REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000

# ESTIMATION OF DISTRIBUTION FUNCTION USING PERCENTILE RANKED SET SAMPLING

Authors: Yusuf Can Sevil ✉*
– The Graduate School of Natural and Applied Sciences, Dokuz Eylul University, Izmir, Turkey (yusuf.sevil@ogr.deu.edu.tr)

Tugba Ozkal Yildiz
– Department of Statistics, Dokuz Eylul University, Izmir, Turkey (tugba.ozkal@deu.edu.tr)

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- The estimation of distribution function has received considerable attention in the literature. Because, many practical problems involve estimation of distribution function from experimental data. Estimating the distribution function makes it possible to do pointwise estimation and to make statistical inference about the distribution of interested population. In this study, we suggested an empirical distribution function (EDF) for percentile ranked set sampling (PRSS). Bias of the EDF estimator is investigated theoretically and numerically. Relative efficiencies of the proposed EDF estimator based on PRSS with respect to EDF estimator based on simple random sampling (SRS) and ranked set sampling (RSS) are obtained. We also considered impact of imperfect rankings on the EDF based on PRSS. According to the results, the proposed EDF estimator is unbiased for the extreme "minimum and maximum" points and center of the distribution. Also, it is clearly appeared that the EDF estimator based on PRSS is more efficient than the EDF based on SRS. Another important result is that the suggested EDF estimator has larger efficiencies than the EDF based on RSS for some special cases of PRSS. In the application, the EDF based on PRSS is used to estimate the proportion of women in obesity class III (BMI$>$ 40).

Key-Words:

- *Percentile ranked set sampling; empirical distribution function; relative efficiency; mean squared error; imperfect ranking; body mass index data.*

AMS Subject Classification:

- 62P10, 62D99, 68U20.

*Corresponding author

## 1. INTRODUCTION

Ranked set sampling (RSS) was introduced by McIntyre [13] as an advantageous alternative to simple random sampling (SRS). McIntyre [13] studied mean estimator based on RSS and showed that this estimator is more efficient than mean estimator using SRS. Then, mathematical theory of RSS was first suggested by Takahasi and Wakimoto [24]. By Dell and Clutter [6], it was proved that mean estimator based on RSS is unbiased and more efficient than mean estimator based on SRS even if ranking is not perfect. In the literature, there are some other estimators based on RSS such as estimation of correlation coefficient [21], estimation of variance [22] and estimation of population proportion [15, 28, 29]. Also, for more extended literature about RSS, see Kaur et al. [11] and Al-Omari and Bouza [3].

The estimation of cumulative distribution function (CDF) with various settings of the RSS has been studied by many authors. Stokes and Sager [23] suggested an unbiased estimator based on RSS for population distribution function. Samawi and Al-Sagheer [19] considered EDF estimator based on extreme ranked set sampling and median ranked set sampling. EDF using double ranked set sampling was investigated by Abu-Dayyeh et al. [1]. Al-Omari [2] studied EDF based on quartile ranked set sampling. Sevil and Yildiz [20] developed estimation of distribution function using RSS based on level-2 sampling design. Also, Kolmogorov Smirnov (KS) test statistic based on RSS was compared with KS test statistic based on SRS by Sevil and Yildiz [20]. EDF estimators using RSS based on three different sampling designs were given by Yildiz and Sevil [25, 26]. Some goodness of fit tests based on these EDF estimators were investigated in their study. Some other distribution function estimators were considered for extreme median ranked set sampling [12], selective order ranked set sampling [4], partially rank-ordered set [16] and pair ranked set sampling [27].

By using percentiles instead of quartiles, more flexible selection procedure named as percentile ranked set sampling (PRSS) was suggested by Muttlak [14]. In Muttlak's study, estimation of mean is investigated using PRSS. Since PRSS is general form of quartile ranked set sampling (QRSS) and extreme ranked set sampling (ERSS), EDF estimators based on QRSS and ERSS can be obtained by using EDF estimator based on PRSS. Moreover, EDF estimator based on median ranked set sampling (MRSS) can be derived by using EDF estimator of PRSS when the set size is even. So, the EDF estimator using PRSS becomes quite useful estimator. Therefore, we considered the performance of EDF estimator using PRSS under perfect and imperfect rankings.

This study is organized as follows. In section two, PRSS procedure is defined. The EDF estimator based on PRSS is given in section three. Also, the properties of the EDF estimator are discussed. In section four, we introduce Frey's one-parameter ranking error model [7] and study imperfect ranking case for proposed EDF estimator. Also, we obtained some results under imperfect

ranking in this section. Some inferences about CDF, $F(x)$, are given in section five. Moreover, body mass index data is used to illustrate the EDF using PRSS. Finally, some conclusion remarks are stated in section six.

## 2. PERCENTILE RANKED SET SAMPLING

Muttlak [14] proposed PRSS as practical sampling scheme according to RSS. In literature, modified versions of PRSS can be seen such as double PRSS [9] and multistage PRSS [10].

In this method, $p$th and $q$th percentile of the sample are selected for full measurement, $0 < p < 1$ and $q = 1 - p$. Before we describe the procedure of PRSS, we give some notations. Let $k$, $l$ and $n$ denote set size, number of cycles and total sample size, respectively. Also, $(X_{11j}, X_{12j}, \cdots, X_{1kj})$, $(X_{21j}, X_{22j}, \cdots, X_{2kj})$,$\cdots$, $(X_{k1j}, X_{k2j}, \cdots, X_{kkj})$ are random sets of size $k$ from $j$th cycle, $j = 1, \cdots, l$. Here, it is assumed that $X_{itj}$ is selected from a population with continuous density function $f(x)$ and CDF $F(x)$. The order statistics of the $i$th set are denoted by $X_{i(1)j}, X_{i(2)j}, \cdots, X_{i(k)j}$, $i = 1, \cdots, k$.

Now, we define the procedure of PRSS. First, $k^2$ units are selected without replacement from the population. These units are divided into the $k$ random sets, each of size $k$. In each set, these units are ranked from the smallest to the largest. If the set size $k$ is odd, PRSS is denoted by $PRSS_O$ and it is obtained by using the following steps.

(i) From the first $(k-1)/2$ sets, the $r$th smallest units are measured, $X_{(r)}$.

(ii) The median ranked unit is measured from the $((k+1)/2)$th set, $X_{(m)}$.

(iii) Then, the $s$th smallest units are measured from the remaining $(k-1)/2$ sets, $X_{(s)}$.

where $r$ and $s$ are the nearest integer value of $p(k+1)$ and $q(k+1)$, respectively. Note that $r = 1$ if $p(k+1) < 0.5$ and $s = k$ if the nearest integer value of $q(k+1)$ is larger than $k$. If the set size $k$ is even, PRSS is denoted by $PRSS_E$ and it is obtained by using the following steps.

(i) From the first $k/2$ sets, the $r$th smallest units are measured, $X_{(r)}$.

(ii) Then, the $s$th smallest units are measured from the remaining $k/2$ sets, $X_{(s)}$.

To obtain $n = lk$ sample observations, these procedures are repeated $l$ times.

$PRSS_O$ and $PRSS_E$ are denoted by

$$PRSS_O = \big\{X_{1(r)j}, X_{2(r)j}, \cdots, X_{\frac{k-1}{2}(r)j}, X_{m(m)j}, X_{\frac{k+3}{2}(s)j}, \cdots, X_{k-1(s)j}, X_{k(s)j}\big\}$$

and

$$PRSS_E = \Big\{X_{1(r)j}, \cdots, X_{\frac{k}{2}(r)j}, X_{\frac{k+2}{2}(s)j}, \cdots, X_{k(s)j}\Big\},$$

respectively, where $m = (k+1)/2$ and $j = 1, \cdots, l$.

As defined in Stokes and Sager [23], $lk$ independent copies $(Y, R)$ are observed as follows: $R$ is first selected at random from $1, \cdots, k$ and $Y$ is observed according to $F_{(i)}(x)$ (the CDF of the $i$th order statistics), then the marginal joint distribution of $Y$'s is the same as that of the SRS. This statement is given in the part (a) of the Theorem 1 by Stokes and Sager [23]. Part (b) of the Theorem 1 capitalizes on this characterization to link RSS with SRS.

Let $T' = (T_1, T_2, \cdots, T_k)$ be a multinomial random vector with $lk$ trials and $P = (\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k})$ be a probability vector. It is supposed that the $lk$ random variables were obtained by first observing $T$ and then selecting $T_i$ units randomly from a population with probability density function (PDF) $f_{(i)}(x)$, $i = 1, \cdots, k$. Also, the obtained $lk$ units are denoted by $Y_1, Y_2, \cdots, Y_{lk}$.

**Theorem 2.1.** *With the same conditions of Theorem 1 in Stokes and Sager [23], we give the following:*

*(1)* *When the set size is odd,* $\{Y_1, Y_2, \cdots, Y_{lk} \mid T = (0, \cdots, 0, t_r = \frac{(k-1)l}{2}, 0, \cdots, 0, t_m = l, 0, \cdots, 0, t_s = \frac{(k-1)l}{2}, 0, \cdots, 0)\}$ *has the same probability structure as* $\{X_{g(r)j}, X_{m(m)j}, X_{h(s)j}; g = 1, 2, \cdots, \frac{k-1}{2}; m = \frac{k+1}{2}; h = \frac{k+3}{2}, \frac{k+5}{2}, \cdots, k; j = 1, 2, \cdots, l\}$ *where ranks of the measured observations could be one of the* $(r, s)$ *pairs,* $\{(1, k), (2, k-1), \cdots, (\frac{k-1}{2}, \frac{k+3}{2})\}$.

*(2)* *When set size is even,* $\{Y_1, Y_2, \cdots, Y_{lk} \mid T = (0, \cdots, 0, t_r = \frac{lk}{2}, 0, \cdots, 0, t_s = \frac{lk}{2}, 0, \cdots, 0)\}$ *has the same probability structure as* $\{X_{g(r)j}, X_{h(s)j}; g = 1, 2, \cdots, \frac{k}{2}; h = \frac{k+2}{2}, \frac{k+4}{2}, \cdots, k; j = 1, 2, \cdots, l\}$ *where ranks of the measured observations could be one of the* $(r, s)$ *pairs,* $\{(1, k), (2, k-1), \cdots, (\frac{k}{2}, \frac{k+2}{2})\}$.

These part (1) and (2) are proved in Appendices.

## 3. EMPIRICAL DISTRIBUTION FUNCTION OF PERCENTILE RANKED SET SAMPLING

In this section, we described the suggested EDF estimator based on PRSS. Also, properties of the EDF estimator are given. Bias and efficiency of the EDF based on PRSS are investigated and compared with distribution function estimators using SRS and RSS. It is assumed that $X_1, X_2, \cdots, X_n$ be a simple random sample. EDF based on SRS is denoted by $\hat{F}_{SRS}(x)$,

$$\hat{F}_{SRS}(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right).$$

where $I(.)$ is indicator function. The EDF based on SRS is unbiased estimator of $F(x)$ for given $x$, with variance $V(\hat{F}_{SRS}(x)) = \frac{1}{n}F(x)(1-F(x))$.

Stokes and Sager [23] proposed $\hat{F}_{RSS}(x)$ for estimating the distribution function $F(x)$. Let $\left\{X_{1(1)j}, X_{2(2)j}, \cdots, X_{k(k)j}\right\}$ be the order statistics that are obtained by using RSS,

$$\hat{F}_{RSS}(x) = \frac{1}{lk}\sum_{j=1}^{l}\sum_{i=1}^{k} I\left(X_{i(i)j} \leq x\right) \tag{3.1}$$

They showed that $\hat{F}_{RSS}(x)$ is unbiased with variance

$$V\left(\hat{F}_{RSS}(x)\right) = \frac{1}{lk^2}\sum_{i=1}^{k} F_{(i)}(x)\left(1 - F_{(i)}(x)\right)$$

where $F_{(i)}(x)$ is distribution function of the $i$th order statistic, and

$$\frac{\hat{F}_{RSS}(x) - E\left(\hat{F}_{RSS}(x)\right)}{\left(V\left(\hat{F}_{RSS}(x)\right)\right)^{1/2}}$$

converges in distribution to standard normal as $l \to \infty$, when $x$ and $k$ are held fixed.

Let $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ are the EDFs of a PRSS data when set size

is odd and even, respectively. If set size is odd,

$$\hat{F}_{PRSS_O}(x) = \frac{1}{lk}\left[\sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} I\left(X_{i(r)j} \leq x\right) + \sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right) + \sum_{j=1}^{l} I\left(X_{m(m)j} \leq x\right)\right] \tag{3.2}$$

and if set size is even,

$$\hat{F}_{PRSS_E}(x) = \frac{1}{lk}\left[\sum_{j=1}^{l}\sum_{i=1}^{\frac{k}{2}} I\left(X_{i(r)j} \leq x\right) + \sum_{j=1}^{l}\sum_{i=1}^{\frac{k}{2}} I\left(X_{\frac{k}{2}+i(s)j} \leq x\right)\right] \tag{3.3}$$

where $r \approx p(k+1)$, $s \approx q(k+1)$ and $m = \frac{k+1}{2}$ is the median ranked unit. Under the perfect ranking, we state the following propositions for some basic properties of these distribution function estimators.

**Proposition 1.** (a) Using $PRSS_O$

i. $E\left(\hat{F}_{PRSS_O}(x)\right) = \left(\frac{1}{2} - \frac{1}{2k}\right)\left(F_{(r)}(x) + F_{(s)}(x)\right) + \frac{1}{k}F_{(m)}(x)$

ii. $V\left(\hat{F}_{PRSS_O}(x)\right) = \frac{1}{lk^2}\left[\begin{array}{c}\left(\frac{k-1}{2}\right)\left(F_{(r)}(x)\left(1-F_{(r)}(x)\right) + F_{(s)}(x)\left(1-F_{(s)}(x)\right)\right) + \\ F_{(m)}(x)\left(1-F_{(m)}(x)\right)\end{array}\right]$

(b) Using $PRSS_E$

i. $E\left(\hat{F}_{PRSS_E}(x)\right) = \frac{1}{2}\left(F_{(r)}(x) + F_{(s)}(x)\right)$

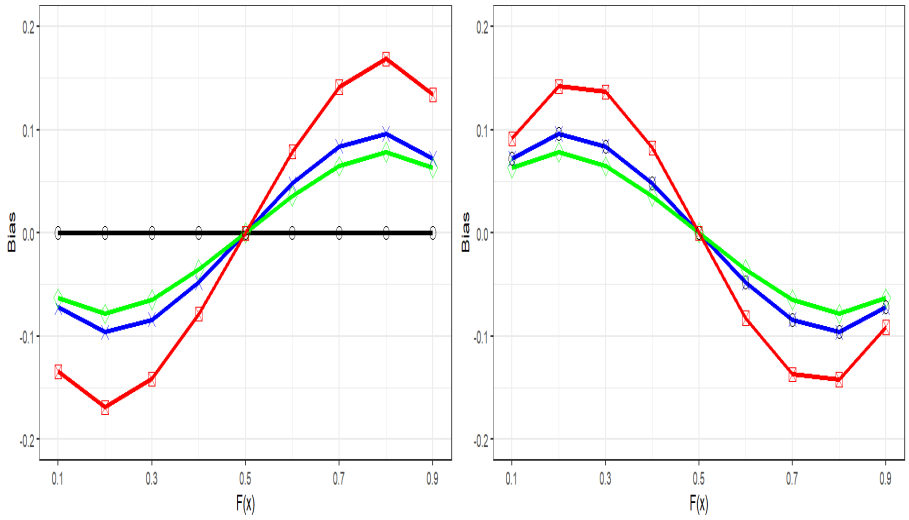
ii. $V\left(\hat{F}_{PRSS_E}(x)\right) = \frac{1}{2lk}\left[F_{(r)}(x)\left(1-F_{(r)}(x)\right) + F_{(s)}(x)\left(1-F_{(s)}(x)\right)\right]$

where $F_{(r)}(x)$, $F_{(s)}(x)$ and $F_{(m)}(x)$ are distribution function of $X_{(r)}$, $X_{(s)}$ and $X_{(m)}$, respectively. Part $(a)$ and part $(b)$ are proved in Appendices. As seen in Proposition 1, $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ are biased estimators for $F(x)$. However, the bias is almost zero as $F(x)$ gets closer to 1, 0.5 and 0 under perfect ranking. Also, the biases of these estimators do not depend on the number of cycles. The biases of these EDFs can be calculated by using following equations.

$$Bias[\hat{F}_{PRSS_O}(x)] = F(x) - E(\hat{F}_{PRSS_O}(x)), \tag{3.4}$$

$$Bias[\hat{F}_{PRSS_E}(x)] = F(x) - E(\hat{F}_{PRSS_E}(x)). \tag{3.5}$$

These biases of $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ are given by Figure 1 when $p = 0.1$ and $p = 0.4$. These EDF estimators are unbiased as $F(x)$ gets closer to 1, 0.5 and 0. The bias increases as $k$ increases except for $F(x) = 0.5$. In the Figure 1(b), the blue and black curves are overlapping. Mean squared error is used as




(a) When $p = 0.1$, the bias of EDFs (b) When $p = 0.4$, the bias of EDFs

Figure 1: Bias for $\hat{F}_{PRSS_O}$ and $\hat{F}_{PRSS_E}$ where black, blue, green and red curves are $k = 3$, $k = 4$, $k = 5$ and $k = 6$, respectively

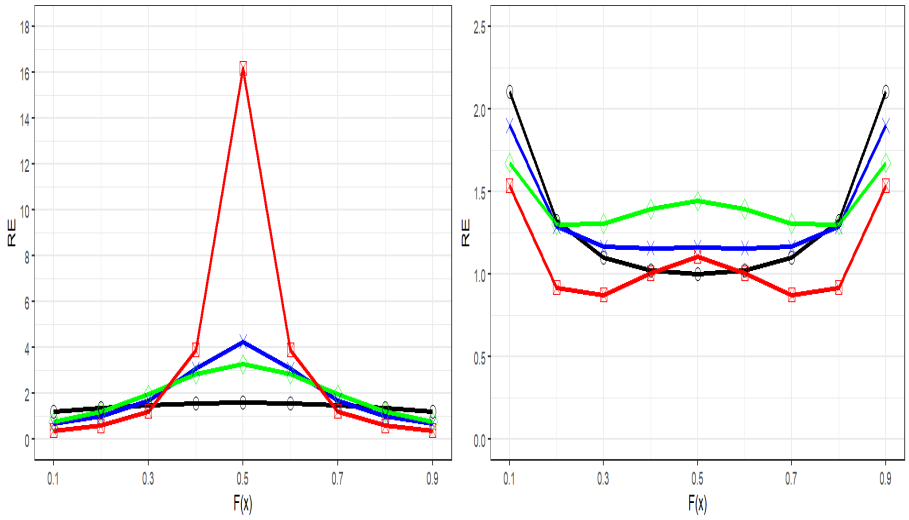
a measure of performance of the proposed estimators. Then, relative efficiencies (RE) of $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ with respect to $\hat{F}_{SRS}(x)$ are described as

$$RE[\hat{F}_{PRSS_O}(x), \hat{F}_{SRS}(x)] = \frac{V(\hat{F}_{SRS}(x))}{MSE(\hat{F}_{PRSS_O}(x))},$$

and

$$RE[\hat{F}_{PRSS_E}(x), \hat{F}_{SRS}(x)] = \frac{V(\hat{F}_{SRS}(x))}{MSE(\hat{F}_{PRSS_E}(x))}.$$

REs are illustrated by the Figure 2. When $p = 0.1$, it is seen that the REs peak on the middle of the distribution function. Even, the EDFs based on PRSS are more efficient than the EDF based on RSS whenever $F(x)$ is close to 0.5 comparing with Stokes and Sager [23]. The REs increase while the set size increases. When $p = 0.4$, Figure 2 shows that the REs are higher on the tails of the distribution function. Whenever $F(x)$ is close to 0.1 (or 0.9) comparing with Stokes and Sager [23], the EDFs based on PRSS are more efficient than the EDF based on RSS. Also, the REs are almost equal to or larger than 1 for any $F(x)$ when $k = 3, 4, 5, 6$ and $p = 0.4$. Table 1 indicates REs of EDFs using PRSS when




(a) When $p = 0.1$, the REs of EDFs (b) When $p = 0.4$, the REs of EDFs

Figure 2: REs for $\hat{F}_{PRSS_O}$ and $\hat{F}_{PRSS_E}$ where black, blue, green and red curves are $k = 3$, $k = 4$, $k = 5$ and $k = 6$, respectively

$F(x) = 0.1$ and $F(x) = 0.5$ relative to RSS. The REs are obtained by using the following equations.

$$RE[\hat{F}_{PRSS_O}(x), \hat{F}_{RSS}(x)] = \frac{V(\hat{F}_{RSS}(x))}{MSE(\hat{F}_{PRSS_O}(x))},$$

and

$$RE[\hat{F}_{PRSS_E}(x), \hat{F}_{RSS}(x)] = \frac{V(\hat{F}_{RSS}(x))}{MSE(\hat{F}_{PRSS_E}(x))}.$$

It can be shown that the EDFs based on PRSS (with $p = 0.4$) have higher performances than the EDF based on RSS when $F(x) = 0.1$. Also, the EDFs using PRSS (with $p = 0.1$) are more efficient than the EDF using RSS when $F(x) = 0.5$.

| | $F(x) = 0.1$ | | $F(x) = 0.5$ | |
|---|---|---|---|---|
| $k$ | $p = 0.1$ | $p = 0.4$ | $p = 0.1$ | $p = 0.4$ |
| 3 | 1.000 | 1.000 | 1.760 | 0.625 |
| 4 | 0.522 | 2.333 | 1.473 | 0.636 |
| 5 | 0.557 | 1.635 | 1.227 | 0.720 |
| 6 | 0.263 | 7.303 | 1.045 | 0.500 |

Table 1: The REs of the EDF estimators based on PRSS with respect to RSS

The following proposition is needed to study some asymtotic inference about the expected value of the estimators, $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$. The Proposition 2 is proved in Appendices.

**Proposition 2.** For fixed $k$ and $l \to \infty$, the following results are obtained.

(a) $\frac{\hat{F}_{PRSS_O}(x) - E\left(\hat{F}_{PRSS_O}(x)\right)}{\sqrt{V\left(\hat{F}_{PRSS_O}(x)\right)}}$ converges in distribution to $N(0,1)$.

(b) $\frac{\hat{F}_{PRSS_E}(x) - E\left(\hat{F}_{PRSS_E}(x)\right)}{\sqrt{V\left(\hat{F}_{PRSS_E}(x)\right)}}$ converges in distribution to $N(0,1)$.

## 4. IMPERFECT RANKING

The efficiency of PRSS is affected by ranking steps. In general, the ranking is performed by subjective judgement or according to concomitant (auxiliary) variable that is correlated to the variable of interest. In the ranking steps, it is assumed that the ranking is completely accurate. However, this is not a realistic assumption. Therefore, one of the interesting topic is ranking error models in the literature. Dell and Clutter [6] proposed adaptive perceptual error model. Bohn and Wolfe [5] suggested ranking error model that constructs the judgement class distributions as a mixture distribution of the actual order statistics. Then, Frey [7, 8] extended the model [5] and introduced new class of models for imperfect ranking. Ozturk [17] estimated the parameters of ranking error models of Bohn and Wolfe [5] and Frey [7, 8]. He proved that one-parameter ranking error model [7, 8] is more efficient than ranking error model [5].

In this section, we investigated the effect of imperfect ranking on PRSS using Frey's one-parameter judgement ranking [7]. It is assumed that $k!$ possible judgment orderings of the true order statistics $X_{(i_1:h)}, \cdots, X_{(i_k:h)}$ selected from a larger set of size $h$, $h \geq k$. Random selection of set of size $k$ yields $\binom{h}{k}$ possible selection of $k$ order statistics out of $h$ order statistics in the larger set and all these selections are equally likely. Let $\boldsymbol{A}(i_1, \cdots, i_k)$ be a doubly stochastic matrix. Frey [7] specified a way to compute the matrix $\boldsymbol{A}(i_1, \cdots, i_k)$,

$$\boldsymbol{A}(i_1, \cdots, i_k) = \frac{1}{k!} \sum_{\pi \in S_k} q(i_{\pi(1)}, \cdots, i_{\pi(k)}) \times Per(\pi(1), \cdots, \pi(k)),$$

where $q(i_{\pi(1)}, \cdots, i_{\pi(k)})$ denotes the probability that corresponds to the ordering of $X_{(i_1:h)} < \cdots < X_{(i_k:h)}$, $Per(\pi(1), \cdots, \pi(k))$ is the permutation matrix whose $(i, \pi(i))$th entry is one for $i = 1, \cdots, k$ and all other entries are zero, and $S_k$ is the set of all permutations. The probabilities $q(i_{\pi(1)}, \cdots, i_{\pi(k)})$ are obtained by selecting an appropriate weight function $w(\pi)$ with $\pi \in S_k$. These weights must be normalized, so these are actually probabilities. A class of weight function was

suggested by Frey [7],

$$w(\pi) = \exp\left\{\delta \sum_{j=1}^{k} j\lambda\left(\frac{i_{\pi(j)}}{h+1}\right)\right\}$$

where $\delta$ is called as power and $\delta \in [0, \infty)$. When $\delta = 0$, a completely random ranking model is constructed. When $\delta$ approaches infinity, the probability $q(i_{\pi(1)}, \cdots, i_{\pi(k)})$ concentrates on the single permutation having the largest value of

$$\sum_{j=1}^{k} j\lambda\left(\frac{i_{\pi(j)}}{h+1}\right)$$

and corresponds to a perfect ranking model. Also, a wide range of imperfect ranking models can be obtained using the other values of $\delta$. Frey [7] proposed three different $\lambda$ function which are $\lambda_1(u) = u$, $\lambda_2(u) = -u^{-1}$ and $\lambda_3(u) = (1-u)^{-1}$ to obtain symmetric, skewed-left and skewed-right imperfect ranking probabilities. Note that these probabilities do not depend on shape of underlying distributions. $\boldsymbol{\Omega}(i_1, \cdots, i_k)$ is a $k \times h$ matrix to exhaust the selection of all possible judgment orderings. In this matrix, the $(i^{'}, i_{i^{'}})$th entry is one for $i^{'} = 1, \cdots, k$ and all other entries are zero. Then, the matrix product

$$\boldsymbol{N}(i_1, \cdots, i_k) = \boldsymbol{A}(i_1, \cdots, i_k)\boldsymbol{\Omega}(i_1, \cdots, i_k)$$

is a $k \times h$ matrix that constructs relation between $\boldsymbol{A}(i_1, \cdots, i_k)$ and the set of independent order statistics $X_{(i_1:h)}, \cdots, X_{(i_k:h)}$ in the larger set of size $h$. The distribution of $X_{[i]}$, conditional on the values of $i_1, \cdots, i_k$ is given by

$$F_{[i]}(x|i_1, \cdots, i_k) = \sum_{\iota=1}^{h} \boldsymbol{N}(i_1, \cdots, i_k)_{i\iota} F_{(\iota)}(x)$$

where $\boldsymbol{N}(i_1, \cdots, i_k)_{i\iota}$ is the $(i, \iota)$th entry of $\boldsymbol{N}(i_1, \cdots, i_k)$. When the contribution of all $\binom{h}{k}$ equally likely choices of the values of $i_1, \cdots, i_k$ the CDF of $X_{[i]}$ can then be written

$$F_{[i]}(x) = \sum_{\iota=1}^{h} p_{k,h}(i, \iota) F_{(\iota)}(x)$$

where $\boldsymbol{P}_{k,h} = (p_{k,h}(i, \iota))$ is the $k \times h$ matrix average

$$\boldsymbol{P}_{k,h} = \binom{h}{k}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le h} \boldsymbol{N}(i_1, \cdots, i_k)_{i\iota}$$

In our study, we assumed that $\boldsymbol{P}_{k,h}$ is a square matrix, so we use $\boldsymbol{P}$ and $p(i, \iota)$ instead of $\boldsymbol{P}_{k,h}$ and $p_{k,h}(i, \iota)$, respectively. For more details about Frey's one-parameter judgement ranking model, see Frey [7]. The matrix $\boldsymbol{P}$ can be estimated by using an R-function that is proposed by Ozturk [17] for any correlation coefficient ($\rho$), the set size ($k$) and the larger set size ($h$). For theoretical backgrounds of the R-function, see Ozturk [17]. In the following example, we illustrate the matrix $\boldsymbol{P}$.

**Example 1.** It is assumed that set size $k = 4$ and the units in the set are ranked perfectly. Then, the matrix $\boldsymbol{P}$ is as follows:

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

If the units in the set are ranked randomly, then the matrix $\boldsymbol{P}$ is as follows:

$$\boldsymbol{P} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

Let us define that $X_{i[1]j}, X_{i[2]j}, \cdots, X_{i[k]j}$ be judgement order statistics in the $i$th set, $i = 1, \cdots, k$ and $j = 1, \cdots, l$. Then, the EDF based on RSS [23] as follows.

$$\hat{F}^*_{RSS}(x) = \frac{1}{lk} \sum_{j=1}^{l} \sum_{i=1}^{k} I\left(X_{i[i]j} \leq x\right) \tag{4.1}$$

On the other hand, the measured units in the steps of the PRSS are denoted by $X_{[r]}$, $X_{[s]}$ and $X_{[m]}$. Thus, the measured units in $PRSS_O$ and $PRSS_E$ are represented by

$$\begin{aligned} PRSS_O = \big\{ & X_{1[r]j}, X_{2[r]j}, \cdots, \\ & X_{\frac{k-1}{2}[r]j}, X_{m[m]j}, X_{\frac{k+3}{2}[s]j}, \cdots, \\ & X_{k-1[s]j}, X_{k[s]j} \big\} \end{aligned}$$

and

$$\begin{aligned} PRSS_E = \Big\{ & X_{1[r]j}, \cdots, X_{\frac{k}{2}[r]j}, \\ & X_{\frac{k+2}{2}[s]j}, \cdots, X_{k[s]j} \Big\}, \end{aligned}$$

respectively, where $m = (k+1)/2$ and $j = 1, \cdots, l$. The CDF estimators based on $PRSS_O$ and $PRSS_E$ are given by

$$\begin{aligned} \hat{F}^*_{PRSS_O}(x) = \frac{1}{lk} \Bigg[ & \sum_{j=1}^{l} \sum_{i=1}^{\frac{k-1}{2}} I\left(X_{i[r]j} \leq x\right) \\ & + \sum_{j=1}^{l} \sum_{i=1}^{\frac{k-1}{2}} I\left(X_{\frac{k+1}{2}+i[s]j} \leq x\right) \\ & + \sum_{j=1}^{l} I\left(X_{m[m]j} \leq x\right) \Bigg] \end{aligned} \tag{4.2}$$

and if set size is even,

$$\hat{F}^*_{PRSS_E}(x) = \frac{1}{lk}\left[\sum_{j=1}^{l}\sum_{i=1}^{\frac{k}{2}} I\left(X_{i[r]j} \leq x\right) + \sum_{j=1}^{l}\sum_{i=1}^{\frac{k}{2}} I\left(X_{\frac{k}{2}+i[s]j} \leq x\right)\right] \tag{4.3}$$

where $r \approx p(k+1)$, $s \approx q(k+1)$ and $m = \frac{k+1}{2}$ is the median ranked unit. The following proposition gives the properties of $\hat{F}^*_{PRSS_O}(x)$ and $\hat{F}^*_{PRSS_E}(x)$.

**Proposition 3.** (a) Using $PRSS_O$

i. $E\left(\hat{F}^*_{PRSS_O}(x)\right) = \left(\frac{1}{2} - \frac{1}{2k}\right)\left(F_{[r]}(x) + F_{[s]}(x)\right) + \frac{1}{k}F_{[m]}(x)$

ii. $V\left(\hat{F}^*_{PRSS_O}(x)\right) = \frac{1}{lk^2}\left[\begin{array}{c}\left(\frac{k-1}{2}\right)\left(F_{[r]}(x)\left(1-F_{[r]}(x)\right) + F_{[s]}(x)\left(1-F_{[s]}(x)\right)\right) + \\ F_{[m]}(x)\left(1-F_{[m]}(x)\right)\end{array}\right]$

(b) Using $PRSS_E$

i. $E\left(\hat{F}^*_{PRSS_E}(x)\right) = \frac{1}{2}\left(F_{[r]}(x) + F_{[s]}(x)\right)$

ii. $V\left(\hat{F}^*_{PRSS_E}(x)\right) = \frac{1}{2lk}\left[F_{[r]}(x)\left(1-F_{[r]}(x)\right) + F_{[s]}(x)\left(1-F_{[s]}(x)\right)\right]$

where

$$F_{[t]}(x) = \sum_{\iota=1}^{k} p(t,\iota)F_{(\iota)}(x), \quad t = \{r,s,m\}.$$

The proof the Proposition 3 is the same as the proof of the Proposition 1. We gave an example in order to illustrate obtaining the distribution of judgement order statistics $F_{[t]}$. Also, we investigated the properties of $\hat{F}^*_{PRSS_O}(x)$ and $\hat{F}^*_{PRSS_E}(x)$ under random ranking case in this example. First, we give the following lemma that is noted by Dell and Clutter [6]. Detailed proof of this lemma was given by Presnell and Bohn [18].

**Lemma 4.1.** $\frac{1}{k}\sum_{i=1}^{k} F_{[i]}(x) = F(x) \qquad \forall x.$

Using this lemma, the results are provided in the following example.

**Example 2.** Let $\{X_{1[r]j}, X_{2[r]j}, \cdots, X_{\frac{k-1}{2}[r]j}, X_{m[m]j}, X_{\frac{k+3}{2}[s]j}, \cdots, X_{k-1[s]j}, X_{k[s]j}\}$ are obtained using $PRSS_O$ under random ranking case. Then, $p(t,\iota) = \frac{1}{k}$ in the matrix $\boldsymbol{P}$ for each $t = \{r,s,m\}$ and $\iota = 1, \cdots, k$. Thus, $F_{[t]}(x)$ is obtained according to Lemma 1.

$$F_{[t]}(x) = \sum_{\iota=1}^{k} \frac{1}{k}F_{(\iota)}(x) = F(x)$$

Straightforwardly, it can be seen that

$$E\left(\hat{F}^*_{PRSS_O}(x)\right) = F(x),$$

$$V\left(\hat{F}^*_{PRSS_O}(x)\right) = \frac{1}{n}F(x)(1-F(x)).$$

Besides, we have to note that the obtained results are not surprising. It means that $\hat{F}^*_{PRSS_O}(x)$ reduce to $\hat{F}(x)$ under random ranking case. Obviously, these results are the same for $\hat{F}^*_{PRSS_E}(x)$ as well.

Now, we investigated the performances of $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ under the imperfect ranking. To construct imperfect ranking schemes, we take the correlation coefficients as $\boldsymbol{\rho} = \{0.90, 0.75, 0.50\}$. The matrix $\boldsymbol{P}_\upsilon$, $\upsilon = 1, 2, 3$ corresponding to each correlation coefficient are estimated using Ozturk's R-function. When $k = 3$, the estimated matrices are

$$\boldsymbol{P}_1 = \begin{bmatrix} 0.841 & 0.151 & 0.008 \\ 0.151 & 0.698 & 0.151 \\ 0.008 & 0.151 & 0.841 \end{bmatrix},$$

$$\boldsymbol{P}_2 = \begin{bmatrix} 0.762 & 0.210 & 0.028 \\ 0.210 & 0.580 & 0.210 \\ 0.028 & 0.210 & 0.762 \end{bmatrix},$$

$$\textit{and} \quad \boldsymbol{P}_3 = \begin{bmatrix} 0.555 & 0.303 & 0.142 \\ 0.303 & 0.395 & 0.303 \\ 0.142 & 0.303 & 0.555 \end{bmatrix}.$$

for $\rho = 0.90$, $\rho = 0.75$ and $\rho = 0.50$, respectively. These matrices are estimated for $k = 4$, $k = 5$ and $k = 6$ as well. Bias for $\hat{F}^*_{PRSS_O}(x)$ and $\hat{F}^*_{PRSS_E}(x)$ are obtained by using Equations (4.4) and (4.5). Figure 3 gives bias for the CDF estimators based on PRSS with $p = 0.1$ and $p = 0.4$, respectively. For any $\rho$, these EDF estimators are unbiased as $F(x)$ gets closer to 1, 0.5 and 0. Also, the bias increases as $k$ increases except for $F(x) = 0.5$. It can be seen that the biases decrease as $\rho$ decreases. This is a result of the Example 2.

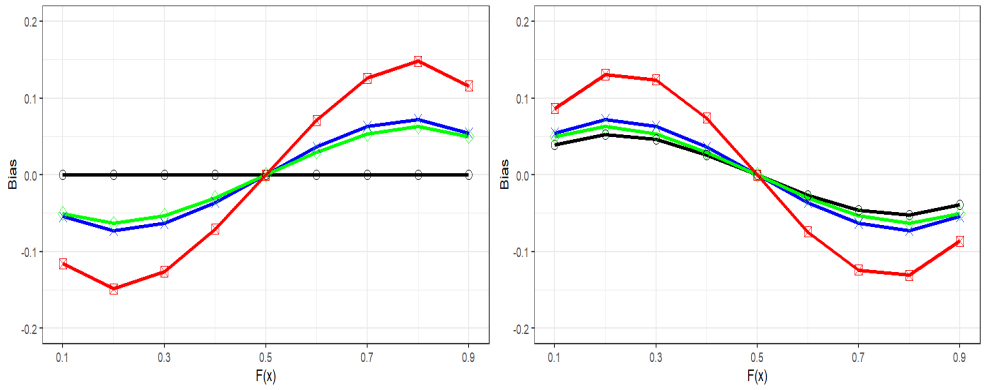
$$Bias[\hat{F}^*_{PRSS_O}(x)] = F(x) - E(\hat{F}^*_{PRSS_O}(x)), \tag{4.4}$$

$$Bias[\hat{F}^*_{PRSS_E}(x)] = F(x) - E(\hat{F}^*_{PRSS_E}(x)). \tag{4.5}$$

Besides, relative efficiencies (RE) of $\hat{F}_{PRSS_O}(x)$ and $\hat{F}_{PRSS_E}(x)$ with respect to $\hat{F}_{SRS}(x)$ are described as

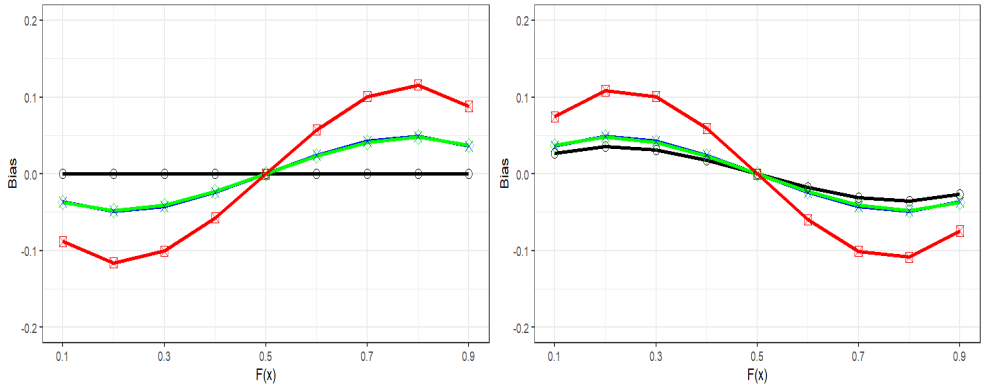
$$RE[\hat{F}^*_{PRSS_O}(x), \hat{F}_{SRS}(x)] = \frac{V(\hat{F}_{SRS}(x))}{MSE(\hat{F}^*_{PRSS_O}(x))},$$

$$RE[\hat{F}^*_{PRSS_E}(x), \hat{F}_{SRS}(x)] = \frac{V(\hat{F}_{SRS}(x))}{MSE(\hat{F}^*_{PRSS_E}(x))}.$$



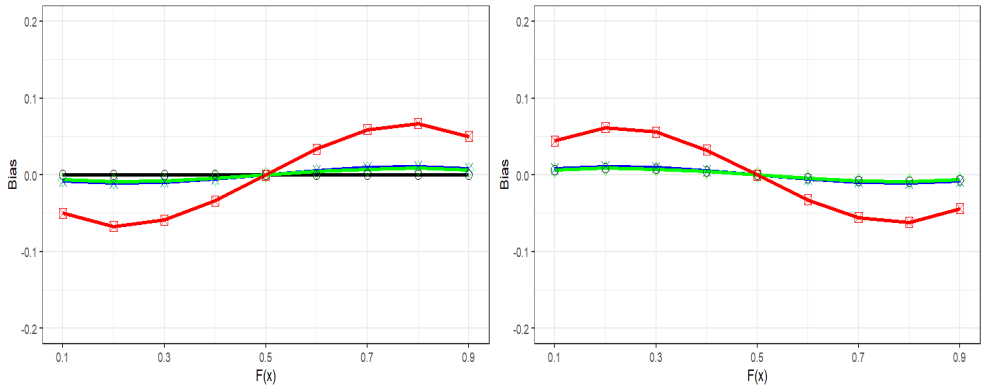

(a) When $p = 0.1$, the bias of EDFs for $\rho = 0.90$

(b) When $p = 0.4$, the bias of EDFs for $\rho = 0.90$




(c) When $p = 0.1$, the bias of EDFs for $\rho = 0.75$

(d) When $p = 0.4$, the bias of EDFs for $\rho = 0.75$

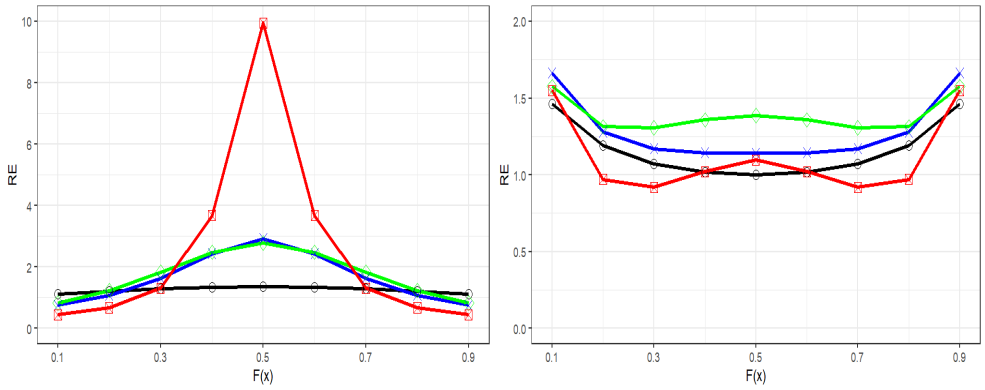



(e) When $p = 0.1$, the bias of EDFs for $\rho = 0.50$

(f) When $p = 0.4$, the bias of EDFs for $\rho = 0.50$

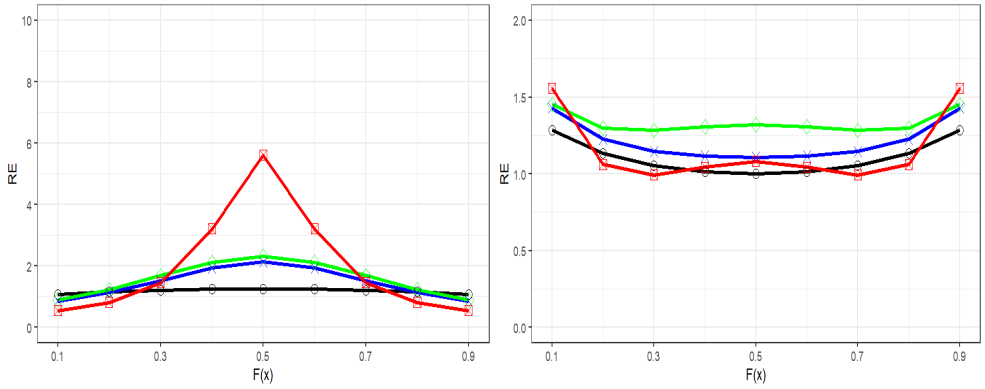
Figure 3: Bias for $\hat{F}_{PRSS_O}$ and $\hat{F}_{PRSS_E}$ where black, blue, green and red curves are $k = 3$, $k = 4$, $k = 5$ and $k = 6$, respectively

REs are given by Figure 4 for $p = 0.1$ and $p = 0.4$, respectively. For any $\rho$, it is seen that the REs peak on the middle of the distribution function when $p = 0.1$.



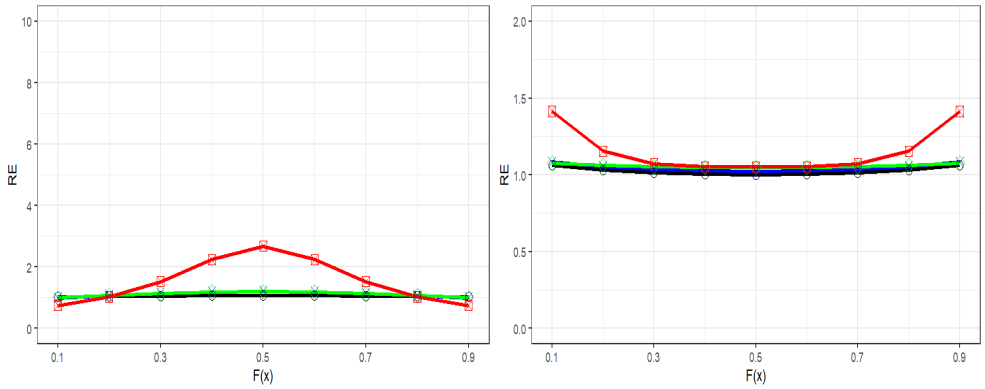

(a) When $p = 0.1$, the REs of EDFs for $\rho = 0.90$

(b) When $p = 0.4$, the REs of EDFs for $\rho = 0.90$




(c) When $p = 0.1$, the REs of EDFs for $\rho = 0.75$

(d) When $p = 0.4$, the REs of EDFs for $\rho = 0.75$




(e) When $p = 0.1$, the REs of EDFs for $\rho = 0.50$

(f) When $p = 0.4$, the REs of EDFs for $\rho = 0.50$

Figure 4: REs for $\hat{F}_{PRSS_O}$ and $\hat{F}_{PRSS_E}$ where black, blue, green and red curves are $k = 3$, $k = 4$, $k = 5$ and $k = 6$, respectively

Also, the REs increase while the set size increases. On the other hand, the REs are higher on the tails of the distribution function when $p = 0.4$. Also, the REs

are almost equal to or larger than 1 for any $F(x)$ and $\rho$ when $k = 3, 4, 5, 6$ and $p = 0.4$.

Table 2 gives REs of EDFs using PRSS when $F(x) = 0.1$ and $F(x) = 0.5$ relative to RSS. The REs are obtained by using the following equations.

$$RE[\hat{F}^*_{PRSS_O}(x), \hat{F}^*_{RSS}(x)] = \frac{V(\hat{F}^*_{RSS}(x))}{MSE(\hat{F}^*_{PRSS_O}(x))},$$

and

$$RE[\hat{F}^*_{PRSS_E}(x), \hat{F}^*_{RSS}(x)] = \frac{V(\hat{F}^*_{RSS}(x))}{MSE(\hat{F}^*_{PRSS_E}(x))}.$$

Table 2 shows that even if $\rho = 0.5$, the gain in efficiency from EDFs using PRSS with $p = 0.4$ (and with $p = 0.1$) are substantial when $F(x) = 0.1$ (and when $F(x) = 0.5$). The Proposition 4 is needed to study some asymptotic inference

| | | $F(x) = 0.1$ | | $F(x) = 0.5$ | |
|---|---|---|---|---|---|
| $\rho$ | $k$ | $p = 0.1$ | $p = 0.4$ | $p = 0.1$ | $p = 0.4$ |
| 0.9 | 3 | 1.000 | 1.312 | 1.000 | 0.740 |
| | 4 | 0.629 | 1.379 | 1.782 | 0.695 |
| | 5 | 0.647 | 1.233 | 1.504 | 0.749 |
| | 6 | 0.323 | 1.128 | 4.784 | 0.527 |
| 0.75 | 3 | 1.000 | 1.185 | 1.000 | 0.798 |
| | 4 | 0.749 | 1.251 | 1.461 | 0.760 |
| | 5 | 0.738 | 1.204 | 1.379 | 0.785 |
| | 6 | 0.434 | 1.231 | 2.993 | 0.578 |
| 0.5 | 3 | 1.000 | 1.036 | 1.000 | 0.936 |
| | 4 | 0.952 | 1.046 | 1.091 | 0.923 |
| | 5 | 0.962 | 1.036 | 1.070 | 0.939 |
| | 6 | 0.651 | 1.231 | 1.775 | 0.696 |

Table 2: The REs of the EDF estimators based on PRSS with respect to RSS

about the expected value of the estimators, $\hat{F}^*_{PRSS_O}(x)$ and $\hat{F}^*_{PRSS_E}(x)$.

**Proposition 4.** For fixed $k$ and $l \to \infty$, the following results are obtained.

(a) $\frac{\hat{F}^*_{PRSS_O}(x) - E\left(\hat{F}^*_{PRSS_O}(x)\right)}{\sqrt{V\left(\hat{F}^*_{PRSS_O}(x)\right)}}$ converges in distribution to $N(0, 1)$.

(b) $\frac{\hat{F}^*_{PRSS_E}(x) - E\left(\hat{F}^*_{PRSS_E}(x)\right)}{\sqrt{V\left(\hat{F}^*_{PRSS_E}(x)\right)}}$ converges in distribution to $N(0, 1)$.

The proof of the Proposition 4 is similar to proof of Proposition 2.

## 5. INFERENCES ABOUT $F(x)$

In this section, we now consider a pointwise estimate of $F(x)$. It supposed that we interest with the proportion, $F(x)$ of population below a specified value $X$. We know that $100(1-\alpha)\%$ confidence interval for $F(x)$ using SRS is as follows:

$$\hat{F}_{SRS}(x) \pm Z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{F}_{SRS}(x))}$$

where $Z_{\frac{\alpha}{2}}$ is the upper quantile of the standard normal distribution and

$$\hat{V}(\hat{F}_{SRS}(x)) = \frac{1}{n-1} \hat{F}_{SRS}(x) \left(1 - \hat{F}_{SRS}(x)\right).$$

Also, Stokes and Sager [23] gave a $100(1-\alpha)\%$ for $F(x)$ using RSS.

$$\hat{F}_{RSS}(x) \pm Z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{F}_{RSS}(x))}$$

where

$$\hat{V}(\hat{F}_{RSS}(x)) = \frac{1}{(l-1)k} \sum_{i=1}^{k} \hat{F}_{(i)}(x) \left(1 - \hat{F}_{(i)}(x)\right).$$

According to Proposition 2, an approximate $100(1-\alpha)\%$ confidence intervals can be constructed when $l$ is larger. For $\hat{F}_{PRSS_O}(x)$, confidence interval of $F(x)$ can be obtained as

$$p\left(Z_{\frac{\alpha}{2}} \leq \frac{\hat{F}_{PRSS_O}(x) - E(\hat{F}_{PRSS_O}(x))}{\sqrt{\hat{V}(\hat{F}_{PRSS_O}(x))}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha, \tag{5.1}$$

where

$$\hat{V}(\hat{F}_{PRSS_O}(x)) = \frac{1}{(l-1)k^2} \left[\left(\frac{k-1}{2}\right) \hat{F}_{(r)}(x)(1-\hat{F}_{(r)}(x)) + \left(\frac{k-1}{2}\right) \hat{F}_{(s)}(x)(1-\hat{F}_{(s)}(x)) + \hat{F}_{(m)}(x)(1-\hat{F}_{(m)}(x))\right]$$

By solving the Equation (5.1) for $E(\hat{F}_{PRSS_O}(x))$, the limits are obtained.

$$\textit{Lower Bound}(LB) = \hat{F}_{PRSS_O}(x) - Z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{F}_{PRSS_O}(x))},$$

and

$$\textit{Upper Bound}(UB) = \hat{F}_{PRSS_O}(x) + Z_{\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{F}_{PRSS_O}(x))}.$$

Thus, $100(1-\alpha)\%$ confidence interval of $F(x)$ can be found by solving the following equations, numerically or any suitable method such as Newton Raphson.

$$\begin{aligned} 2LB &= \frac{1}{k}(k-1)\left(F_{(r)}(x)+F_{(s)}(x)\right)+2F_{(m)}(x) \\ &= \Psi(F), \end{aligned} \tag{5.2}$$

and

$$\begin{aligned} 2UL &= \frac{1}{k}(k-1)\left(F_{(r)}(x)+F_{(s)}(x)\right)+2F_{(m)}(x) \\ &= \Psi(F). \end{aligned} \tag{5.3}$$

For confidence interval of $F(x)$ based on $\hat{F}_{PRSS_E}(x)$,

$$\begin{aligned} p\left(Z_{\frac{\alpha}{2}} \leq \frac{\hat{F}_{PRSS_E}(x)-E(\hat{F}_{PRSS_E}(x))}{\sqrt{\hat{V}(\hat{F}_{PRSS_E}(x))}} \leq Z_{1-\frac{\alpha}{2}}\right) \\ = 1-\alpha, \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} \hat{V}(\hat{F}_{PRSS_E}(x)) = \frac{1}{2(l-1)k}\Big[&\hat{F}_{(r)}(x)(1-\hat{F}_{(r)}(x)) \\ &+\hat{F}_{(s)}(x)(1-\hat{F}_{(s)}(x))\Big] \end{aligned}$$

Thus, the limits are obtained as

$$LB = \hat{F}_{PRSS_E}(x) - Z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{F}_{PRSS_E}(x))},$$

and

$$UB = \hat{F}_{PRSS_E}(x) + Z_{\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{F}_{PRSS_E}(x))}.$$

$100(1-\alpha)\%$ confidence interval of $F(x)$ can be found by solving the following equations,

$$2LB = F_{(r)}(x)+F_{(s)}(x) = \Psi(F), \tag{5.5}$$

and

$$2UL = F_{(r)}(x)+F_{(s)}(x) = \Psi(F). \tag{5.6}$$

Note that $\Psi(F)$ is increasing function in $F(x)$ so the solutions of the Equations (5.2), (5.3), (5.5) and (5.6) should be unique. Similarly, confidence intervals are obtained using $\hat{F}^*_{PRSS_O}(x)$ and $\hat{F}^*_{PRSS_E}(x)$.

## 5.1. A REAL DATA APPLICATION

In the literature, the distribution function estimators are applied to real data such as bilirubin level [19], lung cancer [27] and airquality [26]. The number of case studies can be increased. In the case studies, it can be seen that some quantiles are important hence the probabilities corresponding to them are substantial as well. Thus, if we can estimate the distribution function, these probabilities can also be estimated.

In this section, we consider body mass index data (BMI) to give an illustrative example. BMI is a measure for indicating nutritional status in adults. BMI is frequently used to screen for weight categories that may lead to health problems. A table that includes the weight categories was reported by World Health Organization (WHO), `http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi` and this categories are given by Table 3. According to WHO, the health problems caused

| BMI | Nutritional status |
|---|---|
| Below 18.5 | Underweight |
| $18.5 - 24.9$ | Normal weight |
| $25.0 - 29.9$ | Pre-obesity |
| $30.0 - 34.9$ | Obesity class I |
| $35.0 - 39.9$ | Obesity class II |
| Above 40 | Obesity class III |

Table 3: The weight categories

by obesity are as follows: premature death, cardiovascular diseases, high blood pressure, osteoarthritis, some cancers and diabetes.

Orginal data includes 500 adult people (255 of 500 are women) and four variables such as gender, height ($m$), weight ($kg$) and index (0: extremely weak, 1: weak, 2: normal, 3: overweight, 4: obesity and 5: extreme obesity). This data can be available in `https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex`. However, we assume a population that includes 255 women and their measurements such as height ($m$) and weight ($kg$) in our study. Note that we limited the population size as 255 to give sample observations. Thus, we aimed to illustrate the application, clearly. Also, it is supposed that the proportion of women in the Obesity class III is close to 0.5, $1 - F(40) \approx 0.5$. Therefore, using PRSS with $p = 0.1$ is appropriate in this case. From this population, $n = 100$ observations are selected using PRSS with $p = 0.1$. To obtain PRSS, we take the set size and the number of cycles as $k = 5$ and $l = 20$, respectively.

In the process PRSS, 25 observations are first selected at random among 255 women in $jth$ cycle, $j = 1, \cdots, 20$. Then, the 25 observations are assigned

into 5 sets at random. Ranking the BMI of the 25 observations may be performed by subjective ranking or according to a concomitant variable such as height of the observations. Also, it is assumed that ranking is almost perfect. The ranked sets are given as follows. In the sets, bold faced units represent the measured BMIs of

| Set | Ranked Units | Measured Units |
|---|---|---|
| $S_1$ | $\boldsymbol{X_{1[1]j}} \leq X_{1[2]j} \leq X_{1[3]j} \leq X_{1[4]j} \leq X_{1[5]j}$ | $X_{1[1]j}$ |
| $S_2$ | $\boldsymbol{X_{2[1]j}} \leq X_{2[2]j} \leq X_{2[3]j} \leq X_{2[4]j} \leq X_{2[5]j}$ | $X_{2[1]j}$ |
| $S_3$ | $X_{3[1]j} \leq X_{3[2]j} \leq \boldsymbol{X_{3[3]j}} \leq X_{3[4]j} \leq X_{3[5]j}$ | $X_{3[3]j}$ |
| $S_4$ | $X_{4[1]j} \leq X_{4[2]j} \leq X_{4[3]j} \leq X_{4[4]j} \leq \boldsymbol{X_{4[5]j}}$ | $X_{4[5]j}$ |
| $S_5$ | $X_{5[1]j} \leq X_{5[2]j} \leq X_{5[3]j} \leq X_{5[4]j} \leq \boldsymbol{X_{5[5]j}}$ | $X_{5[5]j}$ |

Table 4: Selected units in PRSS for $jth$ cycle, $j = 1, \cdots, 20$

5 observations among 25 observations. For the first cycle, the measured BMIs are $X_{1[1]1} = 18.52$, $X_{2[1]1} = 12.75$, $X_{3[3]1} = 32.45$, $X_{4[5]1} = 52.89$ and $X_{4[5]1} = 66.66$. These BMIs are given in the first row of Table 5. $1 - \hat{F}_{PRSS_O}(40) = 0.41$ is obtained according to the sample. Also, 95% confidence interval of $1-F(40) \approx 0.5$ is $(0.35, 0.46)$.

## 6. CONCLUSION

In this study, PRSS procedure is considered to estimate the distribution function. Properties of the EDF using PRSS are investigated. We examined how well the estimator performs in comparison with its SRS and RSS counterparts. Finally, we can summarize the following remarks:

1. Whether the ranking is perfect or not, the EDFs based on PRSS are unbiased as $F(x)$ gets closer to 1, 0.5 and 0.
2. Compared with $\hat{F}_{SRS}(x)$, the EDFs based on PRSS are more efficient under perfect and imperfect ranking.
3. If there is a known prior information that the value of $F(x)$ gets closer to 0.1, PRSS with $p = 0.4$ can be preferred instead of RSS whether the ranking is perfect or not.
4. Also, PRSS with $p = 0.1$ can be preferred instead of RSS when $F(x)$ is close to 0.5.
5. As in our application for BMI data, PRSS with $p = 0.1$ is recommended when estimating for the center of the distribution.
6. Also, it is suggested to use PRSS with $p = 0.4$ when estimating the extremes of the distribution.
7. In many studies on EDF estimators based on RSS and its modifications, theoretical results are presented for perfect ranking case while empirical

results are presented for imperfect ranking case. Empirical results are obtained by running Monte Carlo simulations in the studies. Unlike the other studies, the present paper shows that the proposed EDF estimator can be examined theoretically by using Frey [7]'s ranking error model even in the case of imperfect ranking.

As a future work, the moment-based (MB) and maximum likelihood (ML) estimators of the CDF can be considered. A comparable study of the MB, ML and the EDF estimators based on PRSS can be meaningful. The authors continue to work towards this goal.

## ACKNOWLEDGMENTS

The authors thank to Professor Omer Ozturk for providing the R function that is used in computation of ranking error probabilities. Also, the authors are grateful to the referees and the editor for helpful comments.

## APPENDIX

### Proof of Theorem 1

To prove this theorem, we follow the Proof of Lemma 2.1 in Samawi and Al-Sagheer[19] and the Proof of Theorem 1 in Stokes and Sager[23].

(1) Units in $PRSS_O$ are sampled from specific groups. It is assumed that $t_r = \frac{(k-1)l}{2}$ observations comes from $f_{(r)}(x)$, $t_s = \frac{(k-1)l}{2}$ from $f_{(s)}(x)$ and $t_m = l$ from $f_{(m)}(x)$, where $f_{(m)}(x)$ is density function of $m$th order statistic. Note that $t_1 = \cdots = t_{r-1} = 0$, $t_{r+1} = \cdots = t_{m-1} = 0$, $t_{m+1} = \cdots = t_{s-1} = 0$ and $t_{s+1} = \cdots = t_k = 0$. This is accomplished by first randomly select $R$ from $1, \cdots, k$ with replacement and if $r = 1$, $r = k$ or $r = m$ then observe $Y$ according to $F_r(x)$, otherwise reject $r$. In SRS the order in which the groups are sampled is random, by rearranging and relabeling, a realization $(y_1, \cdots, y_{kl})$ of $(Y_1, \cdots, Y_{kl})$ becomes $(Z_{r1}, \cdots, Z_{r\frac{(k-1)l}{2}}, Z_{m1}, \cdots, Z_{ml}, Z_{s1}, \cdots, Z_{s\frac{(k-1)l}{2}})$ the groups $\{Z_{ij}, Z_{mj'}; i = r, s; j = 1, \cdots, \frac{(k-1)l}{2}; j' = 1, \cdots, l\}$. It is necessary to specify a consistent order for the units of the $PRSS_O$ and SRS to compare their distributions

logically. Otherwise, because of the arbitrariness of listing order, a coordinate wise of PDF's or CDF's between $PRSS_O$ and SRS might imply unequal distributions, although the only difference would be a permutation of coordinates. Given

$$\underline{T} = \Bigg(0, \cdots, 0, t_r = \frac{(k-1)l}{2}, 0, \cdots, 0, t_m = l, 0 \cdots, 0, \\ t_s = \frac{(k-1)l}{2}, 0, \cdots, 0\Bigg)$$

and $P(\underline{T} = t_i) = \frac{1}{k}, i = 1, \cdots, k$ then, there are $\frac{(kl)!}{t_r! \cdots t_m! \cdots t_s!} = \frac{(kl)!}{\left(\left(\frac{(k-1)l}{2}\right)!\right)^2 l!}$ rearrangements of $Y$ yielding the same $Z$. So the conditional CDF of $Y$ given $\underline{T} = \underline{t}$ is

$$\frac{1}{P(\underline{T} = \underline{t})} P\Big\{Z_{r1} \leq a_{r1}, \cdots, Z_{r\frac{(k-1)l}{2}} \leq a_{r\frac{(k-1)l}{2}}, \\ Z_{m1} \leq a_{m1}, \cdots, Z_{ml} \leq a_{ml}, \\ Z_{s1} \leq a_{s1}, \cdots, Z_{s\frac{(k-1)l}{2}} \leq a_{s\frac{(k-1)l}{2}}; \underline{T}\Big\}$$

$$= \frac{1}{\frac{(kl)!}{(t_r! \cdots t_m! \cdots t_s!)} \left(\frac{1}{k}\right)^{t_r} \cdots \left(\frac{1}{k}\right)^{t_m} \cdots \left(\frac{1}{k}\right)^{t_s}} \times$$

$$\sum \Bigg[ \prod_{i=1}^{\frac{(k-1)l}{2}} \left(F_{(r)}(a_{ri}) \times \frac{1}{k}\right) \left(F_{(s)}(a_{si}) \times \frac{1}{k}\right) \times \\ \prod_{i'=1}^{l} \left(F_{(m)}(a_{mi'}) \times \frac{1}{k}\right) \Bigg]$$

where the sum is over all rearrangements of $Y$ consistent with $\underline{T} = \underline{t}$. So

$$\sum \frac{\prod_{i=1}^{\frac{(k-1)l}{2}} \left(F_{(r)}(a_{ri}) \times F_{(s)}(a_{si})\right) \prod_{i'=1}^{l} \left(F_{(m)}(a_{mi'})\right)}{\frac{(kl)!}{\left(\left(\frac{(k-1)l}{2}\right)!\right)^2 l!}}$$

$$= \prod_{i=1}^{\frac{(k-1)l}{2}} \left(F_{(r)}(a_{ri}) \times F_{(s)}(a_{si})\right) \prod_{i'=1}^{l} \left(F_{(m)}(a_{mi'})\right)$$

(2) It is assumed that $t_r = \frac{lk}{2}$ observations come from $f_{(r)}(x)$ and $t_s = \frac{lk}{2}$ from $f_{(s)}(x)$, where $f_{(r)}(x)$ and $f_{(s)}(x)$ are density functions of $r$th and $s$th order statistics, respectively. This proof follows from the part (1).

**Proof of Proposition 1**

(*a*) For $\hat{F}_{PRSS_O}(x)$,

*i*.

$$E\left(\hat{F}_{PRSS_O}(x)\right) = \frac{1}{lk}\left[\sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} E\left(I\left(X_{i(r)j} \leq x\right)\right) + \sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} E\left(I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)\right) + \sum_{j=1}^{l} E\left(I\left(X_{m(m)j} \leq x\right)\right)\right]$$

$I\left(X_{i(r)j} \leq x\right)$, $I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)$ and $I\left(X_{m(m)j} \leq x\right)$ have bernoulli distributions with parameters $F_{(r)}(x)$, $F_{(s)}(x)$ and $F_{(m)}(x)$, respectively. Therefore,

$E\left(I\left(X_{i(r)j} \leq x\right)\right) = F_{(r)}(x)$,

$E\left(I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)\right) = F_{(s)}(x)$ and

$E\left(I\left(X_{m(m)j} \leq x\right)\right) = F_{(m)}(x)$.

Thus,
$E\left(\hat{F}_{PRSS_O}(x)\right) = \left(\frac{1}{2} - \frac{1}{2k}\right)\left(F_{(r)}(x) + F_{(s)}(x)\right) + \frac{1}{k}F_{(m)}(x)$.

*ii*.

$$V\left(\hat{F}_{PRSS_O}(x)\right) = \frac{1}{lk}\left[\sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} V\left(I\left(X_{i(r)j} \leq x\right)\right) + \sum_{j=1}^{l}\sum_{i=1}^{\frac{k-1}{2}} V\left(I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)\right) + \sum_{j=1}^{l} V\left(I\left(X_{m(m)j} \leq x\right)\right)\right]$$

Since $I\left(X_{i(r)j} \leq x\right)$, $I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)$ and $I\left(X_{m(m)j} \leq x\right)$ have bernoulli distribution, variance of these indicator functions are given

bellow,

$$V\left(I\left(X_{i(r)j} \leq x\right)\right) = F_{(r)}(x)\left(1 - F_{(r)}(x)\right)$$

$$V(I(X_{\frac{k+1}{2}+i(s)j} \leq x)) = F_{(s)}(x)(1 - F_{(s)}(x))$$

$$V\left(I\left(X_{m(m)j} \leq x\right)\right) = F_{(m)}(x)\left(1 - F_{(m)}(x)\right).$$

Thus, variance of the estimator can be obtained

$$\begin{aligned} V\left(\hat{F}_{PRSS_O}(x)\right) = \frac{1}{lk^2}\Bigg[&\left(\frac{k-1}{2}\right)F_{(r)}(x)\left(1 - F_{(r)}(x)\right) \\ &+ \left(\frac{k-1}{2}\right)F_{(s)}(x)\left(1 - F_{(s)}(x)\right) \\ &+ F_{(m)}(x)\left(1 - F_{(m)}(x)\right)\Bigg] \end{aligned}$$

(*b*) $E\left(\hat{F}_{PRSS_E}(x)\right)$ and $V\left(\hat{F}_{PRSS_E}(x)\right)$ can be proved by using the same steps in Proof (a).

---

**Proof of Proposition 2**

---

Following Samawi and Al-Sagheer[19] and Kim et al.[12],

(*a*) Let $Z_j = \frac{1}{k}\left[\sum_{i=1}^{\frac{k-1}{2}}\left(I(X_{i(r)j} \leq x) + I\left(X_{\frac{k+1}{2}+i(s)j} \leq x\right)\right) + I(X_{m(m)j} \leq x)\right]$, $j = 1, \cdots, l$. Since $Z_j$ are independent and identically with finite mean and variance, then based on Central Limit Theorem

$$\left(\frac{\bar{Z} - E(Z_j)}{\left(\frac{var(Z_j)}{l}\right)^{1/2}}\right) \xrightarrow{D} N(0,1)$$

(*b*) Similarly, this part can be proved by assuming
$Z_j = \frac{1}{k}\sum_{i=1}^{\frac{k}{2}}\left(I(X_{i(r)j} \leq x) + I\left(X_{\frac{k}{2}+i(s)j} \leq x\right)\right)$

**Percentile ranked set sample**

| | PRSS | | | | |
|---|---|---|---|---|---|
| $l$ | $1^{st}$ | $1^{st}$ | $5^{th}$ | $5^{th}$ | $3^{rd}$ |
| 1 | 18.52 | 12.75 | 52.89 | 66.66 | 32.45 |
| 2 | 23.59 | 28.20 | 43.17 | 53.01 | 37.57 |
| 3 | 12.75 | 20.90 | 66.66 | 40.75 | 39.21 |
| 4 | 21.37 | 20.96 | 43.11 | 68.96 | 30.48 |
| 5 | 16.38 | 28.07 | 57.96 | 57.70 | 32.42 |
| 6 | 29.17 | 30.64 | 52.89 | 66.66 | 28.67 |
| 7 | 20.90 | 22.65 | 67.06 | 52.89 | 35.58 |
| 8 | 25.98 | 17.43 | 43.56 | 57.70 | 32.42 |
| 9 | 22.63 | 33.96 | 44.63 | 71.93 | 32.15 |
| 10 | 24.12 | 22.45 | 57.96 | 54.86 | 30.42 |
| 11 | 20.02 | 28.07 | 48.15 | 59.49 | 33.77 |
| 12 | 17.43 | 27.35 | 68.41 | 59.69 | 35.58 |
| 13 | 18.34 | 16.04 | 51.17 | 55.66 | 32.42 |
| 14 | 24.12 | 25.46 | 44.90 | 53.01 | 44.79 |
| 15 | 28.07 | 20.52 | 32.69 | 59.94 | 39.44 |
| 16 | 12.75 | 35.29 | 67.94 | 78.85 | 49.34 |
| 17 | 33.88 | 17.09 | 59.84 | 71.93 | 43.56 |
| 18 | 21.37 | 20.52 | 39.06 | 78.85 | 52.80 |
| 19 | 22.10 | 32.15 | 43.17 | 52.26 | 39.68 |
| 20 | 23.23 | 26.40 | 36.95 | 63.38 | 23.59 |

Table 5: Sample observations that are obtained using PRSS

## REFERENCES


[1] Abu-Dayyeh, W. A., Samawi, H. M. and Bani-Hani L. A. (2002). On distribution function estimation using double ranked set samples with application, *Journal of Modern Applied Statistical Methods*, **1**, 2, 443–451.

[2] Al-Omari, A. I. (2016). Quartile ranked set sampling for estimating the distribution function. *Journal of the Egyptian Mathematical Society*, **24**, 2, 303–308.

[3] Al-Omari, A. I. and Bouza, C. N. (2014) Review of ranked set sampling: modifications and applications. *Revista Investigación Operacional*, **35**, 3, 215–240.

[4] Al-Subh, S. A., Alodat, M. T., Ibrahim, K. and Jemain, A. A. (2009) EDF Goodness of Fit Tests of Logistic Distribution Under Selective Order Statistics. *Pakistan Journal of Statistics*, **25**, 3, 265–274.

[5] Bohn, L. L. and Wolfe, D. A. (1994) The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association*, **89**, 425, 168–176.

[6] Dell, T. R. and Clutter, J. L. (1972) Ranked set sampling theory with order statistics background. *Biometrics*, **28**, 2, 545–555.

[7] Frey, J. (2007a) New imperfect ranking models for ranked set sampling. *Journal of Statistical Planning and Inference*, **137**, 4, 1433–1445.

[8] Frey, J. (2007b) Distribution-free statistical intervals via ranked set sampling. *The Canadian Journal of Statistics*, **35**, 4, 585–596.

[9] Jemain, A. A. and Al-Omari, A. I. (2006) Double percentile ranked set samples for estimating the population mean. *Advances and Applications in Statistics*, **6**, 3, 261–276.

[10] Jemain, A. A. and Al-Omari, A. I. (2007) Multistage percentile ranked set samples. *Advances and Applications in Statistics*, **7**, 1, 127–139.

[11] Kaur, A., Patil, G. P., Sinha, A. K. and Taillie, C. (1995) Ranked set sampling: an annotated bibliography. *Environmental and Ecological Statistics*, **2**, 25–54.

[12] Kim, D. H., Kim, D. W. and Kim, G. H. (2005) On the estimation of the distribution function using extreme median ranked set sampling. *Journal of the Korean Data Analysis Society*, **7** 429–439.

[13] McIntyre, G. A. (1952) A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, **3**, 4, 385–390.

[14] Muttlak, H. A. (2003) Modified ranked set sampling methods. *Pakistan Journal of Statistics*, **19**, 315–323.

[15] Neerchal, N. K., Sinha, B. K. and Lacayo, H. (1998) Ranked set sampling from a dichotomous population. *Journal of Applied Statistical Science*, **11**, 1, 83–90.

[16] Nazari, S., Jafari Jozani, M. and Kharrati-Kopaei, M. (2016) On distribution function estimation with partially rank-ordered set samples: estimating mercury level in fish using length frequency data. *Statistics*, **50**, 6, 1387–1410.

[17] Ozturk, O. (2008) Inference in the presence of ranking error in ranked set sampling. *The Canadian Journal of Statistics*, **36**, 4, 557–594.

[18] Presnell, B. and Bohn, L. L. (1999) U-Statistics and imperfect ranking in ranked set sampling. *Nonparametric Statistics*, **10**, 2, 111–126.

[19] Samawi, H. M. and Al-Sagheer, A. M. (2001) On the Estimation of the Distribution Function Using Extreme and Median Ranked Set Sampling. *Biometrical Journal*, **43**, 3, 357–373.

[20] Sevil, Y. C. and Yildiz, T. O. (2017) Power comparison of the Kolmogorov–Smirnov test under ranked set sampling and simple random sampling. *Journal of Statistical Computation and Simulation*, **87**, 11, 2175–2185.

[21] Stokes, S. L. (1980a) Inferences on the correlation coefficient in bivariate normal populations from ranked set sampling. *Journal of the American Statistical Association*, **75**, 372, 989–995.

[22] Stokes, S. L. (1980b) Estimation of variance using judgment ordered ranked set samples. *Biometrics*, **36**, 1, 35–42.

[23] Stokes, S. L. and Sager, T. W. (1988) Characterization of a ranked-set sample with application to estimating distribution function. *Journal of the American Statistical Association*, **83**, 402, 374–381.

[24] Takahasi, K. and Wakimoto, K. (1968) On unbiased estimates of the population mean based on sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, **20**, 1–31.

[25] Yildiz, T. O. and Sevil, Y. C. (2018) Performances of some goodness-of-fit tests for sampling designs in ranked set sampling. *Journal of Statistical Computation and Simulation*, **88**, 9, 1702–1716.

[26] Yildiz, T. O. and Sevil, Y. C. (2019) Empirical distribution function estimators based on sampling designs in a finite population using single auxiliary variable. *Journal of Applied Statistics*, **46**, 16, 2962–2974.

[27] Zamanzade, E. (2019) EDF-based tests of exponentiality in pair ranked set sampling. *Statistical Papers*, **60**, 6, 2141–2159.

[28] Zamanzade, E. and Mahdizadeh, M (2020) Using ranked set sampling with extreme ranks in estimating the the population proportion. *Statistical Methods in Medical Research*, **29**, 1, 165–177.

[29] Zamanzade, E. and Wang, X. (2018) Proportion estimation in ranked set sampling in the presence of tie information. *Computational Statistics*, **33**, 3, 1349–1366.