FORTHCOMING 62D05-04-22-03

A TWO-STAGE SCRAMBLING PROCEDURE: SIMPLE AND STRATIFIED RANDOM SAMPLING. AN EVALUATION OF COVID19`S DATA IN MEXICO

Carlos N. Bouza-Herrera*, Pablo O. Juárez-Moreno[1]**, Agustín Santiago-Moreno** and José M. Sautto-Vallejo**

*Facultad MATCOM, Universidad de La Habana, Cuba.
**Universidad Autónoma de Guerrero, México

ABSTRACT
This paper proposes a Randomized Response methodology where the report may be produced by one of two procedures for scrambling. The sampler designs a Bernoulli experiment for deciding which scrambling procedure is to be used by the respondent . Then, the response is modeled by a two stage RR procedure. The interviewed selects one of the two scrambling procedures. The respondent gives the response without informing which procedure was selected.
The model is developed for Simple Random Sampling and Stratified designs. The behavior of the proposed models is evaluated using real data of Mexico Covid19.

KEYWORDS: Randomized Response , Scrambling, Simple Random Sampling, Stratified Random Sampling

MSC: 62D05, 62P10

RESUMEN
En este paper se propone una metodología de Muestreo por Rangos Ordenados mediante un procedimiento de Respuestas Aleatorizadas donde el reporte es producido por uno de dos procedimientos alternativos de enmascaramiento. El muestrista diseña un experimento de Bernoulli para decidir cual de los procedimientos debe ser usado por el entrevistado. Entonces la respuesta es modelada por un nuevo procedimiento de Respuestas Aleatorizadas. El entrevistado selecciona uno de los procedimientos de enmascaramiento. Al responder da una respuesta sin informar cual fue seleccionado.
El modelo es desarrollado para el Muestreo Simple Aleatorio y Estratificado. El comportamiento de los modelos propuestos es evaluado usando datos reales de México sobre la Covid19.

PALABRAS CLAVE: Respuesta Aleatorizada, Enmascaramiento, Muestreo Simple Aleatorio, Estratificado

1.   INTRODUCTION

In a sample surveys commonly, some persons do not confide to the interviewer. They provide incorrect answers when the question is potentially sensitive. In such cases they are not answering or giving incorrect answers. This problem generates an evasive answer bias which is difficultly assessed. Warner (1965) proposed a method to reduce the response biases generated by dishonest answers to sensitive questions, see Chaudhuri- Mukerjee (1988). The technique is called randomized response (RR). His model dealt with qualitative variable, but it was extended for handling quantitative sensitive variables. See Greenberg *et al.* (1971) for earlier models. Himmelfarb and Edgell (1980) introduced the idea of scrambling the sensitive variable. A variable X with known distribution was used for scrambling. Eichhorn and Hayre (1983) proposed to request every respondent to report the product of the value of the sensitive variable Y and the scrambling variable X. Many papers have considered estimating the mean of a single sensitive variable. Numerous techniques and models have been developed for implementing RR. Hence RR`s modeling is still placing challenges both for the theory of survey sampling and its application. Challenging new RR methods for dealing with quantitative sensitive variables are the papers of Bouza (2010), Tarray and Singh (2015), Ahmed, Sedory and Singh (2018, 2020) for estimating means. See Chaudhuri et al. (2016) for a broader look to the nowadays problematics in RR.

---

[1] Corresponding author

In the sampling problem is assumed finite population of size N of individuals $\Omega = \{u_i, \ i = 1, \dots, N\}$. A sampling design d is used for choosing randomly n persons from Ω. Y_i denotes the true value of the sensitive quantitative variable of the ith unit of Ω. The proposed RR techniques assume that the respondents answer truthfully when using the randomized response device Thereof, RR protects the privacy persons belonging to stigmatized groups.

This paper proposes a RR methodology where the report may be produced by one of two procedures for scrambling. The sampler designs a Bernoulli experiment for deciding which scrambling procedure is to be used by the respondent. Then, the response is modeled by a two stage RR procedure. The interviewed selects one of the two scrambling procedures. The respondent gives the response without informing which procedure was selected.

Simple random sampling design with replacement is a basic method (SRSWR) using Stratified sampling (SSRSWR) diminishes time and in costs.

The proposed RR procedure is described in Section 2 below. The necessary derivations for generating randomized responses are given. It is also concerned with deriving, explicitly, unbiased estimators of the mean of Y and their errors for SRSWR. Section 3 presents SSRSWR and the two sampling approaches are compared.

2.   A SIMPLE TWO STAGE RR SCRAMBLING PROCEDURE USING SRSWR.

Consider that the interest is estimating the mean of a sensitive variable Y. Some persons of the population carry a stigma and tend to give an incorrect value of Y or to refuse giving a response. The seminal work on RR is due to Warner (1965). It opened a way of dealing with the response bias problem by using the so-called technique of randomized response (RR). The use of RR provides the opportunity of reducing response biases, due to dishonest answers when questioning on Y. This technique protects the privacy of the respondent by granting that belonging to a stigmatized group cannot be detected by the sampler.

In this section is assumed that a sample s of size n is selected from the population U using SRSWR. That is, n units are selected independently by selecting in each draw a $i \in U$ with probability 1/N.

Some simple RR scrambling procedures are described in Chaudhuri-Mukherjee (1988). Consider that the i-th respondent performs an experiment with probability P_1 and generates a value of $A_i \in \{A_1, \dots, A_K\}$ with probability θ_i. Therefore, are known the mean and variance of the random variable A_i : (revise the order of the sentence)

$$\mu_A = \sum_{j=1}^{K} A_j \theta_j; \quad \theta_j \in [0,1]; \ \sum_{j=1}^{K} \theta_j = 1$$

$$\sigma_A^2 = \sum_{j=1}^{K} (A_j - \mu_A)^2 \theta_j.$$

The respondent reports

$$S_i = Y_i + A_i$$

without informing the value of the scrambling variable A. Considering only the randomness of A generated by the RR procedure R_1

$$E(S_i|i) = Y_i + \mu_A$$
$$V(S_i|i) = \sigma_A^2$$

The behavior of R_1 is characterized in the following lemma when the sampling design d is SRSWR

Lemma 2.1. Using R_1 an unbiased estimator of the mean of Y using SRSWR is $\bar{Y}_{R_1} = \bar{S} - \mu_A$ and its sampling error is $V(\bar{S}) = \frac{\sigma_Y^2 + \sigma_A^2}{n}$.

Proof.

The expectation is $E(S_i) = E_d\big(E_{R(1)}(Y_i + A_i)|i\big) = E_d(Y_i + \mu_A) = \mu_Y + \mu_A$. Noting that, as $\hat{Y}_{i(R_1)} = S_i - \mu_A$, hence

$$E(\bar{Y}_{R_1}) = E(\bar{S} - \mu_A) = E_d\left(\frac{1}{n}\sum_{i=1}^{n} E\big(\hat{Y}_{i(R_1)}|i\big)\right) = E_d(\bar{y}) = \mu_Y.$$

The variance of the report of a person is $V(S_i) = \sigma_Y^2 + \sigma_A^2$ because $E_d(V_{R(1)}(S_i|i)) = \sigma_A^2$ and $V_d(E_{R(1)}(S_i|i)) = V_d(y_i) = \sigma_Y^2$. As s is selected using SRSWR is obtained that $V(\bar{Y}_{R_1}) = V(\bar{S}) = \frac{\sigma_Y^2 + \sigma_A^2}{n}$.

Another popular scrambling method, identified by R_2, is described as follows. Each respondent selects randomly and independently values $A_i \in \{A_1, \dots, A_K\}$ and $B_i \in \{B_1, \dots, B_m\}$ with respective probabilities θ_i and π_i and reports

$$T_i = Y_i + B_i A_i.$$

The following lemma characterized the behavior of R_2.

Lemma 2.2. Using R_2 an unbiased estimator of the mean of Y using SRSWR is $\bar{Y}_{R_2} = \bar{T} - \mu_A \mu_B$ and its sampling error is $V(\bar{T}) = \frac{\sigma_Y^2 + \sigma_A^2 \sigma_B^2}{n}$.

Proof.

The expectation of the report under this scrambling procedure is

$$E(T_i) = E_d\big(E_{R(2)}(Y_i + A_i B_i)|i\big) = E_d(Y_i + \mu_A \mu_B) = \mu_Y + \mu_A \mu_B .$$

Note that $\hat{Y}_{i(R_2)} = T_i - \mu_A \mu_B$. The variance of the report is $V(T_i) = \sigma_Y^2 + \sigma_A^2 \sigma_B^2$ because $E_d(V_{R(2)}(T_i|i)) = \sigma_A^2 \sigma_B^2$ and $V_d(E_{R(2)}(T_i|i)) = \sigma_Y^2$.

Then the lemma is proved.

The respondent will be more confident if the scrambling procedure is selected randomly. The interviewed selects the procedure. The sampler does not know which procedure generated the report. That is , the respondent`s confidence is improved, when selecting R_1 with a fixed probability P and R_2 with probability Q=1-P, without informing the scrambling procedure used for reporting. Then the respondent performs a Bernoulli experiment with parameter P and obtains as result γ_i. The report

$$Z_i = \begin{cases} S_i \ if \ \gamma_i = 1 \\ T_i \ if \ \gamma_i = 0 \end{cases}$$

is modeled by

$$Z_i = \gamma_i S_i + (1 - \gamma_i)T_i$$

Denote this procedure as R. The expectation of this report is

$$E(Z_i|i) = P\big(E_{R(1)}(Y_i + A_i)|i\big) + Q\big(E_{R(2)}(Y_i + A_i B_i)|i\big) = Y_i + \mu_A(P + Q\mu_B).$$

Hence , as

$$E(\bar{Z}) = \frac{1}{n}\sum_{i=1}^{n} E(Z_i) = \mu_Y + \mu_A(P + Q\mu_B)$$

an unbiased estimator of the mean of Y is

$$\hat{\mu}_Y = \bar{Z} - \mu_A(P + Q\mu_B)$$

The design variance of the conditional expectation is given by

$$V_d\big(E(Z_i|i)\big) = \sigma_Y^2 .$$

The scrambling procedure variance is

$$V(Z_i|i) = \gamma_i^2\big(V_{R(1)}(Y_i + A_i)|i\big) + (1 - \gamma_i)^2\big(V_{R(2)}(Y_i + A_i B_i)|i\big) = \gamma_i^2 \sigma_A^2 + (1 - \gamma_i)^2 \sigma_A^2 \sigma_B^2$$

Due to the randomness of the selection of the scrambling procedure its expectation is

$$E_d\big(V(Z_i|i)\big) = P\sigma_A^2 + Q\sigma_A^2 \sigma_B^2$$

Hence is proved the following lemma.

Lemma 2.3. Using R and SRSWR the estimator $\hat{\mu}_Y = \bar{Z} - \mu_A(P + Q\mu_B)$ is unbiased for the mean of Y and its variance is $V(\hat{\mu}_Y) = \frac{\sigma_Y^2 + \sigma_A^2(P + Q\sigma_B^2)}{n}$.

Proof .

Previous results support the unbiasedness of $\hat{\mu}_Y$ follows form the fact that $E(\bar{Z}) = \mu_Y + \mu_A(P + Q\mu_B)$. The variance of the estimator is

$$V(\hat{\mu}_Y) = V(\bar{Z}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Z_i) = \frac{1}{n^2}\sum_{i=1}^{n} V_d\big(E(Z_i|i)\big) + E_d\big(V(Z_i|i)\big) = \frac{1}{n^2}\sum \sigma_Y^2 + P\sigma_A^2 + Q\sigma_A^2 \sigma_B^2$$

Remark.2.1. Note that if the sampler fixes a set of value of B satisfying $\sigma_B^2 = 1$ then $V(\hat{\mu}_Y) = \frac{\sigma_Y^2 + \sigma_A^2}{n}$.

Then, the sampling error of R and R_1 are equal, though the respondents would be more confident in being protected by using the proposed two stage scrambling procedure.

3.   STRATIFIED MODEL EXTENSION

In this section, we will work with the two stage RR scrambling procedure seen in the previous section for the stratified random sampling with replacement design (SSRSWR).

In stratified random sampling, a population U with $|U| = N$ is divided into L strata of size $N_1 + N_2 + \cdots + N_L = N$. The sample sizes in each stratum are $n_1 + n_2 + \cdots + n_L = n$. A simple random sample is taken from each stratum.

The Abdelfatah and Mazloum's work (2016), proposed an efficient randomized response model with simple random sampling and its extension to stratified, in the same sense we work here the expansion to stratified., was made

3.1.   R₁ with SSRSWR

In the procedure R_1, each individual ~~of~~ i in each stratum h, must generate $A_{hi} \in \{A_{h1}, \ldots, A_{hK}\}$ with $P[A_{hi}] = \theta_{hi}$, and the involved parameters are

$$\mu_{hA} = \sum_{j=1}^{k} A_{hj}\,\theta_{hj} \;; \qquad \theta_{hj} \in [0,1],\; \sum_{j=1}^{k}\theta_{hj} = 1, \qquad .$$
$$\sigma_{hA}^2 = \sum_{j=1}^{k}(A_{hj} - \mu_{hA})^2\,\theta_{hj} \qquad , \text{ for } h=1, 2, \ldots, L$$

The i-th respondent reports:
$$S_{hi} = Y_{hi} + A_{hi}$$

Considering only the randomness of A in the procedure, we have
$$E(S_{hi}|i) = Y_{hi} + \mu_{hA}$$
$$V(S_{hi}|i) = \sigma_{hA}^2$$

We consider that the sampling design d is SSRSWR. We characterize the behavior of R_1 ~~by:~~ in the next lemma~~.~~

Lemma 3.1. For R_1, an estimator of the mean of Y per stratum using SSRSWR is $\quad \bar{Y}_{h\,(R_1)} = \bar{S}_h - \mu_{hA}$ and its stratified global estimator $\bar{Y}_{ST,R_1} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_1)}$ is unbiased. Its sampling error per stratum is given by $V(\bar{Y}_{h\,(R_1)}) = \frac{\sigma_{hY}^2 + \sigma_{hA}^2}{n_h}$ and its global sampling error by $\quad V(\bar{Y}_{ST,R_1}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2(\sigma_{hy}^2 + \sigma_{hA}^2)}{n_h}$

Proof
$$E(S_{hi}) = E_d\big(E_{R_1}(Y_{hi} + A_{hi})|i\big) = E_d(Y_{hi} + \mu_{hA}) = \mu_{hY} + \mu_{hA}.$$

As $E(S_{hi}|i) = Y_{hi} + \mu_{hA}$, then, $\hat{Y}_{hi\,(R_1)} = S_{hi} - \mu_{hA}$, and, $\bar{Y}_{h\,(R_1)} = \bar{S}_h - \mu_{hA}$.

Let's look at the unbiasedness within stratum.

$$E\,(\bar{Y}_{h\,(R_1)}) = E(\bar{S}_h - \mu_{hA}) = E\left(\frac{1}{n_h}\sum_{i=1}^{n_h} S_{hi} - \frac{1}{n_h}\sum_{i=1}^{n_h}\mu_{hA}\right) = E_d\left(\frac{1}{n_h}\sum_{i=1}^{n_h} E(\hat{Y}_{hi\,(R_1)}|i)\right) = E_d(\bar{y}_h) =$$
$$\mu_{hY}.$$

Let's look at the global unbiasedness.

Due to the general results obtained in the theory on SSRSWR, see Cochran (1977) $\bar{Y}_{h\,(R_1)} = \bar{S}_h - \mu_{hA}$ is unbiased in stratum h, hence, $\bar{Y}_{ST,R_1} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_1)}$ is unbiased.

Let's develop the estimator variance per stratum
$$V(\bar{Y}_{h\,(R_1)}) = \frac{1}{(n_h)^2}\sum_{i=1}^{n_h} V(S_{hi}) = \frac{\sigma_{hY}^2 + \sigma_{hA}^2}{n_h},$$

since
$$V_d(E_{R(1)}(S_{hi}|i)) = V_d(y_{hi}) = \sigma_{hY}^2$$

and
$$E_d(V_{R(1)}(S_{hi}|i)) = \sigma_{hA}^2$$

Now we obtain the global estimator variance

As $V(\bar{Y}_{h\,(R_1)}) = \frac{\sigma_{hY}^2 + \sigma_{hA}^2}{n_h}$ and by Cochran's theorem 5.2, , $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\, V(\bar{y}_h)$, where $\bar{y}_h$ must be an unbiased estimator of $\bar{Y}_h$, which has already been demonstrated. the samples are considered independent. Hence, applying the above mentioned theorem we have,

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

3.2.   R₂ with SSRSWR

Using R_2, each individual of i in each stratum h, must select ~~values~~ randomly and independently $A_{hi} \in \{A_1, \ldots, A_K\}$ and $B_{hi} \in \{B_1, \ldots, B_m\}$ with $P[A_{hi}] = \theta_{hi}$ and $P[B_{hi}] = \pi_{hi}$, the respondent reports:
$$T_{hi} = Y_{hi} + B_{hi}\,A_{hi}$$

Considering the randomness of A and B in this procedure, we have,
$$E(T_{hi}|i) = Y_{hi} + \mu_{hB}\,\mu_{hA}$$
$$V(T_{hi}|i) = \sigma_{hB}^2\,\sigma_{hA}^2$$

Then we obtain a similar result for the sampling design d is SRSWR, under R_2:

Lemma 3.2. For R_2, an unbiased estimator of the mean of Y per stratum using SSRSWR is

$$\bar{Y}_{h\,(R_2)} = \bar{T}_h - \mu_{hB}\,\mu_{hA}$$

and its stratified global estimator $\bar{Y}_{ST,R_2} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{Y}_{h\,(R_2)}$. Its sampling error per stratum is given by

$$V(\bar{Y}_{h\,(R_2)}) = \frac{\sigma_{hY}^2 + \sigma_{hB}^2\,\sigma_{hA}^2}{n_h} \quad \text{and its global sampling error} \quad V(\bar{Y}_{ST,R_2}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\,(\sigma_{hy}^2 + \sigma_{hB}^2\,\sigma_{hA}^2)}{n_h}$$

Proof

$$E(T_{hi}) = E_d\big(E_{R_2}\,(Y_{hi} + B_{hi}\,A_{hi})|i\big) = E_d(Y_{hi} + \mu_{hB}\,\mu_{hA}) = \mu_{hY} + \mu_{hB}\,\mu_{hA}.$$

As $E(T_{hi}|i) = Y_{hi} + \mu_{hB}\,\mu_{hA}$, then, $\hat{Y}_{hi\,(R_2)} = T_{hi} - \mu_{hB}\,\mu_{hA}$, and, $\bar{Y}_{h\,(R_2)} = \bar{T}_h - \mu_{hB}\,\mu_{hA}$.

The unbiasedness of $\bar{Y}_{h\,(R_2)}$ in each stratum is easily derived as follows:

$$E\big(\bar{Y}_{h\,(R_2)}\big) = E(\bar{T}_h - \mu_{hB}\,\mu_{hA}) = E\left(\frac{1}{n_h}\sum_{i=1}^{n_h} T_{hi} - \frac{1}{n_h}\sum_{i=1}^{n_h}\mu_{hB}\,\mu_{hA}\right) = E_d\left(\frac{1}{n_h}\sum_{i=1}^{n_h} E\big(\hat{Y}_{hi\,(R_2)}|i\big)\right) =$$
$$E_d(\bar{y}_h) = \mu_{hY}.$$

For the global unbiasedness we use the same reasoning as in $\bar{Y}_{ST,R_1}$.

On the other hand, the estimator of the variance per stratum is

$$V(\bar{Y}_{h\,(R_2)}) = \frac{1}{(n_h)^2}\sum_{i=1}^{n_h} V(T_{hi}) = \frac{\sigma_{hY}^2 + \sigma_{hB}^2\,\sigma_{hA}^2}{n_h},$$

since

$$V_d(E_{R(2)}(T_{hi}|i)) = V_d(y_{hi}) = \sigma_{hY}^2, \quad E_d(V_{R(2)}(T_{hi}|i)) = \sigma_{hB}^2\sigma_{hA}^2.$$

The global estimator variance is derived applying the same reasoning as in $V(\bar{Y}_{ST,R_1})$.

3.3. R procedure with SSRSWR

Finally, we work the R procedure with SSRSWR. Let be $P(R_{h1}) = P_h$ and $P(R_{h2}) = (1 - P_h) = Q_h$. Then respondent i-th in stratum h, performs a Bernoulli experiment with parameter P_h and obtains γ_{hi} as a result. The report

$$Z_{hi} = \begin{cases} S_{hi} & if \quad \gamma_{hi} = 1 \\ T_{hi} & if \quad \gamma_{hi} = 0 \end{cases}$$

is modeled by

$$Z_{hi} = \gamma_{hi}\,S_{hi} + (1 - \gamma_{hi})\,T_{hi}$$

The expectation of this report is

$$E(Z_{hi}|i) = E(P_h\,S_{hi} + Q_h\,T_{hi}) = P_h\,E_{R_1}\big((Y_{hi} + A_{hi})|i\big) + Q_h\,E_{R_2}\big((Y_{hi} + B_{hi}\,A_{hi})|i\big) =$$
$$Y_{hi}\,(P_h + Q_h) + \mu_{hA}\,(P_h + Q_h\,\mu_{hB}) = Y_{hi} + \mu_{hA}\,(P_h + Q_h\,\mu_{hB}).$$

Hence, as

$$E(\bar{Z}_h) = \frac{1}{n_h}\sum_{i=1}^{n_h} E(Z_{hi}) = E_d(E_{R1}\,P_h(Y_{hi} + A_{hi})|i) + E_d(E_{R2}\,Q_h(Y_{hi} + B_{hi}\,A_{hi})|i)$$
$$= \mu_{hY} + \mu_{hA}\,(P_h + Q_h\,\mu_B).$$

An unbiased estimator of the mean of Y is

$$\hat{\mu}_{h\,Y} = \bar{Z}_h - \mu_{hA}\,(P_h + Q_h\,\mu_B).$$

The design variance of the conditional expectation is given by

$$V_d(E(Z_{hi}|i)) = \sigma_{hY}^2$$

The R procedure variance is

$$V(Z_{hi}|i) = \gamma_{hi}^2\,(V_{R1}\,(Y_{hi} + A_{hi})|i) + (1 - \gamma_{hi})^2(\,V_{R2}(Y_{hi} + B_{hi}\,A_{hi})|i)$$
$$= \gamma_{hi}^2\sigma_{hA}^2 + (1 - \gamma_{hi})^2\sigma_{hB}^2\,\sigma_{hA}^2$$

And its expectation

$$E_d\big(V(Z_{hi}|i)\big) = P_h\,\sigma_{hA}^2 + Q_h\,\sigma_{hB}^2\,\sigma_{hA}^2$$

For this scrambling procedure, we have derived the following lemma

Lemma 3.3. For the procedure R and using SSRSWR, an estimator of the mean of Y is $\hat{\mu}_{h\,Y} = \bar{Z}_h - \mu_{hA}\,(P_h + Q_h\,\mu_{hB})$ and its stratified global estimator $\bar{Y}_{ST,R} = \frac{1}{N}\sum_{h=1}^{L} N_h\,\hat{\mu}_{h\,Y}$. Its sampling error per stratum is given by $V(\hat{\mu}_{h\,Y}) = \frac{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h\,\sigma_{hB}^2)}{n_h}$, and its global sampling error

$$V(\bar{Y}_{ST,R}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2\,(\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h\,\sigma_{hB}^2))}{n_h}$$

Proof

As $E(Z_{hi}|i) = Y_{hi} + \mu_{hA}(P_h + Q_h \mu_{hB})$, then, $\hat{Y}_{hi\,(R)} = Z_{hi} - \mu_{hA}(P_h + Q_h \mu_{hB})$, and, $\bar{Y}_{h\,(R)} = \bar{Z}_h - \mu_{hA}(P_h + Q_h \mu_{hB})$

Now, let's obtain the unbiasedness per stratum.

$$E(\hat{\mu}_{h\,Y}) = E(\bar{Z}_h - \mu_{hA}(P_h + Q_h \mu_{hB})) = E\left(\frac{1}{n_h}\sum_{i=1}^{n_h} z_{hi} - \frac{1}{n_h}\sum_{i=1}^{n_h}\mu_{hA}(P_h + Q_h \mu_{hB})\right) =$$

$$E_d\left(\frac{1}{n_h}\sum_{i=1}^{n_h} E(\hat{Y}_{hi\,(R)}|i)\right) = E_d(\bar{y}_h) = \mu_{h\,Y}.$$

Global unbiasedness is derived using the same reasoning as in the previous cases.

In this case, the variance estimator per stratum is

$$V(\hat{\mu}_{h\,Y}) = \frac{1}{(n_h)^2}\sum_{i=1}^{n_h} V(Z_{hi}) = \frac{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h\,\sigma_{hB}^2)}{n_h}, \text{ since } V_d(E(Z_{hi}|i)) = P_h^2\sigma_{hY}^2 + Q_h^2\sigma_{hY}^2 = \sigma_{hY}^2 \text{ and}$$

$E_d(V(Z_{hi}|i)) = P_h\,\sigma_{hA}^2 + Q_h\,\sigma_{hB}^2\,\sigma_{hA}^2$

We obtain the global estimator variance following the same reasoning as in the previous Lemmas.

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

Let us consider the minimization of the variance in terms of the strata sample sizes for fixed n and cost C. The problem is to solve the optimization problem:

$$ArgMin_{\vec{n}}\{Variance | C = c_0 + \sum c_h\,n_h\}$$

Its dual is :

$$ArgMin_{\vec{n}} C | V = \sum Var_i\}$$

4.1. n_h y n optima para $V(\bar{Y}_{ST,R_1})$

Lemma 4.1. Take the sampling design and the cost function $C = c_0 + \sum c_h\,n_h$, the variance of the estimator of the population mean of the procedure R_1 is minimized when $n_h \propto N_h\sqrt{\sigma_{hy}^2 + \sigma_{hA}^2}\,\frac{1}{\sqrt{c_h}}$.

If the variance $V(\bar{Y}_{ST,R_1}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2(\sigma_{hy}^2 + \sigma_{hA}^2)}{n_h}$ is fixed the cost is minimized when

$$n_h \propto \left(N_h\sqrt{\sigma_{hy}^2 + \sigma_{hA}^2}\,\frac{1}{\sqrt{c_h}}\right)$$

Proof. .

Our objective function is

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

Subject to

$$C = c_0 + \sum c_h\,n_h, n = \sum n_h .$$

Using the method of Lagrange, for determining the optimal sample sizes and taking n_h the Lagrange parameter λ the optimization problem may be rewritten as :

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

The partial derivatives of $f(y,\lambda)$ with respect to the $n_h's$, h=1, 2, ..., L, are

$$\frac{\partial\,h(y,\lambda)}{\partial\,n_1} = -\frac{N_1^2(\sigma_{1y}^2 + \sigma_{1A}^2)}{N^2 n_1^2} + \lambda c_1, \dots, \frac{\partial\,h(y,\lambda)}{\partial\,n_L} = -\frac{N_L^2(\sigma_{Ly}^2 + \sigma_{LA}^2)}{N^2 n_L^2} + \lambda c_L\ h = 1, \dots, L;$$

Say.

$$-\frac{N_h^2(\sigma_{hy}^2 + \sigma_{hA}^2)}{N^2 n_h^2} + \lambda c_h = 0,\ h=1,2, \dots, L.$$

As a result:

$$\lambda c_h = \frac{N_h^2(\sigma_{hy}^2 + \sigma_{hA}^2)}{N^2 n_h^2} \Rightarrow \sqrt{\lambda}\sqrt{c_h} = \frac{\sqrt{N_h^2}\sqrt{(\sigma_{hy}^2 + \sigma_{hA}^2)}}{\sqrt{N^2}\sqrt{n_h^2}} \Rightarrow n_h\sqrt{\lambda} = \frac{N_h\sqrt{(\sigma_{hy}^2 + \sigma_{hA}^2)}}{N\sqrt{c_h}} \quad \dots (4.1.1)$$

Summing them

$$\sum n_h \sqrt{\lambda} = \sum \frac{N_h \sqrt{\left(\sigma_{hy}^2 + \sigma_{hA}^2\right)}}{N \sqrt{c_h}} \Rightarrow n\sqrt{\lambda} = \sum \frac{N_h \sqrt{\left(\sigma_{hy}^2 + \sigma_{hA}^2\right)}}{N \sqrt{c_h}} \ ....$$ (4.1.2)

From (4.1.1) and (4.1.2) is derived the following expressions:

$$\frac{n_h}{n} = \frac{N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}}}{\sum N_h \sqrt{\left(\sigma_{hy}^2 + \sigma_{hA}^2\right)} \ \frac{1}{\sqrt{c_h}}} \quad h = 1, \dots$$ (4.1.3)

The proof of the first result is obtained.

The dual optimal allocation problem is solved similarly. In this case V is fixed. $V\left(\bar{Y}_{ST,R_1}\right)$ may be expressed as:

$$V\left(\bar{Y}_{ST,R_1}\right) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 \left(\sigma_{hy}^2 + \sigma_{hA}^2\right)}{n \frac{N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}}}{\sum N_h \sqrt{\left(\sigma_{hy}^2 + \sigma_{hA}^2\right)} \ \frac{1}{\sqrt{c_h}}}}$$

That is

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

and

$$n_h \propto \left(N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}} \right)$$

sizes will be larger whenever we have a larger variation in the stratum and the cost of sampling is small. This result is the counterpart of the classic optimal allocation theory.

An explicit formula of the optimal allocations are easily derived noting that, see (4.1.3), n_h depends also of the fixed overall sample size $n=n_1+\dots+n_L$.

Doing the needed algebraic manipulations is obtained that, for a fixed cost:

$$n = \frac{(C - c_0) \sum \left(N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}} \right)}{\sum \left(N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \ \sqrt{c_h} \right)}$$

and for a fixed variance

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

4.2 n_h y n optima for $V\left(\bar{Y}_{ST,R_2}\right)$

Lemma 4.2. Take the sampling design SSRSWR and (4.1) as the cost function, the   variance of the estimator of the population mean of the procedure R_2 is minima when $n_h \ \propto \ N_h \sqrt{\sigma_{hy}^2 + \sigma_{hB}^2 \ \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}}$

Proof.

The proof is derived using the also Lagrange multipliers method for R_2

$$\frac{n_h}{n} = \frac{N_h \sqrt{\sigma_{hy}^2 + \sigma_{hB}^2 \ \sigma_{hA}^2} \ \frac{1}{\sqrt{c_h}}}{\sum N_h \sqrt{\left(\sigma_{hy}^2 + \sigma_{hB}^2 \ \sigma_{hA}^2\right)} \ \frac{1}{\sqrt{c_h}}} \quad \dots$$ (4.2.1)

n is an optimum when C is fixed nd

$$n = \frac{(C - c_0) \sum \left(N_h \sqrt{\sigma_{hy}^2 + \sigma_{hB}^2 \, \sigma_{hA}^2} \, \frac{1}{\sqrt{c_h}} \right)}{\sum \left(N_h \sqrt{\sigma_{hy}^2 + \sigma_{hB}^2 \, \sigma_{hA}^2} \, \sqrt{c_h} \right)}$$

The optimum n for a fixed variance is

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

For R_2 the optimal allocation is obtained by substituting in the formulae.

4.3 n_h and n optima for $V(\bar{Y}_{ST,R})$

Lemma 4.3. For SSRSWR and a cost fixed $C = c_0 + \sum c_h \, n_h$, the variance of the estimator for R is minimized for $n_h \propto N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \frac{1}{\sqrt{c_h}}$

Proof.

Doing the same similar manipulations used for proving lemma 4.1 we have:

$$\frac{n_h}{n} = \frac{N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \frac{1}{\sqrt{c_h}}}{\sum N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \frac{1}{\sqrt{c_h}}}$$

For a fixed cost the optimum n is,

$$n = \frac{(C - c_0) \sum \left(N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \frac{1}{\sqrt{c_h}} \right)}{\sum \left(N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \sqrt{c_h} \right)}$$

and for fixed variance

$$n = \frac{1}{N^2 \, V(\bar{Y}_{ST,R})} \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \sqrt{c_h} \right] \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hY}^2 + \sigma_{hA}^2(P_h + Q_h \, \sigma_{hB}^2)} \, \frac{1}{\sqrt{c_h}} \right]$$

.
4.4   Gains in accuracy for the optimal allocation in SSRSWR of $V(\bar{Y}_{ST,R_1})_{opt}$ with respect to $V(\bar{Y}_{ST,R_1})$

From (4.1.3) substituting n_h in $V(\bar{Y}_{ST,R_1})$ we have that $V(\bar{Y}_{ST,R_1})_{opt}$ for fixed n and c_h in for R_1 is :

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

Therefore, the gain of $V(\bar{Y}_{ST,R_1})_{opt}$ compared with $V(\bar{Y}_{ST,R_1})$ is

$$G(ST_{R_1}, ST_{R_1 opt}) = V(\bar{Y}_{ST,R_1}) - V(\bar{Y}_{ST,R_1})_{opt} =$$

$$= \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (\sigma_{hy}^2 + \sigma_{hA}^2)}{n_h} - \frac{1}{N^2 \, n} \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \, \sqrt{c_h} \right] \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \, \frac{1}{\sqrt{c_h}} \right]$$

When the cost of evaluating a unit is equal for all the strata, that is, C = c_0 +c_n, then :

$$V(\bar{Y}_{ST,R_1})_{opt\prime} = \frac{1}{N^2 \, n} \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \right] \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \right]$$

Comparing $V(\bar{Y}_{ST,R_1})_{opt\prime}$ with $V(\bar{Y}_{ST,R_1})$

$$G(ST_{R_1}, ST_{R_1 opt}\prime) = V(\bar{Y}_{ST,R_1}) - V(\bar{Y}_{ST,R_1})_{opt} =$$

$$= \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (\sigma_{hy}^2 + \sigma_{hA}^2)}{n_h} - \frac{1}{N^2 \, n} \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \right] \sum_{h=1}^{L} \left[N_h \sqrt{\sigma_{hy}^2 + \sigma_{hA}^2} \right]$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{S_h^2}{n_h} - \frac{1}{n} \left\{ \sum_{h=1}^{L} \left[\frac{N_h \ S_h}{N} \right] \ \sum_{h=1}^{L} \left[\frac{N_h \ S_h}{N} \right] \right\}$$

$$= \sum_{h=1}^{L} \frac{P_h^2}{n_h}$$

$$- \frac{1}{n} \left[\sum_{h=1}^{L} P_h * \sum_{h=1}^{L} P_h \right] \ ... \tag{4.4.1}$$

Where $S_h^2 = \sigma_{hy}^2 + \sigma_{hA}^2$, $P_h = \frac{N_h \ S_h}{N}$. As the right hand expression of (4.4.1), is divided by n, this difference is always larger or equal to zero.

5.   A SIMULATION STUDY USING COVID´S 19 REAL DATA.

To evaluate the performance of the suggested models for estimating of the population mean of sensitive data, we consider the records of confirmed patients of COVID-19 in México since March-2020, up to May-2021. The variables used were diabetes, hypertension, obesity, age and sex. These variables are of particular interest due to the fact that an infected persons may be discriminated and stigmatized by people in their surrounds Therefore information provided in surveys may be considered sensitive when looking for an employ, for example.

The data consisted of 1 048.575 cases on the risk quantification. The managers considered that the data on Y= risk were strategically sensitive. Min (Y)=0,9, Max (Y)=68, $\mu_Y = 15,85$, $\sigma_Y^2 = 203,51$. The population is naturally divided by the epidemiologists into 10 strata according to the age intervals. This problem is considered a Big Data case and are obtainable at https://coronavirus.gob.mx/

The evaluation of the data was made by quantifying the death-risk due COVID-19 if the person is positive to diabetes, hypertension or obesity as well as age and sex. The quantification of the risk was made using the weighting developed by Pamplona (2020). The accuracy and efficiency of the proposed estimators (R_1, R_2 y R) for SSRS and SSRSWR was used as evaluation measures of the them. The comparison of the accuracy was made by computing in each sample generated s

$$Error(st)_s = \left(\frac{\frac{|\hat{y}_{st} - \bar{Y}_{st}|}{\bar{Y}_{st}}}{\frac{|\hat{y} - \bar{Y}|}{\bar{Y}}} \right)_s$$

the efficiency was measured by

$$E(st)_s = \left(\frac{V(\bar{y}_{st})}{V(\bar{y})} \right)_s$$

We fixed n= 65.000 as the sample size for applying SRSWR. The strata samples sizes were determined proportionally looking for maintaining the proportion $\frac{n}{N}$. The corresponding probabilities P y Q for R were fixed adequately for observing the behavior of the estimators.

The number of simulation runs was 1.000 for each method and design. That is s=1,…, 1.000. The evaluation process used the results of the 1000 runs and they were averaged determining

$$Error(st_\rho) = \frac{\sum_{s=1}^{1000} Error(st)_{s|\rho}}{1000}, \qquad \rho = R_1, R_2, R$$

$$E(st_\rho) = \frac{\sum_{s=1}^{1000} E(st)_{s|\rho}}{1000}, \rho = R_1, R_2, R$$

See in Table 1 that the 3 procedures (R_1, R_2 y R) are more accurate when stratification is used. R_1 scrambling procedures is more accurate than R_2. This conclusion is based in observing that R performs better when the probabilities of using R_1 is P=0,7 are larger than the probability of observing R_2 (Q=0.3).

Table 1. Accuracy of the estimators of the mean in the designs.

	R_1	R_2	R (P=0.3, Q=0.7)	R' (P=0.7, Q=0.3)
SRSWR	0,003035	0,02539	0,04001	0,03747
SSRSWR	0,001296	0,01912	0,02279	0,018519
$Error(st)$	$Error(st_{R1})$ = 0,42701812	$Error(st_{R2})$ = 0,75305238	$Error(st_R)$ = 0,5696076	$Error(st_{R'})$ = 0,49423539

The results in Table 2 are confirming the results observed in Table 1. Then the efficiency of the stratified design is better.

Table 2. Efficiency of the variances of the estimators of the mean in the designs.

	R$_1$	R$_2$	R (P=0,3, Q=0,7)	R' (P=0,7, Q=0,3)
SRSWR	0,003698	0,005282	0,004664	0,003839
SSRSWR	0,0005953	0,004666	0,003445	0,001816
$E(st)$	$E(st_{R1}) = 0,160979$	$E(st_{R2}) = 0,883378$	$E(st_R) = 0,738636$	$E(st_{R'}) = 0,47304$

REFERENCES

[1]   AHMED, S., SEDORY S. A. and SINGH, S. (2018): Simultaneous estimation of means of two sensitive variables*. **Comm. Statist. Theory Methods** 47 , 324–343.

[2]   AHMED, S., SEDORY S. A. and SINGH, S. (2020) Forcibly Re-scrambled randomized response model for simultaneous estimation of means of two sensitive variables. ***Commun. Math. Stat.*** 8, 23–45

[3]   BOUZA, C. (2010): Behavior of a randomized response procedure under unequal selection model of Chaudhuri-Stenger for insensitive variables under ranked set sampling. **Advances and Applications in Statistical Sciences**, 4, 136-44.

[4]   CHAUDHURI A., and MUKERJEE R (1988): **Randomized response: theory and techniques**. Marcel Dekker Inc, New York.

[5]   CHAUDHURI, A. and STENGER, H. (1992): **Sampling Survey**. Marcel Dekker, New York.

[6]   CHAUDHURI, A., CHRISTOFIDES, T. C. and RAO, C. R., (2016): **Handbook of Statistics 34, Data gathering, analysis and protection of privacy through randomized response techniques**. Elsevier, Amsterdam.

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

[8]   COVID19 RISK (2021): Coronavirus Pandemic in Mexico. **https://coronavirus.gob.mx/** (last consulted 20 May, 2021.)

[9]   GREENBERG, B. G., KUBLER, R. R. and HORVITZ, D. G. (1971): Applications of RR technique in obtaining quantitative data. **Journal of the American Statistical Association**, 66, 243-250.

[10]   HIMMELFARB, S. and EDGELL, S.E. (1980): Additive constant model: A randomized response technique for eliminating evasiveness to quantitative response questions. **Psychological Bulletin**, 87, 525-530.

[11]   PAMPLONA, F., (2020): La pandemia de covid-19 en México y la otra epidemia. **Espiral Estudios sobre Estado y Sociedad.** xxvii . 78-79 .

[12]   TARRAY, T. A. and H. SINGH (2015): A general procedure for estimating the mean of a sensitive variable using auxiliary information. **Revista Investigación Operacional**. 36, 268-279

[13]   WARNER, S.L. (1965): Randomized response: a survey technique for eliminating evasive answer bias. **Journal of the American Statistical Association**, 60, 63–69.