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

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

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

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

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 $Y_i$ 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 answer 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 RR procedures 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, which out 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 $\frac{\mu_Y}{\mu_X}$, where $\mu_Y$ and $\mu_X$ 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.

© 2016 Agustin Santiago, Carlos N. Bouza, J. Maclovio Sautto and Amer Ibrahim Al-Omari. This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license.
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 \( i \in U \) 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 \( \mu_A = \frac{\sum A_i}{r} = 0 \), \( \sigma_A^2 = \frac{\sum (A_i - \mu_A)^2}{r} \), \( \mu_B = \frac{\sum B_s}{S} \) and

\[
\sigma_B^2 = \frac{\sum (B_s - \mu_B)^2}{S}.
\]

For each selected a unit \( i \), he/she will not response when \( SRSWR \) is used. The reporter makes random sampling without replacement as design. The question of \( Y_i \) about a sensitive question and report the value of \( Y_i \) using the questionnaire to the reporter. We will describe briefly the RR procedure developed by Chaudhuri and Stenger (1992). It is an illustrating model. For an unit \( i \in U \) 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 \( \mu_A = \frac{\sum A_i}{r} = 0 \), \( \sigma_A^2 = \frac{\sum (A_i - \mu_A)^2}{r} \), \( \mu_B = \frac{\sum B_s}{S} \) and

\[
\sigma_B^2 = \frac{\sum (B_s - \mu_B)^2}{S}.
\]

For each selected a unit \( i \), he/she will not response when \( SRSWR \) is used. The reporter makes random sampling without replacement as design. The question of \( Y_i \) about a sensitive question and report the value of \( Y_i \) using the questionnaire to the reporter.

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

\[
\frac{\sum R_i}{n} = \hat{\mu}_R.
\]

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

\[
\frac{\sum R_i}{n} = \hat{\mu}_R.
\]

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

Due to the independence:

\[
E_m(R_i) = \frac{\sum Y_i}{n}.
\]

And the expected model-error is given by:

\[
E_{E_m(R)} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i - \mu_Y}{\mu_Y} \frac{\sigma_Y^2}{\mu_Y^2} + \frac{\sigma_\varepsilon^2}{\mu_Y} \right) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i - \mu_Y}{\mu_Y} \frac{\sigma_Y^2}{\mu_Y^2} + \frac{\sigma_\varepsilon^2}{\mu_Y} \right)
\]

The variance of the model expectation is:

\[
V(Y_{E_m(R)}) = \frac{\sigma_Y^2}{n} + \left( \frac{\sigma_\varepsilon^2}{\mu_Y} + \frac{\sigma_\varepsilon^2}{\mu_Y} \right)
\]

Note that \( \left( \frac{\sigma_Y^2}{\mu_Y} + \frac{\sigma_\varepsilon^2}{\mu_Y} \right) \) is due to using the RR procedure.

Consider the estimation of \( \zeta \) and take the naïve estimator:

\[
\hat{\zeta}_{naive} = \frac{\bar{\xi}_R}{\bar{x}}
\]

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

\[
\hat{\zeta}_{naive} = \frac{\bar{\xi}_R}{\bar{x}} - \frac{\partial \bar{\xi}_R}{\partial \bar{x}} \bar{x} + \frac{1}{2!} \left( \frac{\partial^2 \bar{\xi}_R}{\partial \bar{x}^2} \right) \bar{x}^2
\]

Its variance is given by:

\[
V(\hat{\zeta}_{naive}) = \frac{1}{\bar{x}^2} \left( \frac{\partial \bar{\xi}_R}{\partial \bar{x}} \right)^2 - 2 \frac{\partial^2 \bar{\xi}_R}{\partial \bar{x}^2} \frac{1}{\bar{x}}
\]

Where:

\[
\text{Cov}(\bar{\xi}_R, \bar{x}) = \text{Cov} \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i, \frac{1}{n} \sum_{i=1}^{n} \xi_i \right)
\]

Note that if \( \mu_x \rightarrow \infty \), then \( \text{Var}(\hat{\zeta}_{naive}) \rightarrow 0 \).

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 implement the procedure, each unit interviewed selects randomly and independently $A_i \in A, B_i \in B$.

The report of the $ith$ ranked sample in the $r$th cycle is:

$$Z_{i(r)} = A_i Y_{(i)} + B_i$$

Then we can compute for each $u_i$:

$$R_{(i)} = Z_{ij} - \frac{\mu_Y}{\mu_A}$$

Its model expectation, for any $t$ and $i$, is $E_{M}(T_{ij}) = Y_{ij}$. The mean of the reports is:

$$\bar{R} = \frac{1}{m} \sum_{i=1}^{m} R_{(i)}$$

For each cycle we have that:

$$E_{M}(\bar{R}) = \frac{1}{m} \sum_{i=1}^{m} \mu_{(i)} = \mu$$

Therefore, we derive easily that an unbiased estimator of $\mu$ is:

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

The model variance of the report is:

$$V_{M}(R_{(i)}) = V_{M}(\frac{Z_{ij} - \mu_{A} + \frac{\mu_{A}}{\mu_{A}}}{\mu_{A}}) = Y_{ij}^{2} \sigma_{x}^{2} + \sigma_{d}^{2}$$

The independence of the involved variables sustains that:

$$E_{M}(V_{M}(\bar{R}_{RSS})) = \frac{1}{rm} \sum_{i=1}^{m} \left( \sigma_{x}^{2} + \frac{\mu_{A}^{2}}{\mu_{A}} \right) \sigma_{x}^{2}$$

Because:

$$E_{M}(Y_{(i)}) = \sigma_{x}^{2} + \mu_{A}^{2}$$

DOI: 10.3844/jmssp.2016.107.114

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).

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

$$F_{X_{(i)}}(x) = \frac{m}{(i-1)!} \int_{0}^{x} v^{i-1}(1-v)^{m-i} dv$$

And the pdf $f_{X_{(i)}}(x)$ is given by:

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

The mean and the variance of $X_{(i)}$ are given by:

$$\mu_{X_{(i)}} = \int_{-\infty}^{\infty} x f_{X_{(i)}}(x) dx$$

and

$$\sigma_{X_{(i)}}^{2} = \int_{-\infty}^{\infty} (x - \mu_{X_{(i)}})^{2} f_{X_{(i)}}(x) dx$$

respectively.

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

$$1 - \text{eff} = \frac{\text{Var}(\tau_{RSS})}{\text{Var}(\tau_{SRS})} \leq \frac{m+1}{2}$$

where $\tau_{RSS} = \frac{1}{n} \sum_{i=1}^{n} \tau_{i}$ and $\tau_{SRS} = \frac{1}{n} \sum_{i=1}^{n} \tau_{i}$ are the sample means using SRS and RSS methods, respectively. Also, they showed that:

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

And:

$$\sigma_{x}^{2} = \frac{1}{m} \sum_{j=1}^{m} \sigma_{x}^{2} + \frac{1}{m} \sum_{j=1}^{m} \left( \mu_{X_{(i)}} - \mu_{*} \right)^{2}$$

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 relation between $\sigma_{y}^{2}$ and $\sigma_{y}^{2}$ allows writing (Chen et al., 2004):

$$\sum_{i=1}^{m} \sigma_{y}^{2} = m \sigma_{y}^{2} - \sum_{i=1}^{m} (\mu_{y} - \mu_{i})^{2} = m \sigma_{y}^{2} - \sum_{i=1}^{m} \Delta_{y}^{2}$$

Hence, we can rewrite (3.6) as:

$$E_{v} \left( \overline{R}_{\text{RSS}} \right) = \frac{\sigma_{y}^{2}}{n \mu_{y}^{2}} \left( m \sigma_{y}^{2} - \sum_{i=1}^{m} \Delta_{y}^{2} \right) + \frac{\sigma_{y}^{2}}{nm \mu_{y}^{2}} \sum_{i=1}^{m} \mu_{y}^{2} + \frac{\sigma_{y}^{2}}{nm}$$

The other term of the error is:

$$E_{v} \left( \overline{R} \right) = \frac{1}{n} \left( \sum_{i=1}^{m} \frac{\mu_{y}^{2} - \Delta_{y}^{2}}{\mu_{i}^{2}} \right)$$

Hence:

$$V \left( \overline{R}_{\text{RSS}} \right) = V_{y} \left( \overline{R} \right) = \frac{1}{n} \left( \frac{\sigma_{y}^{2}}{m \mu_{y}^{2}} \sum_{i=1}^{m} \Delta_{y}^{2} + \frac{\sigma_{y}^{2}}{m} \right)$$

(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 them independent samples and in each cycle. The report of an individual $u_{i}$ is:

$$Z_{i} = \begin{cases} Z_{A_{i}} & \text{if } X = A \\ Z_{B_{i}} & \text{if } X = B \end{cases}$$

Consider the $i$th interviewed in the $i$th cycle and take $Z_{A(i)} = X_{(i)} Y_{(i)} + B_{(i)}$. The computed response variable is $R_{A(i)} = \frac{Z_{A(i)} - \mu_{A(i)}}{\mu_{A(i)}}$ and its model expectation is the value of the sensitive variable $E_{M}(R_{(i)}) = Y_{(i)}$.

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

$$\overline{R}_{\text{RSS}} = \frac{1}{n m} \sum_{i=1}^{m} R_{A(i)}$$

(3.8)

Is an unbiased estimator of $\mu_{y}$ 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 $i$-th order in the cycle $t$ is:

$$V_{v} \left( \overline{R}_{y} \right) = \frac{\sigma_{y}^{2}}{m} + \frac{\sigma_{y}^{2}}{m}$$

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:

$$E_{v} \left( \overline{R}_{y} \right) = \left( \sigma_{y}^{2} + \frac{\sigma_{y}^{2}}{m} \right) \mu_{A(i)}$$

and:

$$E_{v} \left( \overline{R}_{\text{RSS}} \right) = \sum_{i=1}^{m} \left( \sigma_{y}^{2} \mu_{A(i)} + \frac{\sigma_{y}^{2}}{m} \right) + \frac{\sigma_{y}^{2}}{n m} \sum_{i=1}^{m} \left( \frac{1}{nm} \right)$$

As the transformed variable model is unbiased for $Y_{(i)}$:

$$V_{v} \left( \overline{R}_{y} \right) = \frac{1}{n m} \left( \sum_{i=1}^{m} \frac{\sigma_{y}^{2}}{m} \mu_{A(i)}^{2} + \frac{\sigma_{y}^{2}}{m} \right)$$

(3.9)

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

$$V_{v} \left( \overline{R}_{\text{RSS}} \right) = \frac{\sigma_{y}^{2}}{n m} \sum_{i=1}^{m} \left( \frac{1}{\mu_{A(i)}} + 1 \right)$$

(3.10)

$$\left[ \sum_{i=1}^{m} \frac{1}{\mu_{A(i)}} \left( \sigma_{A(i)}^{2} + \frac{\sigma_{A(i)}^{2}}{\mu_{A(i)}} \right) \right]$$

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:

$$\hat{\tau} = \frac{\overline{R}_{\text{RSS}}}{\overline{R}_{\text{RSS}}}$$

(3.11)
The first order Taylor Series expansion of (3.11) is:
\[
\zeta_{RSS} = \frac{\hat{\mu}_i}{\mu_x}(\bar{X}_{RSS} - \mu_x) + \frac{1}{\mu_x}(\hat{R}_{RSS} - \mu_t)
\]  
(3.12)

With variance:
\[
V(\zeta_{RSS}) = \left(\frac{\mu_i}{\mu_x}\right)^2V(\bar{X}_{RSS}) + \frac{1}{\mu_x}V(\hat{R}_{RSS}) - \frac{2\mu_i}{\mu_x}\text{Cov}(\bar{X}_{RSS}, \hat{R}_{RSS})
\]  
(3.13)

Let us consider that the ranking is made using \(B\), the model report is:
\[Z_{h_i} = X_{i}(y) + B_{i(o)}\]

The unscrambled variable is:
\[R_{h_i} = \frac{Z_{h_i} - Y_{i(o)} - \mu_{h_i}}{\mu_x}\]

The model expectation in this case is \(E_M(R_{B[i]}) = Y_{i[o]}\). We estimate unbiasedly \(\mu_t\) is given using the estimator:
\[
\bar{R}_{RSS} = \frac{\sum_{i=1}^{n} \sum_{m=1}^{n} R_{h_i} }{nm}
\]  
(3.14)

This fact follows because \(E_M(Z_{h_i}) = \mu_x Y_{i(o)} + \mu_{h_i}\).

Take for \(B\) the OS’s parameters \(\sigma^2_{h_i}\) and \(\mu_{h_i}\). We have that the design expectation of the error of the proposed estimator is:
\[
V_M(E_M(\bar{R}_{RSS})) = \frac{\sum_{i=1}^{n} \sigma^2_{h_i} }{nm}
\]

And taking \(\Delta_{h(i)} = \mu_{h(i)} - \mu_B\), we have:
\[V_M(Z_{h_i}) = Y_{i(o)}^2 \sigma^2_{h_i} + \sigma^2_{h_i}\]

Then, it is obtained that:
\[E_M V_M(R_{B[i]}) = \left(\frac{\mu_i}{\mu_x} + \frac{\sigma^2_{h_i}}{\sigma^2_{h_i} + \sigma^2_{h_i}}\right)\frac{\sigma^2_{h_i}}{\mu_x}\]

Hence:
\[\zeta_{RSS} = \frac{\hat{\mu}_i}{\mu_x}(\bar{X}_{RSS} - \mu_x) + \frac{1}{\mu_x}(\hat{R}_{RSS} - \mu_t)\]

\[
V(\zeta_{RSS}) = \left(\frac{\mu_i}{\mu_x}\right)^2V(\bar{X}_{RSS}) + \frac{1}{\mu_x}V(\hat{R}_{RSS}) - \frac{2\mu_i}{\mu_x}\text{Cov}(\bar{X}_{RSS}, \hat{R}_{RSS})
\]  
(3.13)

Note that these reclus allow managing the accuracy of this RSS strategy by using an adequate value of \(\mu_i\). Take:
\[
\zeta_{RSS} = \frac{\hat{R}_{RSS}}{\bar{X}_{RSS}}
\]  
(3.16)

Using a Taylor expansion to the first degree of approximation, the estimator in (3.16) will be:
\[
\hat{\zeta}_{RSS} = \frac{\mu_i}{\mu_x} - \frac{\mu_i}{\mu_x}(\bar{X}_{RSS} - \mu_x) + \frac{1}{\mu_x}(\hat{R}_{RSS} - \mu_t)
\]

With variance given by:
\[
Var(\hat{\zeta}_{RSS}) = \left(\frac{\mu_i}{\mu_x}\right)^2Var(\bar{X}_{RSS}) + \frac{1}{\mu_x}Var(\hat{R}_{RSS}) - \frac{2\mu_i}{\mu_x}\text{Cov}(\bar{X}_{RSS}, \hat{R}_{RSS})
\]  
(3.17)

The variance terms are given by (2.3), (3.7), (3.9) and (3.15) and:
\[
Cov(\bar{X}_{RSS}, \hat{R}_{RSS}) = E\left(\bar{X}_{RSS} - \mu_x\right)\left(\hat{R}_{RSS} - \mu_t\right)
\]

**Comparison of the Different Alternatives**

Deriving a measure of the gain in accuracy of the estimators, based on their variance, leads to unmeaningful 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:
\[\varepsilon_q = \varepsilon_q - \zeta_{RSS}, \quad q = msb, A(RSS), B(RSS); s = 1,2,\ldots,5000\]
The empirical evaluation of the estimator was made by computing:

\[ AC_q = \frac{\sum_{i=1}^{5000} E_{m,s} \times q = msa, A(RSS), B(RSS)}{50000} \]

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 \( X = \frac{\text{monthly salary}}{\text{value of the home commodities}} \), the sensitive variables were \( Y_1 = \text{Investments in hardware in the last 5 years}; Y_2 = \text{Number of people involved in his/her unbecoming activities}, N = 75. \)

**Population 2**

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

**Population 3**

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

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

We used as sample size \( n = 3 \times 5 = 15. \) The distribution of \( A^* \) and \( B^* \) were fixed as a Uniform in \((0, 1), U(0, 1); \) the standard normal, \( \text{N}(0,1); \) the standard asymptotical normal, \( \text{AN}(0,1). \) The moments, variances and covariances of the OS’s were computed using the tables developed by Hastings et al. (1947). The OS \( C^*_t, t=1,2, \ C^*_t = A^*, B^*, \) allowed to construct the sets \( C^*_q = \{C^*_0, C^*_1, C^*_2, C^*_3\}, \) where:

\[ C^*_q = \begin{cases} \min \{Y_1\} \times C^*_q, & \text{if } H = 1 \\ \max \{Y_1\} \times C^*_q, & \text{if } H = 2 \end{cases} \]

We compared the estimators by computing:

\[ E(q, w) = \frac{AC_q}{AC_q^*}, q \neq w, msa, RSS, A(RSS), B(RSS) \]

**Table 1.** \( E(q, w) = \frac{AC_q}{AC_q^*}, q \neq w, msa, RSS, A(RSS), B(RSS), A = \{\min \{Y_1\} \times C^*_1\}, B = \{\min \{Y_1\} \times C^*_1\}, t = 1, 2, 3 \)

<table>
<thead>
<tr>
<th>Efficiency measure</th>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
<th>( Y_3 )</th>
<th>( Y_4 )</th>
<th>( Y_5 )</th>
<th>( Y_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.9094</td>
<td>0.7731</td>
<td>0.7113</td>
<td>0.8817</td>
<td>0.8110</td>
<td>0.9011</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>U(0,1)</td>
<td>0.9300</td>
<td>0.7757</td>
<td>0.7081</td>
<td>0.8840</td>
<td>0.8337</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>0.1080</td>
<td>0.1002</td>
<td>0.9967</td>
<td>1.006</td>
<td>1.0236</td>
<td>1.0089</td>
</tr>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.7527</td>
<td>0.7577</td>
<td>0.7108</td>
<td>0.8808</td>
<td>0.8028</td>
<td>0.9001</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>N(0,1)</td>
<td>0.7989</td>
<td>0.7714</td>
<td>0.7023</td>
<td>0.8188</td>
<td>0.8141</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>1.0128</td>
<td>1.0936</td>
<td>0.9955</td>
<td>0.9275</td>
<td>0.9842</td>
<td>1.0076</td>
</tr>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.7683</td>
<td>0.7066</td>
<td>0.7233</td>
<td>0.8088</td>
<td>0.7811</td>
<td>0.9005</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>AN(0,1)</td>
<td>0.7680</td>
<td>0.7110</td>
<td>0.7070</td>
<td>0.8684</td>
<td>0.8087</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>0.9928</td>
<td>1.0276</td>
<td>0.9671</td>
<td>1.0106</td>
<td>1.0165</td>
<td>0.9965</td>
</tr>
</tbody>
</table>

**Table 2.** \( E(q, w) = \frac{AC_q}{AC_q^*}, q \neq w, msa, RSS, A(RSS), B(RSS), A = \{\max \{Y_1\} \times C^*_q\}, B = \{\max \{Y_1\} \times C^*_q\}, t = 1, 2, 3 \)

<table>
<thead>
<tr>
<th>Efficiency measure</th>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
<th>( Y_3 )</th>
<th>( Y_4 )</th>
<th>( Y_5 )</th>
<th>( Y_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.8488</td>
<td>0.7702</td>
<td>0.5766</td>
<td>0.8311</td>
<td>0.8004</td>
<td>0.8914</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>U(0,1)</td>
<td>0.6339</td>
<td>0.6692</td>
<td>0.4393</td>
<td>0.7340</td>
<td>0.7111</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>1.389</td>
<td>1.1250</td>
<td>1.1182</td>
<td>1.2942</td>
<td>1.1139</td>
<td>1.2631</td>
</tr>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.9177</td>
<td>0.8470</td>
<td>0.8124</td>
<td>0.9001</td>
<td>0.9084</td>
<td>0.9372</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>N(0,1)</td>
<td>0.6041</td>
<td>0.6206</td>
<td>0.7006</td>
<td>0.6156</td>
<td>0.72017</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>1.3720</td>
<td>1.2038</td>
<td>1.1926</td>
<td>1.3782</td>
<td>1.1936</td>
<td>1.2178</td>
</tr>
<tr>
<td>( E(A(RSS), msa) )</td>
<td>0.8220</td>
<td>0.6821</td>
<td>0.6301</td>
<td>0.8929</td>
<td>0.8801</td>
<td>0.7731</td>
</tr>
<tr>
<td>( E(B(RSS), msa) )</td>
<td>AN(0,1)</td>
<td>0.7322</td>
<td>0.5100</td>
<td>0.5273</td>
<td>0.7730</td>
<td>0.7772</td>
</tr>
<tr>
<td>( E(A(RSS), B(RSS)) )</td>
<td>1.1890</td>
<td>1.2742</td>
<td>1.1942</td>
<td>1.1860</td>
<td>1.2017</td>
<td>1.0017</td>
</tr>
</tbody>
</table>
The results are given in Table 1 when A and B were generated using $C^*_t$, $C^*_B$ and $B^*$. We have that $\hat{\xi}_{A(RSS)}$ is generally more efficient than $\hat{\xi}_{B(RSS)}$. The construction of $A$ and $B$ fixed that $\text{Cov}(\hat{R}_{A(RSS)}, \hat{R}_{RSS}) - \text{Cov}(\hat{R}_{B(RSS)}, \hat{R}_{RSS}) = 0$. Hence, the comparison of the proposed estimators gives that $\hat{\xi}_{A(RSS)}$ 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 $\hat{\eta}_{B(RSS)}$.

The use of crossed criteria for generating $A$ and $B$ appear in the next tables. Table 3 presents the results for $A = \{\min(Y) \times C^*_t, t = 1,2,3\}$, $B = \{\min(Y) \times C^*_t, t = 1,2,3\}$. 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 $\hat{\eta}_{B(RSS)}$. In this case:

$$\text{Cov}(\hat{R}_{B(RSS)}, \hat{R}_{RSS}) - \text{Cov}(\hat{R}_{A(RSS)}, \hat{R}_{RSS}) > 0$$

Table 4 presents the results for $A = \{\min(Y) \times C^*_t, t = 1,2,3\}$, $B = \{\min(Y) \times C^*_t, t = 1,2,3\}$.

The relationships are changed as $\hat{\xi}_{B(RSS)}$ is not preferred to $\hat{\xi}_{A(RSS)}$ 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 $\hat{\xi}_{B(RSS)}$ is a bad decision because $\text{Cov}(\hat{R}_{B(RSS)}, \hat{R}_{RSS}) - \text{Cov}(\hat{R}_{A(RSS)}, \hat{R}_{RSS}) < 0$.

**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 $Y_{(3)}$ 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.
final issue is an improved 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, Salud y 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


