Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 1, 179–193
DOI: 10.1214/16-BJPS308

Ranked set sampling with scrambled response model to subsample non-respondents

Shakeel Ahmed and Javid Shabbir

Quaid-i-Azam University

Abstract. This paper considers use of the scrambled response model in Ranked Set Sampling (RSS) for collecting information on second call to estimate population mean when non-response is due to sensitivity of the study variable. It also uses Extreme Ranked Set Sampling (ERSS) and Median Ranked Set Sampling (MRSS) to sub-sample the non-respondents. Expressions for variances of different estimators are derived. A Monte Carlo experiment is carried out to observe the efficiency of proposed estimators.

1 Introduction

Non-response is a phenomenon in which complete information for estimation of population parameters is not easily obtainable from the selected sample. In surveys related to human beings, it may happen due to failure to contact the respondents, refusal or unable to answer questions asked by the interviewer. Consequently it estimates population parameters significantly too high or too low. To overcome this difficulty, Hansen and Hurwitz (1946) introduced a method for sub-sampling the non-respondent to estimate the finite population mean. After that many researchers extended this technique in different sampling schemes. When the study character is sensitive in nature then it is difficult to obtain a true response again on second call which result in violation of assumption made by Hansen and Hurwitz (1946) to estimate the finite population mean. In surveys related to sensitive characters, some statistical techniques exist that provide unbiased estimators for population parameters by protecting the confidentiality of respondents. These techniques are known as Randomized Response Techniques (RRT's). Warner (1965) used the RRT to estimate the proportion of persons in population possessing a sensitive character. Later on many authors have worked for improving efficiency of the estimator of population mean or proportion using different RRT's under different sampling designs. Diana and Perri (2011) used a general linear scrambled response model to propose an estimator for population mean of sensitive quantitative character. Later on Diana, Riaz and Shabbir (2014) used this model to sub-sample non-respondent by assuming that the person who do not respond on first call give a scrambled response on the second call. The estimator proposed by Diana, Riaz and Shabbir

Key words and phrases. Non-response, scrambled response, coefficient of variation, ranked set sampling.

Received June 2015; accepted January 2016.

(2014) gives greater privacy protection to respondents but some efficiency loses due to use of scrambled response model on second call.

Ranked set sampling (RSS) is a better alternative to simple random sampling that can offer large improvement in precision in some situations. It was originally developed for estimating the herbage yield in agriculture by McIntyre (1952). It is preferable when actual measurement of a unit is either expensive or time consuming and ranking of a small set of experimental units is cheap and easy. Dell and Clutter (1972) proved that even if ranking is not perfect the ranked set sampling is still unbiased. Many authors including Patil (2002), Muttlak (1996) and Samawi, Abu-Dayyeh and Ahmed (1996) showed that RSS is better than SRSWR in terms of accuracy. Bouza (2009) used RSS with Randomized response technique for estimating the population mean of sensitive quantitative variable to protect confidentiality of respondents. Bouza (2002) proposed unbiased estimator of population mean using RSS in presence of non-response. Bouza (2010) introduced an estimator for population mean using RSS to sub-sample the non-respondents on second call claiming that the first visit allows information on Y for ranking the units in the sub-sample $\acute{S}_2$ from non-response group S_2 and use different RSS methods for selecting sub-sample on second call. Motivating from Diana, Riaz and Shabbir (2014) work, we propose an estimator for finite population mean assuming that non-response is due to sensitivity of the study character and using RSS with randomized response model to sub-sample non-respondents on second call. Use of RSS improves efficiency and use of RRT improves confidentiality so in this way we can obtain these twin objectives simultaneously. Therefore, the proposed estimators perform better than Diana and Perri (2011) estimator, in terms of accuracy, and Bouza (2010) estimator in terms of confidentiality.

2 Estimation of mean in non-response

Let $U = (U_1, U_2, U_3, \dots, U_N)$ be a finite population of size N and y_i be the observed values of the study variable y on the ith unit. We select a sample of size n by using SRSWR. Now suppose that from n sampling units only n_1 units respond on first call and n_2 units don't respond. Consequently whole population divides into two groups G_1 (respondents) with N_1 units and G_2 (non-respondent) with N_2 units such that with $N = N_1 + N_2$. So we select a sub-sample of size $\acute{n}_2 = \frac{n_2}{k}$ $(k > 1)$ from n_2 non-responding units by using SRSWR. Hansen and Hurwitz (1946) estimator in SRSWR, is given by

$$\bar{y}^*_{\text{srs}} = w_1\bar{y}_1 + w_2\acute{\bar{y}}_2, \tag{1}$$

where $w_1 = \frac{n_1}{n}$, $w_1 = \frac{n_2}{n}$, $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ and $\acute{\bar{y}}_2 = \frac{1}{\acute{n}_2}\sum_{i=1}^{\acute{n}_2} y_i$. Also $E(\bar{y}^*_{\text{srs}}) = \mu$ and variance of $\bar{y}^*_{\text{srs}}$ is:

$$V(\bar{y}^*_{\text{srs}}) = \frac{1}{n}\sigma^2 + \frac{W_2(k-1)}{n}\sigma_2^2, \tag{2}$$

where $W_2 = \frac{N_2}{N}$ is fraction of non-respondent in population, σ^2 and σ_2^2 are population variances of the study character for whole population and population of non-respondents respectively. If non-response is due to sensitivity of the study character then it is difficult to obtain a truthful response again on second call. Diana, Riaz and Shabbir (2014) suggested an estimator for population mean using scrambled response model to overcome this deficiency.

Let Z be the scrambled response and A and B are two independent random variables unrelated to Y with known means (μ_A, μ_B) and variances (σ_A^2, σ_B^2), such that:

$$Z = AY + B, \tag{3}$$

where $E_R(Z) = \mu_A Y + \mu_B$ and variance of Z is $V_R(Z) = \sigma_A^2 Y^2 + \sigma_B^2$, here E_R, V_R are expectation and variance with respect to randomization device.

Let $\hat{y}_i$ be transformed scrambled response of the ith unit whose expectation under randomization mechanism equals to true response y_i, that is,

$$\begin{aligned} \hat{y}_i &= \frac{z_i - \mu_B}{\mu_A}, \qquad E_R(\hat{y}_i) = y_i, \\ V_R(\hat{y}_i) &= \frac{\sigma_A^2 y_i^2 + \sigma_B^2}{\mu_A^2}. \end{aligned} \tag{4}$$

Diana, Riaz and Shabbir (2014) proposed following estimator

$$\hat{\bar{y}}^*_{\text{srs}} = w_1 \bar{y}_1 + w_2 \acute{\hat{\bar{y}}}_2, \tag{5}$$

where $\acute{\hat{\bar{y}}}_2 = \frac{1}{\acute{n}_2} \sum_{i=1}^{\acute{n}_2} \hat{y}_i$. It is easy to show that $E(\hat{\bar{y}}^*_{\text{srs}}) = \mu$ using the fact that $E_R(\acute{\hat{y}}_2) = \acute{\bar{y}}_2$. The variance of $\hat{\bar{y}}^*_{\text{srs}}$, is given by

$$V(\hat{\bar{y}}^*_{\text{srs}}) = \frac{1}{n}\sigma^2 + \frac{W_2(k-1)}{n}\sigma_2^2 + \frac{W_2 k}{n}\left[\frac{\sigma_A^2}{\mu_A^2}\{\sigma_2^2 + \mu_2^2\} + \frac{\sigma_B^2}{\mu_A^2}\right]. \tag{6}$$

The estimator in (5) is better than (1) in terms of privacy protection but it is obvious from (2) and (6) that former is less efficient than the later. So our objective is to increase efficiency of the estimator. For this purpose, we use the RSS scheme which gives more efficient result than SRSWR. Bouza (2010) introduced a procedure for selecting sub-sample $\acute{S}_{2(\text{rss})}$ of size $\acute{n}_2$ from S_2 group having size n_2 who don't respond at first call by using RSS. The procedure consist of selecting $\acute{n}_2$ sub-samples by using SRSWR. The units are ranked accordingly with the variable closely related with variable of interest Y. We have $\acute{n}_2$ independent random samples

$$Y_{11}, Y_{12}, \ldots, Y_{1\acute{n}_2}; Y_{21}, Y_{22}, \ldots, Y_{2\acute{n}_2}; \ldots; Y_{\acute{n}_2 1}, Y_{\acute{n}_2 2}, \ldots, Y_{\acute{n}_2\acute{n}_2}.$$

After ranking, we get

$$Y_{(1:1)}, Y_{(1:2)}, \ldots, Y_{(1:\acute{n}_2)}; Y_{(2:1)}, Y_{(2:2)}, \ldots, Y_{(2:\acute{n}_2)}; \ldots; Y_{(\acute{n}_2:1)}, Y_{(\acute{n}_2:2)}, \ldots, Y_{(\acute{n}_2:\acute{n}_2)},$$

where $Y_{(j:t)}$ is the jth order statistics (OS) of the tth sample, $j = 1, 2, \ldots, \acute{n}_2$ and $t = 1, 2, \ldots, \acute{n}_2$. We obtain the following sample,

$$Y_{(1:1)}, Y_{(2:2)}, \ldots, Y_{(\acute{n}_2:\acute{n}_2)}.$$

The estimate of μ_2 is made by using the estimator $\acute{\bar{y}}_{2(\text{rss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2} Y_{(j:j)}$.

Also $E(Y_{(j:j)}|n_2) = \mu_{(j)}$ where $j = 1, 2, \ldots, \acute{n}_2$. Now the estimator introduced by Bouza (2010) using Hansen and Hurwitz (1946) technique, is given by

$$\bar{y}^*_{\text{rss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\acute{\bar{y}}_{2(\text{rss})}, \tag{7}$$

with $E(\bar{y}^*_{\text{rss}}) = \mu$ and variance

$$V(\bar{y}^*_{\text{rss}}) = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2 k}{n}\triangle^2_{2(M)},$$

where $\triangle^2_{2(M)} = E\{\frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2}\triangle^2_{2(j)}\}$

$$V(\bar{y}^*_{\text{rss}}) = V(\bar{y}^*_{\text{srs}}) - \frac{W_2 k}{n}\triangle^2_{2(M)}\,. \tag{8}$$

Since $V(\bar{y}^*_{\text{rss}}) < V(\bar{y}^*_{\text{srs}})$ as $\triangle^2_{2(M)} > 0$. Hence, $\bar{y}^*_{\text{rss}}$ is more efficient than $\bar{y}^*_{\text{srs}}$.

In some cases it is difficult to rank all units, which results in large error. Detecting only some units with distinct ranks may be easier and more accurate. Keeping this point in mind Samawi, Abu-Dayyeh and Ahmed (1996), we use an RSS sampling procedure called Extreme Ranked Set Sampling (ERSS). The procedure includes identification of two extreme values $Y_{(1:j)}$ and $Y_{(\acute{n}_2:j)}$ from the jth sample. The extreme ranked set sampling in case of ranked set sampling works as follow. Select $Y_{2(e:j)}$ such that:

$$Y_{2(e:j)} = \begin{cases} Y_{2(1:j)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ Y_{2(\acute{n}_2:j)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2, \end{cases}$$

where

$$E(Y_{2(e:j)}) = \begin{cases} \mu_{2(1)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ \mu_{2(\acute{n}_2)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2, \end{cases}$$

and

$$V(Y_{2(e:j)}) = \begin{cases} \sigma^2_{2(1)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ \sigma^2_{2(\acute{n}_2)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2, \end{cases}$$

an estimate of μ_2 is:

$$\acute{\bar{y}}_{2(\text{erss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2} Y_{(e:j)} = \frac{Y_{2(1)} + Y_{2(\acute{n}_2)}}{2},$$

$$\text{where } E(\acute{\bar{y}}_{2(\text{erss})}) = \frac{\mu_{2(1)} + \mu_{2(\acute{n}_2)}}{2} \neq \mu_2.$$

Hence, it is a biased estimator of μ_2. It will be unbiased if $\mu_{2(1)} = \mu_{2(\acute{n}_2)}$. It is possible only in case of symmetric distribution. Now Hansen and Hurwitz (1946) estimator in ERSS, is given by

$$\bar{y}^*_{\text{erss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\acute{\bar{y}}_{2(\text{erss})}. \tag{9}$$

The Bias of the estimator $\bar{y}^*_{\text{erss}}$ is given by

$$\text{Bias}(\bar{y}^*_{\text{erss}}) = W_2\frac{(\mu_{2(1)} - \mu_2) + (\mu_{2(\acute{n}_2)} - \mu_2)}{2}, \tag{10}$$

which is almost negligible in case of near to symmetric distribution, its variance is given by

$$V(\bar{y}^*_{\text{erss}}) = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2 k}{n}\triangle^2_{2(e)},$$

where $\triangle^2_{2(e)} = \frac{\triangle^2_{2(1)} + E\triangle^2_{2(n_2)}}{2}$. The above expression can also be written as:

$$V(\bar{y}^*_{\text{erss}}) = V(\bar{y}^*_{\text{srs}}) - \frac{W_2 k}{n}\triangle^2_{2(e)}\,. \tag{11}$$

ERSS will be preferred on RSS if $\triangle^2_{2(e)} > \triangle^2_{2(M)}$.

Bouza (2010) used another modification to RSS that include selecting medians of all ranked set samples. Assuming $\acute{n}_2$ as even, select $Y_{2(m:j)}$ such that:

$$Y_{2(m:j)} = \begin{cases} Y_{2((\acute{n}_2/2):j)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ Y_{2((\acute{n}_2/2)+1:j)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2, \end{cases}$$

where

$$E(Y_{2(m:j)}) = \begin{cases} \mu_{2(\acute{n}_2/2)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ \mu_{2((\acute{n}_2/2)+1)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2 \end{cases}$$

and

$$V(Y_{2(m:j)}) = \begin{cases} \sigma^2_{2(\acute{n}_2/2)} & \text{for } j = 1, \ldots, \frac{\acute{n}_2}{2}, \\ \sigma^2_{2((\acute{n}_2/2)+1)} & \text{for } j = \frac{\acute{n}_2}{2} + 1, \ldots, \acute{n}_2. \end{cases}$$

The estimator for μ_2 using Median Ranked Set Sampling (MRSS) is:

$$\acute{\bar{y}}_{2(\text{mrss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2} Y_{(m:j)} = \frac{Y_{2(\acute{n}_2/2)} + Y_{2((\acute{n}_2+2)/2)}}{2},$$

where

$$E(\acute{\bar{y}}_{2(\text{mrss})}) = \frac{\mu_{2(\acute{n}_2/2)} + \mu_{2((\acute{n}_2+2)/2)}}{2} \neq \mu_2.$$

Hence, it is a biased estimator of μ_2. It will be unbiased if $\mu_{2(\acute{n}_2/2)} = \mu_{2((\acute{n}_2+2)/2)}$. Now Hansen and Hurwitz (1946) estimator in MRSS, is given by

$$\bar{y}^*_{\text{mrss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\acute{\bar{y}}_{2(\text{mrss})}. \tag{12}$$

The Bias of the estimator $\bar{y}^*_{\text{mrss}}$ is given by

$$\text{Bias}(\bar{y}^*_{\text{mrss}}) = W_2 \frac{(\mu_{2(\acute{n}_2/2)} - \mu_2) + (\mu_{2((\acute{n}_2+2)/2)} - \mu_2)}{2}, \tag{13}$$

which is almost negligible in case of approximately symmetric distribution, its variance, is given by

$$V(\bar{y}^*_{\text{mrss}}) = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2 k}{n}\triangle^2_{2(m)},$$

where $\triangle^2_{2(m)} = E\{\frac{\triangle^2_{2(\acute{n}_2/2)} + \triangle^2_{2((\acute{n}_2+2)/2)}}{2}\}$. The above expression can also be written as:

$$V(\bar{y}^*_{\text{mrss}}) = V(\bar{y}^*_{\text{srs}}) - \frac{W_2 k}{n}\triangle^2_{2(m)}. \tag{14}$$

MRSS will be preferred over RSS if $\triangle^2_{2(m)} > \triangle^2_{2(M)}$. Also MRSS will be preferred over ERSS if $\triangle^2_{2(m)} > \triangle^2_{2(e)}$.

3 Proposed estimators

When the study variable is sensitive in nature than non-response occurs due to sensitivity of the character under study, consequently the estimators in (7), (9) and (12) fail to estimate population mean of the study character as it is hard to find a sub-sample on second call. Taking motivation from Diana, Riaz and Shabbir (2014) estimator, we use a randomized response model in RSS, ERSS and MRSS for sub-sampling non-respondents to overcome this difficulty. From (3), a ranked set sampled jth scrambled response in the jth sample is given as follow when ranking is performed on Y:

$$Z_{[j:j]} = A_j Y_{(j:j)} + B_j \qquad (j = 1, 2, \ldots, \acute{n}_2), \tag{15}$$

where $E_R(Z_{[j:j]}) = \mu_A Y_{(j:j)} + \mu_B$ and variance of $Z_{[j:j]}$ is $V_R(Z_{[j:j]}) = \sigma_A^2 Y^2_{(j:j)} + \sigma_B^2$, here E_R, V_R are expectation and variance with respect to randomization device.

Let $\hat{y}_{[j;j]}$ be transformed scrambled response of the jth unit in the jth sample whose expectation under randomization mechanism equals to true response $y_{(j;j)}$.

$$\hat{y}_{[j:j]} = \frac{z_{[j:j]} - \mu_B}{\mu_A}, \qquad E_R(\hat{y}_{[j:j]}) = y_{(j:j)},$$
$$V_R(\hat{y}_{[j:j]}) = \frac{\sigma_A^2 y^2_{[j:j]} + \sigma_B^2}{\mu_A^2}. \tag{16}$$

The estimate of μ_2 using this technique is, $\acute{\bar{\hat{y}}}_{2(\text{rss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2} \hat{y}_{2[j:j]}$. The proposed estimator using this technique to sub-sample non-respondents is given by

$$\hat{\bar{y}}^*_{\text{rss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\acute{\hat{\bar{y}}}_{2(\text{rss})}, \tag{17}$$

with expected value

$$\begin{aligned} E(\hat{\bar{y}}^*_{\text{rss}}) &= E_1E_2[w_1\bar{y}_1 + w_2E_R\{\acute{\hat{\bar{y}}}_{2(\text{rss})}\}] \\ &= E_1E_2[w_1\bar{y}_1 + w_2\acute{\bar{y}}_{2(\text{rss})}] \qquad \text{as } E_R(\acute{\hat{\bar{y}}}_{2(\text{rss})}) = \acute{\bar{y}}_{2(\text{rss})} \\ &= \mu. \end{aligned}$$

Hence, $\hat{\bar{y}}^*_{\text{rss}}$ is an unbiased estimator of μ and variance of $\hat{\bar{y}}^*_{\text{rss}}$, can be derived as following:

$$V(\hat{\bar{y}}^*_{\text{rss}}) = E_1[V_2\{E_R(\hat{\bar{y}}^*_{\text{rss}})\} + E_2\{V_R(\hat{\bar{y}}^*_{\text{rss}})\}]. \tag{18}$$

Take

$$\begin{aligned} E_1[V_2\{E_R(\hat{\bar{y}}^*_{\text{rss}})\}] &= E_1[V_2(\bar{y}^*_{\text{rss}})] \\ &= \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2k}{n}\triangle^2_{2(M)}. \end{aligned} \tag{19}$$

Now for another part,

$$\begin{aligned} V_R(\hat{\bar{y}}^*_{\text{rss}}) &= \frac{w_2^2}{\acute{n}_2^2}\sum_{j=1}^{\acute{n}_2} V_R[\hat{y}_{2[j:j]}] \\ &= \frac{w_2^2}{\acute{n}_2^2}\sum_{j=1}^{\acute{n}_2}\left[\frac{\sigma_A^2 y^2_{(j:j)} + \sigma_B^2}{\mu_A^2}\right], \end{aligned}$$

$$\begin{aligned} E_2\{V_R(\hat{\bar{y}}^*_{\text{rss}})\} &= \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2(\sum_{j=1}^{\acute{n}_2}\sigma^2_{2(j)} + \sum_{j=1}^{\acute{n}_2}\mu^2_{2(j)}) + \acute{n}_2\sigma_B^2}{\mu_A^2}\right] \\ &= \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2\{\acute{n}_2\sigma_2^2 - \sum_{i=1}^{\acute{n}_2}\triangle^2_{2(j)} + \sum_{j=1}^{\acute{n}_2}\mu^2_{2(j)}\} + \acute{n}_2\sigma_B^2}{\mu_A^2}\right]. \end{aligned}$$

Hence

$$E_1E_2\{V_R(\hat{\bar{y}}^*_{\text{rss}})\} = \frac{W_2k}{n}\left[\frac{\sigma_A^2\{\sigma_2^2 + \mu^2_{2(M)}\} + \sigma_B^2}{\mu_A^2} - \frac{\sigma_A^2}{\mu_A}\triangle^2_{2(M)}\right], \tag{20}$$

where $\mu^2_{2(M)} = E\{\frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2}\mu^2_{2(j)}\}$ and $\triangle^2_{2(M)}$ is defined earlier. Substituting (19) and (20) in (18), we get:

$$V(\hat{\bar{y}}^*_{\text{rss}}) = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2k}{n}\triangle^2_{2(M)} + \frac{W_2k}{n}\left[\frac{\sigma_A^2\{\sigma_2^2+\mu^2_{2(M)}\}+\sigma_B^2}{\mu_A^2} - \frac{\sigma_A^2}{\mu_A^2}\triangle^2_{2(M)}\right], \tag{21}$$

$$V(\hat{\bar{y}}^*_{\text{rss}}) = V(\hat{\bar{y}}^*_{\text{srs}}) - \frac{W_2k}{n}\triangle^2_{2(M)}\,\theta,$$

where $\theta = 1 + \frac{\sigma_A^2}{\mu_A^2}$ and $\theta > 1$. Hence,

$$G_{\text{Eff}}(\text{rss}) = \frac{W_2k}{n}\triangle^2_{2(M)}\,\theta > 0,$$

where $G_{\text{Eff}}(\text{rss})$ denotes the gain in efficiency due to RSS.

Taking motivation from Bouza (2010), we propose an estimator of population mean by using ERSS with scrambled response model on second call in situation of non-response. Because it is easier to identify only extreme units from a sample than ranking all units. The scrambled response is given by

$$Z_{[e:j]} = A_jY_{(j:e)} + B_j \qquad (j = 1, 2, \ldots, \acute{n}_2), \tag{22}$$

where $E_R(Z_{[e:j]}) = \mu_AY_{(e:j)} + \mu_B$ and variance of $Z_{[e:j]}$ is $V_R(Z_{[j:e]}) = \sigma_A^2Y^2_{(j:e)} + \sigma_B^2$,

Let $\hat{y}_{[e:j]}$ be transformed scrambled response from extreme units in the jth sample whose expectation under randomization mechanism equals to true response $y_{(e:j)}$.

$$\hat{y}_{[e:j]} = \frac{z_{[e:j]} - \mu_B}{\mu_A}, \qquad E_R(\hat{y}_{[e:j]}) = y_{(e:j)},$$
$$V_R(\hat{y}_{[e:j]}) = \frac{\sigma_A^2y^2_{[e:j]} + \sigma_B^2}{\mu_A^2} = \phi_{[e:j]}. \tag{23}$$

The estimate of μ_2 using this technique is:

$$\hat{\acute{\bar{y}}}_{2(\text{erss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2}\hat{y}_{(e:j)} = \frac{\hat{Y}_{2(1)} + \hat{Y}_{2(\acute{n}_2)}}{2},$$

$$\text{where } E\{E_R(\hat{\acute{\bar{y}}}_{2(\text{erss})})\} = E(\hat{\acute{\bar{y}}}_{2(\text{erss})}) = \frac{\mu_{2(1)} + \mu_{2(\acute{n}_2)}}{2} \neq \mu_2.$$

In case of symmetric distribution it will be unbiased as $\mu_{2(1)} = \mu_{2(\acute{n}_2)}$. The proposed estimator using this technique to sub-sample non-respondents is given by

$$\hat{\bar{y}}^*_{\text{erss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\hat{\acute{\bar{y}}}_{2(\text{erss})}. \tag{24}$$

The Bias of proposed estimator is,

$$E(\hat{\bar{y}}^{*}_{\text{erss}}) = E_1E_2[w_1\bar{y}_1 + w_2E_R\{\acute{\hat{\bar{y}}}_{2(\text{erss})}\}]$$

$$= E_1E_2[w_1\bar{y}_1 + w_2\acute{\bar{y}}_{2(\text{erss})}] \qquad \text{as } E_R(\acute{\hat{\bar{y}}}_{2(\text{erss})}) = \acute{\bar{y}}_{2(\text{erss})} \tag{25}$$

$$\Rightarrow \quad \text{Bias}(\hat{\bar{y}}^{*}_{\text{erss}}) = W_2\frac{(\mu_{2(1)} - \mu_2) + (\mu_{2(\acute{n}_2)} - \mu_2)}{2},$$

which is almost negligible in case of near to symmetric distribution, its variance is given by

$$V(\hat{\bar{y}}^{*}_{\text{erss}}) = E_1[V_2\{E_R(\hat{\bar{y}}^{*}_{\text{erss}})\} + E_2\{V_R(\hat{\bar{y}}^{*}_{\text{erss}})\}]. \tag{26}$$

Consider the first part

$$\begin{aligned} E_1[V_2\{E_R(\hat{\bar{y}}^{*}_{\text{erss}})\}] &= E_1[V_2(\bar{y}^{*}_{\text{erss}})], \\ &= \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2k}{n}\,\triangle^2_{2(e)}\,. \end{aligned} \tag{27}$$

Consider the second part

$$\begin{aligned} &V_R(\hat{\bar{y}}^{*}_{\text{erss}}) \\ &\quad = \frac{w_2^2}{\acute{n}_2^2}\sum_{j=1}^{\acute{n}_2}\left[\frac{\sigma_A^2y^2_{(j:e)} + \sigma_B^2}{\mu_A^2}\right], \\ &E_2\{V_R(\hat{\bar{y}}^{*}_{\text{erss}})\} \\ &\quad = \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2E_2((\acute{n}_2/2)y^2_{2(1)} + (\acute{n}_2/2)y^2_{2(\acute{n}_2)}) + \acute{n}_2\sigma_B^2}{\mu_A^2}\right] \\ &\quad = \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2\{(\acute{n}_2/2)(\sigma^2_{2(1)} + \mu^2_{2(1)}) + (\acute{n}_2/2)(\sigma^2_{2(\acute{n}_2)} + \mu^2_{2(\acute{n}_2)})\} + \acute{n}_2\sigma_B^2}{\mu_A^2}\right] \\ &\quad = \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2\{(\acute{n}_2/2)\tau_1 + (\acute{n}_2/2)\tau_{\acute{n}_2}\} + \acute{n}_2\sigma_B^2}{\mu_A^2}\right], \end{aligned}$$

where $\tau_1 = \sigma_2^2 - \triangle^2_{2(1)} + \mu^2_{2(1)}$ and $\tau_{\acute{n}_2} = \sigma_2^2 - \triangle^2_{2(\acute{n}_2)} + \mu^2_{2(\acute{n}_2)}$. Hence,

$$\begin{aligned} E_1E_2\{V_R(\hat{\bar{y}}^{*}_{\text{erss}})\} = \frac{W_2k}{n}\Bigg[\frac{\sigma_A^2}{\mu_A^2}\Bigg\{&\sigma_2^2 - \frac{1}{2}(\triangle^2_{2(1)} + E_1\triangle^2_{2(\acute{n}_2)}) \\ &+ \frac{1}{2}(\mu^2_{2(1)} + E_1\mu^2_{2(\acute{n}_2)})\Bigg\} + \frac{\sigma_B^2}{\mu_A^2}\Bigg]. \end{aligned} \tag{28}$$

Using (27) and (28) in (26), we get

$$\begin{aligned} V(\hat{\bar{y}}^*_{\text{erss}}) &= \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2 k}{n}\left[\triangle^2_{2(e)} + \frac{\sigma_A^2}{\mu_A^2}\{\sigma_2^2 - \triangle^2_{2(e)} + \mu^2_{2(e)}\} + \frac{\sigma_B^2}{\mu_A^2}\right] \\ &= \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 + \frac{W_2 k}{n}\left[\frac{\sigma_A^2\{\sigma_2^2 + \mu^2_{2(e)}\} + \sigma_B^2}{\mu_A^2}\right] \\ &\quad - \frac{W_2 k}{n}\triangle^2_{2(e)}\,\theta \\ &= V(\hat{\bar{y}}^*_{\text{srs}}) - \frac{W_2 k}{n}\triangle^2_{2(e)}\,\theta, \end{aligned} \tag{29}$$

where $\mu^2_{2(e)} = \frac{1}{2}(\mu^2_{2(1)} + E_1\mu^2_{2(\acute{n}_2)})$ and $\triangle^2_{2(e)}$ is defined earlier in previous section. The gain in efficiency due to ERSS is

$$G_{\text{Eff}}(\text{erss}) = \frac{W_2 k}{n}\triangle^2_{2(e)}\,\theta > 0.$$

Since $\theta > 0$. ERSS will give more efficient result than RSS if:

$$\triangle^2_{2(e)} > \triangle^2_{2(M)}.$$

We propose an estimator of population mean by using scrambled response model in Median of Ranked Set Sampling. The scrambled response is given by

$$Z_{[m:j]} = A_j Y_{(m:j)} + B_j \qquad (j = 1, 2, \ldots, \acute{n}_2), \tag{30}$$

where $E_R(Z_{[m:j]}) = \mu_A Y_{(m:j)} + \mu_B$ and variance of $Z_{[m:j]}$ is $V_R(Z_{[m:j]}) = \sigma_A^2 Y^2_{(m:j)} + \sigma_B^2$.

Let $\hat{y}_{[m:j]}$ be transformed scrambled response from median units in the jth sample whose expectation under randomization mechanism equals to true response $y_{(m:j)}$.

$$\begin{aligned} \hat{y}_{[m:j]} &= \frac{z_{[m:j]} - \mu_B}{\mu_A}, \qquad E_R(\hat{y}_{[m:j]}) = y_{(m:j)}, \\ V_R(\hat{y}_{[m:j]}) &= \frac{\sigma_A^2 y^2_{[m:j]} + \sigma_B^2}{\mu_A^2} = \phi_{[m:j]}. \end{aligned} \tag{31}$$

The estimate of μ_2 using this technique is:

$$\hat{\acute{\bar{y}}}_{2(\text{mrss})} = \frac{1}{\acute{n}_2}\sum_{j=1}^{\acute{n}_2}\hat{y}_{(j:m)} = \frac{\hat{Y}_{2(\acute{n}_2/2)} + \hat{Y}_{2((\acute{n}_2/2)+1)}}{2},$$

$$\text{where } E\{E_R(\hat{\acute{\bar{y}}}_{2(\text{mrss})})\} = E(\hat{\acute{\bar{y}}}_{2(\text{mrss})}) = \frac{\mu_{2(\acute{n}_2/2)} + \mu_{2((\acute{n}_2/2)+1)}}{2} \neq \mu_2.$$

It will be unbiased if $\mu_{2((\acute{n}_2/2)+1)} = \mu_{2(\acute{n}_2/2)}$ which is possible only in case of symmetric distribution. The proposed estimator using this technique to sub-sample non-respondents given bellow.

$$\hat{\bar{y}}^*_{\text{mrss}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\hat{\acute{\bar{y}}}_{2(\text{mrss})}. \tag{32}$$

The Bias of proposed estimator $\hat{\bar{y}}^*_{\text{mrss}}$ is given by

$$\begin{aligned} E(\hat{\bar{y}}^*_{\text{mrss}}) &= E_1E_2[w_1\bar{y}_1 + w_2E_R(\hat{\acute{\bar{y}}}_{2(\text{mrss})})] \\ &= E_1E_2[w_1\bar{y}_1 + w_2(\acute{\bar{y}}_{2(\text{mrss})})] \qquad \text{as } E_R(\hat{\acute{\bar{y}}}_{2(\text{mrss})}) = \acute{\bar{y}}_{2(\text{mrss})} \qquad (33) \\ \Rightarrow \quad \text{Bias}(\hat{\bar{y}}^*_{\text{mrss}}) &= W_2\frac{(\mu_{2(\acute{n}_2/2)} - \mu_2) + (\mu_{2((\acute{n}_2/2)+1)} - \mu_2)}{2}, \end{aligned}$$

which is almost negligible when the distribution tends to symmetric, its variance using law of total variance, is given by

$$V(\hat{\bar{y}}^*_{\text{mrss}}) = E_1[V_2\{E_R(\hat{\bar{y}}^*_{\text{mrss}})\} + E_2\{V_R(\hat{\bar{y}}^*_{\text{mrss}})\}]. \tag{34}$$

The first part of (34) is:

$$E_1[V_2\{E_R(\hat{\bar{y}}^*_{\text{mrss}})\}] = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 - \frac{W_2k}{n}\{\triangle^2_{2(m)}\}. \tag{35}$$

Also the second part of (34) is

$$V_R(\hat{\bar{y}}^*_{\text{mrss}}) = \frac{w_2^2}{\acute{n}_2^2}\sum_{j=1}^{\acute{n}_2}\left\{\frac{\sigma_A^2 y^2_{(j:m)} + \sigma_B^2}{\mu_A^2}\right\},$$

$$E_2\{V_R(\hat{\bar{y}}^*_{\text{mrss}})\} = \frac{w_2^2}{\acute{n}_2^2}\left[\frac{\sigma_A^2\{(\acute{n}_2/2)T_1 + (\acute{n}_2/2)T_2\} + \acute{n}_2\sigma_B^2}{\mu_A^2}\right],$$

where $T_1 = \sigma_2^2 - \triangle^2_{2(\acute{n}_2/2)} + \mu^2_{2(\acute{n}_2/2)}$ and $T_2 = \sigma_2^2 - \triangle^2_{2((\acute{n}_2/2)+1)} + \mu^2_{2((\acute{n}_2/2)+1)}$. Therefore,

$$E_1E_2\{V_R(\hat{\bar{y}}^*_{\text{mrss}})\} = \frac{W_2k}{n}\left[\frac{\sigma_A^2}{\mu_A^2}\{\sigma_2^2 - \triangle^2_{2(m)} + \mu^2_{2(m)}\} + \frac{\sigma_B^2}{\mu_A^2}\right], \tag{36}$$

where $\mu^2_{2(m)} = \frac{1}{2}(\mu^2_{2(\acute{n}_2/2)} + E_1\mu^2_{2((\acute{n}_2/2)+1)}$ and $\triangle^2_{2(m)}$ is defined in previous section. Using (35) and (36) in (34), we get

$$\begin{aligned} V(\hat{\bar{y}}^*_{\text{mrss}}) = \frac{\sigma^2}{n} + \frac{W_2(k-1)}{n}\sigma_2^2 + \frac{W_2k}{n}\left[\frac{\sigma_A^2\{\sigma_2^2 + \mu^2_{2(m)}\} + \sigma_B^2}{\mu_A^2}\right] \\ - \frac{W_2k}{n}\triangle^2_{2(m)}\,\theta, \qquad (37) \end{aligned}$$

$$V(\hat{\bar{y}}^*_{\text{mrss}}) = V(\hat{\bar{y}}^*_{\text{srs}}) - \frac{W_2k}{n}\triangle^2_{2(m)}\,\theta.$$

Gain in efficiency due to MRSS is

$$G_{\text{Eff}}(\text{mrss}) = \frac{W_2 k}{n} \triangle^2_{2(m)}\, \theta > 0.$$

MRSS will give more efficient result than RSS if: $\triangle^2_{2(m)} > \triangle^2_{2(M)}$.

Also MRSS will give more efficient result than ERSS if: $\triangle^2_{2(m)} > \triangle^2_{2(e)}$.

4 A Monte Carlo comparison of the efficiency of the estimators

To compare the efficiency of proposed estimators in RSS to the corresponding estimators in SRSWR, we conduct a simulation study taking motivation from Agarwal, Allende and Bouza (2012). The values of auxiliary variable X are generated by using three different distributions that is, (1) Normal ($\mu = 0, \sigma^2 = 1$), (2) Exponential ($\lambda = 5$) and (3) Uniform ($a = 0, b = 1$). After that, Y is computed such that $Y = rX + e$, taking $r = 0.80$, where r is the coefficient of correlation between X and Y and $e \sim N(0, 1)$ is the error term. Assuming different values of W_2, we identified some unit in the population as non-respondents. We denote each RSS procedure with R:

R = RSS, ERSS, MRSS.

The Monte Carlo experiment works as follows:

Step 1. We select a sample s from the above hypothetical populations and divide it into two groups s_1 with size n_1 and s_2 with size n_2, where $n_2 = n \times W_2$ and $n_1 = n - n_2$. Then calculate the sample mean of Y from s_1 and $\acute{n}_2$ is determined such that $\acute{n}_2 = \frac{n_2}{k}$ $(k > 1)$.

Step 2. Select $\acute{s}_2$ by using R of size $\acute{n}_2$ from s_2 and then compute the sample mean of Y for non-response group. Note that we used scrambled responses (see Section 3) rather than direct responses for computing sample mean on second call to obtain results of last three columns of Table 1.

Step 3. The mean estimator $\hat{\bar{y}}^*_R = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\hat{\bar{y}}_{(2R)}$ is computed for the hth ($h = 1, 2, \ldots, H$) sample corresponding to R.

The cycle is repeated for obtaining $H = 10{,}000$ samples and compute respective sample means. Then the mean and variance for the sample mean of these H samples are calculated and relative efficiencies of these estimators are obtained by using following formula

$$\text{RE} = \frac{\text{Var}(\bar{y}^*)}{\text{Var}(\bar{y}^*_R)} \quad \text{and} \quad \widehat{\text{RE}} = \frac{\text{Var}(\hat{\bar{y}}^*)}{\text{Var}(\hat{\bar{y}}^*_R)}.$$

The results are presented in Table 1.

The first three columns give relative efficiency (RE) of Bouza (2002) estimators w.r.t. Hansen and Hurwitz (1946) estimator and the last three columns of Table 1 give relative efficiency ($\widehat{\text{RE}}$) of proposed estimators with respect to Diana,

Table 1 *The relative efficiency of RSS, ERSS and MRSS*

Distribution	W_2	k	RE			$\widehat{\text{RE}}$		
			RSS	ERSS	MRSS	RSS	ERSS	MRSS
Normal	0.1	2	1.1250	1.0609	1.0935	1.3513	1.2867	1.3946
		4	1.1013	1.1049	1.2241	1.2122	1.2088	1.2862
		6	1.1013	1.1049	1.2241	1.2122	1.2088	1.2862
	0.2	2	1.2728	1.1763	1.2138	2.1318	1.6430	2.2981
		4	1.4389	1.2135	1.3260	1.9032	1.6060	2.1399
		6	1.3291	1.1242	1.3359	1.5354	1.2804	1.7114
	0.3	2	1.4455	1.0917	1.1437	3.1526	1.5960	3.4264
		4	1.6438	1.4535	1.5430	2.7307	1.9072	3.2724
		6	1.6769	1.2640	1.4570	2.2765	1.8958	2.7598
	0.4	2	1.6160	1.2031	1.2515	4.5716	2.1163	5.1349
		4	1.9228	1.6177	1.7300	3.4769	2.1427	4.3755
		6	2.0375	1.2540	1.4467	3.0449	1.5971	3.8519
Exponential	0.1	2	1.1170	1.0543	1.0828	1.6186	1.6780	1.6661
		4	1.1123	1.1098	1.2265	1.3107	1.5030	1.4821
		6	1.1123	1.1098	1.2265	1.3107	1.5030	1.4821
	0.2	2	1.2596	1.1517	1.1909	3.1984	1.8218	3.3736
		4	1.3659	1.1088	1.2214	2.4734	2.4532	2.7109
		6	1.3584	1.1875	1.3860	1.9270	1.7200	2.0568
	0.3	2	1.4439	1.1188	1.1682	5.2756	1.4882	5.6374
		4	1.5923	1.4473	1.5372	3.8759	2.0581	4.4753
		6	1.7389	1.3048	1.5143	3.0151	2.8765	3.4168
	0.4	2	1.6419	1.1481	1.1986	7.7175	1.6036	8.6775
		4	1.9136	1.5654	1.6774	5.0912	1.8970	5.9705
		6	1.9693	1.3389	1.5237	3.9496	1.7195	4.7537
Uniform	0.1	2	1.1138	1.0674	1.0977	1.6765	1.6903	1.7198
		4	1.1204	1.1017	1.2161	1.3577	1.5053	1.4567
		6	1.1204	1.1017	1.2161	1.3577	1.5053	1.4567
	0.2	2	1.2533	1.1726	1.2050	3.2788	1.9403	3.5413
		4	1.3812	1.1515	1.2621	2.4932	2.4591	2.7205
		6	1.3049	1.1039	1.3151	1.8869	1.7240	2.0569
	0.3	2	1.4491	1.1234	1.1726	5.5418	1.6480	6.1193
		4	1.6808	1.5261	1.6343	4.0407	2.2174	4.7350
		6	1.5325	1.2022	1.3913	2.8012	2.8716	3.3526
	0.4	2	1.6165	1.2146	1.2635	8.2002	1.6974	9.1755
		4	1.9746	1.6667	1.7758	5.5181	2.1727	6.3218
		6	2.0436	1.2798	1.4815	4.2700	1.8340	5.0032

Riaz and Shabbir (2014) estimator. The proposed estimators give larger relative efficiency than the Bouza (2002) estimators. We can see that relative efficiency of proposed estimators tend to decreases for larger k. This implies that for small sub-sample size proposed estimators give greater precision. Relative efficiency of proposed estimators are greater for larger values of W_2. It can also be inferred from Table 1 that RSS perform better in term of efficiency for case of Uniform distribution as compared to other two distributions. In all cases, MRSS performs better than RSS under scrambled response model. But relative efficiency of ERSS is smaller than RSS and MRSS for all cases.

5 Conclusion

This article presented a procedure for estimation of population mean in non-response where we can obtain twin objectives of survey sampling; (i) one is to give greater confidentiality to the respondents which results increment of response rate, (ii) another is gain in precision of estimates involve in study. By assuming that non-response is due to sensitivity of the study character we proposed three estimators using scrambled response model by selecting a ranked set sample, extreme ranked set sample and median ranked set sample. It is proved both mathematically and numerically that the estimators of population mean perform better in RSS, ERSS and MRSS than SRSWR. Further work can be extended to give more privacy protection to respondents by applying other randomized response models. Efficiency can be improved by using other ranked set sampling procedure.

Acknowledgments

The authors thank the referees for their valuable comments which allowed improving a previous version.

References

Agarwal, G. K., Allende, S. M. and Bouza, C. (2012). Double sampling with ranked set selection in the second phase with nonresponse: Analytical results and Monte Carlo experiences. *Journal of Probability and Statistics* **2012**, Article ID 214959. MR2897336

Bouza, C. (2002). Estimation of the mean in ranked set sampling with non-responses. *Metrika* **56**, 171–179. MR1932086

Bouza, C. (2009). Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. *Metrika* **6**, 0184–0191. MR2550750

Bouza, C. (2010). Ranked set sampling procedure for the estimation of the population mean under non-response: A comparison. *Revista Investigation Operational* **31**, 140–150. MR2840959

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

Diana, G. and Perri, P. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers* **52**, 633–650. MR2821061

Diana, G., Riaz, S. and Shabbir, J. (2014). Hansen and Hurwitz estimator with scrambled response on the second call. *Journal of Applied Statistics* **41**, 596–611. MR3291275

Hansen, M. H. and Hurwitz, W. (1946). The problem of non response in sample surveys. *Journal of The American Statistical Association* **41**, 517–529.

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

Muttlak, H. (1996). Median ranked set sampling. *Journal of Applied Statistical Science* **6**, 91–98.

Patil, G. (2002). Ranked set sampling. In *Encyclopedia of Enviromentrics* **3** (A. H. El-Shaarawi and W. W. Piegorsch, eds.) 1684–1690. Chichester: Wiley.

Samawi, H., Abu-Dayyeh, W. and Ahmed, S. (1996). Extreme ranked set sampling. *Biometrical Journal* **30**, 577–586.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of The American Statistical Association* **60**, 63–69.

Department of Statistics
Quaid-i-Azam University
Islamabad
Pakistan
E-mail: shakeelatish05@gmail.com