A QUANTITATIVE COMPULSORY RANDOMIZED RESPONSE TECHNIQUE. SIMULATION WITH SENSITIVE DATA ON VIOLENCE IN MEXICO

Carlos N. Bouza-Herrera*, Pablo O. Juárez-Moreno**, Agustín Santiago-Moreno** and José M. Sautto-Vallejo**
*Universidad de La Habana, Cuba.
**Universidad Autónoma de Guerrero, México

ABSTRACT
In the literature we can find a classification of randomized response techniques as compulsory and optional. In this work we present a compulsory randomized response technique with the purpose of having a double random scrambling of the sensitive variable Y through a Bernoulli experiment and a Ri report, and that this translates into a greater protection of the information it provides the interviewee. The document specifies the properties of the population mean of the sensitive variable Y with simple random sampling with replacement, an extension to stratification is made, the optimal allocation and the gain in precision are specified. Finally, simulation is performed to evaluate the accuracy and efficiency of the proposed estimators using real data on the perception of violence in Mexico.
KEYWORDS: Randomized Response, Scrambling, Random Sampling, Violence, México

MSC: 62D05, 62P10

RESUMEN
En la literatura podemos encontrar una clasificación de las técnicas de respuestas aleatorizadas como obligatorias y opcionales. En este trabajo presentamos una técnica de respuesta aleatorizada obligatoria con la finalidad de tener un doble enmascaramiento aleatorio de la variable sensible Y a través de un experimento Bernoulli y un reporte R_i, y que esto, se traduzca en una mayor protección de la información que proporciona el entrevistado. En el documento se especifican las propiedades de la media poblacional de la variable sensible Y con muestreo aleatorio simple con remplazo, se hace extensión a estratificado, se especifica la asignación optima y la ganancia en precisión. Por último, se realiza una simulación para evaluar la precisión y eficiencia de los estimadores propuesto usando datos reales sobre la percepción de la violencia en México.

PALABRAS CLAVE: Respuestas Aleatorizadas, Enmascaramiento, Muestreo Aleatorio, Violencia, México

1. INTRODUCTION

It is usual, when carrying out an investigation using survey sampling to obtain information on characteristics of a population. Therefore, it becomes a priority to access information on that characteristic of the population. This entails a couple of frequent drawbacks in survey sampling, which due to non-response or bias in the information. In addition, they will be greater bigger if the characteristic to be known is of a sensitive type. To solve this, Warner (1965) proposed the methods of randomized responses (RR), which aims to encourage the respondent to provide their true response on issues or information considered sensitive. This action is achieved because the RR procedures are intended not to reveal to the interviewer the personal information that the respondent is providing and thus keeping it private.

To carry out the randomized response (RR) methods, a finite population U of N elements is considered, is assumed that with some design d a random sample s of size n is drawn, with probability $p(s)$, in which, the i-th respondent is of interest to the researcher needing to know his/her sensitive characteristic Y, that for some reason the respondent refuses to answer directly. The true value of Y will not known, but the estimation of the mean, for example, may unbiased. The procedure of the RR methods consists of scrambling the sensitive value of the respondent Y through a random mechanism (variable, experiment or both) M, which will have a distribution θ known to the researcher. This scrambling will generate a report Z for the i-th respondent so it is possible to estimate, through the report, the mean of the sensitive variable Y, which is $E_M(Z_i) = \hat{\mu}_Y$, the estimator variance $V_M(\hat{\mu}_Y) = \sigma_Y^2$ and noting that$C_M = (y_i, y_j) = 0$ for $i \neq j$.

The study of RR techniques has been diversified since Warner's work (1965) to the present day. Based on this, we can find in the literature different uses of RR techniques in different types of works, to mention a few: works of RR with qualitative data, see Abdel-Latif et al. (1967), Horvitz et al. (1967), Huang (2004), Singh et al. (2020), Narjis and Shabbir (2021); works of RR with quantitative data, see Greenberg et al.

(1971), Eriksson et al. (1973), Gupta et al.(2002), Arnab (2018), Bouza et al. (2022); works dealing with sensitive issues such as abortion, see Perri et al. (2016), drug use, see Stubbe et al. (2013), racism, see Krumpal (2012), AIDS, see Bouza (2009), Arnab and Singh (2010); works applied in the Health area, see Bouza (2002) and Murtaza et al. (2020), Social Science, see Pal et al. (2020), Computing, see Rueda et al. (2016); we can also find works where classifications of the RR techniques are made, see Arnab and Rueda (2016), Juárez-Moreno et al. (2023). See Chaudhuri and Mukherjee (1988) y Chaudhuri et al. (2016) for a wide range of RR topics.

In this document, a new compulsory RR technique is proposed specifying the properties of μ_Y when the sampling design is simple random sampling with replacement and its stratified extension. The proportional and optimal allocation is specified, in addition to the gain in precision. Finally, a simulation is performed to evaluate the accuracy and efficiency of the proposed estimators using real data on the perception of violence in Mexico.

2. PROPOSE RR SCRAMBLING PROCEDURE USING SRSWR

The respondent participation to answer or not to a question depends, to a large extent, on how sensitive the question is and how confident he or she is in answering it. Therefore, a good RR technique increases the proportion of respondents who feel confident in answering despite the fact that it is a highly sensitive question. In other words, the RR technique that scramble the true value *Y* of the respondent will be in practice better, since the respondents will trust more, due to the degree the more confident in the scrambling in providing their true answer. Following the Arnab's work (2018), in which he converts two Partial Optional RR techniques to Full Optional RR, in the same way, the Full Optional RR techniques of Arnab (2018) we convert it into a Compulsory RR technique. Therefore, in this section we present a new compulsory RR technique in which the respondent's response is randomly scrambled by first using a Bernoulli experiment with probability *Q* if its sensitive value is scrambled by R_1 is reported or *(1-Q)* if you scramble your sensitive value with the R_2 report, which are:

$$R_1 = Y_i \frac{X_i}{\mu_X} \qquad \text{or} \qquad R_2 = Y_i \frac{X_i}{\mu_X} + T_i$$

Where X and T are independent random variables with mean μ_X and μ_T respectively and variance σ_X^2 and σ_T^2 in the same way. Both random variables are known to the researcher. Hence, the report of i-th respondent will be given by:

$$Q_i = \begin{cases} 1; \text{ the i} - \text{th respondent reports with } R_1 \\ 0; \text{ the i} - \text{th respondent reports with } R_2 \end{cases}$$

and is modeled by

$$Z_i = Q_i\, Y_i \frac{X_i}{\mu_X} + (1 - Q_i)\left[Y_i \frac{X_i}{\mu_X} + T_i\right]$$

which, due to the complex scrambling of the sensitive value Y, it would be "difficult" for the interviewer to deduce the sensitive value Y of the respondent, in addition, that he is also not aware of with which report $R_i (i = 1,2)$ scrambled his value.

With the report Z_i and using SRSWR to select a sample *s* of size *n* from a population *U*, it is of interest to know the population characteristics of the sensitive value *Y*. Below in the following lemmas we present the properties of reports R_1, R_2 and of the model Z.

Lemma 2.1. The estimator of the mean of Y under R_1 is $\bar{Y}_{(R_1)} = \bar{R}_1$ and with variance $V[\bar{R}_1] = \frac{1}{n}[\sigma_Y^2(1 + CV_X^2) + CV_X^2 \mu_Y^2]$, where $CV_X^2 = \frac{\sigma_X^2}{\mu_X^2}$.

Proof.

The conditional expectation of R_1 on the model is $E_{R_1}(R_{1i}|i) = E_{R_1}\left(Y_i \frac{X_i}{\mu_X}|i\right) = Y_i$ and the conditional variance of R_1 under the model is $V_{R_1}(R_{1i}|i) = V_{R_1}\left(Y_i \frac{X_i}{\mu_X}|i\right) = \frac{Y_i^2}{\mu_X^2}\sigma_X^2$. Then R_1 is $E[R_{1i}|i] = E_d\left[E_{R1}\left(Y_i \frac{X_i}{\mu_X}|i\right)\right] = E_d[Y_i] = \mu_Y$, hence, $\bar{R}_1$ is an adequate estimator of $\bar{Y}_{R_1}$.

Unbiasedness of the estimator in R_1

$E[\bar{R}_1] = E_d\left[\frac{1}{n}\sum_{i\in s} E_{R_1}(R_{1i}|i)\right] = E_d\left[\frac{1}{n}\sum_{i\in s} E_{R_1}\left(Y_i\frac{X_i}{\mu_X}|i\right)\right] = E_d\left[\frac{1}{n}\sum_{i\in s} Y_i\right] = \frac{1}{n}\sum_{i\in s} E_d[Y_i] = \mu_Y,$ with this the unbiasedness of the estimator is demonstrated.

Variance of the estimator

$V[\bar{R}_1] = V_d\left[\frac{1}{n}\sum_{i\in s} E_R(R_{1i})\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s} V_R(R_{1i})\right] = V_d\left[\frac{1}{n}\sum_{i\in s} Y_i\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s}\frac{Y_i^2}{\mu_X^2}\sigma_X^2\right] = \frac{1}{n^2}\sum_{i\in s} V_d(Y_i) + \frac{\sigma_X^2}{n^2\mu_X^2}\sum_{i\in s} E_d(Y_i^2) = \frac{1}{n}\sigma_Y^2 + \frac{\sigma_X^2}{n\,\mu_X^2}(\sigma_Y^2+\mu_Y^2) = \frac{1}{n}\left[\sigma_Y^2 + \frac{\sigma_X^2\sigma_Y^2}{\mu_X^2} + \frac{\sigma_X^2\mu_Y^2}{\mu_X^2}\right] = \frac{1}{n}[\sigma_Y^2(1+CV_X^2) + CV_X^2\mu_Y^2].$

Then the lemma is proved.

Lemma 2.2. The estimator of the mean of Y in R_2 *is* $\bar{Y}_{(R_2)} = \bar{R}_2$-$\mu_T$ and with variance $V[\bar{R}_2 - \mu_T] = \frac{1}{n}[\sigma_Y^2(1+CV_X^2) + CV_X^2\mu_Y^2 + \sigma_T^2]$, where $CV_X^2 = \frac{\sigma_X^2}{\mu_X^2}$.

Proof.

The expectation of R_2 under the model is $E_{R_2}(R_{(2)\,i}|i) = E_{R_2}\left(\left(Y_i\frac{X_i}{\mu_X} + T_i\right)|i\right) = Y_i + \mu_T$, the variance of R_2under the model is $V_{R_2}(R_{(2)i}|i) = V_{R_2}\left(\left(Y_i\frac{X_i}{\mu_X} + T_i\right)|i\right) = \frac{Y_i^2}{\mu_X^2}\sigma_X^2 + \sigma_T^2$ and conditional expectation of R_2 is $E[R_{(2)i}|i] = E_d\left[E_{R_2}\left(\left(Y_i\frac{X_i}{\mu_X} + T_i\right)|i\right)\right] = E_d[Y_i] + E_d[\mu_T] = \mu_Y + \mu_T$, hence, $\bar{R}_2$-μ_T is a good estimator of μ_Y.

Unbiasedness of the estimator in R_2

$E[\bar{R}_2 - \mu_T] = E_d\left[\frac{1}{n}\sum_{i\in s} E_{R_2}(R_{(2)i}|i)\right] - E_d[E_{R_2}(\mu_T|i)] = E_d\left[\frac{1}{n}\sum_{i\in s} E_{R_2}(Y_i\frac{X_i}{\mu_X} + T_i)\right] - \mu_T = E_d\left[\frac{1}{n}\sum_{i\in s} Y_i + \mu_T\right] - \mu_T = \mu_Y + \mu_T - \mu_T = \mu_Y$

Variance of the estimator

$V[\bar{R}_2] = V_d\left[\frac{1}{n}\sum_{i\in s} E_R(R_{2i})\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s} V_R(R_{2i})\right] = V_d\left[\frac{1}{n}\sum_{i\in s} Y_i + \mu_T\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s}\frac{Y_i^2}{\mu_X^2}\sigma_X^2 + \sigma_T^2\right] = \frac{1}{n^2}\sum_{i\in s} V_d(Y_i) + \frac{1}{n^2}\sum_{i\in s}\left[\frac{\sigma_X^2}{\mu_X^2}E_d(Y_i^2) + E_d(\sigma_T^2)\right] = \frac{1}{n}\sigma_Y^2 + \frac{1}{n}\left[\left(\frac{\sigma_X^2}{\mu_X^2}(\sigma_Y^2+\mu_Y^2)\right) + \sigma_T^2\right] = \frac{1}{n}\left[\sigma_Y^2 + \frac{\sigma_X^2\sigma_Y^2}{\mu_X^2} + \frac{\sigma_X^2\mu_Y^2}{\mu_X^2} + \sigma_T^2\right] = \frac{1}{n}[\sigma_Y^2(1+CV_X^2) + CV_X^2\mu_Y^2 + \sigma_T^2]$

Then the lemma is proved.

Lastly, we present the report Z, where the properties of the parameter Y are given by the following lemma.
Lemma 2.3. The *Z* report has the following characteristics:

i) $\hat{\mu}_Y = \bar{Z} - \mu_T(1-Q)$, which is the estimator of the population mean of Y.

ii) $V[\hat{\mu}_Y] = \frac{1}{n}\left[\sigma_Y^2 + Q^2\,CV_X^2(\sigma_Y^2+\mu_Y^2) + (1-Q)^2\left((\sigma_Y^2+\mu_Y^2)CV_X^2 + \sigma_T^2\right)\right]$, which is the variance of the estimator, where $CV_X^2 = \frac{\sigma_X^2}{\mu_X^2}$.

iii) $\hat{V}[\hat{\mu}_Y] = \frac{1}{n}\left[\hat{\sigma}_Y^2 + Q^2\,CV_X^2(\hat{\sigma}_Y^2+\hat{\mu}_Y^2) + (1-Q)^2\left((\hat{\sigma}_Y^2+\hat{\mu}_Y^2)CV_X^2 + \sigma_T^2\right)\right]$, which is the estimator of the variance, where $\hat{\sigma}_Y^2 = \frac{S_z^2 - [Q^2\,CV_X^2\,\hat{\mu}_Y^2 + (1-Q)^2(\hat{\mu}_Y^2 CV_X^2 + \sigma_T^2)]}{[1+CV_X^2\,(Q^2+(1-Q)^2)]}$, and $S_z^2 = \frac{\sum_{i\in s}(z_i-\bar{z})^2}{n-1}$

Proof.

The expectation of Z_i under the model is $E_{Z_i}[Z_i|i] = E_{R_1}\left[\left(Q\, Y_i \frac{X_i}{\mu_X}\right)|i\right] + E_{R_2}\left[\left((1-Q)\left(Y_i\frac{X_i}{\mu_X} + T_i\right)\right)|i\right] = Q\, Y_i + (1-Q)(Y_i + \mu_T) = Y_i + \mu_T(1-Q)$. The variance of Z_i under the model is $V_{Z_i}[Z_i|i] = V_{R_1}\left[Q\left(Y_i\frac{X_i}{\mu_X}\right)|i\right] + V_{R_2}\left[(1-Q)\left(Y_i\frac{X_i}{\mu_X} + T_i\right)|i\right] = Q^2\,\frac{Y_i^2}{\mu_X^2}\,\sigma_X^2 + (1-Q)^2\left[\frac{Y_i^2}{\mu_X^2}\,\sigma_X^2 + \sigma_T^2\right] = Q^2\,(Y_i^2 CV_X^2) + (1-Q)^2[Y_i^2 CV_X^2 + \sigma_T^2]$, where $CV_X^2 = \frac{\sigma_X^2}{\mu_X^2}$

The conditional expectation of Z_i is $E[Z_i|i] = E_d\left\{E_{Z_i}\left[\left[Q\, Y_i\frac{X_i}{\mu_X} + (1-Q)\left(Y_i\frac{X_i}{\mu_X} + T_i\right)\right]|i\right]\right\} = E_d[Y_i + \mu_T(1-Q)] = \mu_Y + \mu_T(1-Q)$, hence, $\bar{Z} - \mu_T(1-Q)$ is the estimator of μ_Y.

Unbiasedness of the estimator in Z_i

$E[\hat{\mu}_Y] = E_d\left[E_{Z_i}\left(\bar{Z} - \mu_T(1-Q)\right)|i\right] = E_d\left[\frac{1}{n}\sum_{i\in s} E_{Z_i}(Z_i|i)\right] - \mu_T(1-Q) = \frac{1}{n}\sum_{i\in s} E_d[Y_i + \mu_T(1-Q)] - \mu_T(1-Q) = \mu_Y + \mu_T(1-Q) - \mu_T(1-Q) = \mu_Y$

Variance of the estimator

$V[\hat{\mu}_Y] = V[\bar{Z}] = V_d\left[\frac{1}{n}\sum_{i\in s}\left(E_{Z_i}(Z_i|i)\right)\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s}\left(V_{Z_i}(Z_i|i)\right)\right] = V_d\left[\frac{1}{n}\sum_{i\in s} Y_i + \mu_T(1-Q)\right] + E_d\left[\frac{1}{n^2}\sum_{i\in s}[Q^2\,(Y_i^2 CV_X^2) + (1-Q)^2(Y_i^2 CV_X^2 + \sigma_T^2)]\right] = \frac{1}{n^2}\sum_{i\in s} V_d(Y_i) + \frac{1}{n^2}\sum_{i\in s}[Q^2\,(E_d[Y_i^2]CV_X^2) + (1-Q)^2(E_d[Y_i^2]CV_X^2 + \sigma_T^2)] = \frac{1}{n}\left[\sigma_Y^2 + Q^2\, CV_X^2(\sigma_Y^2 + \mu_Y^2) + (1-Q)^2\left((\sigma_Y^2 + \mu_Y^2)CV_X^2 + \sigma_T^2\right)\right]$

Estimator of the variance

The natural estimator for the variance is: $\hat{V}[\hat{\mu}_Y] = \frac{1}{n}\left[\hat{\sigma}_Y^2 + Q^2\, CV_X^2(\hat{\sigma}_Y^2 + \hat{\mu}_Y^2) + (1-Q)^2\left((\hat{\sigma}_Y^2 + \hat{\mu}_Y^2)CV_X^2 + \sigma_T^2\right)\right]$, where, $\hat{\sigma}_Y^2 = \frac{S_z^2 - [Q^2\, CV_X^2\,\hat{\mu}_Y^2 + (1-Q)^2(\hat{\mu}_Y^2 CV_X^2 + \sigma_T^2)]}{[1 + CV_X^2\,(Q^2 + (1-Q)^2)]}$, and $S_z^2 = \frac{\sum_{i\in s}(z_i - \bar{z})^2}{n-1}$

Then the lemma is proved.

3. STRATIFIED MODEL EXTENSION

In the previous section, under simple random sampling with replacement, we present the characteristics of the estimators of the mean of Y for the reports R_1, R_2 and the model Z, which together make up the new proposed RR technique. Here, in the same way, we present the characteristics of the estimators, under stratified random sampling with replacement (SSRSWR). In a stratified design, the population *U* of size *N* is divided into *H* strata, where $\sum_{i=1}^{H} N_i = N$. Using a random draw, in each of the strata L a sample is chosen in such a way that the sizes n_i of the samples s_i satisfy $\sum_{i=1}^{H} n_i = n$.

The reason for doing this extension from simple random to stratified is to compare the efficiency and precision of both strategies, since theoretically the estimators under stratification are more precise and have minimum variance compared to those with simple random sampling.

3.1. R_1 with SSRSWR

For R_1 and the following reports, using stratified random sampling, the population *U* is divided into *L* strata. Therefore, in the R_1 report, each individual *i* in stratum *h*, must general a random value X_{hi} with probability $P[X_{hi}] = \theta_{hi}$, mean μ_{hX} and variance σ_{hX}^2, both parameters defined by the researcher. The respondent's response *i* in stratum *h* is scrambled by:

$$R_{(1)hi} = Y_{hi}\frac{X_{hi}}{\mu_{hX}}$$

As the previous section, in the following lemma we define the estimator and its properties of R_1 under the SSRSWR design.

Lemma 3.1. For R_1 and using SSRSWR, an estimator of the mean of Y per stratum is $\bar{Y}_{h\,(R_1)} = \bar{R}_{(1)h}$ and its stratified global estimator $\bar{Y}_{ST,R_1} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_1)}$. The variance of the estimator per stratum is $V(\bar{Y}_{h\,(R_1)}) = \frac{1}{n_h}[\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2]$ and the variance of the global estimator is given by $V(\bar{Y}_{ST,R_1}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}{n_h}$

Proof

The expectation and variance of R_1 under the model are $E_{R_1}(R_{(1)hi}|i) = E_{R_1}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}}|i\right) = Y_{hi}$ and $V_{R_1}(R_{(1)hi}|i) = V_{R_1}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}}|i\right) = \frac{Y_{hi}^2}{\mu_{hX}^2}\sigma_{hX}^2$, respectively. The conditional expectation of R_1 is $E[R_{(1)hi}|i] = E_d\left[E_{R1}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}}|i\right)\right] = E_d[Y_{hi}] = \mu_{hY}$, and $\bar{R}_{(1)h}$ is the proposed estimator of $\bar{Y}_{h\,(R_1)}$.

Unbiasedness of the estimator per stratum in R_1

$E[\bar{R}_{(1)h}] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_1}(R_{(1)hi}|i)\right] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_1}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}}|i\right)\right] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}\right] = E_d[Y_{hi}] = \mu_{hY}$. With this, the unbiasedness of the estimator for each stratum is proved.

Stratified global estimator

Using theorem 5.1 in Cochran (1971) and the stratum mean $\bar{Y}_{h\,(R_1)}$, the global estimator for the SSRSWR is $\bar{Y}_{ST,R_1} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_1)}$

Variance of the estimator per stratum

$V[\bar{R}_{(1)h}] = V_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_1}(R_{(1)hi}|i)\right] + E_d\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h} V_{R_1}(R_{(1)hi}|i)\right] = V_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}\right] +$
$E_d\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h}\frac{Y_{hi}^2}{\mu_{hX}^2}\sigma_{hX}^2\right] = \frac{1}{n_h^2}\sum_{i=1}^{n_h} V_d(Y_{hi}) + \frac{\sigma_{hX}^2}{n_h^2\,\mu_{hX}^2}\sum_{i=1}^{n_h} E_d(Y_{hi}^2) = \frac{1}{n_h}\sigma_{hY}^2 + \frac{\sigma_{hX}^2}{n_h\,\mu_{hX}^2}(\sigma_{hY}^2+\mu_{hY}^2) =$
$\frac{1}{n_h}\left[\sigma_{hY}^2 + \frac{\sigma_{hX}^2\sigma_{hY}^2}{\mu_{hX}^2} + \frac{\sigma_{hX}^2\mu_{hY}^2}{\mu_{hX}^2}\right] = \frac{1}{n_h}[\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2]$

Variance of the global estimator. Using the variance of the estimator by stratum $V[\bar{Y}_{h\,(R_1)}]$ and Cochran's theorem 5.2. which is $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\, V(\bar{y}_h)$, where $\bar{y}_h$ must be an unbiased estimator of $\bar{Y}_h$, which has already been proven and the independent sample, applying the previous theorem we have:

$$V(\bar{Y}_{ST,R_1}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\, V(\bar{Y}_{h\,(R_1)}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}{n_h}$$

Then the lemma is proved.

3.2. R_2 with SSRSWR

In this R_2 report, the *i*-th respondent in the *h*-th stratum must generate two values of two random variables. The first is X_{hi} with probability $P[X_{hi}] = \theta_{hi}$, mean μ_{hX} and variance σ_{hX}^2, and the second is T_{hi} with

probability $P[T_{hi}] = \delta_{hi}$, mean μ_{hT} and variance σ_{hT}^2, parameters known by the researcher for both variables. The response of the *i*-th respondent in the *h*-th stratum is generated by:

$$R_{(2)hi} = Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}$$

Lemma 3.2. For R_2 and using SSRSWR, an estimator of the mean of Y per stratum is $\bar{Y}_{h\,(R_2)} = \bar{R}_{(2)h}\text{-}\mu_{hT}$ and its stratified global estimator is $\bar{Y}_{ST,R_2} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_2)}$. The variance of the estimator per stratum is $V(\bar{Y}_{h\,(R_2)}) = \frac{1}{n_h}[\sigma_{hY}^2(1 + CV_{hX}^2) + CV_{hX}^2\,\mu_{hY}^2 + \sigma_{hT}^2]$ and the variance of the global estimator is given by $V(\bar{Y}_{ST,R_2}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\,\mu_{hY}^2+\sigma_{hT}^2)}{n_h}$

Proof

The expectation of R_2 under the model is $E_{R_2}(R_{(2)hi}|i) = E_{R_2}\left(\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)|i\right) = Y_{hi} + \mu_{hT}$ and its variance is $V_{R_2}(R_{(2)hi}|i) = V_{R_2}\left(\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)|i\right) = \frac{Y_{hi}^2}{\mu_{hX}^2}\,\sigma_{hX}^2 + \sigma_{hT}^2$. The conditional expectation of R_2 is $E[R_{(2)i}|i] = E_d\left[E_{R_2}\left(\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)|i\right)\right] = E_d[Y_{hi}] + E_d[\mu_{hT}] = \mu_{hY} + \mu_{hT}$, hence, $\bar{R}_{(2)h}\text{-}\mu_{hT}$ is the estimator of μ_{hY}.

Unbiasedness of the estimator per stratum in R_2

With the next procedure we proof the unbiasedness of the estimator per stratum. $E[\bar{R}_{(2)h} - \mu_{hT}] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_2}(R_{(2)hi}|i)\right] - E_d[E_{R_2}(\mu_{hT}|i)] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_2}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_i|i\right)\right] - \mu_{hT} = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi} + \mu_{hT}\right] - \mu_T = \mu_Y + \mu_T - \mu_T = \mu_{hY}$

Stratified global estimator

As the above report, we have in this R_2 report the mean per stratum $\bar{Y}_{h\,(R_2)}$, so that, the global estimator for SSRSWR design is $\bar{Y}_{ST,R_2} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_2)}$

Variance of the estimator per stratum

$V[\bar{R}_{(2)h}] = V_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{R_2}(R_{(2)hi}|i)\right] + E_d\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h} V_{R_2}(R_{(2)hi}|i)\right] = V_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi} + \mu_{hT}\right] + E_d\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h}\frac{Y_{hi}^2}{\mu_{hX}^2}\,\sigma_{hX}^2 + \sigma_{hT}^2\right] = \frac{1}{n_h^2}\sum_{i=1}^{n_h} V_d(Y_{hi}) + \frac{1}{n_h^2}\sum_{i=1}^{n_h}\left[\frac{\sigma_{hX}^2}{\mu_{hX}^2}\,E_d(Y_{hi}^2) + E_d(\sigma_{hT}^2)\right] = \frac{1}{n_h}\sigma_{hY}^2 + \frac{1}{n_h}\left[\left(\frac{\sigma_{hX}^2}{\mu_{hX}^2}(\sigma_{hY}^2 + \mu_{hY}^2)\right) + \sigma_{hT}^2\right] = \frac{1}{n_h}[\sigma_{hY}^2(1 + CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2 + \sigma_{hT}^2]$

Variance of the global estimator. In the same way as the previous report, we know the variance per stratum $V[\bar{Y}_{h\,(R_2)}]$, therefore, the variance of the global estimator is:

$$V(\bar{Y}_{ST,R_2}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\,V(\bar{Y}_{h\,(R_2)}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2\,(\sigma_{hY}^2(1 + CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2 + \sigma_{hT}^2)}{n_h}.$$

Then the lemma is proved.

3.3. Z procedure with SSRSWR

As a last extension to stratified, we present the Z report. The objective of this report is scrambling the sensitive value of the respondent, in a random way, through either the report R_1 with probability Q_h or R_2

report with probability $(1-Q_h)$. That is, the *i*-th respondent in stratum *h* performs a Bernoulli try and will report his sensitive value Y by:

$$Z_{hi} = \begin{cases} R_{(1)hi} & if \quad \alpha_{hi} = 1 \\ R_{(2)hi} & if \quad \alpha_{hi} = 0 \end{cases}$$

the Z report is modeled by

$$Z_{hi} = Q_{hi}\, R_{(1)hi} + (1-Q_{hi})\, R_{(2)hi}$$

$$= Q_{hi}\, Y_{hi} \frac{X_{hi}}{\mu_{hX}} + (1-Q_{hi}) \left[Y_{hi} \frac{X_{hi}}{\mu_{hX}} + T_{hi} \right]$$

The following lemma presents the properties of the population mean of the sensible value Y.

Lemma 3.3. The *Z* report has the following characteristics:

i) $\hat{\mu}_{hY} = \bar{Z}_h - \mu_{hT}(1-Q_{hi})$, which is the estimator of the mean of Y per stratum.

ii) $\bar{Y}_{ST,Z} = \frac{1}{N} \sum_{h=1}^{L} N_h\, \hat{\mu}_{hY}$, is the global stratified estimator of the population mean.

iii) $V[\hat{\mu}_{hY}] = \frac{1}{n_h}\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]$, which is the variance of the estimator per stratum, where $CV_{hX}^2 = \frac{\sigma_{hX}^2}{\mu_{hX}^2}$.

iv) $\hat{V}[\hat{\mu}_{hY}] = \frac{1}{n_h}\left[\hat{\sigma}_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\hat{\sigma}_{hY}^2 + \hat{\mu}_{hY}^2) + (1-Q_{hi})^2\left((\hat{\sigma}_{hY}^2 + \hat{\mu}_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]$, is the proposed estimator for the variance per stratum, where $\hat{\sigma}_{hY}^2 = \frac{S_{hz}^2 - \left(Q_{hi}^2\, CV_{hX}^2\, \hat{\mu}_{hY}^2 + (1-Q_{hi})^2(\hat{\mu}_{hY}^2 CV_{hX}^2 + \sigma_{hT}^2)\right)}{\left[1 + CV_{hX}^2\,(Q_{hi}^2 + (1-Q_{hi})^2\,)\right]}$, $S_{hz}^2 = \frac{\sum_{i=1}^{n_h}(z_{hi} - \bar{z}_h)^2}{n_h - 1}$

v) $V\left(\bar{Y}_{ST,Z}\right) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h}$, is the global variance of the estimator … (3.3.1)

vi) $\hat{V}\left(\bar{Y}_{ST,Z}\right) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\left[\hat{\sigma}_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\hat{\sigma}_{hY}^2 + \hat{\mu}_{hY}^2) + (1-Q_{hi})^2\left((\hat{\sigma}_{hY}^2 + \hat{\mu}_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h}$, which is a naive estimator of $V\left(\bar{Y}_{ST,Z}\right)$

vii) $V_{pro}\left(\bar{Y}_{ST,Z}\right) = \frac{1}{n}\sum_{h=1}^{L} W_h\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]$, is the proportional variance of the global estimator … (3.3.2)

viii) $V_{opt}(\bar{Y}_{ST,Z}) = \frac{1}{nN^2}\left(\sum_{h=1}^{L} N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}\right)^2$, optimal variance of the global estimator given a fixed n … (3.3.3)

Proof

The expectation of Z_{hi} under the model is $E_{Z_{hi}}[Z_{hi}|i] = E_{R_1}\left[\left(Q_{hi}\, Y_{hi} \frac{X_i}{\mu_X}\right)|i\right] + E_{R_2}\left[\left((1-Q_{hi})\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)\right)|i\right] = Q_{hi}\, Y_{hi} + (1-Q_{hi})\,(Y_{hi} + \mu_{hT}) = Y_{hi} + \mu_{hT}(1-Q_{hi})$ and its variance is $V_{Z_{hi}}[Z_{hi}|i] = V_{R_1}\left[Q_{hi}\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}}\right)|i\right] + V_{R_2}\left[(1-Q_{hi})\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)|i\right] = Q_{hi}^2\,(Y_{hi}^2 CV_{hX}^2) + (1-Q_{hi})^2[Y_{hi}^2 CV_{hX}^2 + \sigma_{hT}^2]$, where $CV_{hX}^2 = \frac{\sigma_{hX}^2}{\mu_{hX}^2}$. Also, we have the conditional expectation of Z_{hi} which is $E[Z_{hi}|i] = E_d\left\{E_{Z_{hi}}\left[\left[Q_{hi}\, Y_{hi}\frac{X_{hi}}{\mu_{hX}} + (1-Q_{hi})\left(Y_{hi}\frac{X_{hi}}{\mu_{hX}} + T_{hi}\right)\right]|i\right]\right\} = E_d[Y_{hi} + \mu_{hT}(1-Q_{hi})] = \mu_{hY} + \mu_{hT}(1-Q_{hi})$, therefore, $\bar{Z}_h - \mu_{hT}(1-Q_{hi})$ is the estimator of μ_{hY}.

Unbiasedness of the estimator in stratum *h* for Z_{hi}.

$E[\hat{\mu}_{hY}] = E_d\left[E_{Z_{hi}}\left(\bar{Z}_h - \mu_{hT}(1-Q_{hi})\right)|i\right] = E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{Z_{hi}}(Z_{hi}|i)\right] - \mu_{hT}(1-Q_{hi}) = E_d[Y_{hi} + \mu_{hT}(1-Q_{hi})] - \mu_{hT}(1-Q_{hi}) = \mu_{hY} + \mu_{hT}(1-Q_{hi}) - \mu_{hT}(1-Q_{hi}) = \mu_{hY}.$ With this, the unbiasedness of the estimator by stratum is demonstrated.

Stratified global estimator. Knowing the mean per stratum, we have as a global estimator for the SSRWR design is $\bar{Y}_{ST,Z} = \frac{1}{N}\sum_{h=1}^{L} N_h\, \hat{\mu}_{hY}$

Variance of the estimator per stratum

$V[\hat{\mu}_{hY}] = V[\bar{Z}_h] = V_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} E_{Z_{hi}}(Z_{hi}|i)\right] + E_d\left[\frac{1}{n_h}\sum_{i=1}^{n_h} V_{Z_{hi}}(Z_{hi}|i)\right] = V_d\left[\frac{1}{n_h}\sum_{i)1}^{n_h} Y_{hi} + \mu_{hT}(1-Q_{hi})\right] + E_d\left[\frac{1}{n_h^2}\sum_{i=1}^{n_h}[Q_{hi}^2\,(Y_{hi}^2 CV_{hX}^2) + (1-Q_{hi})^2(Y_{hi}^2 CV_{hX}^2 + \sigma_{hT}^2)]\right] = \frac{1}{n_h^2}\sum_{i=1}^{n_h} V_d(Y_{hi}) + \frac{1}{n_h^2}\sum_{i=1}^{n_h}[Q_{hi}^2\,(E_d[Y_{hi}^2]CV_{hX}^2) + (1-Q_{hi})^2(E_d[Y_{hi}^2]CV_{hX}^2 + \sigma_{hT}^2)] = \frac{1}{n_h}\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]$

Variance of the global estimator. The deduction of the variance of the global estimator is the same as the previous ones.

$V\left(\bar{Y}_{ST,Z}\right) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\, V(\hat{\mu}_{hY}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h}$.

Then the lemma is proved.

4. OPTIMAL ALLOCATION AND GAINS IN ACCURACY OF MODEL FOR SSRSWR

The researcher, when performing survey sampling with a *p* design, will depend on how robust it can be and the budget allocated to carry it out. To help solving these possible problems and using SSRSWR, it is necessary to determine the best sample size in stratum *h* to minimize the variance (*V*) given a fixed cost ($C = c_0 + \sum c_h\, n_h$) (4.1) or minimize a cost (*C*) given a fixed variance (*V*). This is known as the optimal allocation of the n_h and *n*.

4.1. n_h and *n* optimal for $V\left(\bar{Y}_{ST,R_1}\right)$

Lemma 4.1. Using SSRSWR and a simple cost function $= c_0 + \sum c_h\, n_h$, the variance of the estimator of the population mean of procedure R_1 is a minimized when $n_h \propto N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}$

Proof

We must minimize $V\left(\bar{Y}_{ST,R_1}\right) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2)}{n_h}$, subject to $C = c_0 + \sum c_h\, n_h$.

Using the Lagrange multipliers method, we choose n_h and the multiplier λ to minimize:

$f(y,\lambda) = V\left(\bar{Y}_{ST,R_1}\right) + \lambda\,(\sum c_h\, n_h - C + c_0) = \sum_{h=1}^{L} \frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2)}{N^2\ n_h} + \lambda\,(c_1 n_1 + \cdots + c_L n_L - C + c_0)$.

Partially differentiating $f(y,\lambda)$ with respect to $n_h's$, h=1, 2, …, L, are

$\frac{\partial f(y,\lambda)}{\partial n_1} = -\frac{N_1^2\,(\sigma_{1Y}^2(1+CV_{1X}^2) + CV_{1X}^2\mu_{1Y}^2)}{N^2 n_1^2} + \lambda c_1$, … , $\frac{\partial f(y,\lambda)}{\partial n_L} = -\frac{N_L^2\,(\sigma_{LY}^2(1+CV_{LX}^2) + CV_{LX}^2\mu_{LY}^2)}{N^2 n_L^2} + \lambda c_L$;

we have: $-\frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2) + CV_{hX}^2\mu_{hY}^2)}{N^2 n_h^2} + \lambda c_h = 0$, for h=1,2, …, L. Working the previous expression:

$$\lambda c_h = \frac{N_h^2\,(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}{N^2 n_h^2} \Rightarrow \sqrt{\lambda}\sqrt{c_h} = \frac{\sqrt{N_h^2}\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}}{\sqrt{N^2}\sqrt{n_h^2}}$$

$$\Rightarrow n_h\sqrt{\lambda} = \frac{N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}}{N\,\sqrt{c_h}} \ldots (4.1.1)$$

summing in the L strata in (4.1.1), $\sum_{h=1}^{L} n_h\sqrt{\lambda} = \sum_{h=1}^{L} \frac{N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}}{N\,\sqrt{c_h}} \Rightarrow n\sqrt{\lambda} = \sum_{h=1}^{L} \frac{N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}}{N\,\sqrt{c_h}} \ldots (4.1.2)$

From (4.1.1) and (4.1.2) we have:

$$n_h = n\frac{N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}}{\sum_{h=1}^{L} N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}} \ldots (4.1.3)$$

hence

$$n_h \propto N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}$$

Then the previous lemma is proved.

To complete the optimal allocation, since in (4.1.3) n_h depends on *n*, we have to find the optimal *n* when there is a fixed cost *(C)*, so substituting n_h from (4.1.3) in the cost function (4.1) and working for *n*, it results:

$$n = \frac{(C-c_0)\sum_{h=1}^{L}\left(N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\sqrt{c_h}\right)}$$

On the other hand, if *V* is fixed we substitute n_h in $V\left(\bar{Y}_{ST,R_1}\right)$ and solving for *n* we have:

$$n = \frac{1}{N^2\,V\left(\bar{Y}_{ST,R_1}\right)}\sum_{h=1}^{L}\left[N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\sqrt{c_h}\right]$$

$$\sum_{h=1}^{L}\left[N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2)}\,\frac{1}{\sqrt{c_h}}\right]$$

And with this, the optimal allocation of n_h for R_1 using SSRSWR is completed.

4.2 n_h and *n* optimal for $V\left(\bar{Y}_{ST,R_2}\right)$

Lemma 4.2. Using SSRSWR and the cost function (4.1), the variance of the estimator of the population mean of the procedure R_2 is a minimum when $n_h \propto N_h\,\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}$

Proof.

The proof is similar to the previous method. The resulting proportion for R_2 is:

$$n_h = n\frac{N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}}{\sum_{h=1}^{L}N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}} \quad \ldots(4.2.1)$$

so that, $n_h \propto N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}$

n optimal when there is a fixed cost (*C*),

$$n = \frac{(C-c_0)\sum_{h=1}^{L}\left(N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\sqrt{c_h}\right)}$$

The *n* optimal when you have a fixed variance, is

$$n = \frac{1}{N^2\,V\left(\bar{Y}_{ST,R_2}\right)}\sum_{h=1}^{L}\left[N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\sqrt{c_h}\right]$$

$$\sum_{h=1}^{L}\left[N_h\sqrt{(\sigma_{hY}^2(1+CV_{hX}^2)+CV_{hX}^2\mu_{hY}^2+\sigma_{hT}^2)}\,\frac{1}{\sqrt{c_h}}\right]$$

And with this, the optimal allocation of n_h for R_2 using SSRSWR is completed.

4.3 n_h and *n* optimal for $V\left(\bar{Y}_{ST,Z}\right)$

Lemma 4.3. Using SSRSWR and the cost function (4.1), the variance of the estimator of the population mean of the procedure Z is a minima when $n_h \propto N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\frac{1}{\sqrt{c_h}}$

Proof.

It is proved, by following the reasoning of Lemma 4.1. The resulting proportion for Z is:

$$n_h = n\frac{N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\frac{1}{\sqrt{c_h}}}{\sum_{h=1}^{L}N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\frac{1}{\sqrt{c_h}}} \quad \ldots(4.3.1)$$

hence,

$$n_h \propto N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\frac{1}{\sqrt{c_h}}$$

n optimal when there is a fixed cost (*C*),

$$n = \frac{(C-c_0)\sum_{h=1}^{L}\left(N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\frac{1}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L}\left(N_h\sqrt{\left[\sigma_{hY}^2+Q_{hi}^2\,CV_{hX}^2(\sigma_{hY}^2+\mu_{hY}^2)+(1-Q_{hi})^2\left((\sigma_{hY}^2+\mu_{hY}^2)CV_{hX}^2+\sigma_{hT}^2\right)\right]}\,\sqrt{c_h}\right)}$$

n optimal when you have a fixed variance,

$$n = \frac{1}{N^2\, V(\bar{Y}_{ST,Z})} \sum_{h=1}^{L} \left[N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]} \sqrt{c_h} \right]$$

$$\sum_{h=1}^{L} \left[N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]} \frac{1}{\sqrt{c_h}} \right]$$

The above is the optimal allocation of n_h for Z using SSRSWR.

As and the last results, we substitute $n_h = \frac{nN_h}{N}$ in (3.3.1) to obtain the global proportional variance of the estimator $V_{pro}(\bar{Y}_{ST,Z})$, see (3.3.2). Also, if we substitute n_h of (4.3.1) with unitary cost in (3.3.1) for to get the variance minima with a *n* fix, we have:

$$V(\bar{Y}_{ST,Z}) = \frac{1}{N^2} \sum_{h=1}^{L} \left[\frac{N_h^2 \left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n \dfrac{N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}}{\sum_{h=1}^{L} N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}} \right]$$

$$= \frac{1}{nN^2} \left[\sum_{h=1}^{L} N_h \sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2 \left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]} \right]^2$$

4.4. Gains in accuracy

Next, we will prove the efficiency of using one variance over another as: $V(\hat{\mu}_Y)$ with $V(\bar{Y}_{ST,Z})$, $V(\bar{Y}_{ST,Z})$ with $V_{pro}(\bar{Y}_{ST,Z})$ and $V_{pro}(\bar{Y}_{ST,Z})$ with $V_{opt}(\bar{Y}_{ST,Z})$.

$\boldsymbol{V[\hat{\mu}_Y]}$ with $\boldsymbol{V(\bar{Y}_{ST,Z})}$

First, the variance under SRSWR is will developed to stratified.

Simplifying $V(\hat{\mu}_Y)$, we have $\frac{1}{n}\left[\sigma_Y^2 + \sigma_X^2\sigma_Y^2\mathrm{A} + \sigma_X^2\,\mu_Y^2\mathrm{A} + (1-Q)^2\sigma_T^2\right]$, where $\mathrm{A} = \left[\frac{Q^2}{\mu_X^2} + \frac{(1-Q)^2}{\mu_X^2}\right]$ owing to these variables are fixed by the researcher, then, we express the variance in its stratified form as next,

$\boldsymbol{V(\hat{\mu}_Y)} = \frac{1}{n}\left[\sigma_Y^2 + \sigma_X^2\sigma_Y^2\mathrm{A} + \sigma_X^2\,\mu_Y^2\mathrm{A} + (1-Q)^2\sigma_T^2\right] \Rightarrow nN\,V(\hat{\mu}_Y) =$

$= \sum_{i\in U}(Y_i - \mu_Y)^2 + \mathrm{A}\,\sigma_X^2\,\sum_{i\in U}(Y_i - \mu_Y)^2 + \mu_Y^2\,\mathrm{A}\,\sum_{i\in U}(X_i - \mu_X)^2 + (1-Q)^2\sum_{i\in U}(T_i -$
$\mu_T)^2 \sum_{h=1}^{L}\frac{1}{n}W_h\left(\mu_{T(h)} - \mu_T\right)^2 = \frac{1}{n}\sum_{h=1}^{L}W_h\left[\sigma_{Yh}^2 + \mathrm{A}\,\sigma_X^2\,\sigma_{Yh}^2 + \mu_Y^2\,\mathrm{A}\,\sigma_{Xh}^2 + (1-Q)^2\sigma_{Th}^2\right] +$
$\frac{1}{n}\sum_{h=1}^{L}W_h\left[\left(\mu_{Y(h)} - \mu_Y\right)^2 + \mathrm{A}\,\sigma_X^2\left(\mu_{Y(h)} - \mu_Y\right)^2 + \mu_Y^2\,\mathrm{A}\left(\mu_{X(h)} - \mu_X\right)^2 + (1-Q)^2\left(\mu_{T(h)} - \mu_T\right)^2\right]$

Gain in accuracy of $\boldsymbol{V(\hat{\mu}_Y)}$ with $\boldsymbol{V(\bar{Y}_{ST,Z})}$

$G\left[(V(\hat{\mu}_Y), V(\bar{Y}_{ST,Z})\right] = V(\hat{\mu}_Y) - V(\bar{Y}_{ST,Z})$

$$= \frac{1}{n}\left[\sigma_Y^2 + Q^2\, CV_X^2(\sigma_Y^2 + \mu_Y^2) + (1-Q)^2\left((\sigma_Y^2 + \mu_Y^2)CV_X^2 + \sigma_T^2\right)\right] - \sum_{h=1}^{L} \frac{W_h^2\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h}$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h\left[\sigma_{Yh}^2 + \mathrm{A}\,\sigma_X^2\,\sigma_{Yh}^2 + \mu_Y^2\,\mathrm{A}\,\sigma_{Xh}^2 + (1-Q)^2\sigma_{Th}^2\right] + \frac{1}{n}\sum_{h=1}^{L} W_h\left[\left(\mu_{Y(h)} - \mu_Y\right)^2 + \mathrm{A}\,\sigma_X^2\left(\mu_{Y(h)} - \mu_Y\right)^2 + \mu_Y^2\,\mathrm{A}\left(\mu_{X(h)} - \mu_X\right)^2 + (1-Q)^2\left(\mu_{T(h)} - \mu_T\right)^2\right] - \sum_{h=1}^{L} \frac{W_h^2\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h}$$

This expression will be positive if the means in the strata are heterogeneous. If this happens, it is recommended to use SSRSWR, otherwise the means are almost homogeneous between the strata, the squared difference of the above expressions will be close to zero and therefore, it is recommended to use SRSWR.

4.4.1 Gain in accuracy of $V(\bar{Y}_{ST,Z})$ with $V_{pro}(\bar{Y}_{ST,Z})$

From expressions (3.3.1) and (3.3.2) we have the stratified variance and the proportional stratified variance, respectively. The gain in accuracy is given by:

$$G\left[\left(V(\bar{Y}_{ST,Z}), V_{pro}(\bar{Y}_{ST,Z})\right)\right]$$

$$= \sum_{h=1}^{L} \frac{W_h^2\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n_h} - \sum_{h=1}^{L} \frac{W_h\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{n}$$

$= \sum_{h=1}^{L} \frac{W_h^2\,\ddot{\sigma}_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h\,\ddot{\sigma}_h^2}{n} = \sum_{h=1}^{L}\left[W_h\,\ddot{\sigma}_h^2\left(\frac{W_h}{n_h} - \frac{1}{n}\right)\right]$. The following inequality will hold $\frac{W_h}{n_h} > \frac{1}{n}$, whenever $\frac{n}{N} > \frac{n_h}{N_h}$, hence $V(\bar{Y}_{ST,Z}) > V_{pro}(\bar{Y}_{ST,Z})$.

4.4.2 Gain in accuracy of $V_{pro}(\bar{Y}_{ST,Z})$ with $V_{opt}(\bar{Y}_{ST,Z})$

Using the $V_{pro}(\bar{Y}_{ST,Z})$ in (3.3.2) and the $V_{opt}(\bar{Y}_{ST,Z})$ in (3.3.3), Following the usual procedure, the gain in precision is:

$$G\left[\left(V_{pro}(\bar{Y}_{ST,Z}), V_{opt}(\bar{Y}_{ST,Z})\right)\right]$$

$$= \frac{\sum_{h=1}^{L} N_h\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}{nN} - \frac{\left(\sum_{h=1}^{L} N_h\sqrt{\left[\sigma_{hY}^2 + Q_{hi}^2\, CV_{hX}^2(\sigma_{hY}^2 + \mu_{hY}^2) + (1-Q_{hi})^2\left((\sigma_{hY}^2 + \mu_{hY}^2)CV_{hX}^2 + \sigma_{hT}^2\right)\right]}\right)^2}{nN^2}$$

$= \frac{1}{nN}\left[\sum_{h=1}^{L} N_h\ddot{\sigma}_h^2 - \frac{\left(\sum_{h=1}^{L} N_h\ddot{\sigma}_h\right)^2}{N}\right] = \frac{1}{n}[\sum_{h=1}^{L} W_h\ddot{\sigma}_h^2 - \bar{\sigma}^2] = \frac{1}{n}\sum_{h=1}^{L} W_h(\ddot{\sigma}_h^2 - \bar{\sigma})^2$, hence, $V_{pro}(\bar{Y}_{ST,Z}) > V_{opt}(\bar{Y}_{ST,Z})$. Where $\bar{\sigma} = \sum_{h=1}^{L} N_h\ddot{\sigma}_h / N$

As a partial conclusion to using stratified sampling to minimize variance, it is better to use SSRSWR when the standard deviations are more different between strata.

5. A SIMULATION STUDY.

To evaluate and visualize the performance of the proposed estimators, both in SRSWR and SSRSWR, a simulation was carried out in terms of precision and efficiency. For the simulation, real data obtained from the National Survey of Victimization and Perception of Public Security (2021) developed by INEGI were considered, in which *Y* was chosen as a sensitive variable (question): "In terms of crime, tell me, How safe do you feel walking alone at night around your home?" with an ordinal scale of : Very Secure = 1, Secure = 2, Insecure = 3 y Very Insecure = 4, with $N = 80508$, $\mu = 2.678$ and $\sigma^2 = 0.642$. The selection of this sensitive variable was due to the national context of Mexico, where crime has a negative impact in both the social and economic aspects and, therefore, it is of interest to characterize this type of information. In SSRS, the total population was stratified using the age range of the respondents as a criterion, resulting in 7 strata. To evaluate the precision of the estimator of the mean of the sensitive variable *Y* we have (4.1), which is the ratio of the relative errors under SRSWR and SSRSWR. To evaluate the efficiency we present (4.2), which is the ratio of two estimators of the variance of the estimated mean. $Error\left[RE(d_j)/RE(d_k)\right]_s = \left[\left(\frac{|\hat{y}_j - \bar{Y}_j|}{\bar{Y}_j}\right)_{d_j} \Big/ \left(\frac{|\hat{y}_k - \bar{Y}_k|}{\bar{Y}_k}\right)_{d_k}\right]_s$ ... (4.1), where d is a design, $j \neq k$ and RE is the error relative with respect to μ_Y. $E[V_l/V_m]_s = \left(\frac{V(\bar{y}_l)}{V(\bar{y}_m)}\right)_s$ ... (4.2), where $l \neq m$.

For the simulation process, a sample size of *n*= 9081 was calculated, given a population sampling error for SRSWR. Fixing this sample size, the sample size n_h was calculated for each stratum proportionally following Cochran (1971) for the global variance (3.3.1) and the proportional global variance (3.3.2). From (4.3.1) the optimal sample for the optimal global variance (3.3.3) was calculated. A simulation of 1000 iterations was carried out. Following other works, Greenberg et al. (1971) and Bouza et al. (2022), similar values to the population values Y_i were set for the auxiliary variables X_i and T_i.Two simulations were performed to compare the accuracy and efficiency of the estimators doing greater use of the scrambling model R_1 or R_2. The first simulation was assigned probability *Q*=0.7 to select the scrambling model R_1 and the second simulation was assigned *Q*=0.3 to select the same model.

The simulated results of the statistics to evaluate the precision and efficiency of the proposed estimators are presented in Table 1 and Table 2.

	***Q*=0.7**	***Q*=0.3**
$Error\left(\frac{SRS}{SSRS}\right) =$	1.0611	1.0071

Table 1. Accuracy of the estimators of SRSWR and SSRSWR

	***Q*=0.7**	***Q*=0.3**
$E\left(\frac{V(SRS)}{V(SSRS)}\right) =$	0.60034	0.62126
$E\left(\frac{V(SSRS)}{V_{pro}(SSRS)}\right) =$	1.00896	1.00001
$E\left(\frac{V_{pro}(SSRS)}{V_{opt}(SSRS)}\right) =$	1.00019	1.00018

Table 2. Efficiency of the estimators of the mean

The precision of the estimators under each design is shown in Table 1, in which it is observed that when the probability of using the scrambling model R_1 is 0.7, it is more precise to use SSRS to estimate the population mean of Y. Similarly, when there is a greater probability of using R_2, it is more accurate to use SSRS than SRS. Between the models $R_1$and R_2, it is better to assign a higher probability to use R_1since the estimator is of higher precision than R_2. The results in Table 2 indicate that the variance of the estimator under SRS is smaller than the variance using SSRS regardless of which scrambling model is used. This could be explained with the sensitive variable Y used for the simulation, in which the population means of the strata are very similar, with the consequence that the use of SSRS loses efficiency. Finally, $V(SSRS) > V_{pro}(SSRS)$ and $V_{pro}(SSRS) > V_{opt}(SSRS)$, although with a very minimal difference.

To visualize the behavior of the estimators $\hat{\mu}_{y_s}$, $Error\left[RE(d_j)/RE(d_k)\right]_s$ and $E[V_l/V_m]_s$ were simulated under SRS sample sizes which were increased with $n = 250, 500, \ldots, 10000$. Using SSRS, the sample sizes for each stratum were proportional, resulting in a total sample with $n = 250, 506, \ldots, 10073$.

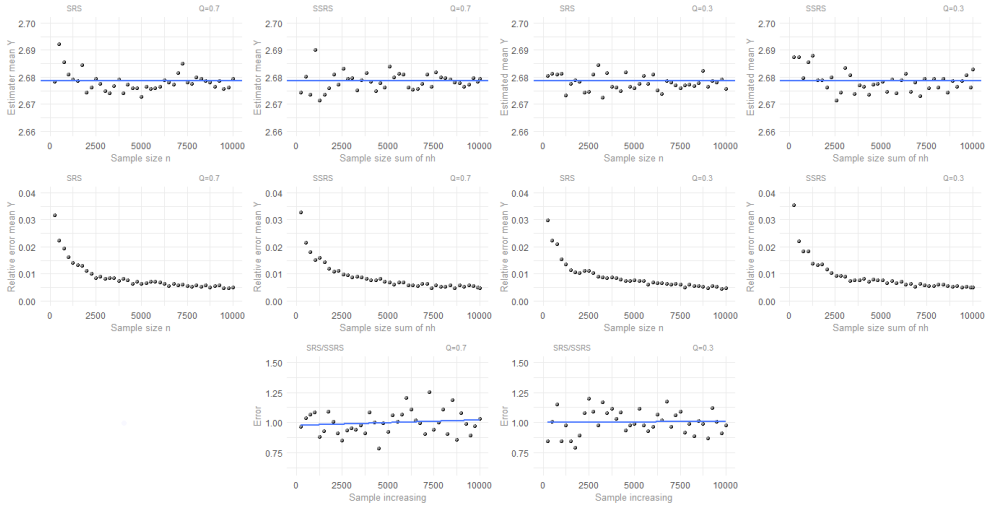

Figure 1. Estimators under SRS and SSRS, when Q=0.7 and Q=0.3

In the previous figure, the graphs presented in the first row show the values taken by the estimator of the mean of Y, when the sample increases when using SRS and SSRS for Q=0.7 or Q=0.3. In the second row, the graphs show is the relative error with respect to μ_Y under the same previous conditions. The last line shows the precision when comparing SRS with SSRS with the Error statistic (4.1) for Q=0.7 or Q=0.3. In these last graphs, a regression line is drawn with which we can visualize that the more the sample size grows, the more precise it is to use SSRS than SRS since the straight line exceeds one.

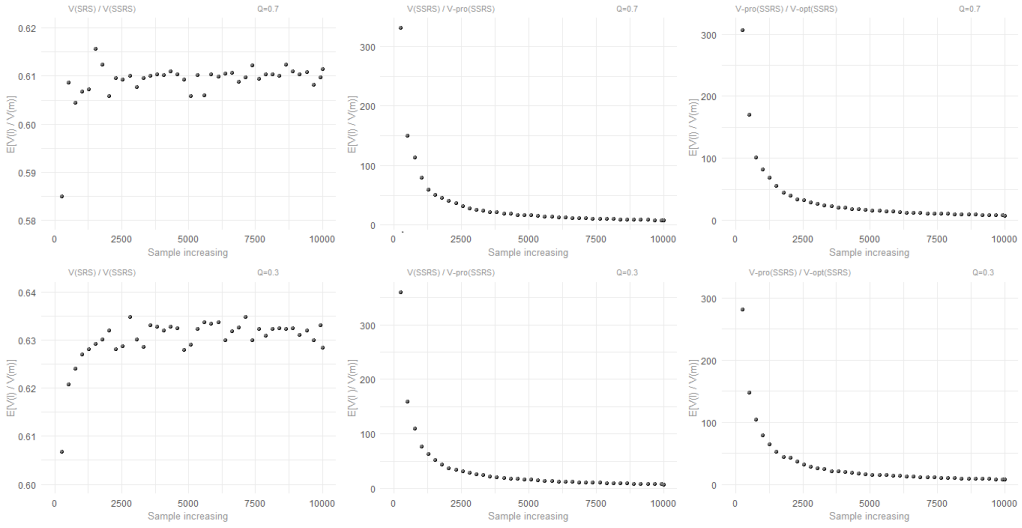

Figure 2. Comparation efficiency

To observe under which design and which report presents minimum variance, the ratios between $V(SRS)$, $V(SSRS)$, $V_{pro}(SSRS)$ and $V_{opt}(SSRS)$ were calculate, when the sample size n increased until it reached 10000. In the first column of Figure 2, whether assigning Q=0.7 or Q=0.3, $V(SRS)$ is smaller than $V(SSRS)$ as seen in the numerical results; in the graphs in column two, the smaller the sample, the smaller the variance when using $V_{pro}(SSRS)$ than $V(SSRS)$, but as the sample increases, the efficiency of $V_{opt}(SSRS)$ is reduced over $V_{pro}(SSRS)$. The same happens in the last graphic column, where $V_{opt}(SSRS)$ is smaller than $V_{pro}(SSRS)$.

As a conclusion of the simulation, it is better to use SRS when you have a small sample size or when in the stratification design the strata have homogeneous means. It is more accurate to use R_1, but if greater scrambling of the respondent's sensitive value Y is desired, it is better to use R_2.

RECEIVED: DECEMBER, 2023.
REVISED: SEPTEMBER, 2024

REFERENCES

[1] ABDEL-LATIF A., ABUL-ELA, BERNARD, GREENBERG and DANIEL G. HORVITZ. (1967): A Multi-Proportions Randomized Response Model. **Journal of the American Statistical Association**, 62, 319, 990-1008.

[2] ARNAB, R. (2018): Optional randomized response techniques for quantitative characteristics. **Communications in Statistics - Theory and Methods**, DOI: 10.1080/03610926.2018.1489554.

[3] ARNAB, R., and M. RUEDA. (2016): Optional randomized response: A critical review. **In Vol. 34 of Hand book of statistics**, ed. A. Chaudhuri, T. C. Christofides and C. R. Rao, 253–71. Oxford, UK: Elsevier.

[4] ARNAB, R. and SINGH, S. (2010): Randomized response techniques: An application to the Botswana AIDS impact survey, **J. Statist. Plann. Inference** 140, pp. 941–953.

[5] BOUZA, C.N. (2002): Estimation of the mean in ranked set sampling with non responses. **Metrika** 56: 171–179.

[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.

[7] BOUZA-HERRERA, C., N., JUÁREZ-MORENO, P.O., SANTIAGO-MORENO, A. and SAUTTO-VALLEJO, J.M. (2022): A Two-Stage Scrambling Procedure: Simple and Stratified Random Sampling. An Evaluation of COVID 19's data in Mexico. **Revista Investigación Operacional**. Vol 43, No. 4, 421-430.

[8] CHAUDHURI, A. and MUKHERJEE, R. (1988): **Randomized Response: Theory and Techniques**. Marcel Dekker, New York.

[9] CHAUDHURI, A., CHRISTOFIDES, T.C. and RAO, C.R. (2016): **Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits**. North Holland. Amsterdam.

[10] COCHRAN, W, G. (1971): **Técnicas de muestreo**. John Willey and Sons. Inc.

[11] ERIKSSON, S.A., (1973): A new model for randomized response. **Int. Stat. Rev**. 41, 40–43.

[12] GREENBERG, B. G., KUEBLER, R. R., JR., ABERNATHY, J. R., and HORVITZ, D. G. (1971): Application of the Randomized Response Technique in Obtaining Quantitative Data. **J Journal of the American Statistical Association**, 66, 334, 243-250.

[13] GUPTA, S., GUPTA, B., and SINGH, S. (2002): Estimation of sensitivity level of personal interview survey question. **Journal of Statistical Planning and Inference**. 100, 239-247.

[14] HORVITZ, D. G., SHAH, B. V. and SIMMONS, WALT, R. (1967): The unrelated question randomized response model. **Social Statistics Section Proceedings of the American Statistical Association**, 65-72.

[15] HUANG, K.C. (2004): A survey technique for estimating the proportion and sensitivity in dichotomous finite population**. Statistica Neerlandica** 58, 75-82.

[16] Juárez-Moreno, P.O., Bouza-Herrera, C. N., Santiago-Moreno, A. y Sautto-Vallejo, J.M. (2023). Procedures for scrambling sensitive quantitative variables: an updated review. **Investigación Operacional**, 44 (3), 438 – 449

[17] KRUMPAL, I. (2012): Estimating the prevalence of xenophobia and anti-semitism in Germany: A comparison of the randomized response technique and direct questioning. **Social Science Research**. 41, pp. 1387–1403.

[18] MURTAZA, M. SINGH, S. and HUSSAIN, Z. (2020): Use of correlated scrambling variables in quantitative randomized response technique. **Biometrical Journal**.1–14. DOI: 10.1002/bimj.201900137.

[19] NARJIS, G. and SHABBIR, J. (2021). An improved two-stage randomized response model for estimating the proportion of sensitive attribute. **Sociological Methods & Research**. 1-21. DOI: 10.1177/00491241211009950.

[20] NATIONAL SURVEY OF VICTIMIZATION AND PERCEPTION OF PUBLIC SECURITY. (2021): Instituto Nacional De Estadística Y Geografía. Available at: https://bit.ly/3P9qEmW. (Last consulted 06 July, 2022.).

[21] PAL, S., CHAUDHURI, A. and PATRA, D. (2020): How privacy may be protected in optional randomized response surveys. **Statistics in transition new series**, Vol. 21, No. 2, pp. 61–87.

[22] PERRI, P.F, PELLE, E. and STRANGES, M. (2016): Estimating induced abortion and foreign irregular presence using the randomized response crossed model. **Social Indicators Research** 129, pp. 601–618.

[23] RUEDA, M., COBO B., ARCOS, A. and ARNAB, R. (2016): Software for Randomized Response Techniques. **In Vol. 34 of Hand book of statistics**, ed. A. Chaudhuri, T. C. Christofides and C. R. Rao, 155–164. Oxford, UK: Elsevier.

[24] SINGH, G., N., SINGH, C. and SUMAN, S. (2020): Randomized response model to alter the nuisance effect of non-response due to stigmatized issues in survey sampling. **Journal of Statistical Computation and Simulation.** https://doi.org/10.1080/00949655.2020.1777295.
[25] STUBBE, J.H., CHORUS, A.M.J., FRANK, L. E., DE HON, O. and VAN DER HEIJDEN, P., G., M. (2013): Prevalence of use of performance enhancing drugs by fitness center members**. Drug test and analysis**, 6, pp. 434–438.
[26] WARNER, S.L. (1965): Randomized response: a survey technique for eliminating evasive answer bias. **Journal of the American Statistical Association**, 60, 63–69.