Journal of Mathematics and Statistics

Original Research Paper

Randomized Response Procedure for the Estimation of the Population Ratio using Ranked Set Sampling

[1]Agustin Santiago, [2]Carlos N. Bouza, [1]J. Maclovio Sautto and [3]Amer Ibrahim Al-Omari

[1]Facultad de Matemáticas, Universidad Autónoma de Guerrero, Acapulco, Guerrero, México
[2]Facultad de Matemáticay Computación, Universidad de La Habana, La Habana, Cuba
[3]Department of Mathematics, Faculty of Science, Al Al-Bayt University, Jordan

Article history
Received: 09-10-2015
Revised: 25-04-2016
Accepted: 27-04-2016

Corresponding Author:
Agustín Santiago
Facultad de Matemáticas,
Universidad Autónoma de Guerrero, Acapulco, Guerrero, México
Email: asantiago@uagro.mx

Abstract: In this study we deal with the estimation of the population ratio, when a Randomized Response (RR) procedure is used for collecting responses and Ranked Set Sampling (RSS) is the selection method. The variances of the suggested estimators are calculated. Comparisons between different estimators are presented.

Keywords: Randomized Response, Population Ratio, Ranked Set Sampling, Simple Random Sampling Without Replacement

Introduction

The common model considers that we are interested in the study of Y, a sensitive variable evaluated in a finite population $U = \{u_1, u_2,\ldots,u_N\}$, u_i, is an identifiable unit. Some values Y, identifies having a stigma. Hence stigmatized individuals will tend to give incorrect reports on Y or to refusing to give an answer. A solution is introducing the use of a Random Response (RR) query. The seminal work on RR is due to Warner (1965). Warner's method consisted in including two alternatives questions: The question associated with the stigma and other insensitive questions. The interviewed chooses at random one of the questions and gives an answers without revealing which question he/she has selected. In the case of a quantitative variable a similar reasoning can be used. Chaudhuri and Stenger (2006), for example, for a discussion on RRprocedures when we deal with a quantitative character.

RR models are in development due both to their practical and theoretical interest. Give a look to the papers of (Singh and Singh, 1993; Christofides, 2003), for example. Commonly the authors considered the behavior of their proposals when simple random sampling is the design used for selecting the samples. Rueda and González (2004; Singh and Tarray, 2014a; 2014b), for a comprehensive look at this problem.

RSS is a relative new sampling design, whichout performs Simple Random Sampling With Replacement (SRSWR). The seminal paper is due to McIntyre; see Chen *et al*. (2004). The units may be ranked cheaply and then an Order Statistics (OS) is selected from each provisionally selected sample. The provisional samples are selected using SRSWR. It has been proved that RSS generally supports an increase in accuracy of the estimators.

Some interesting recently published results are: Al-Saleh and Al-Omari (2002), who suggested multistage ranked set sampling for estimating the population mean; Bouza (2010) who considered the estimation of the mean of a sensitive quantitative character in RSS using auxiliary variables for RR procedures; Chen and Lim (2011) who considered the estimation of variances of strata in RSS. Patil (2002; Patil *et al*., 1994; 1999; Bouza and Al-Omari, 2010; Al-Omari, 2011; Jemain and Al-Omari, 2006; Chen *et al*., 2004) for a detailed discussion on RSS.

In this study, we considered the ratio estimation problem. Let X be a known variable highly correlated with Y which is used both for selecting the ranked sample and for computing estimation of the ratio, $\zeta = \frac{\mu_Y}{\mu_X}$, where μ_X and μ_Y are the population means of X and Y, respectively.

The remaining part of the paper is organized as follows: In section 2, is concerned with a model based RR responses procedure when is used SRSWR. A RSS with RR procedures is developed in section 3. Comparison between different estimators is conducted in section 4. In section 5, an empirical comparison of the proposed estimators is presented. Conclusions are given in section 6.

A Scrambled Variable RR Procedure under SRSWR

We will describe briefly the RR procedure developed by Chaudhuri and Stenger (1992). It is an illustrating model. For an unit $u_i \in U$ the sampler determines the sets of known variables A = $\{A_1, A_2,\ldots,A_T\}$ and $B = \{B_1, B_2,\ldots,B_S\}$. Once they are fixed, we calculate $\mu_A = \frac{\sum_{t=1}^{T} A_t}{T} \neq 0$, $\sigma_A^2 = \frac{\sum_{t=1}^{T}(A_t - \mu_A)^2}{T}$, $\mu_B = \frac{\sum_{s=1}^{S} B_s}{S}$ and $\sigma_B^2 = \frac{\sum_{s=1}^{S}(B_s - \mu_B)^2}{S}$.

For each selected a $u_i \in U$, he/she will not response to the sensitive question and report the value of Y_i. The unit (individual) performs a random experiment and selects independently $a \in A$ and $b \in B$, say (A_i, B_i). The report made by the questioned is:

$$Z_i = A_i Y_i + B_i$$

A "prediction" of Y_i is:

$$R_i = \frac{Z_i - \mu_B}{\mu_A}$$

The model expectation and variance of the "prediction" are:

$$E_M(R_i) = Y_i, V_M(R_i) = \frac{Y_i^2\sigma_A^2 + \sigma_B^2}{\mu_A^2} = V_i,$$

The selection of a sample of size n using simple random sampling without replacement as design generates the reports, $R_1, R_2,...,R_n$. The RR procedure generates the data $D(R) = \{(u_i, Y_i, A_i, B_i)| u_i, \in s, A_i \in A, B_i \in B\}$. Then the sample mean of the computed R_is are used for estimating the mean of the sensitive variable:

$$\bar{R} = \frac{\sum_{i=1}^{n} R_i}{n} \quad (2.1)$$

As $E_M(R_i) = Y_i$, the expectation of (2.1) is the sample mean of Y. Therefore when SRSWR is used

$E_d E_M(\bar{R}) = E_d(\bar{y}) = \mu_Y$.

Due to the independence:

$$V_M(\bar{R}) = \frac{\sum_{i=1}^{n} V_i}{n^2}$$

And the expected model-error is given by:

$$E_d V_M(\bar{R}) = \frac{1}{n^2}\sum_{i=1}^{n} E_d\left(\frac{Y_i^2\sigma_A^2 + \sigma_B^2}{\mu_A^2}\right) = \left(\mu_Y^2 + \sigma_Y^2\right)\frac{\sigma_A^2}{n\mu_A^2} + \frac{\sigma_B^2}{n\mu_A^2} \quad (2.2)$$

The variance of the model expectation is:

$$V_d E_M(\bar{R}) = V_d\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\sigma_Y^2}{n}$$

Therefore, the expected error of (2.1) is:

$$V(\bar{R}) = \frac{\sigma_Y^2}{n} + \left(\mu_Y^2 + \sigma_Y^2\right)\frac{\sigma_A^2}{n\mu_A^2} + \frac{\sigma_B^2}{n\mu_A^2} \quad (2.3)$$

Note that $\left(\mu_Y^2 + \sigma_Y^2\right)\frac{\sigma_A^2}{n\mu_A^2} + \frac{\sigma_B^2}{n\mu_B^2}$ is due to using the RR procedure.

Consider the estimation of ζ and take the naïve estimator:

$$\hat{\zeta}_{msa} = \frac{\bar{R}}{\bar{x}} \quad (2.4)$$

A Taylor series expansion for the first order of (2.4) yields the approximation:

$$\hat{\zeta}_{msa} = \frac{\mu_Y}{\mu_X} - \frac{\mu_Y}{\mu^2{}_X}(\bar{x} - \mu_X) + \frac{1}{\mu_X}\left(\bar{R} - \mu_Y\right) \quad (2.5)$$

Its variance is given by:

$$V\left(\hat{\zeta}_{msa}\right) = \frac{1}{\mu^2{}_X}\left[\frac{1}{n}\begin{pmatrix}\zeta^2\sigma_X^2 + \sigma_Y^2 \\ +\left(\mu_Y^2 + \sigma_Y^2\right)\frac{\sigma_A^2}{\mu_A^2} + \frac{\sigma_B^2}{\mu_A^2}\end{pmatrix}\right] - 2Cov\left(\bar{R}, \bar{x}\right) \quad (2.6)$$

Where:

$$Cov\left(\bar{R}, \bar{x}\right) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} R_i, \frac{1}{n}\sum_{i=1}^{n} x_i\right)$$
$$= E\left(\frac{1}{n^2}\sum_{i=1}^{n} R_i \sum_{i=1}^{n} x_i\right) - E\left(\bar{R}\right)E(\bar{x})$$

Note that if $\mu_X^2 \to \infty$, then $\mathrm{Var}\left(\hat{\zeta}_{msa}\right) \to 0$.

RSS for the RR Procedure

For implementing the selection of a RSS, we use SRSWR for choosing independently m samples of size m. The units in each sample are ranked using some additional information on Y. Commonly a highly correlated covariate X. Take for example as covariate:

- The known salary of the functionaries allows establishing a ranking of variables related with the money obtained by briberies once that the homes of them are visited
- The size of the network of people with whom an infected AIDS patient has had sex is known. It is correlated with different interest sensitive variables
- The area of a farm is correlated with variables associated with the production of it. Consider the study of the evasion of tax variables. The magnitude of undeclared production, sells and other economic issues is sensitive. Ranking the area permits to derive an adequate ranking of sensitive variables

The unit occupying the place *i* in the ranked sample $S_{(i)}$ is included in the ranked sample. Then a sample of size m is obtained. When we need a sample of size n we apply the procedure independently $r \geq 1$ times (cycles). Then we have $n = mr$ sample units. David and Levine (1972) developed a study of the effect of ranking judgmental errors. They proved that the errors do not affect the properties of RSS. We will use this fact in the sequel and we work with judgmental order statistic.

Let $X_{i(1)j}, X_{i(2)j},\ldots,X_{i(\mathrm{m})j}$ be the order statistics of the *i*th sample $X_{i1j}, X_{i2j}, \ldots, X_{imj}$, for $i = 1, 2,\ldots,m$ in the *j*th cycle, $j = 1, 2,..,r$. Then, $X_{1(1)j}, X_{2(2)j},\ldots, X_{m(m)j}$, denote the measured RSS. The cdf $F_{X(i)}(x)$ of the *i*th order statistics $X(i)$, is given by:

$$F_{X(i)}(x) = \frac{m!}{(i-1)!(m-i)!}\int_0^{F(x)} v^{i-1}(1-v)^{m-i}\,dv \tag{3.1}$$

And the pdf $F_{X(i)}(x)$ is given by:

$$f_{X(i)}(x) = \frac{m!}{(i-1)!(m-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{m-i} f(x) \tag{3.2}$$

The mean and the variance of $X_{(i)}$ are given by $\mu_{X(i)} = \int_{-\infty}^{\infty} x f_{X(i)}(x)dx$ and $\sigma^2_{X(i)} = \int_{-\infty}^{\infty}\left(x-\mu_{X(i)}\right)^2 f_{X(i)}(x)dx$, respectively.

Takahasi and Wakimoto (1968) showed that the efficiency of *RSS* relative to *SRS* is:

$$1 \leq eff\left(\bar{x}_{RSS}, \bar{x}_{SRS}\right) = \frac{\mathrm{Var}\left(\bar{x}_{SRS}\right)}{\mathrm{Var}\left(\bar{x}_{RSS}\right)} \leq \frac{m+1}{2} \tag{3.3}$$

where, $\bar{x}_{(SRS)} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{x}_{(RSS)} = \frac{1}{n}\sum_{i=1}^{n} x_{i(i)}$ are the sample means using *SRS* and *RSS* methods, respectively. Also, they showed that:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_{X(i)}(x),\ \mu = \frac{1}{m}\sum_{i=1}^{m}\mu_{X(i)}$$

And:

$$\sigma^2_X = \frac{1}{m}\sum_{i=1}^{m}\sigma^2_{X(i)} + \frac{1}{m}\sum_{i=1}^{m}\left(\mu_{X(i)} - \mu_X\right)^2 \tag{3.4}$$

Bouza (2009; Hussain and Shabbir 2011; Bouza, 2010; Agarwal *et al*., 2012) for more insights on these issues. We assumed that the ranking is made on *Y*. For implementing the procedure, each u_i interviewed selects randomly and independently $A_i \in A$, $B_i \in B$. The report of the *i*th ranked sample in the *t*th cycle is:

$$Z_{[i]t} = A_i Y_{[i]t} + B_i$$

Then we can compute for each u_i:

$$R_{[i]t} = \frac{Z_{[i]t} - \mu_B}{\mu_A}$$

Its model expectation, for any *t* and *i*, is $E_M(T_{[i]t}) = Y_{[i]t}$. The mean of the reports is:

$$\bar{R}_t = \frac{1}{m}\sum_{i=1}^{m} R_{[i]t}$$

For each cycle we have that:

$$E_d E_M\left(\bar{R}_t\right) = \frac{1}{m}\sum_{i=1}^{m}\mu_{[i]} = \mu_Y$$

Therefore, we derive easily that an unbiased estimator of μ_Y is:

$$\bar{R}_{(RSS)} = \frac{1}{rm}\sum_{t=1}^{r}\sum_{i=1}^{m} R_{[i]t} \tag{3.5}$$

The model variance of the report is:

$$V_M\left(R_{[i]t}\right) = V_M\left(\frac{Z_{[i]t} - \mu_B}{\mu_A}\right) = \frac{Y^2_{[i]t}\sigma^2_A + \sigma^2_B}{\mu^2_A}$$

The independence of the involved variables sustains that:

$$E_d V_M\left(\bar{R}'_{(RSS)}\right) = \sum_{i=1}^{m}\frac{\left(\sigma^2_{[i]} + \mu^2_{[i]}\right)\sigma^2_A + \sigma^2_B}{rm^2\mu^2_A} \tag{3.6}$$

Because:

$$E_d\left(Y^2_{[i]t}\right) = \sigma^2_{[i]} + \mu^2_{[i]},$$

The relation between $\sigma_{[i]}^2$ and σ_Y^2 allows writing (Chen *et al*., 2004):

$$\sum_{i=1}^{m}\sigma_{[i]}^2 = m\sigma_Y^2 - \sum_{i=1}^{m}\left(\mu_{[i]}-\mu_Y\right)^2 = m\sigma_Y^2 - \sum_{i=1}^{m}\Delta_{Y_{[i]}}^2$$

Hence, we can rewrite (3.6) as:

$$E_d V_M\left(\bar{R}_{(RSS)}\right) = \frac{\sigma_A^2}{n\mu_A^2}\left(m\sigma_Y^2 - \sum_{i=1}^{m}\Delta_{Y_{[i]}}^2\right) + \frac{\sigma_A^2}{nm\mu_A^2}\sum_{i=1}^{m}\mu_{[i]}^2 + \frac{\sigma_B^2}{n\mu_A^2}$$

The other term of the error is:

$$V_d E_M\left(\bar{R}_t\right) = V_d\left(\frac{\sum_{t=1}^{r}\sum_{i=1}^{m}Y_{[i]t}}{rm}\right) = \frac{\sigma_Y^2}{n} - \frac{\sum_{i=1}^{m}\Delta_{Y_{[i]}}^2}{nm}$$

Hence:

$$V\left(\bar{R}_{(RSS)}\right) = V_{(Y)} = \frac{1}{n}\left[\begin{array}{l}\sigma_Y^2 + \dfrac{\sigma_B^2}{\mu_A^2} - \dfrac{\sum_{i=1}^{m}\Delta_{Y_{[i]}}^2}{m} + \dfrac{\sigma_A^2}{\mu_A^2} \\ \left(\sigma_Y^2 - \dfrac{\sum_{i=1}^{m}\Delta_{Y_{[i]}}^2}{m}\right) + \dfrac{\sigma_A^2}{m\mu_A^2}\sum_{i=1}^{m}\mu_{[i]}^2\end{array}\right] \qquad (3.7)$$

We implement the ranking of the selected individuals using the information provided by the selected auxiliary variable *X*. The persons included in each sample select randomly the corresponding insensitive variables *A* and *B*. We will consider the cases in which *A* or *B*are equal to *X*. The RSS procedure is used in the*m* independent samples and in each cycle. The report of an individual u_i is:

$$Z_{[i]} = \begin{cases} Z_{A_{[i]}} = X_{(i)}Y_{[i]t} + B_i & if\ X = A \\ Z_{B_{[i]}} = X_i Y_{[i]t} + B_{(i)} & if\ X = B \end{cases}$$

Consider the *i*th interviewed in the *t*th cycle and take $Z_{A[i]t} = X_{(i)t}\ Y_{[i]t} + B_{it}$. The computed response variable is $R_{A_{[i]t}} = \dfrac{Z_{A_{[i]t}} - \mu_B}{\mu_{X_{[i]}}}$ and its model expectation is the value of the sensitive variable $E_M(R_{A[i]}) = Y_{[i]t}$.

Therefore, to average the reports generates a model unbiased estimation of the mean of *Y*. Hence:

$$\bar{R}_{A(RSS)} = \frac{\sum_{t=1}^{r}\sum_{i=1}^{m}R_{A_{[i]t}}}{rm} \qquad (3.8)$$

Is an unbiased estimator of μ_Y as the reports are model unbiased for the corresponding sensitive variable and the arithmetic mean is design unbiased. Its model variance for the OS of the *i*-*t*h order in the cycle *t* is:

$$V_M\left(R_{A_{[i]t}}\right) = \frac{Y_{[i]t}^2\sigma_{A(i)}^2 + \sigma_B^2}{\mu_{A(i)}^2}$$

where, $\sigma_{A(i)}^2$ and $\mu_{A(i)}$ are the variance and mean of the OS of $A_{(i)}$. Then, the design expectation of the model error for the *i*th OS is:

$$E_d V_M\left(R_{A_{[i]t}}\right) = \frac{\left(\sigma_{[i]}^2 + \mu_{[i]}^2\right)\sigma_{A(i)}^2 + \sigma_B^2}{\mu_{A(i)}^2}$$

and:

$$E_d V_M\left(\bar{R}_{A(RSS)}\right) = \sum_{i=1}^{m}\left(\sigma_{[i]}^2 + \mu_{[i]}^2\right)\left(\frac{\sigma_{A(i)}^2}{nm\mu_{A(i)}^2}\right) + \sigma_B^2\sum_{i=1}^{m}\left(\frac{1}{nm\mu_{A(i)}^2}\right)$$

As the transformed variable model is unbiased for $Y_{[i]}$:

$$V_d E_M\left(\frac{\sum_{t=1}^{r}\sum_{i=1}^{m}R_{A_{[i]t}}}{rm}\right) = V_d\left(\frac{\sum_{t=1}^{r}\sum_{i=1}^{m}Y_{[i]t}}{rm}\right) = \frac{\sum_{i=1}^{m}\sigma_{[i]}^2}{nm}$$

The variance of (3.8) is given by:

$$V\left(\bar{R}_{A(RSS)}\right) = \frac{1}{nm}\left[\begin{array}{l}\sum_{i=1}^{m}\sigma_{[i]}^2\left(\dfrac{\sigma_{A(i)}^2}{\mu_{A(i)}^2}+1\right) \\ +\sum_{i=1}^{m}\dfrac{\sigma_{A(i)}^2\mu_{[i]}^2}{\mu_{A(i)}^2} + \sigma_B^2\sum_{i=1}^{m}\dfrac{1}{\mu_{A(i)}^2}\end{array}\right] \qquad (3.9)$$

Noting that $\sigma_{A(i)}^2 = \sigma_A^2 - \Delta_{A(i)}^2$, where $\Delta_{A(i)} = \mu_{A(i)} - \mu_A$, this relation between the variance of an OS and the population variance permits to rewrite (3.9) as:

$$\begin{array}{l} V\left(\bar{R}_{A(RSS)}\right) = \dfrac{\sigma_Y^2\sigma_A^2}{n}\sum_{i=1}^{m}\left(\dfrac{1}{\mu_{A(i)}^2}+1\right) \\ -\dfrac{1}{nm}\left[\begin{array}{l}\sum_{i=1}^{m}\dfrac{1}{\mu_{A(i)}^2}\left(\dfrac{\sigma_A^2\Delta_{Y_{[i]}}^2}{\mu_{A(i)}^2} + \dfrac{\sigma_Y^2\Delta_{A(i)}^2}{\mu_{A(i)}^2}\right) \\ -\dfrac{1}{\mu_{A(i)}^2}\sum_{i=1}^{m}\Delta_{A(i)}^2\sum_{i=1}^{m}\Delta_{Y_{[i]}}^2\end{array}\right] \\ +\dfrac{1}{nm}\sum_{i=1}^{m}\dfrac{\sigma_{A(i)}^2\mu_{Y_{[i]}}^2 + \sigma_B^2}{\mu_{A(i)}^2}\end{array} \qquad (3.10)$$

Note that from (3.10), is clearly indicated the effect of using RSS on the accuracy with respect to SRSWR strategy. Then, the estimator of the ratio of the population is given by:

$$\hat{\zeta}_{A(RSS)} = \frac{\bar{R}_{A(RSS)}}{\bar{x}_{(RSS)}} \qquad (3.11)$$

The first order Taylor Series expansion of (3.11) is:

$$\hat{\zeta}_{A(RSS)} = \frac{\mu_Y}{\mu_X} - \frac{\mu_Y}{\mu_X^2}\left(\bar{x}_{(RSS)} - \mu_X\right) + \frac{1}{\mu_X}\left(\bar{R}_{A(RSS)} - \mu_Y\right) \tag{3.12}$$

With variance:

$$V\left(\hat{\zeta}_{A(RSS)}\right) = \left(\frac{\mu_Y}{\mu_X^2}\right)^2 V\left(\bar{x}_{(RSS)}\right) + \frac{1}{\mu_X^2}V\left(\bar{R}_{A(RSS)}\right) - \frac{2\mu_Y}{\mu_X^3}Cov\left(\bar{x}_{(RSS)}, \bar{R}_{A(RSS)}\right) \tag{3.13}$$

Let us consider that the ranking is made using *B*. the model report is:

$$Z_{B_{[i]}} = X_{(i)}Y_{[i]t} + B_{(i)}$$

The unscrambled variable is:

$$R_{B_{[i]t}} = \frac{Z_{B_{[i]t}}Y_{[i]t} - \mu_{B_{(i)}}}{\mu_A}$$

The model expectation in this case is $E_M(R_{B[i]t}) = Y_{[i]t}$. The we estimate unbiasedly μ_Y is given using the estimator:

$$\bar{R}_{B(RSS)} = \frac{\sum_{t=1}^{r}\sum_{i=1}^{m}R_{B_{[i]t}}}{rm} \tag{3.14}$$

This fact follows because $E_M\left(Z_{B_{(i)}}\right) = \mu_A Y_{[i]t} + \mu_{B_{(i)}}$.

Take for *B* the OS´s parameters $\sigma^2_{B_{(i)}}$ and $\mu_{B_{(i)}}$. We have that the design expectation of the error of the proposed estimator is:

$$V_d E_M\left(\bar{R}_{B(RSS)}\right) = \frac{\sum_{i=1}^{m}\sigma^2_{[i]}}{nm}$$

And taking $\Delta_{B(i)} = \mu_{B(i)} - \mu_B$, we have:

$$V_M\left(Z_{B_{[i]}}\right) = Y^2_{[i]t}\sigma^2_{X_{(i)}} + \sigma^2_{B_{(i)}}$$

Then, it is obtained that:

$$E_d V_M\left(R_{B[i]t}\right) = \frac{\left(\mu^2_{[i]} + \sigma^2_{[i]}\right)\sigma^2_{X_{(i)}} + \sigma^2_{B_{(i)}}}{\mu^2_A}$$

Hence:

$$V\left(\bar{R}_{B(RSS)}\right) = \frac{\sum_{i=1}^{m}\sigma^2_{[i]}}{nm} + \frac{\sum_{i=1}^{m}\left(\mu^2_{[i]} + \sigma^2_{[i]}\right)\sigma^2_{X_{(i)}} + \sigma^2_{B_{(i)}}}{nm\mu^2_A},$$

$$V\left(\bar{R}_{B(RSS)}\right) = \frac{\sigma^2_Y}{n}\left(1 + \frac{\sigma^2_X}{m\mu^2_A}\right) + \frac{\sum_{i=1}^{m}\mu^2_{[i]}\sigma^2_{X_{(i)}} + \sigma^2_{B_{(i)}}}{nm\mu^2_A} - \frac{1}{nm\mu^2_A}\left[\sigma^2_X\sum_{i=1}^{m}\Delta^2_{[i]} + \sigma^2_Y\sum_{i=1}^{m}\Delta^2_{X_{(i)}} - \sum_{i=1}^{m}\Delta^2_{[i]}\sum_{i=1}^{m}\Delta^2_{X_{(i)}}\right] \tag{3.15}$$

Note that these reclus allow managing the accuracy of this RSS strategy by using an adequate value of μ_A. Take:

$$\hat{\zeta}_{B(RSS)} = \frac{\bar{R}_{B(RSS)}}{\bar{x}_{(RSS)}} \tag{3.16}$$

Using a Taylor expansion to the first degree of approximation, the estimator in (3.16) will be:

$$\hat{\zeta}_{B(RSS)} = \frac{\mu_Y}{\mu_X} - \frac{\mu_Y}{\mu_X^2}\left(\bar{x}_{(RSS)} - \mu_X\right) + \frac{1}{\mu_X}\left(\bar{R}_{B(RSS)} - \mu_Y\right)$$

With variance given by:

$$Var\left(\hat{\zeta}_{B(RSS)}\right) = \left(\frac{\mu_Y}{\mu_X^2}\right)^2 Var\left(\bar{x}_{(RSS)}\right) + \frac{1}{\mu_X^2}Var\left(\bar{R}_{B(RSS)}\right) - 2\frac{\mu_Y}{\mu_X^3}Cov\left(\bar{x}_{(RSS)}, \bar{R}_{B(RSS)}\right) \tag{3.17}$$

The variance terms are given by (2.3), (3.7), (3.9) and (3.15) and:

$$Cov\left(\bar{x}_{(RSS)}, \bar{R}_{Q(RSS)}\right) = E\left[\begin{pmatrix}\bar{x}_{(RSS)} - \mu_X\end{pmatrix}\begin{pmatrix}\bar{R}_{Q(RSS)} - E\left(\bar{R}_{Q(RSS)}\right)\end{pmatrix}\right], Q = A, B$$

Comparison of the Different Alternatives

Deriving a measure of the gain in accuracy of the estimators, based on their variance, leads to un-meaningful expressions, for deciding which is the best alternative. These expressions do not allow fixing values of the controllable parameters and establishing which the expected behavior of the sampling errors is. We considered a series of data bases designed Monte Carlo experiments. We planned the experiments for obtaining insight on the behavior of the distance between parameters and their estimations. Simulation experiments were conducted and the performance of the estimators were measured by calculating, for each generated sample:

$$\varepsilon_{q,s} = \left|\zeta_q - \right|\zeta_s, \quad q = msa, A(RSS), B(RSS); s = 1,2,\ldots,5000$$

The empirical evaluation of the estimator was made by computing:

$$AC_q = \frac{\sum_{s=1}^{5000} \varepsilon_{q,s}}{50000\zeta},\ q = msa, A(RSS), B(RSS)$$

As we know, dealing with sensitive information poses a problem for obtaining real data with the true values of the variables related to a stigma. We have 3 small populations where the true value of the sensitive information was known. The individuals were confronted with 6 pairs of sets of values of A and B. The values were compatible with the values of the variables. The populations were:

Population 1

Employees convicted of briberies. The ranking variable was $X = \frac{monthly\ salary}{value\ of\ the\ home\ commodities}$ the sensitive variables were Y_1= Investments in hardware in the last 5 years; Y_2 = Number of people involved in his/her unbecoming activities, N = 75.

Population 2

Patients of AIDS. The ranking variable was $X = \frac{number\ of\ persons\ infested\ in\ the\ sexual\ activities}{number\ of\ sexual\ partners}$, the sensitive variables was Y_3= Number of homosexual relations previous to the detection of the illness, N = 43.

Population 3

Farmers selling products directly in the market. The ranking variable was $X = \frac{reported\ cultivated\ area\ of\ the\ farm}{total\ area\ of\ the\ farm}$, the sensitive variables were Y_4 = Unreported income derived from selling their products in the last 6 months; Y_5 = Real cultivated area; Y_6 = Income from unauthorized services, N = 52.

Note that in all the cases, $X \in [0,1]$.

We used as sample size n = 3×5 = 15. The distribution of A^* and B^* were fixed as a Uniform in (0,1), $U(0,1)$; the standard normal, N(0,1); the standard asymptotical normal, $AN(0,1)$. The moments, variances and covariances of the OS´s were computed using the tables developed by Hastings *et al*. (1947). The OS $C_t^*, t = 1,2$, $C^* = A^*, B^*$, allowed to construct the sets $C_H^* = \left\{C_{(1)H}^*, C_{(2)H}^*, C_{(3)H}^*\right\}$, where:

$$C_{(t)H}^* = \begin{cases} \min\left(Y_j\right) \times C_{(t)}^* & if\ H = 1 \\ \max\left(Y_j\right) \times C_{(t)}^* & if\ H = 2 \end{cases}$$

We compared the estimators by computing:

$$E(q,w) = \frac{AC_q}{AC_w}, q \neq w, msa, RSS, A(RSS), B(RSS)$$

Table 1. $E(q,w) = \frac{AC_q}{AC_w}$, $q \neq w$, msa, RSS, $A(RSS)$, $B(RSS)$, $A = \left\{\min\left(Y_j\right) \times C_{(t)}^*\right\}$, $B = \left\{min(Y_j) \times C_{(t)}^*,\right\}, t = 1,2,3$

Efficiency measure		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
E[A(RSS),msa]		0,9094	0,7731	0,7113	0,8817	0,8110	0,9011
E[B(RSS),msa]	U(0,1)	0,9300	0,7757	0,7081	0,8840	0,8337	0,9090
E[A(RSS),B(RSS)]		1,0180	1,0012	0,9967	1,006	1,0236	1,0089
E[A(RSS),msa]		0,7527	0,7577	0,7108	0,8808	0,8028	0,9001
E[B(RSS),msa]	N(0,1)	0,7989	0,7714	0,7023	0,8188	0,8141	0,9029
E[A(RSS),B(RSS)]		1,0128	1,0936	0,9955	0,9275	0,9842	1,0076
E[A(RSS),msa]		0,7683	0,7066	0,7233	0,8088	0,7811	0,9005
E[B(RSS),msa]	AN(0,1)	0,7680	0,7110	0,7070	0,8864	0,8087	0,9004
E[A(RSS),B(RSS)]		0,9928	1,0276	0,9671	1,0106	1,0165	0,9965

Table 2. $E(q,w) = \frac{AC_q}{AC_w}$, $q \neq w$, msa, RSS, $A(RSS)$, $B(RSS)$, $A = \left\{\max\left(Y_j\right) \times C_{(t)}^*\right\}$, $B = \left\{\max\left(Y_j\right) \times C_{(t)}^*,\right\}, t = 1,2,3$

Efficiency measure		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
E[A(RSS),msa]		0,8488	0,7702	0,5766	0,8311	0,8004	0,8914
E[B(RSS),msa]	U(0,1)	0,6339	0,6692	0,4393	0,7340	0,7111	0,6004
E[A(RSS),B(RSS)]		1,389	1,1250	1,1182	1,2942	1,1139	1,2631
E[A(RSS),msa]		0,9177	0,8470	0,8124	0,9001	0,9084	0,9372
E[B(RSS),msa]	N(0,1)	0,6041	0,6206	0,7006	0,6156	0,72017	0,7219
E[A(RSS),B(RSS)]		1,3720	1,2038	1,1926	1,3782	1,1936	1,2178
E[A(RSS),msa]		0,8220	0,6821	0,6301	0,8929	0,8801	0,7731
E[B(RSS),msa]	AN(0,1)	0,7322	0,5100	0,5273	0,7730	0,7772	0,7520
E[A(RSS),B(RSS)]		1,1890	1,2742	1,1942	1,1860	1,2017	1,0017

Table 3. $E(q,w)=\frac{AC_q}{AC_w}, q\neq w, msa, RSS, A(RSS), B(RSS), A=\left\{\min\left(Y_j\right)\times C^*_{(t)}\right\}, B=\left\{\max\left(Y_j\right)\times C^*_{(t)}\right\}, t=1,2,3$

Efficiency measure		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
E[*A*(*RSS*),*msa*]		0,7391	0,6361	0,4939	0,7525	0,7474	0,7230
E[*B*(*RSS*),*msa*]	U(0,1)	0,7317	0,5720	0,4261	0,7328	0,7345	0,7089
E[*A*(*RSS*),*B*(*RSS*)]		1,0317	1,0138	1,1593	1,0245	1,0273	1,0142
E[*A*(*RSS*),*msa*]		0,7197	0,5971	0,4807	0,7107	0,7017	0,6958
E[*B*(*RSS*),*msa*]	N(0,1)	0,7023	0,5856	0,4087	0,6805	0,6156	0,5728
E[*A*(*RSS*),*B*(*RSS*)]		1,0286	1,0195	1,1706	1,0444	1,1112	1,2461
E[*A*(*RSS*),*msa*]		0,6663	0,5721	0,4325	0,6066	0,6611	0,6623
E[*B*(*RSS*),*msa*]	AN(0,1)	0,6360	0,5630	0,4166	0,5650	0,6066	0,5920
E[*A*(*RSS*),*B*(*RSS*)]		1,0472	1,0162	1,0378	1,0702	1,0901	1,1866

Table 4. $E(q,w)=\frac{AC_q}{AC_w}, q\neq w, msa, RSS, A(RSS), B(RSS), A=\left\{\max\left(Y_j\right)\times C^*_{(t)}\right\}, B=\left\{\min\left(Y_j\right)\times C^*_{(t)}\right\}, t=1,2,3$

Efficiency measure		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
E[*A*(*RSS*),*msa*]		0,8063	0,7531	0,6643	0,8642	0,8304	0,8085
E[*B*(*RSS*),*msa*]	U(0,1)	0,9011	0,8896	0,8187	0,9189	0,9095	0,9615
E[*A*(*RSS*),*B*(*RSS*)]		0,8911	0,9322	0,7386	0,9623	0,9328	0,9003
E[*A*(*RSS*),*msa*]		0,8088	0,8824	0,8112	0,8245	0,8083	0,8019
E[*B*(*RSS*),*msa*]	N(0,1)	0,6920	0,6435	0,5229	0,6269	0,6399	0,6830
E[*A*(*RSS*),*B*(*RSS*)]		0,8388	0,7416	0,6942	0,8090	0,8431	0,8194
E[*A*(*RSS*),*msa*]		0,5959	0,7998	0,7828	0,5512	0,5589	0,5792
E[*B*(*RSS*),*msa*]	AN(0,1)	0,6504	0,6675	0,8995	0,6046	0,7262	0,6042
E[*A*(*RSS*),*B*(*RSS*)]		0,9527	0,7974	0,8953	0,9239	0,8084	0,9145

The results are given in Table 1 when A and B were generated using C_1^*, $C^* = A^*$, B^*. We have that $\hat{\zeta}_{A(RSS)}$ is generally more efficient than $\hat{\zeta}_{B(RSS)}$. The construction of A and B fixed that $Cov\left(\bar{R}_{B(RSS)},\bar{x}_{RSS)}\right)-Cov\left(\bar{R}_{A(RSS)},\bar{x}_{RSS)}\right)=0$. Hence, the comparison of the proposed estimators gives that $\hat{\zeta}_{A(RSS)}$ is to be preferred when $A = B$. The use of $AN(0,1)$ is the best procedure. We consider that this results are supported by the fact that the means and standard deviations of the involved OS of $AN(0,1)$ are more similar than the parameters of the other two distributions.

The results of the analysis when A and B were generated using the maxima are given in Table 2. They are similar to those of Table 1, but the preferred estimator is $\hat{\zeta}_{B(RSS)}$.

The use of crossed criteria for generating A and B appear in the next tables.

Table 3 presents the results for $A=\left\{min\left(Y_j\right)\times C^*_{(t)} t=1,2,3\right\}, B=\left\{min\left(Y_j\right)\times C^*_{(t)} t=1,2,3\right\}$. The discussion on the relationships among the results for the distributions gives rise to similar comments and $AN(0,1)$ has the best behavior. It is remarkable that it is preferable using $\hat{\zeta}_{B(RSS)}$. In this case:

$$Cov\left(\bar{R}_{B(RSS)},\bar{x}_{RSS)}\right)-Cov\left(\bar{R}_{A(RSS)},\bar{x}_{RSS)}\right)>0$$

Table 4 presents the results for $A=\left\{mix\left(Y_j\right)\times C^*_{(t)} t=1,2,3\right\}, B=\left\{min\left(Y_j\right)\times C^*_{(t)} t=1,2,3\right\}$. The relationships are changed as $\hat{\zeta}_{B(RSS)}$ is not preferred to $\hat{\zeta}_{A(RSS)}$ in any case. The use of the $N(0,1)$ generates larger gains than in the previous experiments with respect to $U(0,1)$ and the use of $AN(0,1)$ is not associated to large gains in efficiency when compared with the $N(0,1)$. It is remarkable that using $\hat{\zeta}_{B(RSS)}$ is a bad decision because $Cov\left(\bar{R}_{B(RSS)},\bar{x}_{RSS)}\right)-Cov\left(\bar{R}_{A(RSS)},\bar{x}_{RSS)}\right)<0$.

Conclusion

From the derived results is obtained that the RSS models are more efficient than using the classic SRSWR estimators. The use of sets of auxiliary variables related with $Y_{(1)}$ or $Y_{(N)}$ allows determining which estimator is to be preferred and the expected gain in accuracy. The best method for generating them is to use $AN(0,1)$. A recommendation to practitioners is to fix the bound to the values of the sensitive variable Y, it is feasible to construct A and B accordingly using $U(0,1)$, $N(0,1)$ or $AN(0,1)$ and to decide which is the more efficient estimator.

Acknowledgement

The authors thanks the suggestions made by anonymous referees on a first version of the paper. This

final issue is animproved version benefited from the suggestions of them. The research of one of the authors was supported by the PNCB "Modelos Matemáticos para el Estudio de Medio Ambiente, Saludy Desarrollo Humano".

Author's Contributions

Agustín Santiago Moreno: Development of the idea of estimating the proportions using RSS randomized response method. Writing and editing the paper, work overall coordination and integration of ideas.

Carlos N. Bouza Herrera: He contributed estimators developed in previous work and some proofs of theorems. Also he contributed in drafting the final version of the entire document.

José Maclovio Sautto Vallejo: He programmed simulations algorithms, helped in the interpretation of results.

Amer Ibrahim Al-Omari: He contributed demonstrations RSS related estimators, in partial wording of paper, contributed central ideas and writing the paper in English.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Agarwal, G.K., S.M. Allende and C.N. Bouza, 2012. Double Sampling with ranked set selection in the second phase with nonresponse: Analytical results and monte carlo experiences. J. Probab. Stat. DOI: 10.1155/2012/214959

Al-Saleh, M.F. and A.I. Al-Omari, 2002. Multistage ranked set sampling. J. Stat. Plann. Inference. 102: 273-286. DOI: 10.1016/S0378-3758(01)00086-6

Bouza, C.N., 2009. Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character. Metrika, 70: 267-277. DOI: 10.1007/s00184-008-0191-6

Bouza, C.N., 2010. Ranked set sampling using auxiliary variables of a randomized response procedure for estimating the mean of a sensitive quantitative character. J. Modern Applied Stat. Methods, 9: 248-254.

Bouza, C.N. and A.I. Al-Omari, 2012. Estimating the population mean in the case of missing data using simple random sampling. Stat. J. Theoretical Applied Stat., 46: 279-290. DOI: 10.1080/02331888.2010.505654

Chaudhuri, A. and H. Stenger, 1992. Sampling survey. Springer, New York.

Chen, M. and J. Lim, 2011. Estimating variances of strata in ranked set sampling. J. Stat. Plann. Inference, 141: 2513-2518. DOI: 10.1016/j.jspi.2010.11.043

Chen, Z., Z. Bai and B.K. Sinha, 2004. Ranked Set Sampling: Theory and Applications. 1st Edn., Springer Science and Business Media, New York, ISBN-10: 0387402632, pp: 227.

Christofides, T.C., 2003. A generalized randomized response technique. Metrika, 57: 195-200. DOI: 10.1007/s001840200216

David, H.A. and D.W. Levine, 1972. Ranked set sampling in the presence of judgment error. Biometrics, 28: 553-555.

Hastings, C., F. Mosteller, J.W. Tukey and C.P. Windsor, 1947. Low moments for small samples: A comparative study of order statistics. Ann. Math. Stat., 18: 413-426. DOI: 10.1214/aoms/1177730388

Hussain, Z. and J. Shabbir, 2011. Improved estimation of mean in randomized response models. Hacettepe J. Math. Stat., 40: 91-104.

Jemain, A.A. and A.I. Al-Omari, 2006. Double quartile ranked set samples. Pak. J. Stat., 22: 217-228.

Patil, G.P., A.K. Sinha and C. Taillie, 1994. 5 Ranked set sampling. Handbook Stat., 12: 167-200. DOI: 10.1016/S0169-7161(05)80007-0

Patil, G.P., A.K. Sinha and C. Taillie, 1999. Ranked set sampling: A bibliography. Environ. Ecol. Stat., 6: 91-98. DOI: 10.1023/A:1009647718555

Patil, G.P., 2002. Ranked set sampling. Encyclopedia Enviro., 3: 1684-1690. DOI: 10.1002/9780470057339.var015.pub2

Rueda, M. and S. González, 2004. Missing data and auxiliary information in surveys. Comput. Stat., 19: 551-567. DOI: 10.1007/BF02753912

Singh, H.P. and T.A. Tarray, 2014a. An alternative to stratified Kim and Warde's randomized response model using optimal (Neyman) allocation. Model Assisted Stat. Applic., 9: 37-62. DOI: 10.3233/MAS-130277

Singh, H.P. and T.A. Tarray, 2014b. An improvement over Kim and Elam stratified unrelated question randomized response model using Neyman allocation. Sankhya-B, 77: 91-107. DOI: 10.1007/s13571-014-0088-5

Singh, S. and R. Singh, 1993. Generalized franklin's model for randomized response sampling. Commun. Stat. Theory Meth., 22: 741-755. DOI: 10.1080/03610929308831052

Takahasi, K. and K. Wakimoto, 1968. On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Insti. Stat. Math., 20: 1-31. DOI: 10.1007/BF02911622

Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Associat., 60: 63-69. DOI: 10.1080/01621459.1965.10480775