doi:10.21608/esju.2024.281318.1031 Research Article

The Egyptian Statistical Journal (ESJ), 68 (1): 59-77

Homepage: https://esju.journals.ekb.eg/

The Egyptian Statistical Journal

Print ISSN 0542-1748– Online ISSN 2786-0086

Valid estimation of the distribution function using dual-rank ranked set sampling: Missing data approach

Mohamed S. Abdallah*

Received 03 April 2024; revised 12 May 2024; accepted 20 May 2024

Keywords

Cumulative distribution function; Dual ranked set sampling; EM algorithm; Linear interpolation; Concomitant variable.

Abstract

In this study, two nonparametric estimators of estimating distribution function under dual-rank ranked set sampling (DRSS) are discussed and compared. The first estimator incorporates the information supported by the relationship between ranks of measured sampling items and information from unmeasured sampling items. This estimator is constructed through an iterative algorithm. The second one depends on using the information generated by a concomitant variable. This estimator is derived by the linear interpolation method. Under both perfect and imperfect ranking situations, a series of simulation studies are carried out, and the proposed estimators are then compared with their counterparts using traditional ranked set sampling (RSS). It is concluded that the proposed procedures have dramatically better performance than their competitors at points in the center of the parent distribution, specifically in the case of an appropriate level of ranking quality. Finally, an illustrative example using a real dataset is also used to investigate the performance of the proposed procedures.

1. Introduction

Ranked set sampling (RSS) was initially introduced by McIntyre (1952) to increase the precision of estimating the average pasture and forage yields. This technique is operated by randomly selecting sets of k items each from the parent population. Importantly, it is assumed that the actual quantification of the sample items is not easy, costly, or time-consuming relative to their ranking. The ranking process can be performed by either eye inspection or an auxiliary variable known as a concomitant variable, X, which is affordable, easily obtainable, and has a reasonable correlation with the variable under consideration. One can draw a RSS sample of size n by getting k sets, each with k sampling items, ranking individually the k sets, and selecting the lowest sampling item ranked from the first set for actual quantification. However, the second lowest of the sampling items was quantified based on the second set. Following the same manner, the selection process is performed until the highest-ranked sampling item is exactly considered from the k^{th} set of k sampling items. This enables us to get k-independent measured values. To obtain a larger sample, one could repeat this process m independent times (cycles). Therefore, one can get $n = km$

✉ Corresponding author*: mohamed_abdallah@com.aswu.edu.eg
Department of Quantitative Techniques, Faculty of Commerce, Aswan University. Egypt.

independent measured values. It should be stressed that if the number of measured items varies across the k sets, RSS is termed unbalanced RSS. Our focus here is on balanced RSS.

Let $Y_{i(1)j}, Y_{i(2)j}, \dots, Y_{i(k)j}$ be the order statistics of the i^{th} sample $(i = 1,2,\dots,k)$ in the j^{th} cycle $(j = 1,2,\dots,m)$ drawn from a population of interest with a probability density function (pdf) $f(y)$ and cumulative distribution function (CDF) $F(y)$. Then, $\{Y_{i(i)j}:\ i = 1,2,\dots,k\ ;\ j = 1,2,\dots,m\}$ is a RSS of size n. It is important to highlight that as long as the sampling items are ranked with free errors, then this situation is termed as a perfect ranking situation, and the brackets take a round symbol in $Y_{i(i)j}$'s. In contradiction, if it is expected to get error ranks during the ranking mechanism, this case is called imperfect ranking, and the brackets will become a square symbol, $Y_{i[i]j}$'s.

It is well known that the pdf and the CDF corresponding to $Y_{(i:k)}$ are respectively, given by (see David and Nagaraja (2003)):

$$f_{(i:k)}(y) = \frac{k!}{(i-1)!\,(k-i)!}[F(y)]^{i-1}[1-F(y)]^{k-i}f(y) = b_{i,k-i+1}\big(F(y)\big),$$

and

$$F_{(i:k)}(y) = \sum_{j=i}^{k}\binom{k}{j}[F(y)]^{j-1}[1-F(y)]^{k-j} = B_{i,k-i+1}\big(F(y)\big),$$

where $b_{a,b}(y)$ and $B_{a,b}(y)$ are respectively the pdf and the CDF of the Beta distribution with parameters a and b at the point y.

Stokes and Sager (1988) first used the empirical distribution function to introduce the CDF estimator under RSS given by:

$$\hat{F}_R(t) = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{k} I\big(Y_{i(i)j} \le t\big),$$

where $t \in \mathbb{R}$. Additionally, they introduced a rigorous proof indicating that $\hat{F}_R(t)$ leads to more efficient population CDF estimates than its counterpart in simple random sample (SRS) design.

In the same sequel, other studies utilized the missing data approach for improving CDF estimation based on RSS. Generally speaking, this approach can be grouped into two categories. The first category constructs its procedures by combining the imputed information based on the unmeasured units with those supported by the measured items; typically, an iterative algorithm, such as the EM algorithm, is adopted for performing this mission. Examples of research in this category are Kvam and Samaniego (1994) and Frey and Zhang (2023). On the other hand, the second category of research prefers to raise the CDF estimation efficiency by incorporating the concomitant variable,X, information, using the linear interpolation method. Zamanzade and Mahdizadeh (2018) and Ashour and Abdallah (2020) are famous studies in this category. Motivated by these publications, we sought to extend the work addressed by Kvam and Samaniego (1994) and Zamanzade and Mahdizadeh (2018) to a new variation of RSS, recently proposed, known as dual-rank ranked set sampling (DRRSS). To the best of our knowledge, the problem of CDF estimation based on DRRSS using either the unmeasured item information or concomitant variable information has not yet been discussed in the literature.

Several studies have also addressed the efficiency of RSS extensions in estimating different population parameters. To name a few of these studies, Göçoğlu and Demirel (2020) examined the superiority of various RSS designs for estimating the population proportion. Hassan et al. (2021) considered the stress–strength model under median RSS. Abdallah (2023) investigated the efficiency of paired RSS in estimating the ROC curve. Al-Saleh and Ahmad (2023) recently adopted RSS and some other RSS schemes for estimating the common mean of two normal distributions. Zamanzade et al. (2024) addressed the mean residual lifetime under RSS. The rest of this work is structured as follows: Section 2 discusses the CDF estimators proposed by Kvam and Samaniego (1994) and Zamanzade and Mahdizadeh (2018) in detail. Section 3 explains the dual-rank RSS (DRRSS) mechanism and introduces the two proposed CDF estimators under DRRSS. The comparison studies of the two proposed procedures for their analogues under perfect ranking and imperfect ranking situations are provided in Section 4. Section 5 presents an illustrative case study of the two novel estimators. Lastly, Section 6 shows some overall remarks and potential guideline points for future studies.

2. The CDF estimation in RSS using auxiliary information

Kvam and Samaniego (1994) were the first to introduce a novel idea of estimating $F(t)$ based on RSS using a missing data approach. Their efforts are to use, assuming the perfectness situation, the ranking information generated by the measured sampling items for estimating the expected count of the unmeasured sampling items less than t.

For instance, at a certain t, if we observed that the sampling item $Y_{i(i)j}$ actually is less than t, then it is logical to think about that:

$$P\big(Y_{i(l)j} < t\big) = 1 \qquad\qquad l = 1,2 \dots i-1. \tag{1}$$

while for the remaining items:

$$P\big(Y_{i(l)j} < t \big| Y_{i(l)j} > Y_{i(i)j}\big) = \frac{\int_{y_{i(i)j}}^{t} f(y)\,dy}{\int_{y_{i(i)j}}^{\infty} f(y)\,dy} = \frac{F(t) - F\big(y_{i(i)j}\big)}{1 - F\big(y_{i(i)j}\big)} \qquad l = i+1, i+2 \dots k. \tag{2}$$

Oppositely, if we observed that the sampling item $Y_{i(i)j}$ actually is greater than t, then it is logical to think about that:

$$P\big(Y_{i(l)j} < t\big) = 0, \qquad\qquad l = i+1, i+2 \dots k. \tag{3}$$

yet for the remaining items:

$$P\left(Y_{i(l)j} < t \middle| Y_{i(l)j} < Y_{i(i)j}\right) = \frac{\int_{-\infty}^{t} f(y)\,dy}{\int_{-\infty}^{y_{i(i)j}} f(y)\,dy} = \frac{F(t)}{F\left(y_{i(i)j}\right)}\,, \qquad l = 1,2 \ldots i-1. \tag{4}$$

Putting $(1-4)$ together, the CDF estimator based on the ranking information generated by the measured sampling as well as the unmeasured sampling can be formulated as:

$$\hat{F}_{R1}^{*}(t) = \frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(i + (k-i)\frac{F(t) - F\left(y_{i(i)j}\right)}{1 - F\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} < t\right)\right] + \left[(i-1)\left(\frac{F(t)}{F\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} > t\right)\right].$$

One can easily observe that $\hat{F}_{R1}^{*}(t)$ is not itself an estimable function since it simply depends on unknown quantities, therefore Kvam and Samaniego (1994) decided to solve this dilemma by using EM algorithm, supported by Dempster et al. (1977), whose steps are listed below:

1- Set $r = 0$.

2- Estimate the unknown CDFs in $\hat{F}_{R1}^{*}(t)$ with the estimator based on the empirical distribution function proposed by Stokes and Sager (1988), i.e. compute the following equation:

$$\hat{F}_{R1}^{*(r)}(t) = \frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(i + (k-i)\frac{\hat{F}_R(t) - \hat{F}_R\left(y_{i(i)j}\right)}{1 - \hat{F}_R\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} < t\right)\right] + \left[(i-1)\left(\frac{\hat{F}_R(t)}{\hat{F}_R\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} > t\right)\right],$$

3- Replace t in $\hat{F}_{R1}^{*(r)}(t)$ with all the sampling items, i.e. compute the following equation:

$$\hat{F}_{R1}^{*(r)}\left(y_{l(l)w}\right) = \frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(i + (k-i)\frac{\hat{F}_R\left(y_{l(l)w}\right) - \hat{F}_R\left(y_{i(i)j}\right)}{1 - \hat{F}_R\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} < y_{l(l)w}\right)\right]$$
$$+ \left[(i-1)\left(\frac{\hat{F}_R\left(y_{l(l)w}\right)}{\hat{F}_R\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} > y_{l(l)w}\right)\right],$$

$l = 1,2,..k$ and $\mathrm{w} = 1,2,..m$.

4- Set $r = r + 1$.

5- Obtain $\hat{F}_{R1}^{*(r)}(t)$ by using the following recursive equation:

$$\hat{F}_{R1}^{*(r)}(t) = \frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(i + (k-i)\frac{\hat{F}_{R1}^{*(r-1)}(t) - \hat{F}_{R1}^{*(r-1)}\left(y_{i(i)j}\right)}{1 - \hat{F}_{R1}^{*(r-1)}\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} < t\right)\right]$$
$$+ \left[(i-1)\left(\frac{\hat{F}_{R1}^{*(r-1)}(t)}{\hat{F}_{R1}^{*(r-1)}\left(y_{i(i)j}\right)}\right) I\left(Y_{i(i)j} > t\right)\right],$$

6- Repeat steps (4-5) until the following stopping rule satisfy:

$$\left|\hat{F}_{R1}^{*(r)}(t) - \hat{F}_{R1}^{*(r-1)}(t)\right| < .001$$

7- Finally, the suggested estimator denoted by $\hat{F}_{R1}(t)$ given by:

$$\hat{F}_{R1}(t) = \hat{F}_{R1}^{*(r)}(t)$$

Kvam and Samaniego (1994) investigated analytically the statistical properties of $\hat{F}_{R1}(t)$ and proved that the suggested algorithm converges to a unique consistent solution as long as the initial CDF value used at step 2 in the above algorithm is a consistent estimator. Using a Monte Carlo simulation study, they concluded that performing the above algorithm will considerably improve the efficiency of CDF estimation under RSS.

On the other hand, Zamanzade and Mahdizadeh (2018) also preferred to adopt the missing data mechanism, however, by taking into account the information supported by the concomitant sampling items. Their idea is based on imputing the unmeasured items relevant to the interested variable in the light of the corresponding measured items associated with the concomitant variable. In order to estimate the CDF under RSS based on missing data mechanism, Zamanzade and Mahdizadeh (2018) decided to implement the linear interpolation technique whose steps are given by:

1- Combining $y_{i(i)j}$ and their associated sampling items of $x_{i(i)j}$ into two new variables $(y_z^*, x_z^*\ , z = 1 \dots n)$ respectively.

2- Sorting ascending (y_z^*, x_z^*) according to x^* items leading to $(y_{[z]}^*, x_{(z)}^*)$.

3- Compute the isotonized values, see Ozturk (2007), for $I\left(y_{[z]}^* \leq t\right)$ and retain these values in $\hat{F}_z^{iso}(t)$.

4- For each $x_{lj(i)}$, calculate the corresponding $\hat{F}_z^{iso}(t)$ by implementing the linear interpolation formula given by:

$$\hat{F}_x(t) = \begin{cases} \hat{F}_1^{iso}(t) & x \leq x_{(1)}^* \\ \hat{F}_z^{iso}(t) + \frac{\hat{F}_{z+1}^{iso}(t) - \hat{F}_z^{iso}(t)}{x_{(z+1)}^* - x_{(z)}^*}\left[x - x_{(z)}^*\right] & x_{(z)}^* \leq x < x_{(z+1)}^* \quad z = 1 \dots n-1. \\ \hat{F}_n^{iso}(t) & x \geq x_{(n)}^* \end{cases}$$

5. Finally, the suggested estimator denoted by $\hat{F}_{R2}(t)$ given by:

$$\hat{F}_{R2}(t) = \frac{1}{nk} \sum_{j=1}^{m} \sum_{i=1}^{k} \sum_{l=1}^{k} \hat{F}_{x_{l(i)j}}(t).$$

It is pertinent to mention here that we assume the positivity between Y and X. However, if there is a negative relation between Y and X, then the sorting process is carried out in a descending way in step 2. Zamanzade and Mahdizadeh (2018) concluded, based on a numerical study, that $\hat{F}_{R2}(t)$ outperforms $\hat{F}_R(t)$ at almost the considered cases, particularly when the rankings are done perfectly, i.e., Y and X are linearly correlated variables.

3. The CDF estimation in DRSS

In this part, the main components of this study are described. At the beginning, the DRRSS will be briefly presented. Following, the proposed CDF estimators based on DRRSS are derived and explained.

3.1 The DRRSS scheme

DRRSS is a novel scheme recently suggested by Taconeli (2023) as another variation of the traditional RSS. He argued that adopting DRRSS can allow scholars to obtain more representative samples from the parent population at the same number of sampling items, kn, needed for producing RSS. In contrast, RSS and DRRSS depend on the same number of wasted measurement units during the ranking mechanism. They have a strong distinction between the two designs. The DRRSS technique requires selecting the sampling units through two ranking stages rather than one ranking step, which is carried out in the RSS setting to get an extra representation of the population of interest. One practical deficiency concerning DRRSS is that it is more prone to ranking errors than RSS due to repeated performing the raking stage. Thus, it may be advisable to adopt DRSS if the ranking quality is reasonable enough. The steps of balanced DRRSS can be described as follows:

1- Assign randomly k^2 items into equal k sets from the interested population.
2- Sort the k sampling items within each set without exact quantification by judgment or any cheap ranking rule.
3- Resort the sampling items in Step 2 separately across the judgmental order statistics $i = 1,2, \dots, k$, based on the same ranking rule carried out in the first ranking step.
4- Measure exactly the median units across i^{th} judgment order statistics for k odd, i.e., consider the items in position $\left(\frac{k+1}{2}\right)^{th}$ for $i = 1,2, \dots, k$. while for k even, measure the judgment order statistics in position $\left(\frac{k}{2}\right)$ across $1^{st}, 3^{rd}, \dots (k-1)^{th}$ judgment order statistics and, on the other hand, $\left(\frac{k}{2}+1\right)$ across the remaining judgment order statistics.
5- For a larger sample, the above steps can be repeated m cycles to obtain a DRRSS with size $n = km$.

One can easily conclude that the sampling items measured by DRRSS are no longer independent due to the second-ranking stage. Let $Y_{(z)(i)j}$ denotes the item position in i^{th} judgment order statistics during the first ranking stage and in the position z^{th} judgment order statistics in the second-ranking stage from j^{th} cycle. Mathematically, the DRRSS measured sampling units can be expressed as:

$$\begin{cases} \left\{Y_{(\frac{k+1}{2})(i)j};\quad i = 1,2, \dots, k\,, j = 1,2, \dots, m\right\}, & \text{if k is odd} \\ \left\{\left(Y_{(\frac{k}{2})(2i-1)j}; Y_{(\frac{k}{2}+1)(2i)j}\right);\quad i = 1,2, \dots, \frac{k}{2}, j = 1,2, \dots, m\right\}, & \text{if k is even} \end{cases}$$

Taconeli (2023) investigated numerically that DRRSS enables us to estimate the population mean with higher efficiency and precision than RSS, specifically when the parent distribution is symmetric. Abdallah and Al-Omari (2024) used DRRSS to propose a new CDF estimator based on the empirical distribution function given by:

$$\hat{F}_D(t) = \begin{cases} \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{k} I(Y_{(\frac{k+1}{2})(i)j} \leq t), & \text{if } k \text{ is odd} \\ \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{\frac{k}{2}}\left(I(Y_{(\frac{k}{2})(2i-1)j} \leq t) + I(Y_{(\frac{k}{2}+1)(2i)j} \leq t)\right)\right), & \text{if } k \text{ is even.} \end{cases}$$

Abdallah and Al-Omari (2024) discussed the consistency property and the asymptotic distribution of $\hat{F}_D(t)$. Moreover, based on a series of comparison studies, they found that $\hat{F}_D(t)$ performs better than $\hat{F}_R(t)$ for most of the considered cases, regardless of the ranking quality.

3.2 The proposed CDF estimators using auxiliary information

Proceeding the same way described by Kvam and Samaniego (1994), one can utilize the ranking information generated by the measured sampling items for the construction of the first proposed CDF estimator under DRRSS corresponding to $\hat{F}_{R1}(t)$. Our idea can be summarized as follows:

- If k is odd

For a certain t, if we remarked that the sampling item $Y_{(\frac{k+1}{2})(i)j}$ actually is less than t, then one can claim that:

$$P\left(Y_{(l)(i)j} < t\right) = 1 \qquad i = 1,2, \ldots k;\ l = 1,2 \ldots \frac{k-1}{2}. \qquad (5)$$

while for the remaining items:

$$P\left(Y_{(l)(i)j} < t \middle| Y_{(l)(i)j} > Y_{(\frac{k+1}{2})(i)j}\right) = \frac{F(t) - F\left(y_{(\frac{k+1}{2})(i)j}\right)}{1 - F\left(y_{(\frac{k+1}{2})(i)j}\right)}$$

$$i = 1,2, \ldots k;\ l = \frac{k+3}{2}, \frac{k+5}{2} \ldots k. \qquad (6)$$

In contrast, if we observed that the sampling item $Y_{(\frac{k+1}{2})(i)j}$ actually is greater than t, then it is logical to think about that:

$$P\left(Y_{(l)(i)j} < t\right) = 0, \qquad i = 1,2, \ldots k;\ l = \frac{k+3}{2}, \frac{k+5}{2} \ldots k. \qquad (7)$$

yet for the remaining items:

$$P\left(Y_{i(l)j} < t \middle| Y_{i(l)j} < Y_{(\frac{k+1}{2})(i)j}\right) = \frac{F(t)}{F\left(Y_{(\frac{k+1}{2})(i)j}\right)}, \qquad \forall\, l = 1,2 \ldots \frac{k-1}{2}. \tag{8}$$

Putting $(5-8)$ together, the proposed CDF estimator under DRRSS for odd k can be formulated as:

$$\hat{F}^{*}_{DO1}(t) = \frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(\left(\frac{k+1}{2}\right)+\left(\frac{k-1}{2}\right)\frac{F(t)-F\left(y_{(\frac{k+1}{2})(i)j}\right)}{1-F\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j} < t\right)\right] + \left[\left(\frac{k-1}{2}\right)\left(\frac{F(t)}{F\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j} > t\right)\right]$$

- If k is even

$$\begin{cases} Y_{(\frac{k}{2})(i)j} < t & i = 1,3,\ldots k-1 \\ Y_{(\frac{k}{2}+1)(i)j} < t & i = 2,4,\ldots k \end{cases}$$

then one can impute the remaining items as:

$$\begin{cases} \begin{cases} P\left(Y_{(l)(i)j} < t\right) = 1 \quad l = 1,2 \ldots \frac{k}{2} - 1 \\ P\left(Y_{(l)(i)j} < t \middle| Y_{(l)(i)j} > Y_{\left(\frac{k}{2}\right)(i)j}\right) = \frac{F(t)-F\left(y_{\left(\frac{k}{2}\right)(i)j}\right)}{1-F\left(y_{\left(\frac{k}{2}\right)(i)j}\right)} l = \frac{k+2}{2}, \frac{k+4}{2} \ldots k \end{cases} & i = 1,3,\ldots k-1 \\ . & \\ \begin{cases} P\left(Y_{(l)(i)j} < t\right) = 1 \qquad l = 1,2 \ldots \frac{k}{2};. \\ P\left(Y_{(l)(i)j} < t \middle| Y_{(l)(i)j} > Y_{\left(\frac{k}{2}+1\right)(i)j}\right) = \frac{F(t)-F\left(y_{\left(\frac{k}{2}+1\right)(i)j}\right)}{1-F\left(y_{\left(\frac{k}{2}+1\right)(i)j}\right)} \quad l = \frac{k}{2}+2, \frac{k}{2}+3 \ldots k. \end{cases} & i = 2,4,\ldots k \end{cases} .(9)$$

Whereas if:

$$\begin{cases} Y_{(\frac{k}{2})(i)j} > t & i = 1,3,\ldots k-1 \\ Y_{(\frac{k}{2}+1)(i)j} > t & i = 2,4,\ldots k \end{cases}$$

then one can impute the remaining items as:

$$
\begin{cases}
\begin{cases}
P\left(Y_{(l)(i)j} < t\right) = 0 \quad l = \frac{k+2}{2}, \frac{k+4}{2} \dots k \\
P\left(Y_{(l)(i)j} < t \middle| Y_{(l)(i)j} < Y_{\left(\frac{k}{2}\right)(i)j}\right) = \frac{F(t)}{F\left(y_{\left(\frac{k}{2}\right)(i)j}\right)} \quad l = 1,2 \dots \frac{k}{2} - 1 \qquad i = 1,3,\dots k-1
\end{cases} \\
. \\
\begin{cases}
P\left(Y_{(l)(i)j} < t\right) = 0 \quad l = \frac{k}{2} + 2, \frac{k}{2} + 3 \dots k\,. \\
P\left(Y_{(l)(i)j} < t \middle| Y_{(l)(i)j} < Y_{\left(\frac{k}{2}+1\right)(i)j}\right) = \frac{F(t)}{F\left(y_{\left(\frac{k}{2}+1\right)(i)j}\right)} \quad l = 1,2 \dots \frac{k}{2}\,. \qquad i = 2,4, \dots k
\end{cases}
\end{cases}
\tag{10}
$$

In the light of $(9 - 10)$, the proposed CDF estimator under DRRSS for even k can be formulated as:

$$
\begin{aligned}
\hat{F}^{*}_{DE1}(t) = \frac{1}{nk} \sum_{j=1}^{m} \sum_{i=1}^{\frac{k}{2}} & \left[\left(\left(\frac{k}{2}\right) + \left(\frac{k}{2} - 1\right) \frac{F(t) - F\left(y_{\left(\frac{k}{2}\right)(2i-1)j}\right)}{1 - F\left(y_{\left(\frac{k}{2}\right)(2i-1)j}\right)} \right) I\left(Y_{\left(\frac{k}{2}\right)(2i-1)j} < t\right) \right] \\
& + \left[\left(\frac{k}{2} - 1\right) \left(\frac{F(t)}{F\left(y_{\left(\frac{k}{2}\right)(2i-1)j}\right)} \right) I\left(Y_{\left(\frac{k}{2}\right)(2i-1)j} > t\right) \right] \\
& + \left[\left(\left(\frac{k}{2} + 1\right) + \left(\frac{k}{2}\right) \frac{F(t) - F\left(y_{\left(\frac{k}{2}+1\right)(2i)j}\right)}{1 - F\left(y_{\left(\frac{k}{2}+1\right)(2i)j}\right)} \right) I\left(Y_{\left(\frac{k}{2}+1\right)(2i)j} < t\right) \right] \\
& + \left[\left[\left(\frac{k}{2}\right) \left(\frac{F(t)}{F\left(y_{\left(\frac{k}{2}+1\right)(2i)j}\right)} \right) I\left(Y_{\left(\frac{k}{2}+1\right)(2i)j} > t\right) \right] \right].
\end{aligned}
$$

The problem of $\hat{F}^{*}_{DO1}(t)$ essentially boils down to estimation of $F(t)$ and $F\left(y_{\left(\frac{k+1}{2}\right)(i)j}\right)$. Similar to the steps of EM algorithm proceeded in the preceding section, one can overcome this problem by using the following steps:

1- Set $r = 0$.

2- Estimate the unknown CDFs in $\hat{F}^{*}_{DO1}(t)$ with the estimator proposed by Abdallah and Al-Omari (2024), $\hat{F}_D(t)$, i.e. compute the following equation:

$$
\hat{F}^{*(r)}_{DO1}(t) = \frac{1}{nk} \sum_{j=1}^{m} \sum_{i=1}^{k} \left[\left(\left(\frac{k+1}{2}\right) + \left(\frac{k-1}{2}\right) \frac{\hat{F}_D(t) - \hat{F}_D\left(y_{\left(\frac{k+1}{2}\right)(i)j}\right)}{1 - \hat{F}_D\left(y_{\left(\frac{k+1}{2}\right)(i)j}\right)} \right) I\left(Y_{\left(\frac{k+1}{2}\right)(i)j} < t\right) \right] + \left[\left(\frac{k-1}{2}\right) \left(\frac{\hat{F}_D(t)}{\hat{F}_D\left(y_{\left(\frac{k+1}{2}\right)(i)j}\right)} \right) I\left(Y_{\left(\frac{k+1}{2}\right)(i)j} > t\right) \right],
$$

3- Replace t in $\hat{F}^{*(r)}_{DO1}(t)$ with all the sampling items, i.e. compute the following equation:

$$\hat{F}^{*(r)}_{DO1}\left(y_{(\frac{k+1}{2})(l)w}\right)=\frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(\left(\frac{k+1}{2}\right)+\left(\frac{k-1}{2}\right)\frac{\hat{F}_D\left(y_{(\frac{k+1}{2})(l)w}\right)-\hat{F}_D\left(y_{(\frac{k+1}{2})(i)j}\right)}{1-\hat{F}_D\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j}<y_{(\frac{k+1}{2})(l)w}\right)\right]$$
$$+\left[\left(\frac{k-1}{2}\right)\left(\frac{\hat{F}_D\left(y_{(\frac{k+1}{2})(l)w}\right)}{\hat{F}_D\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j}>y_{(\frac{k+1}{2})(l)w}\right)\right],$$

$l=1,2,..k$ and $w=1,2,..m.$

4- Set $r=r+1$.

5- Obtain $\hat{F}^{*(r)}_{DO1}(t)$ by using the following recursive equation:

$$\hat{F}^{*(r)}_{DO1}(t)=\frac{1}{nk}\sum_{j=1}^{m}\sum_{i=1}^{k}\left[\left(\left(\frac{k+1}{2}\right)+\left(\frac{k-1}{2}\right)\frac{\hat{F}^{*(r-1)}_{DO1}(t)-\hat{F}^{*(r-1)}_{DO1}\left(y_{(\frac{k+1}{2})(i)j}\right)}{1-\hat{F}^{*(r-1)}_{DO1}\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j}<t\right)\right]$$
$$+\left[\left(\frac{k-1}{2}\right)\left(\frac{\hat{F}^{*(r-1)}_{DO1}(t)}{\hat{F}^{*(r-1)}_{DO1}\left(y_{(\frac{k+1}{2})(i)j}\right)}\right)I\left(Y_{(\frac{k+1}{2})(i)j}>t\right)\right],$$

6- Repeat steps (4-5) until stopping rule satisfies. i.e.

$$\left|\hat{F}^{*(r)}_{DO1}(t)-\hat{F}^{*(r-1)}_{DO1}(t)\right|<.001$$

7- Finally, the suggested estimator denoted by $\hat{F}_{DO1}(t)$ given by:

$$\hat{F}_{DO1}(t)=\hat{F}^{*(r)}_{DO1}(t)$$

By applying the same idea explained above, $\hat{F}^{*}_{DE1}(t)$ can become as an estimable function denoted by $\hat{F}_{DE1}(t)$.

The second proposed CDF estimator can be obtained, firstly, by defining the items of concomitant-based DRRSS as:

$$\begin{cases}\left\{\left(Y_{(\frac{k+1}{2})(i)j},X_{(\frac{k+1}{2})(i)j}\right);\ \ i=1,2,\dots,k\ ,j=1,2,\dots,m\right\}, & \text{if k is odd}\\ \left\{\left(\left(Y_{(\frac{k}{2})(2i-1)j},X_{(\frac{k}{2})(2i-1)j}\right);\left(Y_{(\frac{k}{2}+1)(2i)j},X_{(\frac{k}{2}+1)(2i)j}\right)\right);\ \ i=1,2,\dots,\frac{k}{2},j=1,2,\dots,m\right\}, & \text{if k is even}\end{cases},$$

where $\left\{X_{(h)(i)j};\ \ h=1,\dots,k;i=1,\dots,k;j=1,\dots,m\right\}$ be the set of all concomitant variable values used to impute the wasted measurement sampling units of Y. Secondly, by implementing the same manner previously reported, adopted for getting $\hat{F}_{R2}(t)$ which can be listed as: 1- Create new variables $(\mathcal{y}^{*}_{z},x^{*}_{z}\ ,z=1\dots n)$, including all the values of the interested variable and their corresponding concomitant sampling items. 2- Sorting ascending $(\mathscr{y}^*_z, x^*_z)$ according to x^* items leading to $(\mathscr{y}^*_{[z]}, x^*_{(z)})$. 3- Calculate the isotonized values for $I\left(\mathscr{y}^*_{[z]} \le t\right)$ and keep these values in $\hat{\mathcal{F}}^{iso}_z(t)$. For each $\left\{x_{(h)(i)j};\ h = 1, \dots, k; i = 1, \dots, k; j = 1, \dots, m\right\}$, estimate the corresponding $\hat{\mathcal{F}}^{iso}_z(t)$ by implementing the linear interpolation formula given by:

$$\hat{\mathcal{F}}_x(t) = \begin{cases} \hat{\mathcal{F}}^{iso}_1(t) & x \le x^*_{(1)} \\ \hat{\mathcal{F}}^{iso}_z(t) + \frac{\hat{\mathcal{F}}^{iso}_{z+1}(t) - \hat{\mathcal{F}}^{iso}_z(t)}{x^*_{(z+1)} - x^*_{(z)}} \left[x - x^*_{(z)}\right] & x^*_{(z)} \le x < x^*_{(z+1)} \quad z = 1 \dots n-1. \\ \hat{\mathcal{F}}^{iso}_n(t) & x \ge x^*_{(n)} \end{cases}$$

6. Finally, the suggested estimator denoted by $\hat{F}_{D2}(t)$ given by:

$$\hat{F}_{D2}(t) = \frac{1}{nk} \sum_{j=1}^{m} \sum_{i=1}^{k} \sum_{h=1}^{k} \hat{\mathcal{F}}_{x_{(h)(i)j}}(t).$$

It is interesting to remark that the preceding steps described for getting $\hat{\mathcal{F}}_{D2}(t)$ does not depend on whether k is either an odd or even number, as opposed to what was done early in deriving our first proposed CDF estimator.

4. Monte Carlo Comparisons

In this part, we examine to what extent the proposed estimators perform well relative to their analog under RSS using a comprehensive simulation experiment. The relative efficiency (RE) criterion is used for comparison purposes defined as:

$$RE(t)_1 = \begin{cases} \frac{MSE\left(\hat{F}_{R1}(t)\right)}{MSE\left(\hat{F}_{DO1}(t)\right)} & \text{if k is odd} \\ \frac{MSE\left(\hat{F}_{R1}(t)\right)}{MSE\left(\hat{F}_{DE1}(t)\right)} & \text{if k is even} \end{cases} \quad \text{and} \quad RE(t)_2 = \frac{MSE\left(\hat{F}_{R2}(t)\right)}{MSE\left(\hat{F}_{D2}(t)\right)},$$

Where MSE refers to mean square error. Dell and Clutter's (1972) model with controlling parameter ρ is adopted for assessing the quality of ranking such that $\rho \epsilon \{1, 0.9, 0.5\}$, which $\rho = 1$ for perfect ranking, $\rho = 0.9$ for good ranking, and $\rho = 0.5$ for weak ranking. Standard normal, standard uniform and standard exponential will be assumed hereafter concerning the distribution of the parent population. This enables us to investigate the properties of the proposed estimators under both symmetric and asymmetric populations. Different set sizes and cycle sizes are considered: $(k, m) = (3,5), (3,10), (5,3), (5,6)$, $(4,6), (4,10), (6,4)$ and $(6,5)$. The values of $RE(t)$ are computed for $t = Q_p$, $p = 0.05, 0.25, 0.5, 0.75$ and 0.90. where Q_p is the p^{th} quantile of the underlying distribution. Using R language and environment for statistical computing supported by R Core Team (2020), the code is available upon request from the author, a total of 5000 samples were simulated under various combinations of p, k, m and ρ at each distribution as displayed in Tables $(1-6)$.

One can observe that the proposed estimators have a very promising performance in all the cases, for values of t lie at the center of the underlying distribution. A sizeable efficiency gain can be obtained by improving the ranking quality. Moving t for the upper / lower tail of the distribution leads to a loss in the efficiency of all the proposed estimators. The proposed estimators have a very competitive performance for small sample sizes with a few exceptions. The pattern of the parent distribution does not have a pronounced effect on the behavior of the proposed estimators. Generally speaking, the behavior of the proposed estimators is better for symmetric distributions. Overall examination of the simulation results reported in Tables $(1-6)$, shows that none of the presented estimators can uniformly beat the others. The results are reasonably good in favor of the DRSS-based estimators when the true value of $F(t)$ is near the center of the underlying parent distribution and the ranking quality is good. This fact enjoys the advantage of being almost robust to parent distribution. This may be justified by the fact that DRSS selects the median item during the implementation of the second stage. Further, DRSS requires performing the ranking process twice to be more sensitive to the error ranking relative to RSS. It is worth mentioning that additional comparisons are performed between the proposed procedures and $\hat{F}_D(t)$ proposed by Abdallah and Al-Omari (2024). It turns out that the proposed procedures uniformly outperform the latter. Thus, these results are omitted for space considerations.

5. Empirical study

In this part, the performance of the proposed estimators is assessed based on a real data set. This data set was found by Chen (2004), and it was of size 396 sampling items with seven variables. Hereafter, two variables will only be considered: "the entire height in feet" denoted by the interested variable, and the "diameter in centimeters at breast height," denoted by the concomitant variable. It is easy to explore that ρ between "the entire height in feet" and "diameter in centimeters at breast height" equals 0.91, implying that the ranking quality expectedly will be reasonably perfect. For the same values of k , m, and p used in section 4, 5,000 samples are generated using RSS and DRRSS mechanisms. To guarantee that the independent condition of the generated samples is satisfied, all samples are drawn with replacement. Again, the procedures discussed are calculated for each of the generated samples. The $RE(t)_1$ and $RE(t)_2$ values are obtained as reported in Table $(7-8)$. t reveals that the proposed estimators are the winner and have satisfactory precision relative to the RSS-based estimators if t lies away from the boundaries. Moreover, in many cases, better efficiency can be obtained by increasing the set size rather than the cycle size. Thus, there is a strong consistency between the remarks concluded based on Tables $(1-6)$ and the notes observed from Tables $(7-8)$. Further evidence of comparing the proposed CDF estimators and their competitors is exhibited by an additional detailed particular case that visualizes the actual population CDF besides the proposed CDF estimators and their competitors using a tree dataset in the case of $k=5$ and $m=6$. It is apparent that the lines of the proposed CDF estimators become closer to the line of the true CDF at most of the middle points and sometimes for the upper tail point regarding to $\hat{F}_{DO1}(t)$.

Table 1: Estimated $RE_1(t)$ based on standard normal distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	0.97	0.75	0.85	0.69	0.67	0.59	0.68	0.55
	0.90	0.85	0.79	0.85	0.78	0.64	0.59	0.69	0.71
	0.50	0.79	0.88	0.76	0.79	0.66	0.72	0.75	0.75
0.25	1.00	1.22	1.26	1.54	1.33	1.15	1.05	1.36	1.57
	0.90	1.11	1.13	1.11	1.05	1.14	1.14	1.16	1.07
	0.50	1.18	0.92	1.03	0.85	1.01	0.93	1.03	0.98
0.50	1.00	1.64	1.65	2.19	2.15	1.19	1.28	3.01	2.92
	0.90	1.35	1.46	1.52	1.62	1.33	1.18	1.53	1.41
	0.50	1.13	1.16	1.28	1.11	1.23	1.10	1.30	1.22
0.75	1.00	1.15	1.29	1.58	1.37	1.08	0.97	1.34	1.01
	0.90	1.15	1.15	1.19	0.98	1.07	1.11	1.11	1.04
	0.50	1.15	0.99	1.01	0.92	1.02	0.96	1.02	1.05
0.90	1.00	0.92	0.91	0.73	0.82	0.54	0.78	0.66	0.87
	0.90	0.88	0.78	0.91	0.82	0.94	0.71	0.67	0.78
	0.50	0.98	0.79	0.88	0.84	0.89	0.81	0.65	0.73

Table 2: Estimated $RE_2(t)$ based on standard normal distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	0.88	0.71	0.79	0.57	0.88	0.98	0.78	0.76
	0.90	0.75	0.65	0.87	0.65	0.99	1.01	0.91	0.85
	0.50	0.79	0.85	0.77	0.69	0.95	1.01	0.95	0.85
0.25	1.00	1.11	1.15	1.29	1.11	1.21	1.03	0.99	1.05
	0.90	1.17	1.12	1.13	1.22	1.01	0.99	1.12	1.02
	0.50	1.24	1.05	1.11	0.98	1.03	1.01	1.14	1.10
0.50	1.00	1.05	1.11	1.15	1.17	1.05	1.09	1.25	1.09
	0.90	1.11	1.09	1.09	1.28	1.22	1.10	1.08	1.05
	0.50	1.17	1.15	1.15	1.06	1.09	1.07	1.11	1.06
0.75	1.00	1.09	1.13	1.31	1.20	1.22	0.98	0.99	0.98
	0.90	1.13	1.15	1.11	1.19	0.95	1.05	1.10	1.01
	0.50	1.20	1.01	1.10	1.02	1.01	1.04	1.01	1.08
0.90	1.00	0.87	0.75	0.74	0.75	1.03	0.99	0.84	0.92
	0.90	0.89	0.88	0.92	1.02	0.97	1.02	1.01	0.99
	0.50	0.99	0.84	0.85	0.95	0.99	1.03	1.01	0.93

Table 3: Estimated $RE_1(t)$ based on standard exponential distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	1.01	0.78	0.79	0.69	0.69	0.66	0.69	0.76
	0.90	0.96	0.79	0.79	0.81	0.65	0.66	0.68	0.66
	0.50	0.76	0.74	0.86	0.81	0.87	0.69	0.86	0.64
0.25	1.00	1.11	1.07	1.60	1.10	1.15	1.95	1.85	1.62
	0.90	0.96	1.12	1.16	1.02	1.14	1.21	1.10	1.01
	0.50	0.96	1.05	1.05	1.02	1.12	1.02	0.99	0.89
0.50	1.00	1.67	1.70	2.14	2.12	1.18	1.11	2.01	2.91
	0.90	1.33	1.48	1.56	1.61	1.19	1.12	1.65	1.67
	0.50	1.19	1.20	1.25	1.27	1.16	1.06	1.16	1.15
0.75	1.00	1.17	1.07	1.37	1.03	1.02	0.95	1.16	0.99
	0.90	1.07	0.99	1.04	0.99	0.78	0.79	0.98	0.69
	0.50	1.14	1.02	1.07	0.99	0.78	0.66	0.89	0.78
0.90	1.00	0.79	0.81	0.72	0.89	0.85	0.74	0.66	0.65
	0.90	0.86	0.89	0.79	0.82	0.76	0.65	0.78	0.56
	0.50	0.89	0.88	0.88	0.77	0.78	0.55	0.69	0.55

Table 4: Estimated $RE_2(t)$ based on standard exponential distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	0.89	0.67	0.69	0.59	0.92	0.90	0.77	0.71
	0.90	0.92	0.69	0.75	0.79	1.01	1.11	0.90	0.78
	0.50	0.70	0.62	0.89	0.68	1.02	0.98	1.02	0.75
0.25	1.00	1.05	1.02	1.27	1.11	1.13	0.99	1.06	0.93
	0.90	1.01	1.21	1.18	1.08	1.09	1.02	0.98	1.02
	0.50	1.02	1.12	1.08	1.10	1.11	0.96	1.08	1.02
0.50	1.00	1.13	1.09	1.09	1.12	1.08	1.13	1.14	1.19
	0.90	1.13	1.08	1.10	1.09	1.17	1.06	1.10	1.12
	0.50	1.25	1.28	1.14	1.18	0.98	0.98	1.03	1.02
0.75	1.00	1.09	0.98	1.28	1.14	1.03	0.97	0.91	0.82
	0.90	1.13	1.05	1.15	1.02	0.89	0.94	1.01	1.01
	0.50	1.23	1.14	1.20	1.07	0.97	0.92	1.10	1.02
0.90	1.00	0.69	0.87	0.76	0.84	0.99	1.25	1.13	1.20
	0.90	0.90	0.87	0.89	0.99	1.09	1.19	1.14	1.11
	0.50	0.80	0.80	0.94	0.86	1.11	1.17	1.08	1.02

Table 5: Estimated $RE_1(t)$ based on standard uniform distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	0.95	0.81	0.95	0.66	0.58	0.72	0.76	0.86
	0.90	0.87	0.86	0.81	0.86	0.69	0.61	0.77	0.69
	0.50	0.88	0.74	0.77	0.91	0.74	0.70	0.71	0.78
0.25	1.00	1.33	1.07	1.46	1.29	2.91	2.71	1.52	1.55
	0.90	0.96	1.08	1.12	0.99	1.35	1.37	1.16	0.99
	0.50	0.94	1.01	1.02	0.93	1.07	0.95	1.01	0.96
0.50	1.00	1.70	1.71	2.01	2.45	1.14	1.21	3.12	2.74
	0.90	1.30	1.35	1.55	1.59	1.27	1.22	1.54	1.51
	0.50	1.16	1.22	1.20	1.14	1.08	1.15	1.38	1.28
0.75	1.00	1.29	1.10	1.32	1.20	2.99	2.41	1.49	1.47
	0.90	1.01	1.06	1.10	1.01	1.31	1.30	1.11	0.98
	0.50	0.97	1.03	1.03	0.98	1.04	0.98	1.04	1.02
0.90	1.00	0.89	0.79	0.78	0.87	0.65	0.82	0.78	0.57
	0.90	0.90	0.86	0.99	0.89	0.88	0.74	0.66	0.68
	0.50	1.03	0.89	0.86	0.76	0.71	0.99	0.89	0.85

Table 6: Estimated $RE_2(t)$ based on standard uniform distribution

					(k,m)				
$F(t)$	$\rho =$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	1.00	0.89	0.68	0.79	0.69	0.97	0.95	0.86	0.76
	0.90	0.76	0.69	0.80	0.77	1.01	1.02	0.91	0.99
	0.50	0.79	0.67	0.79	0.78	1.04	0.97	0.85	0.81
0.25	1.00	1.16	0.99	1.25	1.04	1.01	1.04	1.02	1.02
	0.90	1.01	1.14	1.20	1.11	1.03	1.01	1.08	1.01
	0.50	1.05	1.12	1.18	1.12	1.11	1.03	1.11	1.08
0.50	1.00	1.20	1.09	1.09	1.17	1.06	1.12	1.17	1.08
	0.90	1.16	1.06	1.11	1.04	1.24	1.11	1.04	1.11
	0.50	1.19	1.23	1.12	1.10	0.98	1.05	1.21	1.11
0.75	1.00	1.05	1.04	1.21	1.08	1.05	1.03	1.01	0.99
	0.90	0.99	1.05	1.16	1.08	1.04	0.99	1.05	1.02
	0.50	1.03	1.10	1.15	1.09	1.15	1.02	1.09	1.07
0.90	1.00	0.74	0.69	0.77	0.86	1.02	1.02	0.99	1.02
	0.90	0.90	0.86	1.03	1.03	1.03	1.05	1.01	1.05
	0.50	0.91	0.88	0.88	0.79	1.05	0.99	1.02	1.05

Table 7: Estimated $RE_1(t)$ based on Tree dataset

	(k,m)							
$F(t)$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	0.78	0.76	0.67	0.70	0.65	0.65	0.76	0.87
0.25	1.23	1.17	1.20	1.25	1.17	1.06	1.25	1.38
0.50	1.55	1.45	1.43	1.75	1.25	1.16	2.01	1.93
0.75	1.06	1.13	1.35	1.12	0.89	0.76	1.05	0.94
0.90	0.97	0.87	0.63	0.79	0.98	0.89	$\hat{F}_{DO1}(t)$ 99	0.67

Table 8: Estimated $RE_2(t)$ based on Tree dataset

	(k,m)							
$F(t)$	(3,5)	(3,10)	(5,3)	(5,6)	(4,6)	(4,10)	(6,4)	(6,5)
0.05	0.95	0.88	0.90	0.98	1.02	1.03	1.03	0.95
0.25	1.08	1.13	1.05	1.13	1.06	1.02	1.19	0.99
0.50	1.19	1.10	1.04	1.11	1.21	1.21	1.31	1.20
0.75	1.10	1.21	1.15	1.10	1.11	0.90	0.95	1.15
0.90	1.06	0.90	0.87	1.01	1.17	0.99	0.89	0.99

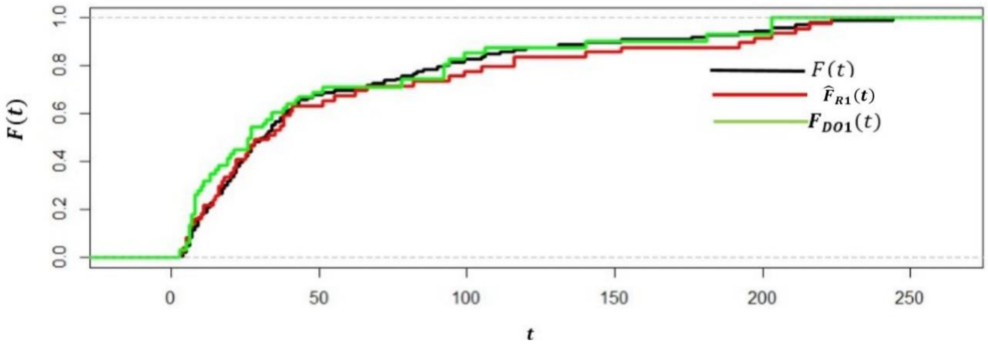

Figure 1: The population CDF and EM algorithm-based estimators based on tree data set

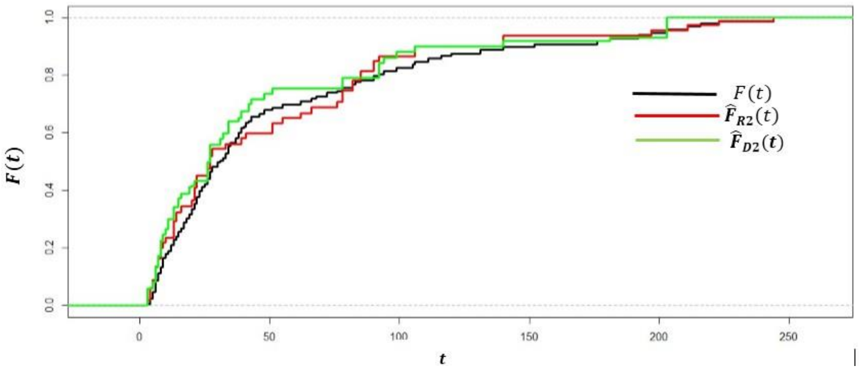

Figure 2: The population CDF and concomitant-based estimators based on tree data set

6. Conclusion

Here, we addressed the problem of CDF estimation in light of the missing data approach under a new RSS-based design known as DRRSS. Since this issue has not yet been adequately studied in the RSS literature, this study can be considered a first trial. Two novel CDF estimators are introduced. The first proposed estimator is based on the relationship between the ranks of measured sampling items and information from unmeasured sampling items. The second one depends essentially on the information generated by a concomitant variable. A series of simulated datasets enables us to identify and explore the precision of our proposed estimator relative to Kvam and Samaniego (1994) and Zamanzade and Mahdizadeh (2018). It is concluded that a considerable efficiency gain is observed if the actual CDF value at the center of the underlying distribution and the quality of rankings is reasonable enough. The performance of the proposed estimators is also examined based on an empirical dataset. Furthermore, additional studies are needed to address the theoretical properties of the proposed estimators. Moreover, examining the parametric inference, see Hassan et al. (2023) and the nonparametric inference, see Ghamsari et al. (2023) under DRRSS, can be considered a possible topic for future studies. The author seeks to take on these academic issues shortly.

References

Abdallah, M. S. (2023). More efficient estimators of the area under the receiver operating characteristic curve in paired ranked set sampling. *Statistical Methods in Medical Research*, 32(6), 1217-1233. doi:org/10.1177/09622802231167434

Abdallah, M. S., & Al-Omari, A. I. (2024). An Efficient CDF Estimator Based on Dual-Rank Ranked Set Sampling with an Application to Body Mass Index Data. *Journal of the Indian Society for Probability and Statistic*s, 1-18. doi:10.1007/s41096-023-00171-8

Al-Saleh, M. F., & Al-Ta'ani, L. A. K. (2023). Estimation of Population Size Using Ranked Set Sampling and Some of its Variations. *Journal of Probability and Statistical Science,* 21(2). doi:org/10.37119/jpss2023.v21i2.699

Ashour, S. and Abdallah, M. (2020). Estimation of distribution function based on ranked set sampling: missing data approach. *Thailand Statistician*, 18. 27-42.

Chen, Z., Bai, Z., and Sinha, B.K. (2004). Ranked set sampling: Theory and Applications, Springer. New York, 2004.

David H.A. & Nagaraja H.N. (2003). Order Statistics. 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Dell, T. R. and Clutter, J. L. (1972). Ranked set sampling theory with an order statistics background. *Biometrics*, 28. 545-555. doi:org/10.2307/2556166

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series* B. 39. 1–38.

Frey, J., & Zhang, Y. (2023). Nonparametric maximum likelihood estimation of the distribution function using ranked-set sampling. *Journal of the Korean Statistical Society,* 52(4), 901-920. doi: 10.1007/s42952-023-00229-0

Göçoğlu, A., & Demirel, N. (2019). Estimating the population proportion in modified ranked set sampling methods. *Journal of Statistical Computation and Simulation*, 89(14), 2694-2710. doi: 10.1080/00949655.2019.1631315

Hassan, A. S., Al-Omari, A., & Nagy, H. F. (2021). Stress–strength reliability for the generalized inverted exponential distribution using MRSS. *Iranian Journal of Science and Technology, Transactions A: Science*, 45(2), 641-659. doi:10.1007/s40995-020-01033-9

Hassan, A. S., Alsadat, N., Elgarhy, M., Chesneau, C., & Mohamed, R. E. (2023). Different classical estimation methods using ranked set sampling and data analysis for the inverse power Cauchy distribution. *Journal of Radiation Research and Applied Sciences,* 16(4), 100685. doi:org/10.1016/j.jrras.2023.100685

Kvam, P. H., & Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association*, 89(426): 526-537. doi:10.1080/01621459.1994.10476777

Ozturk, O. (2007). Statistical inference under a stochastic ordering constraint in ranked set sampling. *Nonparametric Statistics*, 19. 131–144. doi:10.1080/10485250701437232

Stokes, S.L. & Sager, T.W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association,* 83(402), 374–381. doi:org/10.2307/2288852

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Taconeli, C. A. (2023). Dual-rank ranked set sampling. Journal of Statistical Computation and Simulation, 94(1), 29-49. doi:10.1080/00949655.2023.2229472

Zamanzade, E. & Mahdizadeh, M. (2018). Distribution function estimation using concomitant-based ranked set sampling. *Hacettepe Journal of Mathematics and Statistics.* 47(3). 755-761. 01.06.2018

Zamanzade, E. Mahdizadeh, M. & Samawi, H. (2024). Nonparametric estimation of mean residual lifetime in ranked set sampling with a concomitant variable. *Journal of Applied Statistics.* 1-17.