Pak. J. Statist.
2024 Vol. 40(2), 225-242

AN IMPROVED QUALITATIVE RANDOMIZED RESPONSE MODEL UNDER RANKED SET SAMPLING

Azhar Mehmood Abbasi[§] and **Amber Asghar**
Department of Statistics, Virtual University of Pakistan
Lahore, Pakistan
[§] Corresponding author Email:azhar.mehmood@vu.edu.pk

ABSTRACT

This study introduces a new qualitative randomized response technique for gathering trust-worthy sensitive data using a perfect (imperfect) ranked set sampling design. This model takes into account the sizes (weights) of various balls on the randomizing device to choose one of the two questions (out of which one is sensitive); so the respondent feels, liberty in choosing balls. In this way, the likelihood of an honest response increases. It is established that the procedure is easy to apply and the estimate under the proposed model is more precise than the original Warner's randomized response model. In addition, an application to real medical data is considered. Finally, a cost analysis of the proposed model is also presented.

KEYWORDS

Randomized response; ranked set sampling; imperfect ranked set sampling; percentage relative efficiency.

Subject classifications: 62D05; 62F10

1. INTRODUCTION

In survey sampling, gathering of reliable data becomes problematic when it pertains to stigmatizing characteristics (e.g., questions about private matters). In such surveys, some respondents may refuse to respond or provide falsified answers. Thus, a direct inquiry of the respondent fails to collect reliable data relating to sensitive topics. To this end, Warner (1965) launched a randomized response technique (RRT) model using a simple random sampling (SRS) design. The novelty of this model is that it fully protects the respondent's privacy. This model involves a probabilistic approach to asking sensitive questioning. An interviewee uses the randomizing device to select one of the two questions (one of which is sensitive) that is to be answered by “Yes” or “No”. The interviewer, standing in front of the respondent, does not know the result of a randomizing device; hence the interviewee can honestly respond without privacy concerns. As a parameter of the randomizing device is known to the interviewer, the response gives some information relating to stigmatizing question.

After development of the first randomized response model, several variants have been proposed by different researchers to obtain more reliable estimates of the population sensitive proportion, say π, by increasing respondent's degree of privacy. A detailed literature will not be demonstrated. However, some interesting theories established under SRS design can be seen, for example, in Horvitz et al. (1967), Greenberg et al. (1969), Kuk (1990) and the studies referred to therein. Much of the RRT literature have been proposed in SRS design. But little attention has been paid to address this problem in RSS design, which is superior to its SRS counterpart in estimating population parameters.

Ranked set sampling (RSS) was introduced by McIntyre (1952), as an efficient alternative to simple random sampling (SRS), for estimation of pasture and forage yields. However, nowadays, all standard statistical problems are being addressed in RSS design, see Zamanzade and Mahdizadeh (2017). The RSS employs ranking of the small sets of units visually or via a concomitant information before selecting a final sample for actual quantification. For example, suppose a medical researcher is interested in estimating the prevalence of human immunodeficiency virus (HIV) in the population; in other words, the population sensitive proportion having HIV positive. Before using the RRT model, a medical expert can easily rank (by visual inspection) a group of people from the population according to the severity of the disease. Moreover, this would be cheaper and require least amount of effort compared to, for instance, using a costly, but more reliable process such as Nucleic acid test (NAT) to determine viral load in the body. Lastly, a gynecologist might be able to order his patients based on the perceived likelihood of a particular sexual disease by their facial expressions or asking some medical non-sensitive questions. More detail and application of RSS design can be explored in the studies carried out by Terpstra (2004), Al-Nasser (2007),Al-Omari (2011), Haq et al. (2014), Zamanzade and Vock(2015), Abbasi and shad (2017), Zamanzade et al. (2020a), Mahdizadeh and Zamanzade(2021),Bhushan and Kumar (2022), Abbasi and Shad (2022), Bhushan and Kumar (2022a), Bhushan and Kumar (2022b) and Mahdizadeh and Zamanzade (2022).

Recently, Abbasi and Shad (2021) have presented a modified version of Warner's model using a concomitant-based RSS design. They showed that their estimator is unbiased, and its variance is less than that of Warner's model. In this study, we introduce a new RRT model under RSS design without using a concomitant variable, and taking into account a new chance device. The idea is that, sometimes, auxiliary information is not available and, hence, units are to be ranked on the basis of study variable. To increase confidence of the respondent, we use balls of the same color but different predetermined weights (in grams), ranging from a to b i.e., $[a, b]$, in a randomizing device (weight-machine) instead of using only two types of balls/playing cards, as in Warner (1965) model, for selection of sensitive(non-sensitive) question or statement. The details of the new RRT model is provided in the next section.

2. DESCRIPTION OF THE MODEL

The chance device used in the new model is displayed in Figure 1. A number of balls of different predetermined weights (grams), ranging from a to b, are placed in a bottle. To proceed, define a constant $c \in [a, b]$. The proportion of balls, say w, having weight $\leq c$ is known. A respondent selected by RRS design is asked to select one ball at random from the bottle by turning it upside down and put the selected ball on the weight-machine (randomizing device). If weight lies in the interval $[a, c]$, he/she is required to answer the sensitive question (e.g., do you have HIV positive), otherwise, he would respond to the non-sensitive question (e.g., I don't have HIV positive). The respondent will accordingly respond in either “Yes” or “No”, depending on whether he/she belongs to that group or not. An interviewer standing opposite to the respondent does not know what weight appeared in the screen of the device, and, therefore, do not know if the respondents have experienced the sensitive (non-sensitive) event in question. That is, the privacy of a respondent is fully protected.

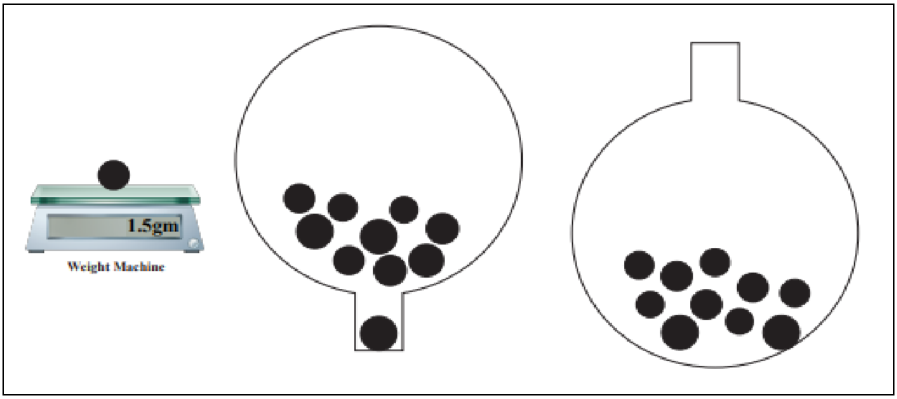

Figure 1: Proposed Randomizing Device

3. ESTIMATION OF THE PARAMETERS

To select a sample of size k using the RSS design, the experimenter begins by randomly choosing $k^2\,(k>1)$ units from the target population and then randomly arrange them into k small groups each of size k. Thereafter, for each set of size k individuals, an expert ranks (by visually or any cheapest way) the units from smallest to largest. For example, the ranking procedure could be founded on the perceived likelihood of the attribute of interest, as discussed in the introduction. Next, the subject with rank one is retained from the first set of k units, the subject with rank two is retained for the second set of individuals and so on, until the subject with rank k is determined for the last set of k units for actual quantification. Thus, this process yields a set of k independent observations. The whole process can be repeated n times to get a sample of size $m=nk$, say $\{y_{(i)ij};\ i=1,2,...,k;\ j=1,2,...,n\}$, where $Y_{(i)ij}$ denotes the *i*th order statistic of *i*th set obtained in *j*th cycle of size k units. In short, we can denote $Y_{(i)ij}$ by $Y_{(i)j}$. Note that, (.) denotes perfect ranking. The RRS method, for $k=3$ and $n=2$, is displayed in Table 1.

Table 1
A Layout of RSS for $k = 3$ and $n = 2$

Cycle(n)	**Set(m)**	**Set Units**			**Acquired Data**
	1	$\boxed{Y_{(1)11}}$	$Y_{(2)11}$	$Y_{(3)11}$	$Y_{(1)1} = 1$
1	2	$Y_{(1)21}$	$\boxed{Y_{(2)21}}$	$Y_{(3)21}$	$Y_{(2)1} = 1$
	3	$Y_{(1)31}$	$Y_{(2)31}$	$\boxed{Y_{(3)31}}$	$Y_{(3)1} = 1$
	1	$\boxed{Y_{(1)12}}$	$Y_{(2)12}$	$Y_{(3)12}$	$Y_{(1)2} = 1$
2	2	$Y_{(1)22}$	$\boxed{Y_{(2)22}}$	$Y_{(3)22}$	$Y_{(2)2} = 1$
	3	$Y_{(1)32}$	$Y_{(2)32}$	$\boxed{Y_{(3)32}}$	$Y_{(3)2} = 1$

Recall that in case of direct questioning when under study characteristic is non-sensitive, the probability of “Yes” response i.e., $p_{ki}(\pi)$ of the ith respondent is given by

$$p_{ki}(\pi) = \sum_{r=k-i+1}^{k} \binom{k}{r} \pi^r (1-\pi)^{k-r} \quad \text{for } i = 1,2,\ldots,k;\ 0 \le \pi \le 1$$

For details, see Arnold (1992a). As this study involves the RRT model for acquiring data for sensitive question, the respondents are instructed to choose one of the two above-mentioned statements using the given randomizing device. The respondent reports “Yes” (“No”) according to the outcomes of the randomizing device and his/her actual status. If w denotes the proportion (probability) that a chance device selects a sensitive question, then, $\lambda_{(i)}$, the probability of a “Yes” response of ith unit is modeled as

$$\lambda_{(i)} = w\, p_{ki}(\pi) + (1-w)\ (1\text{-}p_{ki}(\pi)) \tag{1}$$

Let $Y_{(i)1}, Y_{(i)2}, \ldots, Y_{(i)n}$ are n independent and identically distributed (IID) Bernoulli randomized responses out of which $z_{(i)} = \sum_{j=1}^{n} y_{(i)j}$ report “Yes” responses. Since the sampling process is Bernoulli with parameter $\lambda_{(i)}$, the likelihood function of $p_{ki}(\pi)$ for the given data $Y_{(i)j};\ j = 1,2,\ldots,n$ is

$$L = \lambda_{(i)}^{z_{(i)}} \left(1-\lambda_{(i)}\right)^{n-z_{(i)}} \quad \text{for } 1-w \le \lambda_{(i)} \le w,\ 0 < w < 1$$

In terms of $p_{ki}(\pi)$

$$L = \left[(2w-1)\, p_{ki}(\pi) + (1-w)\right]^{z_{(i)}} \left[w-(2w-1)\, p_{ki}(\pi)\right]^{n-z_{(i)}} \quad \text{for } 0 \le p_{ki}(\pi) \le 1$$

The log likelihood function gives

$$L = z_{(i)} \log\left[(2w-1)\, p_{ki}(\pi) + (1-w)\right] + (n-z_{(i)}) \log\left[w-(2w-1)\, p_{ki}(\pi)\right].$$

It is easy to obtain the estimate of $p_{ki}(\pi)$ from the equation $\partial \log L / \partial p_{ki}(\pi) = 0$, and after simple algebra, we have $\hat{p}_{ki}(\pi) = \left(\hat{\lambda}_{(i)} - (1-w)\right)/(2w-1),\ \hat{\lambda}_{(i)} = z_{(i)}/n$. Now, combining these individual proportions, using the relation $\pi = \sum_{i=1}^{k} p_{ki}(\pi)/k$, see Abbasi and shad (2021), the MLE of π, denoted by $\hat{\pi}_N$, is given by

$$\hat{\pi}_N = \left(\tfrac{1}{2w-1}\right)\left(w - 1 + \tfrac{1}{m}\sum_{i=1}^{k} z_{(i)}\right), \tag{2}$$

Since $z_{(i)}$ is binomial $(n, z_{(i)})$, therefore, $\hat{\pi}_N$ is an unbiased estimator and its associated variance is

$$\mathrm{Var}(\hat{\pi}_N) = \tfrac{1}{m}\left(k^{-1}\sum_{i=1}^{k} p_{ki}(\pi)\left(1 - p_{ki}(\pi)\right) + \tau\right), \tag{3}$$

$\tau = w(1-w)/(2w-1)^2$. Note that, as expected, the variance given in (3) is greater than that of the estimator suggested by Terpstra (2004a) due to indirect inquiry method.

Now, we consider the asymptotic distribution of $\hat{\pi}_N$ for later use in comparison of the estimates under RSS and SRS designs. Since $\hat{\lambda}_{(i)} = z_{(i)}/n$ is an estimate of Bernoulli proportion $\pi_{(i)}$, and for large n, $\hat{\lambda}_{(i)}$ will follow normal distribution. Moreover, the sum of the normal variates is also normal, $\sum_{i=1}^{k} z_{(i)}$ will also follow normal distribution. It is also well known that any linear combination of normal distribution is also normal, hence, the following result holds true for $\hat{\pi}_N$, being linear combination of $z_{(i)}$, when k is fixed and n is infinitely large.

$$\sqrt{nk}(\hat{\pi}_N - \pi) \xrightarrow{d} \text{Normal}\left(0,\ k^{-1}\sum_{i=1}^{k} p_{ki}(\pi)\left(1 - p_{ki}(\pi)\right) + \tau\right). \tag{4}$$

Note that, for $k = 1$, (4) simplifies to the conventional Warner's result, as given by

$$\sqrt{n}(\hat{\pi}_W - \pi) \xrightarrow{d} \text{Normal}\left(0,\ \pi(1-\pi) + \tau\right). \tag{5}$$

A closer look at the expression (4) shows some other interesting outcomes. For example, when $w = 1$ i.e., the likelihood of selection of a stigmatizing characteristic is 1 and the interviewee’s privacy goes to zero, (4) reduces to Terpstra (2004a) direct method of inquiry under RSS design. Whereas the choice $w = 0$ i.e., the likelihood of selection of a stigmatizing characteristic is zero, it also reduced (4) to Terpstra (2004a) procedure. Furthermore, if both w and k are equal to 1, (4) takes the usual method of direct inquiry under SRS design. For obtaining precise and reliable estimate, the conditions $k \geq 2$ and

$0<w<0.5$ are required. If $w=0.5$, the estimate fails. It is notable that the variance, in (4), is central symmetric i.e., it gives the same value in the parametric space $\{(\pi,w);(1-\pi,w);(\pi,1-w);(1-\pi,1-w)\}$.

3.1 An Application: HIV in Pakistan

This section continues the example from Section 1 in which a medical researcher is taking interest in estimating the prevalence of HIV-positive patients in a population. Following Abbasi (2021), a small scale survey is conducted to collect the primary data set of 150 patients from different hospitals located in Islamabad/Rawalpindi, Pakistan. In this survey, each patient was investigated about the attribute of interest using our randomizing device for $w=0.3$. The purpose of this exercise was to make known the population sensitive proportion π. That is 0.33. An interested reader can get the data from the corresponding author.

Assuming the above population data of 150 patients, we took ranked set samples of size $k = 2,3,4,5$ using R-package and computed the estimates, standard errors and confidence intervals (CI) of $\hat{\pi}_W$ and $\hat{\pi}_N$, for each value of k. For illustration, the method for selection of the patients and necessary calculations are manually replicated for $k = 4$ and $n = 25$. To this end, $k^2 = 16$ patients are selected by SRS method and arranged them into 4 sets each of size 4. Furthermore, the patient in each set are ranked (visually) and then ith patient is selected from the ith set $(i = 1,2,3,4)$ to estimate π. The same process is repeated 25 times to get a sample of size $m = 100$. Next, suppose both samples i.e., RSS and SRS have 30% HIV-positive patients. Note that, this permits us for a direct evaluation between the different estimators. Further, assuming $w = 0.3$, we obtained $(z_{(1)},z_{(2)},z_{(3)},z_{(4)})=(10,18,16,14)$. Thus,

$$\hat{\lambda}_N=\frac{z_{(1)}+z_{(2)}+\ldots,z_{(k)}}{m}=\frac{10+18+16+14}{100}=0.58$$

Therefore, $\hat{\pi}_N=0.30$. Likewise, other estimates, along with their standard errors and corresponding 95 percent confidence intervals, based on (4) and (5), are given in Table 2.

It is pertinent to highlight that we can also apply the bootstrap method to compute CIs for π. For example, to construct a 95% CI, choose the constants d_1 and d_2 such that the probability of $d_1<\hat{\pi}_h-\pi<d_2; h=W,N \text{ } is \text{ } 95\%$. To obtain the estimate of d_1 and d_2, one can proceed as follows. Select a ranked set sample of size k from the target population and compute $\hat{\pi}$ and repeat this process a number of times, say B, to get the estimates $\hat{\pi}_1,\hat{\pi}_2,\hat{\pi}_3,\ldots,\hat{\pi}_B$. Thereafter, estimate lower and upper population quantile points $B\left(\frac{\alpha}{2}\right)$ and $B\left(1-\frac{\alpha}{2}\right)$ respectively. For example, B =1000, sort the above estimates as $\hat{\pi}_{(1)},\hat{\pi}_{(2)},\hat{\pi}_{(3)},\ldots,\hat{\pi}_{(1000)}$ and use $\hat{\pi}_{(25)}$ and $\hat{\pi}_{(975)}$ as estimate of d_1 and d_2, respectively. Using this information, we can easily obtain a 95% CI for a particular sensitive proportion. For more details, see Taconeli (2022).

From Table 2, it can be seen that all intervals contained the true proportion of HIV cases i.e., $\pi = 0.33.$ As expected, the length of the confidence intervals (CI) based on the RSS design are narrow (shorter) than those based on the SRS design. That is, a substantial gain in precision has been obtained by adopting the RSS protocol.

Furthermore, for the RSS estimates, the standard errors are much smaller than those based on SRS for larger set sizes. These results are aligned with the RSS theory given in Dell (1972).This gain in precision obtained from RSS may be classified in terms of the number of units desired in both designs for obtaining the same estimation precision. For example, from Hettmansperger (1998), Terpstra (2004), the Percentage relative efficiency (PRE) between two estimation procedures can be interpreted as the ratio of sample sizes needed in order to obtain the same estimation precision i.e., $\text{PRE}(\hat{\pi}_N, \hat{\pi}_W) = K_1 / K_2$. Now, suppose we want to use an SRS and a 95 per cent confidence interval to estimate the proportion of HIV positive in the population. Assuming true sensitive proportion of HIV-positive cases as 0.33 and the desired margin of error (e) as 5 per cent. Then, classical sample size formula, as given below, suggests that $K_1 = 2356$ patients would need medical check-up.

Table 2
An Empirical Example Results

Set Size	Estimate	Standard Error	95 percent CI	Width Reduction
2	$\hat{\pi}_W = 0.30$	0.0951	(0.1136, 0.4864)	-
	$\hat{\pi}_N = 0.30$	0.0875	(0.1285, 0.4715)	8.00%
3	$\hat{\pi}_W = 0.30$	0.0776	(0.1479, 0.4521)	-
	$\hat{\pi}_N = 0.30$	0.0684	(0.1659, 0.4340)	11.87%
4	$\hat{\pi}_W = 0.30$	0.0672	(0.1682, 0.4317)	-
	$\hat{\pi}_N = 0.30$	0.0575	(0.1873, 0.4127)	14.46%
5	$\hat{\pi}_W = 0.30$	0.0614	(0.1822, 0.4178)	-
	$\hat{\pi}_N = 0.30$	0.0502	(0.2016, 0.3984)	16.46%

$$K_1 = \left(\frac{z_{\alpha/2}}{e}\right)^2 \left(\pi(1-\pi) + \tau\right),\ 0 < \pi < 1$$

On the other hand, suppose we decide to use RSS procedure with $k = 3$; Table 1 suggests that $\text{PRE}(\hat{\pi}_N, \hat{\pi}_W) = (0.0776 / 0.0684)^2 \times 100 = 128.7\%$. Then, for the RSS case, the experimenter would only require $K_2 = (1/1.287)2356 = 1831$ patients to be diagnosed. This gives a measurement savings of 525 units. Similarly, when $k = 5,\ K_2 = 1574$ i.e., measurement saving increases with an increase to k. Here, it is important to recall that a measurement is based on an expensive NAT test.

3.2 Relative Efficiency Comparison

In this section, we compare the estimators via Percentage relative efficiency (PRE), as given by

$$\text{PRE}(\hat{\pi}_N, \hat{\pi}_W) = \frac{\text{Var}(\hat{\pi}_W)}{\text{Var}(\hat{\pi}_N)} \times 100$$

Thus, an PRE larger than one implies that $\hat{\pi}_N$ is more efficient than $\hat{\pi}_W$. In Table 3, we have computed its values against different π and w for $k = 2,3,4,5$. Clearly, $\hat{\pi}_N$ is uniformly more efficient than $\hat{\pi}_W$. A closer look on the table, shows that PRE is an increasing function of k and π. Note also that the greatest gains in efficiency occur when π is close to 0.5. To more easily look into the trend of PRE values, the results are displayed in Figure 2 for different parametric values (π, w, k). From Figure 2, it can be seen that the PRE values have a decreasing trend when w approaches to 0.5. Here, it may be recalled that the variance of the proposed model is central symmetric as explained above in Section 3. That is, when $w > 0.5$, PRE values will again get increasing trend. It is also worthwhile that the PRE results given in Table 3 are smaller than those calculated by Terpstra (2004a) due to the RRT procedure.

Table 3
PRE ($\hat{\pi}_N, \hat{\pi}_W$) for Different Values of π, w and k

True Sensitive Population Proportion (π)	**Size of Ranked Set Sample (k)**	**Probability of Selecting Sensitive Question (w)**			
		0.1	**0.2**	**0.3**	**0.4**
0.1	2	103.6	101.5	100.6	100.1
	3	105.4	102.2	100.8	100.2
	4	109.6	103.9	101.4	100.3
	5	112.1	104.9	101.8	100.4
0.2	2	109.3	104.4	101.7	100.4
	3	116.7	107.6	103.0	100.7
	4	122.6	110.1	103.9	100.9
	5	127.5	112.0	104.6	101.1
0.3	2	114.4	107.2	102.9	100.7
	3	124.8	111.9	104.8	101.1
	4	132.7	115.2	106.0	101.4
	5	144.8	119.8	107.7	101.8
0.4	2	117.8	109.2	103.8	100.9
	3	130.0	114.7	105.9	101.4
	4	138.8	118.4	107.3	101.7
	5	146.0	121.2	108.3	102.0

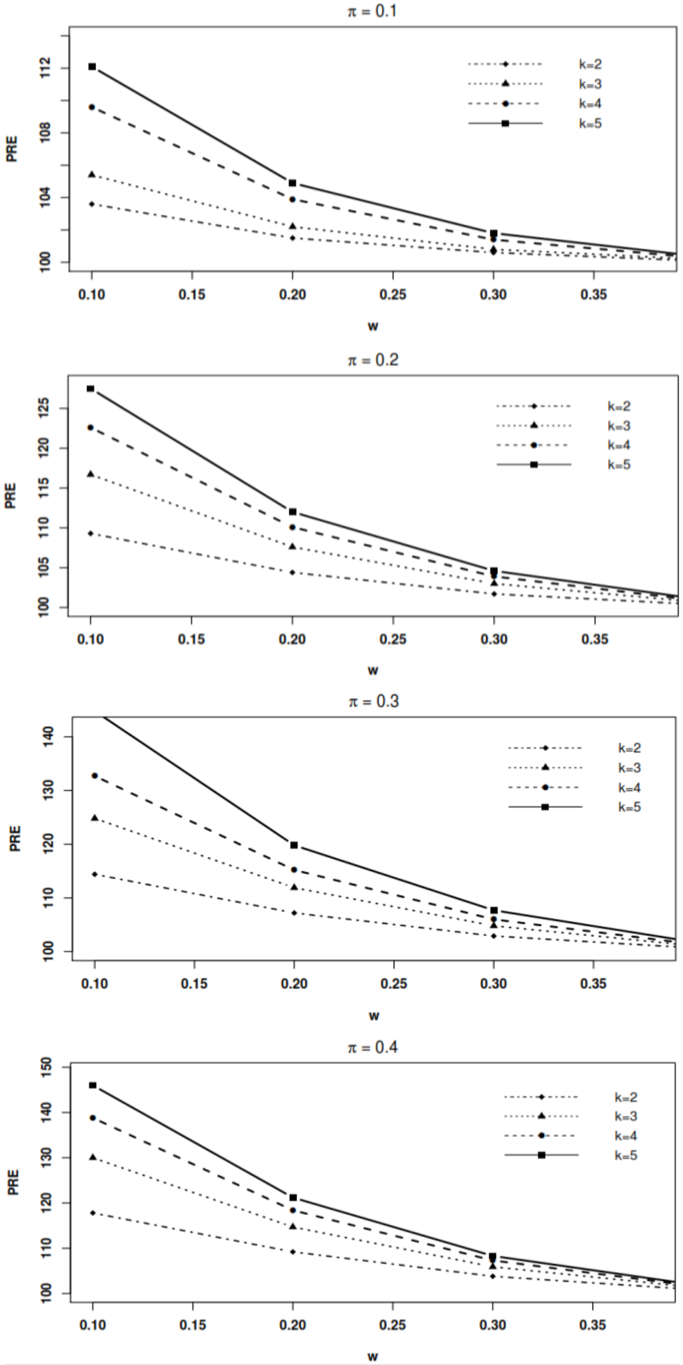

Figure 2: PRE $\hat{\pi}_N$ vs $\hat{\pi}_W$ for different π,w and k

4. SIMULATION STUDY

To validate the theoretical finding given in Table 3, we carried out a comprehensive simulation study. The simulated results, against different parameters such as $w = 0.1, 0.2, 0.3, 0.4$, $k = 2, 3, 4, 5$ and $n = 20$, 40, 80, 200 for $\pi = 0.1, 0.2, 0.3, 0.4$, are given in the Tables 4-7. For each combination (w, k, n, π), the results were simulated 10000 times. The PRE values given in Tables 4-7 are defined in Section 3 wherein we have used empirical variance instead of asymptotic variance. As expected, the results in Tables 4-7 are similar to those found in Table 3 or Figure 2. This validates our theoretical results.

Table 4
PRE ($\hat{\pi}_N, \hat{\pi}_W$) for different Values of w and k when $\pi = 0.1$

Number of Cycles (n)	**Size of Ranked Set Sample (k)**	**Probability of Selecting Sensitive Question (w)**			
		0.1	**0.2**	**0.3**	**0.4**
20	2	101.7	101.0	100.1	100.1
	3	103.6	101.4	101.0	100.5
	4	106.2	102.2	100.4	100.1
	5	109.3	103.3	101.2	100.1
40	2	101.6	101.4	101.0	100.3
	3	104.0	103.1	101.1	100.2
	4	104.8	102.6	100.7	100.3
	5	110.0	103.1	101.2	100.2
80	2	104.1	102.1	101.3	100.8
	3	103.4	102.3	100.2	100.1
	4	110.0	102.7	102.3	101.0
	5	113.4	103.7	102.4	101.2
200	2	102.8	102.2	101.0	100.5
	3	104.8	102.1	101.2	100.2
	4	107.9	104.3	101.2	100.0
	5	113.0	105.2	102.8	100.8

Table 5
PRE ($\hat{\pi}_N, \hat{\pi}_W$) for different Values of w and k when $\pi = 0.2$

Number of Cycles (n)	Size of Ranked Set Sample (k)	Probability of Selecting Sensitive Question (w)			
		0.1	**0.2**	**0.3**	**0.4**
20	2	110.4	102.8	103.1	100.8
	3	118.2	105.4	102.5	101.2
	4	124.0	111.1	102.4	101.2
	5	128.7	113.1	103.2	101.5
40	2	104.5	103.7	102.2	101.0
	3	115.8	106.8	104.3	101.0
	4	123.3	109.2	102.1	101.1
	5	128.2	110.8	103.9	101.3
80	2	110.3	103.7	102.4	100.2
	3	115.2	104.9	102.4	100.2
	4	120.0	111.1	104.3	101.3
	5	124.4	111.6	103.5	100.7
200	2	108.1	103.3	102.2	100.0
	3	117.0	105.7	102.3	101.0
	4	121.2	111.2	103.3	101.0
	5	128.3	110.1	103.9	100.4

Table 6
PRE ($\hat{\pi}_N, \hat{\pi}_W$) for different Values of w and k when $\pi = 0.3$

Number of Cycles (n)	Size of Ranked Set Sample (k)	Probability of Selecting Sensitive Question (w)			
		0.1	**0.2**	**0.3**	**0.4**
20	2	115.6	105.9	103.3	100.8
	3	125.1	112.2	105.0	100.3
	4	133.1	113.8	105.4	100.8
	5	145.5	121.0	108.1	102.0
40	2	113.2	108.0	103.0	101.0
	3	123.6	112.4	105.2	100.8
	4	133.5	116.3	107.2	102.1
	5	143.4	118.6	108.5	102.2
80	2	115.0	106.6	104.3	101.0
	3	124.4	110.6	105.1	100.2
	4	133.2	116.3	105.0	100.8
	5	143.9	118.9	106.7	102.7
200	2	113.3	106.6	103.2	101.0
	3	125.1	110.6	105.0	100.2
	4	130.5	114.4	107.1	102.1
	5	145.6	118.4	108.4	101.2

Table 7
PRE ($\hat{\pi}_N , \hat{\pi}_W$) for different Values of w and k when $\pi = 0.4$

Number of Cycles (n)	**Size of Ranked Set Sample** (k)	**Probability of Selecting Sensitive Question (w)**			
		0.1	**0.2**	**0.3**	**0.4**
20	2	118.4	110.0	102.9	101.5
	3	129.6	113.9	106.6	102.1
	4	137.6	117.6	106.7	102.2
	5	148.1	122.4	107.1	102.7
40	2	118.8	108.4	104.8	101.3
	3	131.1	115.0	104.0	100.8
	4	139.6	119.1	106.7	101.4
	5	147.0	120.8	107.4	101.6
80	2	118.1	110.0	102.8	101.1
	3	129.6	115.5	104.4	101.2
	4	139.7	117.8	106.6	102.2
	5	146.8	120.8	107.7	101.8
200	2	118.9	108.8	102.8	101.1
	3	129.4	115.4	104.8	102.1
	4	139.7	119.0	108.1	100.8
	5	147.4	121.2	107.9	101.7

5. IMPERFECT RANKING MODEL

In the preceding work, we assumed that the experimental units are perfectly ordered (ranked) according to the characteristics of interest. However, in real life problems, it is sometimes difficult to perfectly rank the units, and we have to rely on judgment ranking; that causes error in the ranking process. This section covers the effects of error in relation to PRE by demonstrating a RRT model for imperfect ranking situations. The method will be illustrated with the help of an example.

In the foregoing sections, $Y_{(i)}$ has a Bernoulli distribution with parameter $p_{ki}(\pi)$. However, for the case of errors in rankings, the value of Y against the *i*th judgment ranking, say $Y_{[i]}$, is not usually equal to $Y_{(i)}$. For details, see Terpstra (2004a). To further proceed with the study, we need to specify the distribution of $Y_{[i]}, \ i = 1, 2, \ldots, k$. From Terpstra (2004a), the *i*th order statistic and the event that it receives judgment rank *j* are independent. Thus, it follows that the distribution of $Y_{[i]}$ is a blend of k order Bernoulli variate with parameter, denoted by $p^*_{ki}(\pi)$, is given by

$$p^*_{ki}(\pi) = \sum_{l=1}^{k} t_{il}\, p_{kl}(\pi) \text{ for } i = 1, 2, \ldots, k$$

Let t_{il} is transition probability that ith order statistic is judged to be the lth $(l=1,2,\ldots,k)$ order statistics i.e., if $Y_{[l]}=Y_{(i)}$. If $T=[t_{il}]$ is a transition matrix, then $\sum_{l=1}^{k} t_{il}=1$; $i=1,2,\ldots,k$. Some examples of T, for $k = 2$, are:

$$\mathrm{T}_1=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \mathrm{T}_2=\begin{pmatrix}1/2 & 1/2\\ 1/3 & 2/3\end{pmatrix} \text{ and } \mathrm{T}_3=\begin{pmatrix}1/3 & 2/3\\ 2/3 & 1/3\end{pmatrix}$$

Using the new randomizing device, $\lambda^*_{[i]}$, the probability of “Yes” response is given by

$$\lambda^*_{[i]}=w p^*_{ki}(\pi)+(1-w)\left(1-p^*_{ki}(\pi)\right). \tag{6}$$

or

$$\hat{p}^*_{ki}(\pi)=(\hat{\lambda}^*_{[i]}-1+w)/(2w-1),$$

or

$$\sum_{l=1}^{k} t_{il}\,\hat{p}_{kl}(\pi)=\hat{\delta}_{(i)}, \tag{7}$$

$\hat{\delta}_{(i)}=(\hat{\lambda}^*_{[i]}-1+w)/(2w-1).$ In matrix notation, (7), can be written as

$$TP=\Delta, \tag{8}$$

where

$$T=\begin{pmatrix}t_{11} & t_{12} & \cdots & t_{1k}\\ t_{21} & t_{22} & \cdots & t_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ t_{k1} & t_{k2} & \cdots & t_{kk}\end{pmatrix}, \; P=\begin{pmatrix}\hat{p}_{k1}\\ \hat{p}_{k2}\\ \cdots\\ \hat{p}_{kk}\end{pmatrix} \text{ and } \Delta=\begin{pmatrix}\hat{\delta}_1\\ \hat{\delta}_2\\ \cdots\\ \hat{\delta}_k\end{pmatrix}$$

If $z_{[i]}=\sum_{j=1}^{n} Y_{[i]j}$ represents “Yes” reports out of n respondents, then values of P, from (8), can be obtained as

$$P=\begin{pmatrix}\hat{p}_{k1} & \hat{p}_{k2} & \cdots & \hat{p}_{kk}\end{pmatrix}'=T^{-1}\Delta=S\Delta,$$

where $S=T^{-1}$, provided that T is non-singular.

Since the $Y_{[i]j}; j=1,2,\ldots,n$ are IID and follow Bernoulli distribution and hence $\hat{\lambda}^*_{[i]}=z_{[i]}/n$ follow Binomial distribution with parameters $(n,\lambda^*_{[i]})$. Now, the dispersion matrix of Δ is $\mathrm{disp}(\Delta)=\mathrm{diag}(V_1,V_2,\ldots,V_k)$, $V_i=\lambda^*_{[i]}(1-\lambda^*_{[i]})/n(2w-1)^2$. Hence, the estimate of π is given by

$$\hat{\pi}=k^{-1}IP=k^{-1}IS\Delta, \tag{9}$$

where $\mathrm{I}=(11\ldots1)$ is a row matrix. It follows that dispersion of $\hat{\pi}$ is given by

$$\mathrm{disp}(\hat{\pi})=k^{-2}IS\,\mathrm{diag}(V_1,V_2,\ldots,V_k)S'I' \tag{10}$$

5.1 An Example

To illustrate and compare the imperfect ranking based RRT model with that of perfect ranking based model, let $\pi=0.4,\ w=0.1,\ k=2,\ n=20,\ \hat{\lambda}_1=8/20$ and $\hat{\lambda}_2=12/20$. Furthermore, if $T=T_3$, then $\mathrm{S}=\begin{pmatrix}-1 & 2\\ 2 & -1\end{pmatrix}$, $\Delta=\begin{pmatrix}0.625\\ 0.375\end{pmatrix}$ and $\mathrm{P}=\begin{pmatrix}0.125\\ 0.875\end{pmatrix}$

From the above information, we have, $\hat{\pi}_N=(0.125+0.875)/2=0.50$ and variance of $\hat{\pi}_N$ is 0.915%, and similarly the dispersion in Warner's estimator $\hat{\pi}_W$ is given by 0.951%. Hence, $\mathrm{PRE}(\hat{\pi}_N,\hat{\pi}_W)=(0.951/0.915)\times100$ is equal to 103.9%. As expected, this gain is less than that of under perfect ranking, see the value against $w=0.1,\ \pi=0.4$ and $k=2$ in Table 3. These results are consistent with the theory, given by Dell (1972); RSS based results are better than corresponding SRS results even under imperfect ranking situation. To further explore the behavior of imperfect RSS (IRSS) under other parametric values, we need to calculate estimated PRE.

To proceed with IRSS, we first assume transition matrices such $T_2,\ T_3,\ T_4$ and T_5 as

$$T_2=\begin{pmatrix}1/3 & 2/3\\ 2/3 & 1/3\end{pmatrix},\ T_3=\begin{pmatrix}1/3 & 1/3 & 1/3\\ 1/3 & 1/3 & 1/3\\ 1/3 & 1/3 & 1/3\end{pmatrix},\ T_4=\begin{pmatrix}2/6 & 1/6 & 2/6 & 1/6\\ 1/6 & 1/6 & 2/6 & 2/6\\ 3/6 & 1/6 & 1/6 & 1/6\\ 1/6 & 2/6 & 2/6 & 1/6\end{pmatrix}$$

and $$T_5=\begin{pmatrix}2/7 & 1/7 & 1/7 & 2/7 & 1/7\\ 1/7 & 1/7 & 2/7 & 1/7 & 2/7\\ 1/7 & 2/7 & 1/7 & 2/7 & 1/7\\ 3/7 & 1/7 & 1/7 & 0 & 2/7\\ 2/7 & 1/7 & 1/7 & 2/7 & 1/7\end{pmatrix}$$

The results of estimated PRE under IRSS are obtained on the same lines as we proceeded in the Section-3 except that we have adjusted transition matrices. The results so obtained are displayed in the Table 8. As expected, it can be observed that PRE values are less than those obtained in the Table 3. All these results are allied with Dell (1972) theory of imperfect ranked set sampling. That is, the results under IRSS are also better than their SRS counterparts.

Table 8
PRE ($\hat{\pi}_N, \hat{\pi}_W$) under IRSS for different Values of w, k and π

True Sensitive Population Proportion (π)	Size of Ranked Set sample (k)	Transition Matrix (T)	Probability of Selecting Sensitive Question (w)			
			0.1	0.2	0.3	0.4
0.1	2	T_2	101.3	100.7	100.2	100.1
	3	T_3	103.3	101.0	100.3	100.1
	4	T_4	104.2	100.9	100.4	100.1
	5	T_5	106.0	101.1	100.2	100.1
0.2	2	T_2	103.4	101.5	100.7	100.1
	3	T_3	105.4	102.5	100.4	100.2
	4	T_4	109.2	104.0	100.5	100.1
	5	T_5	112.2	104.6	102.4	100.1
0.3	2	T_2	107.5	102.3	100.7	100.3
	3	T_3	111.9	104.6	101.7	100.1
	4	T_4	119.4	107.7	101.8	101.4
	5	T_5	129.2	109.9	104.3	100.3
0.4	2	T_2	103.9	103.0	101.3	100.1
	3	T_3	120.3	104.4	101.0	101.2
	4	T_4	128.0	111.3	102.3	101.1
	5	T_5	136.0	110.2	101.2	100.1

6. COST ANALYSIS

In the previous sections, we have not taken into account the cost of ranking the units by supposing that it is cost-free. In fact, for the selection of an appropriate sampling design, we need to consider factors such as cost, time and accuracy or precision. Among these factors, cost of sampling units is, generally, the main focus of all sampling methods.

Following Dell (1972), we construct a cost model to evaluate the performance of the proposed estimator with respect to Warner's model. Let c_s be the cost of stratification to adhere to each quantified unit in RSS. Generally, this is the cost of choosing k-1 items and completing judgment ordering of the k units of a sample. Similarly, c_q denotes the cost of choosing and measuring a item without ranking. Now, the PRE is defined as the ratio of the variance of the estimator under SRS to that of under RSS by assuming that the total sampling cost, say C, is the same for both designs. Now, the PRE of $\hat{\pi}_N$ with respect to $\hat{\pi}_W$ is given by

$$\text{PRE}(\hat{\pi}_N, \hat{\pi}_W) = \frac{c_q}{c_q + c_s} \frac{\text{Var}(\hat{\pi}_W)}{\text{Var}(\hat{\pi}_N)} \times 100 \tag{11}$$

From (11), it appears that for fixed c_q, when c_s increases, the value of PRE decreases. The PRE attains it maximum value when $c_s = 0$. For example, when $c_q = 5$ and $c_s = 0.4$, the graph of (11) is given in Figure 3. From Figures 2-3, it can be observed that PRE decreases as $c_s > 0$. However, the proposed model still remains superior.

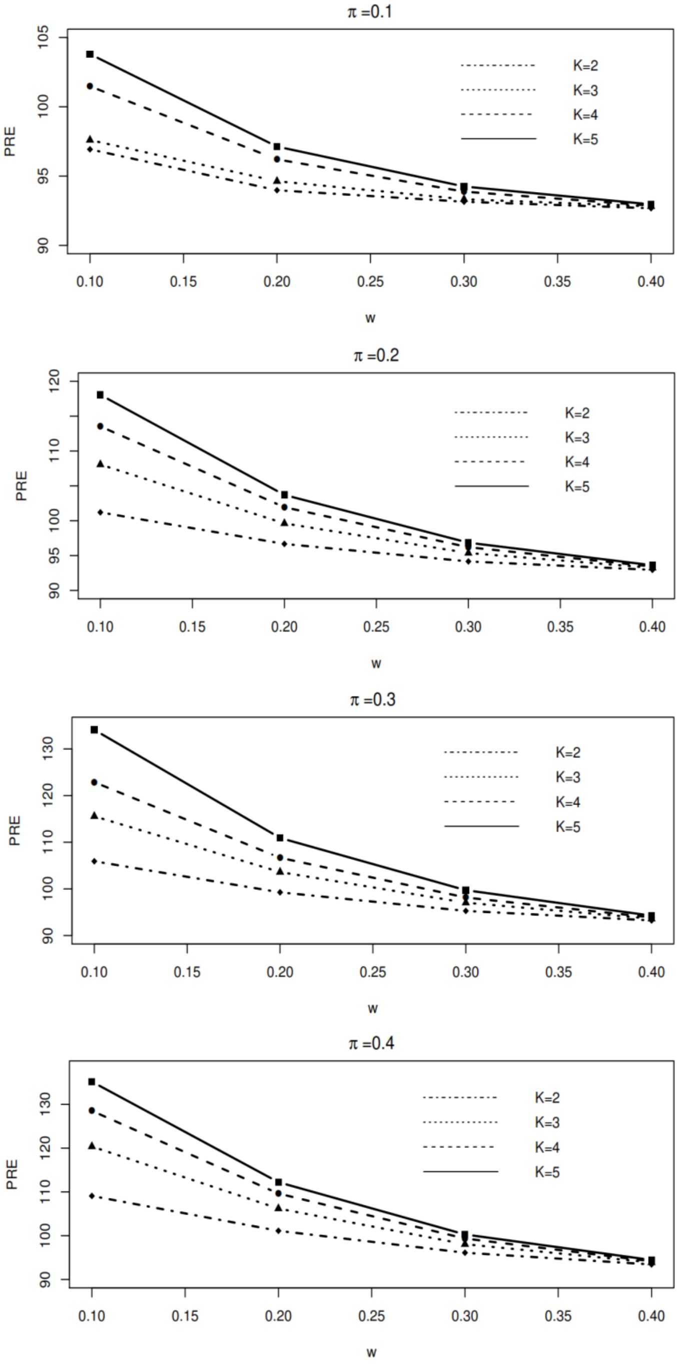

Figure 3: PRE $\hat{\pi}_N$ vs $\hat{\pi}_W$ for different π, *w* and *k* when $c_q = \$5$ and $c_s = \$0.4$

7. CONCLUSION

The paper has highlighted the importance of ranked-set sampling design in estimating the population-sensitive proportion. In addition, we have used a different randomizing device that takes into account the weight of balls on a chance device for selecting one of the two statements (out of which one is sensitive). Since balls of different sizes and weights are used, the respondent feels liberty in choosing balls from the device, ultimately, the probability of honest response increases.

Furthermore, it was shown that the proposed model asymptotically follows normal distribution. The confidence interval (CI) estimates, using real medical data, were derived for different k, and found that these are shorter or narrower than their usual counterparts. Moreover, the percentage relative efficiency (PRE) of the proposed model is better than the Warner's model. In addition, simplicity and feasibility are other key points of the model. The cost analysis also supports the new model.

This study can be extended by using other variants of RSS design. For example, RSS with tie information or judgment post-stratification.

REFERENCES

1. Abbasi, A.M. and Shad, M.Y. (2017). Manipulation-based ranked set sampling scheme. *Pakistan Journal of Statistics and Operation Research*, 775-798.
2. Abbasi, A.M. and Shad, M.Y. (2021). Sensitive proportion in ranked set sampling. *PloS one*, 16(8), e0256699.
3. Abbasi, A.M. and Shad, M.Y. (2022). Estimation of population proportion using concomitant based ranked set sampling. *Communications in Statistics-Theory and Methods*, 51(9), 2689-2709.
4. Al-Nasser, A.D. (2007). L ranked set sampling: A generalization procedure for robust visual sampling. *Communications in Statistics Simulation and Computation*, 36(1), 33-43.
5. Al-Omari, A.I. (2011). Estimation of mean based on modified robust extreme ranked set sampling. *Journal of Statistical Computation and Simulation*, 81(8), 1055-1066.
6. Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N. (2008). A first course in order statistics. *Society for Industrial and Applied Mathematics*, Vol. 54.
7. Bhushan, S. and Kumar, A. (2022a). Novel log type class of estimators under ranked set sampling. *Sankhya* B, 84(1), 421-447.
8. Bhushan, S. and Kumar, A. (2022b). On optimal classes of estimators under ranked set sampling. *Communications in Statistics-Theory and Methods*, 51(8), 2610-2639.
9. Bhushan, S., Kumar, A. and Lone, S.A. (2022). On some novel classes of estimators using ranked set sampling. *Alexandria Engineering Journal*, 61(7), 5465-5474.
10. Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Bio-metrics*, 545-555.
11. Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326), 520-539.
12. Haq, A., Brown, J., Moltchanova, E. and Al-Omari, A.I. (2014). Mixed ranked set sampling design. *Journal of Applied Statistics*, 41(10), 2141-2156.

13. Hettmansperger, T.P. and McKean, J.W. (2010). *Robust nonparametric statistical methods*. CRC Press.
14. Simmons, W.R., Horvitz, D.G. and Shah, B.V. (1967). The unrelated question randomized response model. Proceedings in the Social Statistics Section. *ASA*, 64, 520-539.
15. Kuk, A.Y. (1990). Asking sensitive questions indirectly. *Biometrika*, 77(2), 436-438.
16. Mahdizadeh, M. and Zamanzade, E. (2021). Smooth estimation of the area under the roc curve in multistage ranked set sampling. *Statistical Papers*, 62(4), 1753-1776.
17. Mahdizadeh, M. and Zamanzade, E. (2022). On estimating the area under the roc curve in ranked set sampling. *Statistical Methods in Medical Research*, 31(8), 1500-1514.
18. McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3(4), 385-390.
19. Taconeli, C.A. and Lara, I.A.R. de. (2022). Improved confidence intervals based on ranked set sampling designs within a parametric bootstrap approach. *Computational Statistics*, 37(5), 2267-2293.
20. Terpstra, J. (2004). On estimating a population proportion via ranked set sampling. *Biometrical Journal: Journal of Mathematical Methods in Bio Sciences*, 46(2), 264-272.
21. Terpstra, J.T. and Liudahl, L.A. (2004). Concomitant-based rank set sampling proportion estimates. *Statistics in Medicine*, 23(13), 2061-2070.
22. Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69.
23. Zamanzade, E. and Mahdizadeh, M. (2017). A more efficient proportion estimator in ranked set sampling. *Statistics & Probability Letters*, 129, 28-33.
24. Zamanzade, E., Mahdizadeh, M. and Samawi, H.M. (2020). Efficient estimation of cumulative distribution function using moving extreme ranked set sampling with application to reliability. *AStA Advances in Statistical Analysis*, 104(3), 485-502.
25. Zamanzade, E. and Vock, M. (2015). Variance estimation in ranked set sampling using a concomitant variable. *Statistics & Probability Letters*, 105, 1-5.