Journal of King Saud University – Science 31 (2019) 171–176

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Subsampling rules for item non response of an estimator based on the combination of regression and ratio

Carlos N. Bouza-Herrera [a,*], Mir Subzar [b]

[a] Universidad de La Habana, Cuba
[b] Division of Agricultural Statistics, SKUAST-Kashmir, Srinagar 190025, India

1. Introduction

Survey sampling models assume the existence of a finite population $U=\{u_1,\ldots,u_N\}$, where the units are perfectly identifiable, and a sample s of size $n \leq N$ is selected from U. Another assumption is that the variable of interest Y is measured in each selected unit. Unfortunately, in real life, surveys should deal with the existence of some missing observations. The existence of non-response suggests that the population U is divided into two strata: U_1, where are grouped the units that give a response at the first visit, and U_2, which contains the rest of the individuals. This is the so called 'response strata' model and was the framework proposed by Hansen and Hurwitz (1946), see text books as Arnab (2017), Singh (2003), and Lohr (2010).

The behavior of estimators based on the use of subsampling depends heavily on the used sub-sampling rule. Alternative sampling rules to Hansen-Hurwitz's rule have been proposed; see for example Srinath (1971) and Bouza (1981).

The quality of the inquiries depends of the rate of responses. A question is how many non-respondents should be subsampled for having a good response rate. If sufficient information is available from prior rounds or other sources, the decision can be made on the basis of the experience of the sampler who analyzes the response rate, design effect, costs etc. The role of non-response in the accuracy of estimation is still generating discussions among statisticians, see an enlightening discussion in Särndal and Lundquist (2014). In a seminal paper Hansen and Hurwitz (1946) suggested subsampling non-respondents for alleviating the effect of having missing data.

Many sampling models consider increasing the survey's precision by utilizing information on an auxiliary variable. That is the case of ratio, regression and product estimators Singh and Kumar (2009) developed a general class of estimators for the population mean of the interest variable Y, by using information on two auxiliary variables, when missing observations are present. The class includes some well-known estimators.

In this paper we consider the case in which we have missing information on the interest variable but is available the information on the auxiliary variables in the sample, as well as their population means. Singh and Kumar (2009) considered the subsampling rule proposed by Hansen and Hurwitz (1946). This paper extends their results studying the effect of using the rules of Srinath (1971) and Bouza (1981), when dealing with estimators of the class. The behavior the estimators in the class is analyzed in terms of accuracy and cost for each rule. The approximate errors are given.

Section 2 presents the basic issues on the nonresponse procedures. Section 3 is devoted to analyzing the statistical properties of the estimators. In this section are developed comparisons, of the effects of using the three sub-sampling rules in the variance and cost function are developed. We presented also a numerical study, using real life studies, where the rules are evaluated. Finally, in Section 4 some concluding remarks are given.

2. The non-response problem

It is increasingly common to subsample non-respondents for increasing the response rates at a reduced cost. The usual theory of survey sampling is developed assuming that the finite population $U = \{u_1, \cdots, u_N\}$ is composed by individuals that can be perfectly identified. Assume that a sample s of size $n \leq N$ is selected using simple random sampling with replacement (SRSWR). The variable of interest Y is to be measured in each selected unit. Real-life surveys should deal with the existence of missing observations. This fact establishes that the population is divided into two strata:

$$U_1 = \{u \in U u\, gives\, a\, response\, in\, the\, first\, visit\},$$

$$U_2 = \{u \in U\, u\, does\, not\, give\, a\, response\, in\, the\, first\, visit\}.$$

Then we may distinguish the strata parameters

$$\bar{Y}_t = \frac{1}{N_t}\sum_{u_i \in U_t} Y_i,\ \sigma_t^2 = \frac{1}{N_t}\sum_{u_i \in U_t}(Y_i - \bar{Y}_t)^2,\ N_t = ||U_t|| t = 1,\ 2$$

* Corresponding author.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2018.10.006

as well as the population ones

$$\bar{Y} = \frac{1}{N}\sum_{u_i \in U} Y_i,\ \sigma_Y^2 = \frac{1}{N_t}\sum_{u_i \in U}(Y_i - \bar{Y})^2,\ N = N_1 + N_2$$

There are three solutions to cope with this fact: to ignore the non-respondents, to impute the missing values or to subsample the non-respondents. Rarely, ignoring the non-responses is a good solution, as Y may be related with having very different values, in the units belonging to U_2 with respect to U_1. Imputation of the missing data depends on having an adequate model of the non-responses mechanism and reliable information for predicting Y for each non-respondent. Subsampling the non-respondents is a conservative solution. Theoretically, dealing with subsampling the non-respondents stratum is a particular case of Double Sampling (DS), see Bouza et al. (2011) for a motivating discussion on the subject. It was proposed firstly by Hansen and Hurwitz (1946). Its use increases the costs but provides the confidence of estimating using information on U_2 Deciding which subsampling procedure is to be used is of practical importance, see Thompson and Washington (2013), Torres van Grinsven et al. (2014), Andridge and Thompson (2015) and Heffetz and Reeves (2016). Then it makes sense analyzing the behavior of alternative sampling rules to Hansen-Hurwitz's rule. In the literature are reported the rules of Srinath (1971) and Bouza (1981), as other rules for determining the sub sample size. They fix the size of the subsample to be drawn from the set of non-respondents.

Let us present a general algorithm for implementing one of the subsampling procedures.

2.1. Subsampling algorithm

Step 1. Select a sample s from U using simple random sampling with replacement (SRSWR).

Step 2. Evaluate Y among the subsample of the respondents $s_1 \subset s$, determine $\{y_i,\ i \in s_1,\ ||s_1|| = n_1\}$ and compute

$$\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1} \tag{2.1}$$

Step 3. Determine $s_2 = \{u_i \in s_2 = s - s_1\}, s_2 = n_2$

Step 4. Fix $n_2^* = \theta n_2, \theta \leq 1$

Step 5. Select using SRSWR a sub-sample $s_2^* \subset s_2$ of size n_2^*

Step 6. Evaluate Y among the units in s_2^* and compute

$$\bar{y}_2^* = \frac{\sum_{i=1}^{n_2^*} y_i}{n_2^*} \tag{2.2}$$

Step 7. Compute the estimate of $\bar{Y}$

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_2^*,\ w_j = \frac{n_j}{n}, j = 1,2 \tag{2.3}$$

As s_1 is a subsample of U_1, (2.1) is an unbiased estimator of the mean of the response stratum, that is $E(\bar{y}_1|s) = \bar{Y}_1$. The subsample selected from the non-respondents provides (2.2) which is a conditionally unbiased estimator of $\bar{y}_2$ because $E(\bar{y}_2^*,|s) = \bar{y}_2$.Therefore, as $s_2 \subset U_2$, $EE(\bar{y}_2^*|s) = \bar{Y}_2$. The usual analysis of the behavior of (2.3) is based on studying the expression

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_2^* = (w_1\bar{y}_1 + w_2\bar{y}_2) + w_2(\bar{y}_2^* - \bar{y}_2), \tag{2.4}$$

The first term is the sample mean of s, hence $EE(w_1\bar{y}_1 + w_2\bar{y}_2) = \bar{Y}$,We have that $E(\bar{y}_2^* - \bar{y}_2|s) = 0$. Therefore, $\bar{y}^*$ is an unbiased estimator of the population mean.

We have that the expected variance of the first term is

$$E((V(w_1\bar{y}_1 + w_2\bar{y}_2)|s)) = \frac{\sigma_Y^2}{n} \tag{2.5}$$

The conditional variance of the second term is given by

$$V(w_2(\bar{y}_2^* - \bar{y}_2)|s) = w_2^2\ \sigma_{2Y}^2\left(\frac{1}{\theta n_2} - \frac{1}{n_2}\right) = \frac{n_2(1-\theta)}{\theta n^2}\sigma_{2Y}^2$$

Note that n_2 is a Binomial random variable, hence

$$E[V(w_2(\bar{y}_2^* - \bar{y}_2)|s)] = \frac{W_2(1-\theta)}{\theta n}\sigma_{2Y}^2$$

Due to the independence the expectation of the cross product is zero and is deduced the well-known expression

$$EV(\bar{y}^*|s) = \frac{\sigma_Y^2}{n} + \frac{W_2(1-\theta)}{\theta n}\sigma_{2Y}^2.$$

The value of θ determines the value of the second term in the expected error. The subsampling rules deal with determining θ. The existing particular rules fix the value of θ. They are:

Hansen and Hurwitz (1946): $\theta = \frac{1}{K}, K \geq 1$

Srinath (1971): $\theta = \frac{n_2^2}{Hn+n_2}, H \geq 0$

Bouza (1981): $\theta = w_2 n_2$

The rules of Hansen-Hurwitz and Srinath depend on the decision of the sampler for fixing θ. Bouza's rule is determined as proportional to the proportion of non-responses. Hence, it is a randomized rule and fixing the sub-sampling size does not depend on the expertise of the sampler.

Having auxiliary information X the use of ratio estimators is commonly used. Under nonresponse we have the knowledge of the population mean of X, $\bar{X}$, and are computed the estimators:

$$\bar{x}^* = w_1\bar{x}_1 + w_2\bar{x}_2^*,\ \bar{x} = w_1\bar{x}_1 + w_2\bar{x}_2$$

$$s_x^{*2} = \frac{1}{n-1}\left(\sum_{i=1}^{n_1} x_i^2 + \frac{n_2}{n_2^*}\sum_{j=1}^{n_2^*} x_j^2 - n\bar{x}^*\bar{x}\right)$$

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{i=1}^{n_1} x_i y_i + \frac{n_2}{n_2^*}\sum_{j=1}^{n_2^*} x_j y_j - n\bar{y}^*\bar{x}\right)$$

The ratio estimator in this case is given by

$$\bar{y}_{ratio}^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{X}.$$

The regression estimator is

$$\bar{y}_{reg} = \bar{y}^* + b_{yx}^*(\bar{X} - \bar{x}^*),\ b_{yx}^* = \frac{s_{xy}^*}{s_x^{*2}}$$

In the next section we will consider the estimation problems.

3. The estimation problem

3.1. The class of estimators of Singh-Kumar

Singh and Kumar (2009) developed a class of estimators for $\bar{Y}$ when auxiliary information on two variables X and Z is available and non-responses are present. The sampling design analyzed was a DS one. They derived expressions of the mean squared error (MSE) for the estimators of the proposed class. Take

$$\bar{Q}^- = \sum_{j=1}^{N}\frac{Q_j}{N},\ \bar{q} = \sum_{j=1}^{n}\frac{q_j}{n},\ \bar{q}_t = \sum_{j=1}^{n_t}\frac{q_j}{n_t},\ \bar{q}_2^* = \sum_{j=1}^{n_2^*}\frac{q_j}{n_2^*},$$

$$Q = X, Y, Z,\ q = x, y, z,\ t = 1,2$$

Consider that we deal with missing information in the variable of interest (item non-response). Hence, we have responses in x and

z when a unit belongs to s. Following the model, we are going to estimate the mean of Y using (2.4). We may compute

$$\bar{g} = w_1\bar{g}_1 + w_2\bar{g}_2,\ g = x, z.$$

Take α as a fixed scalar. The estimators of this class are characterized by the general formula

$$\bar{y}_\alpha = \left(\bar{y}^* + \beta^*\left(\bar{X} - \bar{x}\right)\right)\frac{\bar{Z}}{\bar{Z} + \alpha(\bar{z} - \bar{z})}$$

$$\beta^* = \frac{s^*_{xy}}{s^2_x},\ s^*_{xy} = \sum_{i=1}^{n_1+n_2^*}\frac{(x_i - \bar{x})(y_i - \bar{y}^*)}{n_1 + n_2^* - 1},\ s^2_x = \sum_{i=1}^{n_1+n_2^*}\frac{(x_i - \bar{x})^2}{n_1 + n_2^* - 1}$$

It is considered that we know $\bar{Q},\ Q = X, Z$.

Let us use a Taylor Series development for (2.4). Take

$$\bar{y}^* = \bar{y}(1 + \varepsilon_y),\ \bar{g} = \bar{G}(1 + \varepsilon_g),\ g = x,\ z\ s^*_{xy} = \sigma_{xy}(1 + \varepsilon_{xy}),\ s^2_x = \sigma^2_x(1 + \varepsilon_{2x})$$

For accepting the validity of the development in Taylor Series is necessary that $|\alpha\varepsilon_z| < 1$ *and* $|\varepsilon_{2x}| < 1$ hold.

From the results of Singh and Kumar (2009) we may write

$$y_\alpha = \frac{\bar{Y}(1 + \varepsilon_y) - \bar{X}\varepsilon_x\frac{\sigma_{xy}(1+\varepsilon_{xy})}{\sigma^2_x(1+\varepsilon_{2x})}}{1 - \alpha\varepsilon_z} = \bar{Y}\left[\frac{1 + \varepsilon_y - \frac{\beta R_{xy}\varepsilon_x(1+\varepsilon_{xy})}{(1+\varepsilon_{2x})}}{(1 + \alpha\varepsilon_z)}\right] \tag{3.1}$$

where $\beta = \frac{\sigma_{xy}}{\sigma^2_x}, R_{xy} = \frac{\bar{X}}{\bar{Y}}$. Using the corresponding development for expanding (3.1) we have that

$$\bar{y}_\alpha = \bar{Y}\left[\frac{1 + \varepsilon_y - \frac{\beta R_{xy}\varepsilon_x(1+\varepsilon_{xy})}{(1+\varepsilon_{2x})}}{(1 + \alpha\varepsilon_z)}\right] \tag{3.2}$$

The approximation of the expectation, variances and covariances of the errors are developed considering that the terms of order larger than 2 are negligible. Then we may write:

$$E\left(\varepsilon^t_q\varepsilon^h_u\right) = \begin{cases} V_{1q}\ if\ q = u,\quad q = x, z;\ t = h = 1 \\ V_{2qu}\ if\ q = x, y\ and\ q \neq u = x, y, z, t = h = 1 \\ V_3\ if\ q = y\ and\ u = xy \\ V_4\ if\ q = s^2_x\ and\ u = x \end{cases} \tag{3.3}$$

where

$$V_{1q} = \frac{C^2_q}{n}, C^2_q = \frac{\sigma^2_q}{\bar{Q}^2}, \mathrm{q} = \mathrm{x}, \mathrm{y}, \mathrm{z}$$

$$V_{2qu} = \rho_{qu}C_qC_u, \rho_{qu} = \frac{\sigma_{qu}}{\sigma_q\sigma_u},$$

$$V_3 = \frac{N\mu_{21}}{(N-2)n\bar{X}\sigma_{xy}},$$

$$\mu_{21} = \sum_{i=1}^{N}\frac{(x_i - \bar{X})(y_i - \bar{Y})^2}{N},$$

$$V_4 = \frac{N\mu_{30}}{(N-2)n\bar{X}\sigma^2_x},\ \mu_{30} = \sum_{i=1}^{N}\frac{(x_i - \bar{X})^3}{N},$$

Let us look for the approximate bias and variance of the estimator. Considering again that the terms of order larger than 2 are negligible, we have the expansion

$$\bar{y}_\alpha - \bar{Y} \cong \bar{Y}\left(\varepsilon_y + \beta R_{xy}\varepsilon_x(\alpha\varepsilon_z - 1) + \alpha\varepsilon_z(\alpha\varepsilon_z - \varepsilon_y - 1) + \beta R_{xy}\varepsilon_x(\varepsilon_x - \varepsilon_{xy})\right)$$

Its expectation is equal to

$$Bias\ (\bar{y}_\alpha) = E(\bar{y}_\alpha - \bar{Y}) \cong \bar{Y}\left(\alpha\beta R_{xy}V_{2xy} + \alpha^2V_{1z} + \alpha V_{2zy} + \beta R_{xy}V_{1x}\right)$$

Note that only ε_y is affected by the existence of missing observations. Squaring both terms and calculating the variance is obtained

$$E\left[\left(\bar{y}_\alpha - \bar{Y}\right)^2|s\right] \cong \frac{1}{n}\left[\sigma^2_y\left(1 - \rho^2_{xy}\right) + \alpha R_{yz}(\alpha R_{yz} - 2A)\sigma^2_z + w_2\sigma^2_{2y}\right] = V[\bar{y}_\alpha|s] \tag{3.5}$$

$$R_{yz} = \frac{\bar{Y}}{\bar{Z}},\ A = \frac{\sigma_{yz}}{\sigma^2_z} - \frac{\sigma_{yx}}{\sigma^2_x}\frac{\sigma_{xz}}{\sigma^2_z}$$

The value of α determines a particular member of the class. An optimal estimator may be determined looking for the minimization of (3.5) by determining its optimum value. T is given by

$$\alpha_0 = \frac{A}{R_{yz}}$$

which depends of unknown population parameters, see Singh and Kumar (2009) for a detailed discussion on the members of this class.

3.2. A comparison of estimators

It is well known that to the first degree of approximation of the Taylor Series the conditional variances are

$$V(\bar{y}^*_{ratio}|s) \cong \frac{1}{n}\left(\sigma^2_y + \sigma^2_xR(R - 2B_{yx}) + \frac{n^*_2}{n}\sigma^2_{2y}\right)$$

For the regression estimator the conditional variance is

$$V(\bar{y}_{reg}|s) \cong \frac{1}{n}\left(\sigma^2_y\left(1 - \rho^2_{xy}\right) + \frac{n^*_2}{n}\sigma^2_{2y}\right)$$

Noting that

$$V(\bar{y}^*|s) = \frac{\sigma^2_Y}{n} + \frac{n^*_2}{n^2}\sigma^2_{2Y}$$

we have that if $\alpha_0 = \frac{A}{R_{yz}}$ is known

$$G(\bar{y}^*_{ratio}, \bar{y}_{\alpha_0}) = V(\bar{y}^*_{ratio}|s) - V[\bar{y}_{\alpha_0}|s] = \frac{1}{n}(A - R_{yz})^2\sigma^2_z$$

$$G\left(\bar{y}^*_{reg}, \bar{y}_{\alpha_0}\right) = V(\bar{y}^*_{reg}|s) - V[\bar{y}_{\alpha_0}|s] = \frac{A^2\sigma^2_z}{n}$$

$$G(\bar{y}^*, \bar{y}_{\alpha_0}) = V(\bar{y}^*|s) - V[\bar{y}_{\alpha_0}|s] = \frac{1}{n}\left(A^2\sigma^2_z + \rho^2_{xy}\sigma^2_y\right)$$

Hence $\bar{y}_{\alpha_0}$ is better than the other estimators. Singh and Kumar (2011) pointed out that, even if α_0 is unknown, $\bar{y}_\alpha$ is to be preferred, if the sampler evaluates for which feasible values of α it behaves better.

An evaluation of the magnitude of the gain in accuracy due to the use of $\bar{y}_{\alpha_0}$ was developed using real life data.

We developed the numerical analysis using population data obtained in three studies. A brief description of them is the following

Problem 1. 793 factories contaminate a source of water. They were inspected and was obtained

X = percent of samples with an index superior to the permitted level.

The historical report of this percent was also known

Z = historical percent of samples with an index superior to the permitted level.

The managers improved the collection of solid contaminants and a sample an reported

Y = percent of samples reported with an index superior to the permitted level.

$N_2 = 104$ factories did not send the report. They were visited and Y was obtained.

The parameters of interest are

$$\bar{X} = 24,7\ \sigma_x^2 = 31,4;\ \bar{Z}^- = 10,3\ \sigma_z^2 = 9,34;\ \bar{Y} = 18,7\ \sigma_y^2 = 7,78$$

$$\sigma_{yz} = -7,96;\ \sigma_{yx} = 10,20;\ \sigma_{xz} = 19,62;\ R_{yz} = 1,82,\ A = -1,19$$

Problem 2. 120 persons with VIH were included in an experiment with a new drug. The levels of hemoglobin were one of the measurements made to them. The variables involved were

X = measurement of hemoglobin before starting with the treatment.
Z = first measurement of hemoglobin after starting with the treatment.
Y = measurement of hemoglobin 6 months after starting with the treatment.

$N_2 = 51$ patients did not visit the hospital. They were visited and Y was obtained.

The parameters of interest are

$$\bar{X} = 6,60\ \sigma_x^2 = 1,43;\ \bar{Z}^- = 9,90\ \sigma_z^2 = 2,21;\ \bar{Y} = 8,06\ \sigma_y^2 = 3,08$$

$$\sigma_{yz} = 0,64;\ \sigma_{yx} = 1,01;\ \sigma_{xz} = 0,82;\ R_{yz} = 0,81,\ A = 0,03$$

Problem 3. 1840 farmers increased the area of their farms. The interest was to evaluate the tax to be pay. The variables involved were

X = initial area of the farms in hectares
Z = Actual area of the farms in hectares.
Y = Harvested area in hectares.

$N_2 = 176$ farmers did not return eth form of the tax to be pay patients. They were visited and Y was obtained.

The parameters of interest are

$$\bar{X} = 23,35\ \sigma_x^2 = 60,46;\ \bar{Z}^- = 34,86\ \sigma_z^2 = 88,75;\ \bar{Y} = 26,72\ \sigma_y^2 = 49,33$$

$$\sigma_{yz} = 15,58;\ \sigma_{yx} = -22,67;\ \sigma_{xz} = 46,93;\ R_{yz} = 0,77,\ A = 0,02$$

The resulting Gains in accuracy obtained are presented in Table 1.

Table 1
Gains in accuracy in 3 real life problems to the use of $\bar{y}_{\alpha_0}$.

Problem	$G(\bar{y}^*_{ratio}, \bar{y}_{\alpha_0})$	$G(\bar{y}^*_{reg}, \bar{y}_{\alpha_0})$	$G(\bar{y}^*, \bar{y}_{\alpha_0})$
1	84,06	13,16	16,47
2	1,34	0,20	0,91
3	49,92	0,04	0,41

Table 1 suggests that the improvements in accuracy due to the use of $\bar{y}_{\alpha_0}$ are very large when compared with $\bar{y}^*_{ratio}$. The error is decreased a little in the studies of VIH patients and farmers for the other estimators.

3.3. A comparison of the subsampling rules performance

From the above discussion is clear that the preference for a certain subsampling rule does not affect in the comparison of the estimators. Note that the effect of using a certain rule is important when we calculate the expected variance. Then we are interested in evaluating the behavior of the expectations under each rule. That is to compare the different expectations of $E\left[\frac{w_2(1-\theta)}{\theta n}\right]$.

The use of Hansen-Hurwitz's rule, HH, fixes that $\theta = 1/K$, $K \geq 1$. Then its use yields that

$$E\left[\frac{w_2(1-\theta)}{\theta n}\right] = \frac{W_2(K-1)}{n}, \tag{3.6}$$

When we use the rule of Srinath (1971), S, we have that

$$\theta = \frac{n_2}{Hn + n_2}$$

Doing some calculus is derived that $\frac{1-\theta}{\theta} = \frac{Hn}{n_2}$. Substituting in the conditional variance, we have

$$E\left[V\left(w_2(\bar{y}^*_2 - \bar{y}_2)|s\right)\right] = E\left(\frac{Hw_2\sigma^2_{2Y}}{nn_2}\right) \cong \frac{H}{n}\sigma^2_{2Y} \tag{3.7}$$

Comparing this term with (3.5), we should prefer HH to S whenever

$$\frac{W_2(K-1)}{n} \leq \frac{H}{n}$$

That is if $K \leq \frac{H+W_2}{W_2} = \frac{H}{W_2} + 1$.

A similar analysis of the use of the rule of Bouza (1981) needs of considering the new structure. Due to the randomness of θ we have that the conditional variance is

$$V\left(w_2(\bar{y}^*_2 - \bar{y}_2)|s\right) = \frac{(1 - n_2/n)}{n}\sigma^2_{2Y} = \frac{n_1}{n^2}\sigma^2_{2Y}$$

Its expectation is

$$E\left[V\left(w_2(\bar{y}^*_2 - \bar{y}_2)|s\right)\right] = \frac{W_1}{n}\sigma^2_{2Y} \tag{3.8}$$

Note that HH is to be preferred to B when

$$\frac{W_2(K-1)}{n} \leq \frac{W_1}{n}$$

That is if $K \leq \frac{W_1}{W_2} + 1 = \frac{1}{W_2}$ as $\frac{W_1}{W_2} \geq 0$ and $K \geq 1$ we may fix a value of K that satisfies this relationship.

Comparing S and B we have that the former generates a smaller coefficient if is satisfied the inequality

$$H \leq W_1$$

This relationship suggests that we S may be preferred if is used values of H smaller than 1.

Considering the costs, we may use the cost function

$$C = c_0 + c_1 n + c_2 n^*_2$$

Its expectation depends of the subsampling rule. The results are

$$E(C_{HH}) = c_0 + c_1 n + \frac{c_2 n W_2}{K} \tag{3.9}$$

Accepting that $E(n_2 - n^*_2)^t \cong 0$, $t > 2$, a development in Taylor Series allows deriving that

$$E(C_S) \cong c_0 + c_1 n + \frac{c_2 n W_2}{H + W_2} \tag{3.10}$$

We have that for B

$$E(C_B) = c_0 + c_1 n + c_2 (n W_2^2 + W_1 W_2) \tag{3.11}$$

In terms of the expected costs, we may look for the preference of the rules. We have:

$$HH \precsim SifK > H + W_2$$

$$HH \precsim BifK > \frac{nW_2}{nW_2^2 + W_1 W_2}$$

A comparison with S yields the preference rule,

$$S \precsim B \text{ if } H > \frac{n\left(1 - W_2^2\right) - W_1}{nW_2 + W_1 W_2}$$

It is easily derived that none of the rules may be more accurate and cheaper with respect to any of the other two simultaneously.

We used the results reported with four populations in the paper of Azeem and Hanif (2017) for establishing adequate values of the parameters of the subsampling rules. In the next table we have that N is the total number of units in the population questioned, N_1 the number of units responding the survey questions, N_2 the number of units which do not respond, σ_y^2 is the population variance of Y and σ_y^2 is the variance of Y for non-respondents part of the population (Table 2).

Note that for the populations 1 and 2 the variance of non-respondents is similar to the overall variance. Populations 3 and 4, have a considerably larger variance of the non-respondent strata than the population variance. We will use the weights observed in the inquiries of the different non-respondent stratum, $W_2 = \frac{N_2}{N}$.

We fixed a set of values of H in Table 3 for comparing HH with S in terms of their accuracy.

Note that for H = 0,1 fixing a value of K, for which HH is to be preferred, implies using large subsample sizes. An increase in the non-respondent's stratum determine also the need of using larger values of the sub sample size for preferring HH's rule.

Table 4 illustrates that for small subsampling sizes HH may have a better accuracy than the rule of Bouza. S will have the same behavior for relatively small values of H.

The analysis of the costs is presented in the next 2 tables.

We prefer using HH by using $K > H + W_2$. Analyzing Table 4 we have that in the population analyzed the relation holds for K > 1,20, which may be easily satisfied in practice.

The analysis of the costs associated with B needs to take into account the sample size. We consider the commonly used sampling fractions 0,01, 0,05 and 0,1 for illustrating. Note that the results in the Table 5 suggest that the sub sampling rule HH will have smaller expected costs than B if K > 9,82. The subsample parameter H should be very large for preferring S to B in terms of costs (Table 6).

Table 3
Selected Values of the lower bound $\frac{H}{W_2} + 1$ for accepting that HH is less variable than S.

	H				
W_2	0,1	0,3	0,5	0,7	1
0,1	2	4,00	6,00	8,00	11,00
0,15	1,67	3,00	4,33	5,67	7,67
0,2	1,50	2,5	3,5	4,50	6,00

Table 4
Selected Values of the lower bound for accepting that HH or S are less variable than B.

	$1 + W_1/W_2$	W_1
W_2		
0,1	10	0,90
0,15	6,67	0,85
0,2	5,00	0,80

Table 5
Selected Values of the upper bound $H + W_2$ for accepting that HH is less costly than S.

	H				
W_2	0,1	0,3	0,5	0,7	1
0,10	0,20	0,40	0,60	0,80	1,10
0,15	0,35	0,45	0,65	0,85	1,15
0,20	0,30	0,60	0,70	0,90	1,20

Table 6
Selected Values of the upper bounds of K and H for accepting that HH and S are less costly than B.

	HH $\frac{nW_2}{nW_2^2 + W_1 W_2}$			S $\frac{n\left(1-W_2^2\right)-W_1}{nW_2+W_1W_2}$		
W_2/n	50	250	500	50	250	500
0,10	8,47	6,53	9,82	82,37	64,36	95,48
0,15	6,05	6,38	6,45	39,36	41,54	41,97
0,20	4,63	4,92	4,96	22,18	23,53	23,82

4. Conclusions

Non-responses are present in the practice of survey research. Deciding to sub sample the non-respondents poses the need of deciding which will be the size of the sub-sample. The sampler must select a sub-sampling rule and fix a value of K or H or using instead a randomized rule. We developed a study of this problem when dealing with the class of estimators proposed by Singh and Kumar (2009).

Evaluating the preference of one of the rules may be performed by analyzing the effect of them in the corresponding expected

Table 2
Population Data.

Population	N	N1	N2	σ_Y^2	σ_{2Y}^2	Population	N	N1	N2	σ_Y^2	σ_{2Y}^2
1	5000	4500	500	102.007	99.99174	3	5000	4500	500	101.2633	5000
1	5000	4250	750	102.007	100.8224	3	5000	4250	750	101.2633	5000
1	5000	4000	1000	102.007	103.2349	3	5000	4000	1000	101.2633	5000
2	5000	4500	500	97.1206	94.5457	3	5000	4500	500	25.441	5000
2	5000	4250	750	97.1206	98.2761	4	5000	4250	750	25.441	5000
2	5000	4000	1000	97.1206	96.0935	4	5000	4000	1000	25.441	5000

variance or cost. The numerical study developed in Section 3 illustrated how a simple procedure allows deciding on the convenience of using one of the rules. The evaluation of the subsampling rules does not necessarily conveys to preferring one of them both in terms of accuracy and cost.

This study may be extended to other estimation procedures, as the product of a ratio and regression estimators proposed by Singh and Kumar (2011).

Acknowledgments

We heartily appreciate the suggestions of the referees which allowed improving the presentation of the results. One of the authors thanks CYTED and "A Cuban-Flemish Training and Research Program in Data Science and Big Data Analysis" projects for supporting his research.

References

Arnab, R., 2017. Survey Sampling Theory and Applications. Academic Press, Elsevier.

Azeem, M., Hanif, M., 2017. Joint influence of measurement error and non response on estimation of population mean. Commun. Statist. Theory Methods 4, 1679–1693.

Andridge, R.R., Thompson, K.J., 2015. Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models. Int. Statist. Rev. 83, 472–492.

Bouza, C.N., 1981. Sobre el problema de la fracción de submuestreo para el caso de las no respuestas. Trabajos de Estadística y de Investigación Operativa 32, 30–36.

Bouza, C.N., Covarrubias, D., Fernandez, Z., 2011. Handling with missing observations in simple random sampling and ranked set sampling. In: Lovric, Miodrag (Ed.), International Encyclopedia of Statistical Science. Springer-Verlag, Berlin, pp. 621–622. Part 8.

Hansen, M.H., Hurvitz, W.N., 1946. The problem of non-responses in survey sampling. J. Am. Statist. Assoc. 41, 517–523.

Heffetz, O., Reeves, D.B., 2016. Difficulty to Reach Respondents and Nonresponse Bias: Evidence from Large Government Surveys. NBER Working Paper No. 22333. (Consulted January 10-2018. http://www.nber.org/papers/w22333).

Lohr, S.L., 2010. Sampling: Design and Analysis. Brooks/Cole, Boston.

Särndal, C., Lundquist, P., 2014. Accuracy in estimation with nonresponse: a function of degree of imbalance and degree of explanation. J. Survey Statist. Methodol. 2, 361–3087.

Singh, S., 2003. Advanced Sampling Theory with Applications. Kluwer Academic, Dordrecht, The Netherlands.

Singh, H.P., Kumar, S., 2009. A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. Statist. Oper. Res. Trans. 33, 71–84.

Singh, H.P., Kumar, S., 2011. Combination of regression and ratio estimate in presence of nonresponse. Braz. J. Probable. Stat. 25, 205–217.

Srinath, K.P., 1971. Multiphase sampling in nonresponse problems. J. Am. Statist. Assoc. 66, 583–658.

Thompson, K.J., Washington, K.T., 2013. Challenges in the treatment of unit nonresponse for selected business surveys: a case study. Survey Methods: Insights from the Field. Last Consulted December 30, 2017 Available at: http://surveyinsights.org/?p=2991.

Torres van Grinsven, V., Bolko, I., Bavdaž, Villund, O., 2014. Comparing Subsample Approaches. Presentation to the 9th Workshop on Labor Force Survey Methodology Rome.