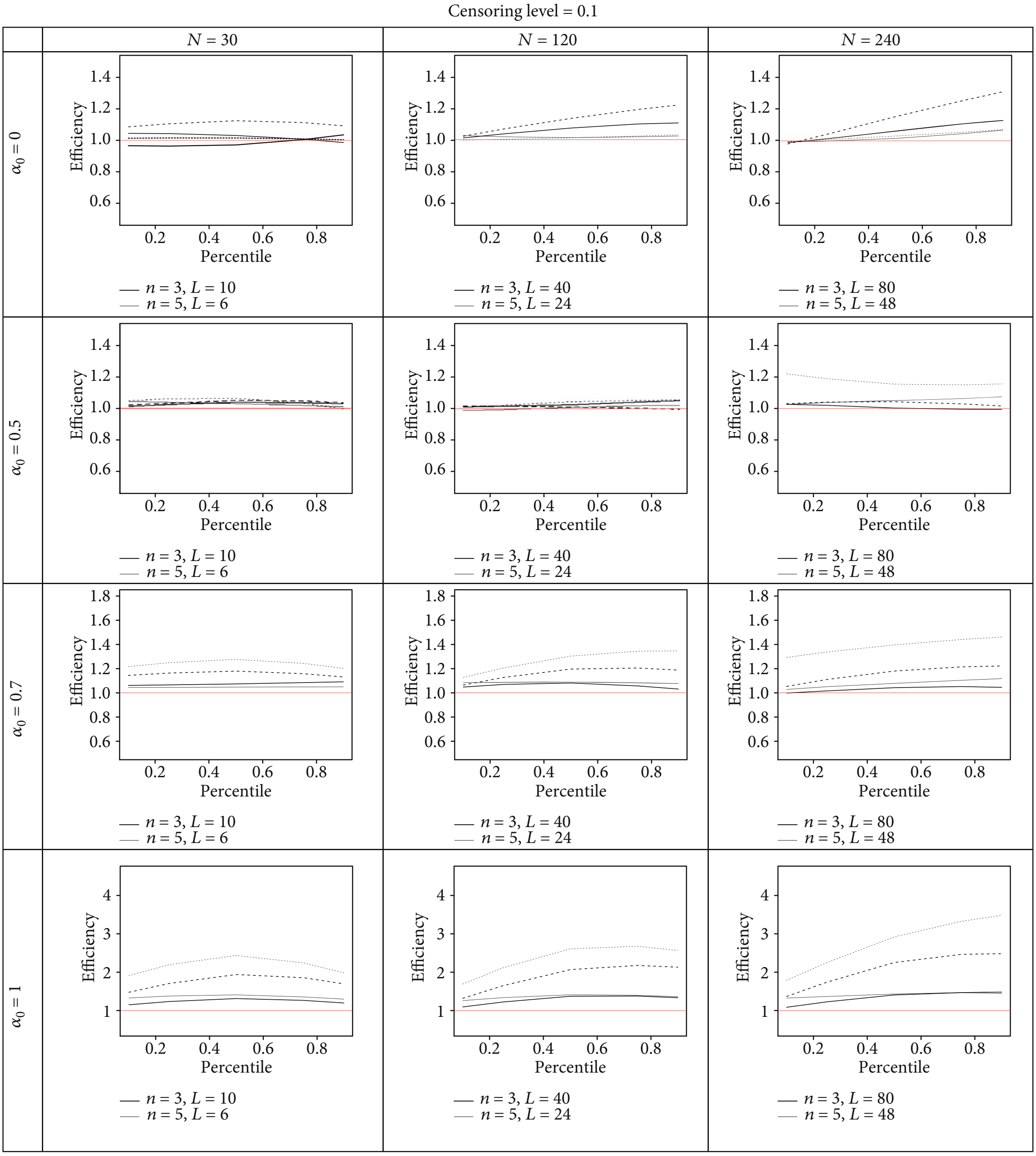

Figure 1: The efficiency of the KM estimator based on PROS with respect to RSS (solid line) and SRS (dashed line) counterparts at different percentiles.

and RSS designs with similar sample sizes. In all figures, we consider $m = 3$ for the PROS design.

In addition, we compared these three sampling methods using a mean integrated squared error (MISE) indicator, defined as

$$\text{MISE} = \int_{-\infty}^{+\infty} E\left\{\widehat{F}_n(t) - F(t)\right\}^2 dF(t). \qquad (31)$$

From Table 2, we can conclude that most of the time, PROS has less MISE than the RSS and SRS sampling methods with similar sample sizes, especially for a large α_0. In addition, we observe that as the level of censored data increases, the amount of the MISE value increases as well in both models. It should be mentioned that in the low level of censorship, the log-normally distributed (model B) has lower MISE than the exponentially distributed (model A), but at the high level of censorship, model B has larger MISE than

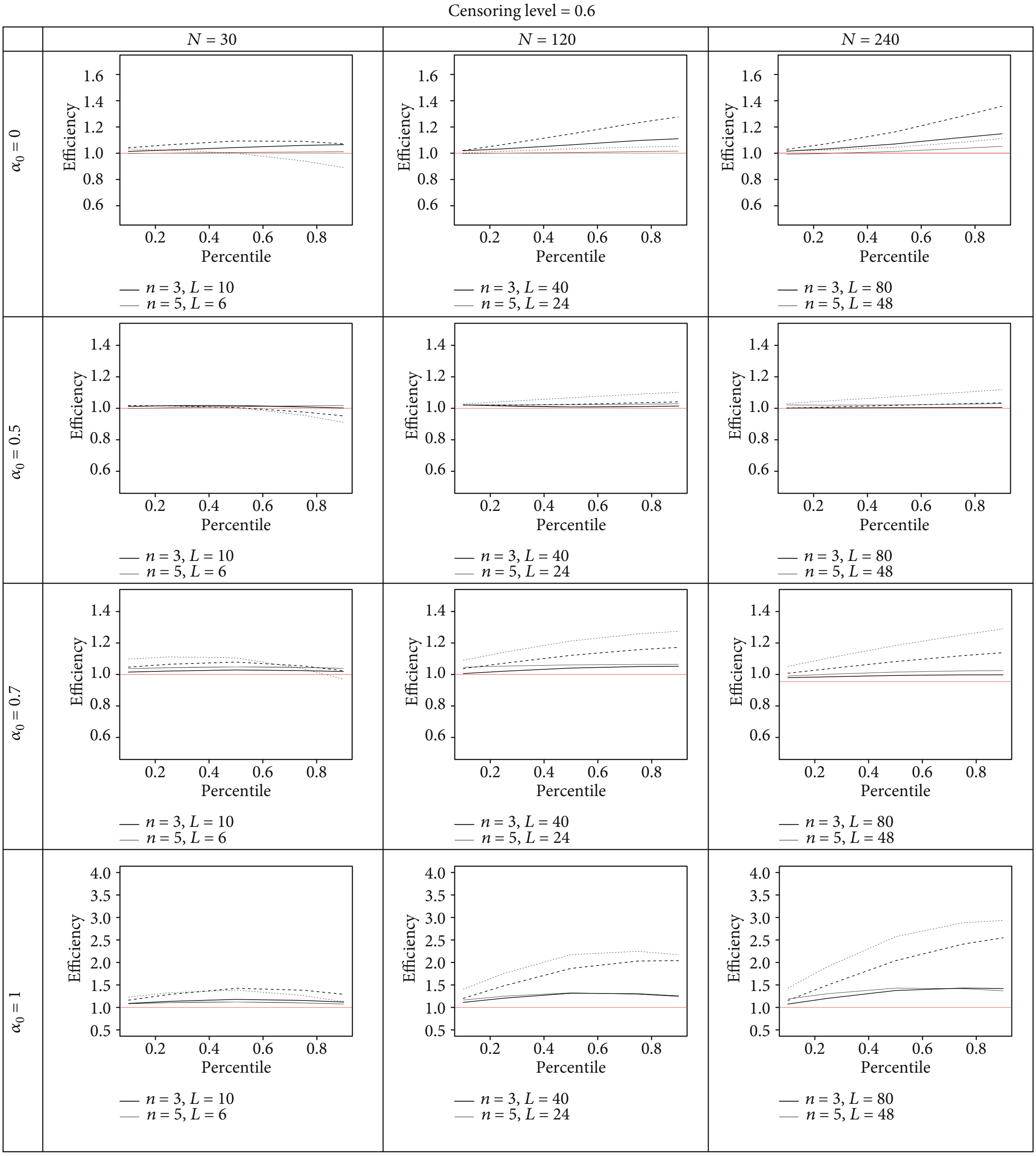

Figure 2: The efficiency of KM estimator based on PROS with respect to RSS (solid line) and SRS (dashed line) counterparts at different percentiles.

model A, for the same values of n and m and the subsetting probabilities α_{d_i,d_j}. As we expect, increasing the sample size reduces the MISE.

The results show that when $\alpha_0 = 0.5$, 0.7, and 1 in a smaller sample size with a low percentage of censored data, the larger n leads to the smaller MISE of the estimators, but with a high percentage of censored data, the MISE value increases as n increases. However, in larger sample sizes, the MISE of the estimator decreases as the n goes up in all censoring levels.

In Table 2, as the misplacement probabilities decrease, the superiority of the PROS estimator compared to the RSS and SRS estimators becomes more obvious. The MISE values of the KM estimator derived from perfect PROS and perfect RSS sampling designs are smaller than those in imperfect methods. Note that the KM estimator based on the SRS

Table 2: Estimated MISE of Kaplan Meier estimator, $N = 30$, 120, and 240.

		Censoring	$\alpha_0 = 0$			$\alpha_0 = 0.5$			$\alpha_0 = 0.7$			$\alpha_0 = 1$		
		level	PROS	RSS	SRS	PROS	RSS	SRS	PROS	RSS	SRS	PROS	RSS	SRS
Model A	$N = 30$ ($n = 3, L = 10$)	0.1	0.0062	0.0065	0.0071	0.0066	0.0069	0.0069	0.0059	0.0064	0.0071	0.0037	0.0048	0.0070
		0.6	0.0620	0.0653	0.0664	0.0672	0.0679	0.0653	0.0628	0.0644	0.0655	0.0492	0.0563	0.0661
	$N = 30$ ($n = 5, L = 6$)	0.1	0.0069	0.0068	0.0071	0.0065	0.0068	0.0069	0.0055	0.0058	0.0071	0.0030	0.0042	0.0070
		0.6	0.0713	0.0717	0.0664	0.0687	0.0699	0.0653	0.0642	0.0668	0.0655	0.0539	0.0590	0.0661
	$N = 120$ ($n = 3, L = 40$)	0.1	0.0021	0.0022	0.0024	0.0024	0.0025	0.0024	0.0021	0.0022	0.0025	0.0012	0.0016	0.0024
		0.6	0.0493	0.0531	0.0581	0.0569	0.0574	0.0584	0.0511	0.0534	0.0578	0.0305	0.0395	0.0581
	$N = 120$ ($n = 5, L = 24$)	0.1	0.0023	0.0023	0.0024	0.0022	0.0023	0.0024	0.0018	0.0020	0.0025	0.0009	0.0013	0.0024
		0.6	0.0565	0.0569	0.0581	0.0546	0.0559	0.0584	0.0474	0.0502	0.0578	0.0274	0.0352	0.0581
	$N = 240$ ($n = 3, L = 80$)	0.1	0.0014	0.0015	0.0017	0.0017	0.0016	0.0017	0.0014	0.0015	0.0017	0.0007	0.0011	0.0017
		0.6	0.0469	0.0514	0.0569	0.0555	0.0557	0.0567	0.0491	0.0515	0.0571	0.0260	0.0363	0.0569
	$N = 240$ ($n = 5, L = 48$)	0.1	0.0016	0.0017	0.0017	0.0015	0.0016	0.0017	0.0012	0.0013	0.0017	0.0006	0.0008	0.0017
		0.6	0.0539	0.0553	0.0569	0.0524	0.0536	0.0567	0.0445	0.0477	0.0571	0.0215	0.0301	0.0569

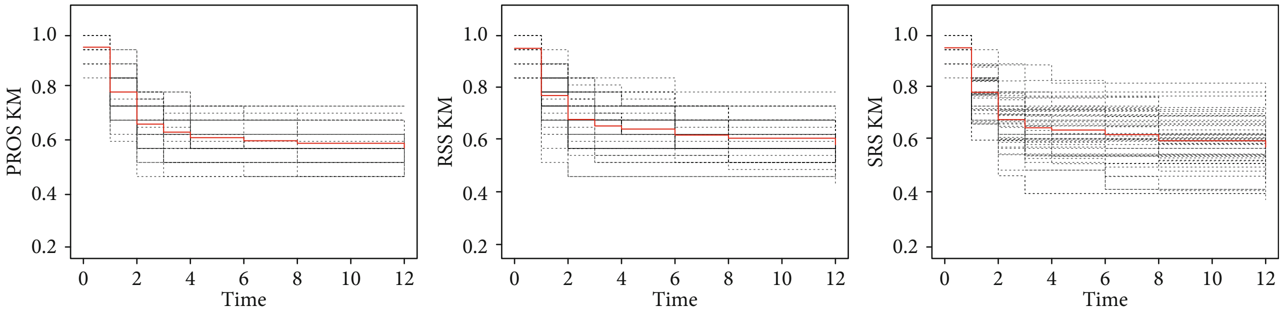

Figure 3: The KM estimates in different time for $n = 5$, $L = 3$, and $m = 3$ under 50 PROS, RSS, and SRS samples. The red solid line shows the mean of KM estimators.

sampling design has a smaller MISE value than the one based on the imperfect rank-based sampling designs for some cases in small sample sizes and high censorship percentage.

Note that the RSS KM estimator can have a lower MISE than the PROS one, when we consider a similar set size and fixed sample size.

6. Real Data Application

In this section, we use the information of children under 18 years of age with nonhematological disorders such as Beta-Thalassemia and Idiopathic Thrombocytopenic Purpura (ITP) and children with hematological malignancies including various types of lymphoma and Acute Lymphocytic Leukemia (ALL), registered in the Amir Medical Oncology Center during May 2014 to August 2017. The dataset contains the survival information of 61 patients. We provide KM estimates of Y which is the survival time (in months) as the variable of interest by using Z which is the white blood cells as the concomitant variable, which are used for ranking purpose. The correlation coefficient between Z and Y is 0.455 and is significant (p value = 0.0001); also, we should add that 50.8% of people are censored. We considered the perfect PROS and RSS sampling designs. In order to estimate the KM estimator of survival time, we regarded this data set as a target population and extract PROS, RSS, and SRS samples (with replacement) of size $N = nL$ from the population. We considered design parameter $D = \{d_1, d_2, d_3, d_4, d_5\}$. At the first step, we randomly selected $nm = 15$ patients from the target population and then partitioned these patients into subsets d_1, d_2, d_3, d_4, and d_5 based on their WBC values. At the next step, we randomly selected a unit from subset d_1 and observed its survival time. Again, we randomly selected 15 patients and assigned them to d_1, d_2, d_3, d_4, and d_5 and randomly drew a member from subset d_2 and repeated these steps until we selected a unit from subset d_5; these observations constitute one cycle of PROS; in this real data, we considered 3 cycles, and finally, we have 15 survival time observations from patients.

In RSS, we randomly selected 5 patients from the target population and ranked them based on their WBC values, then we selected the patient with the smallest WBC and observed its survival time. This procedure continued until the survival time of the 5th ranked unit in the 5th set of units measured. These 5 observations constitute one cycle of RSS; in this example, we considered 3 cycles, and finally, we observed the survival time of 15 patients.

For each sampling design, the KM estimator was calculated in different time points. Then, this process was repeated M times. We took $(n, m, L, M) = (5, 3, 3, 50)$. These 50 KM charts under the three sampling designs are shown in Figure 3. Figure 3 shows that the variation of the KM estimators in each fixed time under the PROS sampling design is less than the variation of the RSS and SRS counterparts. We conclude that in this real data, the PROS estimate performs better than the RSS and SRS designs. We uploaded the raw data as a supplementary material (available here).

7. Summary and Concluding Remarks

In numerous medical fields, the exact measurement of the desired variable is expensive or time-consuming. Rank-based sampling designs such as PROS can help overcome this difficulty by ranking a small number of sampling units based on a concomitant variable. These sampling designs can be used to obtain samples that are more informative and also result in more accurate inference about the parameters of interest.

In this paper, we considered the problem of the KM estimator that is a proper and commonly used technique in survival analysis associated with an imperfect PROS sampling design. PROS is a new sampling design that avoids ranking all units in a given set. Furthermore, we developed asymptotic distributional properties of the new KM estimator based on a proposed sampling method. We showed how well this estimator performs in comparison with its RSS and SRS counterparts. The simulation results recommend that under both perfect and imperfect subsetting assumptions, the efficiency of the estimator based on the PROS sampling design is higher than the efficiency of the estimator based on the two other sampling methods with the same sample sizes. It is noteworthy that, by increasing the set size in RSS while keeping the sample size fixed in both designs, the RSS KM estimator can have smaller values of MSE than the PROS one. Finally, we applied all the introduced sampling designs to a real data set. We believe that it would be appealing to apply the proposed methodology to useful statistical models, for example, a Cox regression model for analyzing time to event data that is applicable to the majority of medical fields.

Finally, we will recommend the use of recently proposed sampling designs to extend this study, for example, even order ranked set sampling (EORSS) [22] and quartile pair ranked set sampling (QPRSS) [23] designs that have recently received attention by some researchers.

Appendix

A. The Proof of Lemma 1

We have

$$\begin{aligned}
1 - H_{[d_j]}(t) &= 1 - P\left(\min\left(X_{[d_j]i}, C_{[d_j]i}\right) \le t\right) \\
&= P\left(\min\left(X_{[d_j]i}, C_{[d_j]i}\right) > t\right) \\
&= P\left(X_{[d_j]i} > t, C_{[d_j]i} > t\right) \\
&= P\left(X_{[d_j]i} > t\right) P\left(C_{[d_j]i} > t\right) \\
&= \left[1 - F_{[d_j]}(t)\right][1 - G(t)].
\end{aligned} \tag{A.1}$$

B. Algorithm of Simulation Scenarios

The steps of simulation study algorithm are as follows:

Step 1: Perform data generation in the following ways:

(i) *Generate 1000 random event time observations from the desired distribution (X)*

(ii) *Generate 1000 random censored time observations from the desired distribution (C)*

(iii) *Observe the status variables* $(\delta = I(X < C))$

(iv) *Calculate survival time variable* $(T = \min\ (X, C))$

Step 2: Perform sampling in the different studied designs:

(i) *Generate PROS, RSS, and SRS samples from the target population. For PROS and RSS, we generate the samples based on different values for subsetting error matrices, set sizes, and cycle sizes*

Step 3: Estimate the desired estimators:

(i) *estimate the KM estimator using the corresponding formula coding*

Step 4: calculate comparison criteria:

(i) *Compute the MSE of the KM estimator in different percentile points*

(ii) *Compute the MISE values for KM estimators under the three different sampling designs*

Step 5: Repeat all the above steps 5000 times.

Step 6: Compute the mean of 5000 calculated MSE and MISE and report them.

Data Availability

In the present study, we used the information about children under 18 years of age with non-hematological disorders such as Beta-Thalassemia and Idiopathic Thrombocytopenic Purpura (ITP) and also children with hematological malignancies including various types of lymphoma and Acute Lymphocytic Leukemia (ALL), registered in Amir Medical Oncology Center during May 2014 to August 2017, as a population of interest.

Conflicts of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Acknowledgments

The present paper was extracted from the Ph.D. dissertation of Mrs. Samane Nematolahi and was supported by Shiraz University of Medical Sciences, Shiraz, Iran (Grant no. 97-01-01-16858).

Supplementary Materials

This supplementary file includes the data for the Section 6 (real data) example in the paper. This file contains the information of children under 18 years of age with nonhematological disorders such as Beta-Thalassemia and Idiopathic Thrombocytopenic Purpura (ITP) and children with hematological malignancies including various types of lymphoma and Acute Lymphocytic Leukemia (ALL), registered in Amir Medical Oncology Center during May 2014 to August 2017. The dataset contains the survival information of 61 patients. *(Supplementary Materials)*

References

[1] G. McIntyre, "A method for unbiased selective sampling, using ranked sets," *Australian Journal of Agricultural Research*, vol. 3, no. 4, pp. 385–390, 1952.

[2] J. Frey and T. G. Feeman, "Efficiency comparisons for partially rank-ordered set sampling," *Statistical Papers*, vol. 58, no. 4, pp. 1149–1163, 2017.

[3] A. Hatefi, M. J. Jozani, and O. Ozturk, "Mixture model analysis of partially rank-ordered set samples: age groups of fish from length-frequency data," *Scandinavian Journal of Statistics*, vol. 42, no. 3, pp. 848–871, 2015.

[4] O. Ozturk, "Sampling from partially rank-ordered sets," *Environmental and Ecological Statistics*, vol. 18, no. 4, pp. 757–779, 2011.

[5] O. Ozturk, "Combining multi-observer information in partially rank-ordered judgment post-stratified and ranked set samples," *Canadian Journal of Statistics*, vol. 41, no. 2, pp. 304–324, 2013.

[6] J. Frey, "Nonparametric mean estimation using partially ordered sets," *Environmental and Ecological Statistics*, vol. 19, no. 3, pp. 309–326, 2012.

[7] S. Nazari, M. Jafari Jozani, and M. Kharrati-Kopaei, "Nonparametric density estimation using partially rank-ordered set samples with application in estimating the distribution of wheat yield," *Electronic Journal of Statistics*, vol. 8, no. 1, pp. 738–761, 2014.

[8] O. Ozturk, "Estimation of a finite population mean and total using population ranks of sample units," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 21, no. 1, pp. 181–202, 2016.

[9] S. Nazari, M. Jafari Jozani, and M. Kharrati-Kopaei, "On distribution function estimation with partially rank-ordered set samples: estimating mercury level in fish using length frequency data," *Statistics*, vol. 50, no. 6, pp. 1387–1410, 2016.

[10] A. Hatefi and M. J. Jozani, "Information content of partially rank-ordered set samples," *AStA Advances in Statistical Analysis*, vol. 101, no. 2, pp. 117–149, 2017.

[11] Q. Zaman and K. P. Pfeiffer, "Survival analysis in medical research," *Interstat*, vol. 17, no. 4, pp. 1–36, 2011.

[12] J. W. Song and K. C. Chung, "Observational studies: cohort and case-control studies," *Plastic and Reconstructive Surgery*, vol. 126, no. 6, pp. 2234–2242, 2010.

[13] E. Strzalkowska-Kominiak and M. Mahdizadeh, "On the Kaplan–Meier estimator based on ranked set samples," *Journal of Statistical Computation and Simulation*, vol. 84, no. 12, pp. 2577–2591, 2013.

[14] P. L. H. Yu and C. Y. C. Tam, "Ranked set sampling in the presence of censored data," *Environmetrics*, vol. 13, no. 4, pp. 379–396, 2002.

[15] L. Zhang, X. Dong, and X. Xu, "Nonparametric estimation for random censored data based on ranking set sampling," *Communications in Statistics - Simulation and Computation*, vol. 43, no. 8, pp. 2004–2015, 2014.

[16] M. Mahdizadeh and E. Strzalkowska-Kominiak, "Resampling based inference for a distribution function using censored ranked set samples," *Computational Statistics*, vol. 32, no. 4, pp. 1285–1308, 2017.

[17] D. A. Wolfe, "Ranked set sampling: an approach to more efficient data collection," *Statistical Science*, vol. 19, no. 4, pp. 636–643, 2004.

[18] W. Stute and J. L. Wang, "The strong law under random censorship," *The Annals of Statistics*, vol. 21, no. 3, pp. 1591–1607, 1993.

[19] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[20] P. Major and L. Rejto, "Strong embedding of the estimator of the distribution function under random censorship," *The Annals of Statistics*, vol. 16, no. 3, pp. 1113–1132, 1988.

[21] W. Stute, "The Central Limit Theorem under random censorship," *The Annals of Statistics*, vol. 23, no. 2, pp. 422–439, 1995.

[22] M. Noor-ul-Amin, M. Tayyab, and M. Hanif, "Mean estimation using even order ranked set sampling," *Punjab University Journal of Mathematics*, vol. 51, no. 1, pp. 91–99, 2019.

[23] M. Tayyab, M. Noor-ul-Amin, and M. Hanif, "Quartile pair ranked set sampling: development and estimation," *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, vol. 18, pp. 1–6, 2019.