An improved class of estimators in RR surveys

Journal:	*Mathematical Methods in the Applied Sciences*
Manuscript ID	MMA-16-8827.R1
Wiley - Manuscript type:	Special Issue Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Rueda, Maria del Mar; University of Granada, Statistics and O.R.; Cobo, Beatriz; University of Granada Arcos, Antonio; University of Granada, Statistics and O.R.
Keyword:	Auxiliary information, Randomized Response Technique, Sampling theory, sample surveys

Research Article

Mathematical Methods in the Applied Sciences

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000
MOS subject classification: 62D05

An improved class of estimators in RR surveys

M. Rueda B. Cobo A. Arcos

This work proposes a general class of estimators for the population total of a sensitive variable using auxiliary information. Under a general randomized response model, the optimal estimator in this class is derived. Design based properties of proposed estimators are obtained. A simulation study reflects the potential gains from the use of the proposed estimators instead of the customary estimators. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: Auxiliary information; Randomized Response Technique; Horvitz-Thompson estimator

1. Introduction

Linear estimation parameters in a population is done through surveys. An example is the number of voters to a particular party in an election poll.

In many surveys it becomes necessary to probe into areas considered sensitive and potentially embarrassing. The validity of self-reports of sensitive attitudes and behaviors suffers from the tendency of individuals to distort their responses towards their perception of what is socially acceptable. As a consequence, studies self-report measures consistently underestimate the prevalence of undesirable attitudes or behaviors and overestimate the prevalence of desirable attitudes or behaviors. In an attempt to reduce this bias, [41] developed the randomized response technique (RRT). His idea spawned a vast volume of literature, see, for instance [5], [9], [16], [10], [34].

[22] and [21] have extended Warners model to the case where the responses to the sensitive question are quantitative rather than a simple yes or no. The respondent selects, by means of a randomization device, one of the two questions: one being the sensitive question, the other being unrelated. There are several difficulties that arise when using this unrelated question method ([37]). These difficulties are no longer present in the scrambled randomized response method introduced by [17]. In Eichhorn and Hayre model each respondent scrambles their response y by multiplying it by a random variable S and then reveals only the scrambled result $z = yS$ to the interviewer, thus, the scrambled randomized response model maintains the privacy of the respondents.[32] discussed the use of scrambled responses based on both multiplicative and additive model which involve the respondent adding and multiplying the answer to the sensitive question by two random number. [8] proposed a method that generalizes the Eichhorn and Hayre model which introducing a design parameter controlled by the researcher and used for randomizing the responses. Other important RR models are proposed by [18], [16] and by [20].

Most research into RRT techniques deals exclusively with the interest variable and does not make explicit use of auxiliary variables in the construction of estimators. Examples of these auxiliary variables in election polls could be sex, age, educational level or taxes. [14] pointed out that in sampling practice direct techniques for collecting information about non-sensitive characteristics make massive use of auxiliary variables to improve sampling design and to achieve higher precision in population parameter estimates. Nevertheless, very few procedures have been suggested to improve randomization technique performance using supplementary information. Regression estimators for scrambled variables are defined in [35], [15], [29] and [38]. [40] introduced the calibration of scrambled responses and find the conditional bias and variance of the proposed estimator. [36] proposed an empirical log-likelihood estimator for estimating the population mean of a sensitive variable in the presence of an auxiliary variable. [16] discussed the use of auxiliary information to estimate the population mean of a sensitive variable when data are perturbed by means of three scrambled response devices, namely the additive, the multiplicative and the mixed model. [25] proposed exponential-type estimators using one and two auxiliary variables.

From a mathematical point of view, a process of seeking an optimal estimator in a class of estimators for the total of sensitive characteristic arises; under a general model for the scrambling response and in presence of additional information.

Facultad de Ciencias, Avd. Fuente Nueva, 18071, Granada, Spain.
[*] *Correspondence to: M. Rueda. email: mrueda@ugr.es*

In this paper we suggest a class of estimators for a finite population total when the population totals of the auxiliary variables are known. In Section 2, we introduce the problem of estimating the total of the target population when there are scrambled variables. In section 3, we propose a general class of estimators for the population total. Proposed estimators are based upon auxiliary variables and assume that observations on the variable of interest are obtained using a Randomized Response Technique. We present particular estimators of the proposed class of estimators and we derive the asymptotic properties of these estimators. Using a real population, the proposed estimators are evaluated empirically in Section 4, and they are compared to alternative estimators. Finally, some conclusions are drawn.

2. Estimation of the population total in RRT

Consider a finite population U, consisting of N different individuals. Let y_i, $i = 1, \ldots, N$ be the value of the sensitive aspect under study for the ith population element. Our aim is to estimate the finite population total $Y = \sum_{i=1}^{N} y_i$ of the variable of interest y or the population mean $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$.

Assume that a sample s of individuals is chosen according to a non informative sampling design p with first order inclusion probabilities $\pi_i = \sum_{s \ni i} p(s), i \in U$ and second order inclusion probabilities $\pi_{ij} = \sum_{s \ni i,j} p(s), i,j \in U$. Let us assume that the operators E_d and V_d denote expectation and variance with respect to the sampling design (see [7]), and that the first and second order inclusion probabilities are positive.

If the value y_i is known exactly by observing the i-th individual, then the standard Horvitz and Thompson (HT) estimator of the total Y can be used:

$$\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

with variance:

$$V_{HT}(\hat{Y}) = \frac{1}{2} \sum_{i \neq} \sum_{j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 ,$$

which can be unbiasedly estimated as

$$\hat{V}_{HT}(\hat{Y}) = \frac{1}{2} \sum_{i \neq} \sum_{j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

Let y be the variable under study, a sensitive variable which can not be observed directly. In order to consider a wide variety of RR procedures, we consider the unified approach given by [5]. The interviews of individuals in the sample s are conducted in accordance with a RR model. Since y_i is not directly available from the respondent, y_i is estimated through the randomized response obtained from the ith respondent. Suppose that the ith respondent has to conduct a RR trial independently and z_i is the randomized response (or scrambled response) for the trial. For each $i \in s$ the RR induces a revised randomized response r_i such as $E_R(r_i) = y_i$ and $V_R(r_i) = \phi_i$ where the operators E_R and V_R denote expectation and variance with respect to randomization procedure RR.

As usual in the design-based approach to RR techniques, it is assumed that the sampling design and the randomization stage are independent of each other (see, e.g., [7]), and that the randomization stage is performed on each selected individual independently. In this general set-up, the Horvitz-Thompson type estimator for the population total of the sensitive characteristic y given by

$$\hat{Y}(r) = \sum_{i \in s} \frac{r_i}{\pi_i}$$

is an unbiased estimator since:

$$E(\hat{Y}(r)) = E_d(E_R(\sum_{i \in s} \frac{r_i}{\pi_i})) = Y$$

The variance of $\hat{Y}(r)$ can be obtained from

$$V(\hat{Y}(r)) = V_d(E_R(\hat{Y}(r)) + V_R(E_d(\hat{Y}(r))) = \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 + \sum_{i \in U} \frac{\phi_i}{\pi_i} \right] = V_{HT} + \sum_{i \in U} \frac{\phi_i}{\pi_i}$$

being V_{HT} the variance of the HT estimator based on the y_i's. An estimator of $V(\hat{Y}(r))$ is given by

$$\hat{V}(\hat{Y}(r)) = \left[\frac{1}{2} \sum_{i \neq} \sum_{j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j} \right)^2 + \sum_{i \in s} \frac{\hat{\phi}_i}{\pi_i} \right] .$$

This estimator is an unbiased estimator of $V(\hat{Y}(r))$ if $\hat{\phi}_i$ is an RR-unbiased for ϕ_i.

3. Estimators in the presence of auxiliary information

3.1. A general class of estimators for the total

The proposed estimators consider k auxiliary variables $x_1, \ldots, x_k$, for which the population totals $X_1, \ldots, X_k$, are known. We assume that the values of auxiliary variables can be observed directly in the sample. Our goal is to estimate the population parameter Y by using observations of the variables $r, x_1, \ldots, x_k$ in the sample s, and the known population values $X_1, \ldots, X_k$ associated with the auxiliary variables. We note by $\widehat{X}_h$ the Horvitz-Thompson estimator of the total X_h $(h = 1, ..., k)$.

Motivated by [39], we suggest the class of estimators of Y

$$\widehat{Y}_g^{(r)} = \{G(\hat{Y}(r), u_1, ..., u_k)\}, \qquad (1)$$

where $G(\cdot)$ is a function of $u_h = \widehat{X}_h / X_h$, continuous in a closed convex sub-space, $P \subset \mathbb{R}^{k+1}$, containing the point $(Y, 1, ..., 1) = (Y, \mathbf{1})$, and such that

(A1) $G(Y, \mathbf{1}) = Y$
(A2) $G'_0(Y, \mathbf{1}) = 1$ where $G'_0(Y, \mathbf{1})$ denoting the first partial derivative of $G(\cdot)$ with respect to $\hat{Y}(r)$, and
(A3) The first and second order partial derivatives of $G(\cdot)$ exist and are also continuous and bounded in P.

Now we studies some asymptotic design-based properties of $\widehat{Y}_g^{(r)}$. We consider the asymptotic framework of [23], in which the finite population U and the sampling design p are embedded into a sequence of such populations and designs indexed by N, $\{U_N, p_N\}$, with $N \to \infty$. We assume that $N_N \to \infty$ and $n_N \to \infty$, $n_N / N_N \to f \in (0, 1)$, as $N \to \infty$. Subscript N may be dropped for ease of notation, although all limiting processes are understood under the above cited conditions. Stochastic order $O_p(\cdot)$ is with respect to the aforementioned sequence of designs.

Theorem 1

Any estimator into the class (1) is asymptotically unbiased for Y.

Proof

By expanding G about the point $(Y, \mathbf{1})$ in a first order Taylor series, it is found that

$$\widehat{Y}_g^{(r)} = G(Y, \mathbf{1}) + (\hat{Y}(r) - Y) G'_0(Y, \mathbf{1}) + \sum_{h=1}^{k} G'_h|_{(Y,1)} (u_h - 1) + O_p(n^{-1}) \qquad (2)$$

where G'_h denotes the first order partial derivative with respect to u_h.

By taking expectations on both sides in (2) we obtain

$$E[\widehat{Y}_g^{(r)}] \simeq Y + E[\hat{Y}(r)] - Y + \sum_{h=1}^{k} E(\widehat{X}_h - X_h) \frac{G'_h|_{(Y,1)}}{X_h}.$$

We have $E[\hat{Y}(r)] = E_d E_R(\hat{Y}(r)) = Y$, $E[\widehat{X}_h] = E_d(\widehat{X}_h) = X_h$. Thus $E[\widehat{Y}_g^{(r)}] = Y + O(n^{-1})$ so the bias is of order n^{-1}.

Theorem 2 An approximation of the bias of the proposed class of estimators is given by:

$$B[\widehat{Y}_g^{(r)}] = \sum_{h<t} \frac{Cov(\widehat{X}_h, \widehat{X}_t)}{X_h X_t} G''_{ht}|_{(Y,1)} + \frac{1}{2} \sum_{h=1}^{k} \frac{V(\widehat{X}_h)}{X_h^2} G''_{hh}|_{(Y,1)}$$

$$+ \frac{1}{2} \frac{V(\hat{Y}(r))}{Y} G''_{00}|_{(Y,1)} + \frac{1}{2} \sum_{h=1}^{k} \frac{Cov(\widehat{X}_h, \hat{Y}(r))}{X_h} G''_{0h}|_{(Y,1)}$$

where G''_{ht} denote the second order partial derivative with respect to u_h and u_t, G''_{0h} is the second order partial derivative with respect to Y and u_h, and G''_{00} is second order partial derivative respect to Y.

Proof.

By expanding G about the point $(Y, \mathbf{1})$ in a second order Taylor series,

$$\widehat{Y}_g^{(r)} = Y + (\hat{Y}(r) - Y) + \sum_{h=1}^{k} G'_h|_{(Y,1)} (u_h - 1) +$$

$$\sum_{h<t} (u_h - 1)(u_t - 1) G''_{ht}|_{(Y,1)} + \frac{1}{2} \sum_{h=1}^{k} (u_h - 1)^2 G''_{hh}|_{(Y,1)} +$$

$$\frac{1}{2} \sum_{h=1}^{k} (u_h - 1)(\hat{Y}(r) - Y) G''_{0h}|_{(Y,1)} + \frac{1}{2} (\hat{Y}(r) - Y)^2 G''_{00}|_{(Y,1)} + O_p(n^{-2})$$

Taking expectations in the above second degree approximation we obtain the approximate bias (of order $O(n^{-2})$) of the proposed estimator.

Note that, under a general sampling design, the variances and covariances in Theorem 2, can be computed as:

$$V(\widehat{X}_h) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{x_{hj}}{\pi_j}\right)^2$$

and

$$Cov(\widehat{X}_h, \widehat{X}_t) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{x_{tj}}{\pi_j}\right)^2,$$

$V(\hat{Y}(r))$ is given in section 2. Then, we only need to obtain the $Cov(\widehat{X}_h, \hat{Y}(r))$. For this, using the covariance theorem, we have:

$$Cov(\widehat{X}_h, \hat{Y}(r)) = E_d(cov_R(\widehat{X}_h, \hat{Y}(r)) + cov_d(E_R(\widehat{X}_h), E_R(\hat{Y}(r))) =$$

$$E_d(0) + cov_d\left(\widehat{X}_h, \sum_{i\in s}\frac{y_i}{\pi_i}\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{x_{hi}}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Note 1. In deriving the expected value of $\widehat{Y}_g^{(r)}$ we assumed that the contribution of terms involving powers higher that the second is negligible. One can retain the terms up to and including degree third and four and proceed to obtain a better approximation to the expected value of $\widehat{Y}_g^{(r)}$. Unless n is small, the contribution of the third and fourth degree terms to the relative bias can be considered to be negligible. For appreciable large n, say 30 or larger, the approximation to $O(n^{-1})$ may be considered as adequate (see [4].)

Theorem 3 The asymptotic variance of any estimator into the class (1) verifies:

$$AV(\widehat{Y}_g^{(r)}) \geq V(\hat{Y}(r)) - \sigma'\,\Sigma^{-1}\sigma,$$

where $\Sigma = (a_{ht})_{(k\times k)}$ with $a_{hh} = V(\widehat{X}_h)$, $a_{ht} = Cov(\widehat{X}_h, \widehat{X}_t)$ and $\sigma = (Cov(\widehat{X}_1, \hat{Y}(r)), \ldots, Cov(\widehat{X}_k, \hat{Y}(r)))'$.

Proof.

By squaring both sides in expression (2), taking expectations and neglecting higher order terms we obtain the following approximation

$$V(\widehat{Y}_g^{(r)}) = E[\widehat{Y}_g^{(r)} - Y]^2 \simeq E\left[\hat{Y}(r) + \sum_{h=1}^{k} G'_h|_{(Y,1)}(u_h - 1) - Y\right]^2. \tag{3}$$

On differentiating (3) and equating to zero, we obtain the optimum values of the parameters as

$$(G'_1|_{(Y,1)}, \ldots, G'_k|_{(Y,1)})' = D^{-1}b,$$

where $D = (d_{ht})$, $b = (b_1, \ldots, b_k)'$ and

$$d_{ht} = \frac{Y^2 Cov(\widehat{X}_h, \widehat{X}_t)}{X_h X_t} \quad ; \quad b_h = \frac{Y Cov(\widehat{X}_h, \hat{Y}(r))}{X_h}.$$

On substituting the optimum values into (3) we obtain the minimum first order approximation for the variance

$$AV_{\min}(\widehat{Y}_g^{(r)}) = V(\hat{Y}(r)) - \sigma'\Sigma^{-1}\sigma = V(\hat{Y}(r))(1 - R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}),$$

where $R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}$ is the multiple correlation coefficient. This proofs the Theorem 3.

Note2. The above expression emphasizes the role of the auxiliary variables in improving the accuracy of the estimates. $(1 - R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k})$ denotes the reduction in the variance due to the use of auxiliary variables. We observe that the multiple correlation coefficient increases with the number of secondary variables and with the number of auxiliary parameters, hence the variance of proposed estimators is a monotone decreasing function of the number of secondary variables.

Note 3. In practice the value of $R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}$ is unknown, and this fact is even more complicated in this case as y, being sensitive, makes difficult making some guess on the value of the $AV_{\min}$. If we consider the generalized randomised response procedure given in [3] the revised values are given by $r_i = \frac{z_i - a}{b}$ being a and b constants, thus an idea of the multiple correlation coefficient $R^2_{\hat{Y}(r),\widehat{X}_1,\ldots,\widehat{X}_k}$ can be obtained from the correspondent correlation coefficient using the scrambled responses z_i.

The proposed class of estimator can be used to obtain an optimal difference type estimator using the idea proposed in [26] and [27]. Specifically, let us now consider a choice within the class G of the type

$$G(\hat{Y}(r), u_1, \ldots, u_k) = \hat{Y}(r) + \sum_{h=1}^{k} d_h(u_h - 1)X_h,$$

which yields to the difference estimator

$$\widehat{Y}_{gD} = \hat{Y}(r) + \sum_{h=1}^{k} d_h(X_h - \widehat{X}_h) \qquad (4)$$

The optimum d_h values are: $(d_1, ..., d_k)' = \Sigma^{-1}\sigma$ and

$$V(\widehat{Y}_{gD}) = V(\hat{Y}(r)) - \sigma'\Sigma^{-1}\sigma = AV_{\min}(\widehat{Y}_g^{(r)}).$$

It is interesting to note that the lower bound of the asymptotic variance of $\widehat{Y}_g^{(r)}$ is the variance of the difference estimator $\widehat{Y}_{gD}$ with the optimum d_h values. Thus, $\widehat{Y}_{gD}$ is, asymptotically, an optimal estimator into the class in the sense that it has a lower asymptotic variance, but is not unique. Any other estimator which attains the minimum variance bound is optimum as well, thus to the first order of approximation, i.e., up to terms O(n1), these estimators will be equivalent to the optimal difference estimator $\widehat{Y}_{gD}$. For d_h, with $h = 1, \dots, k$, known, this estimator has the advantage of providing exact results for the unbiasedness and the variance of the estimator of the total.

The optimum values d_h, with $h = 1, \dots, k$, depend on population values, which are generally unknown in practice, hence the optimal difference estimator $\widehat{Y}_{gD}$ cannot be used in general. Population values can be estimated by using sample values or using some replication methods.

After replacing Σ and σ by their estimators $\widehat{\Sigma}$ and $\widehat{\sigma}$, we obtain the difference type estimator

$$\widehat{Y}_{gd} = \hat{Y}(r) + \left(\Theta - \widehat{\Theta}\right)' \widehat{\Sigma}^{-1}\widehat{\sigma}, \qquad (5)$$

where $\widehat{\Theta} = (\widehat{X}_1, \dots, \widehat{X}_k)'$ and $\Theta = (X_1, \dots, X_k)'$.

3.2. Application to simple random sampling

Some asymptotic properties under simple random sampling are derived in this section.

Theorem 4

Assuming simple random sampling, the estimators $\widehat{Y}_g^{(r)}$, $\widehat{Y}_{gD}$ and $\widehat{Y}_{gd}$ are asymptotically unbiased and normally distributed.

Proof.

The asymptotic unbiasedness of $\widehat{Y}_{gD}$ and $\widehat{Y}_g^{(r)}$ is easily derived from its linear expression (6), and using the fact that $\hat{Y}(r)$ and $\widehat{X}_h$, with $h = 1, \dots, k$, are unbiased of their respective parameters. Similarly, since $\hat{Y}(r)$ and $\widehat{X}_h$ are asymptotically normal, the estimators $\widehat{Y}_g^{(r)}$ and $\widehat{Y}_{gD}$ are also asymptotically normal.

Results derived from [30] can be used to show that $\widehat{Y}_{gd}$ has asymptotically the same distribution than

$$\widehat{Y}_{gD} = \hat{Y}(r) + \left(\Theta - \widehat{\Theta}\right)' \Sigma^{-1}\sigma.$$

Following [30], the proposed difference estimator can be expressed as $\widehat{Y}_{gd} = T_n(\widehat{\Sigma}^{-1}\widehat{\sigma})$, whereas $\widehat{Y}_{gD}$ can be expressed as $\widehat{Y}_{gD} = T_n(\Sigma^{-1}\sigma)$, where $T_n(\widehat{\Sigma}^{-1}\widehat{\sigma})$ is a function of the data and uses the estimator $\widehat{\Sigma}^{-1}\widehat{\sigma} = (\widehat{d}_1, ..., \widehat{d}_{kl})'$, which is also a function of the data, consistently estimating the vector parameter $\Sigma\ ^{-1}\sigma$.

Let γ be a k dimensional vector of variables. By replacing the estimator $\widehat{\Sigma}^{-1}\widehat{\sigma}$ into $T_n(\cdot)$ by γ, which is denoted by $T_n(\gamma)$, the limiting mean of $T_n(\gamma)$ can be obtained when the actual parameter value is $\Sigma\ ^{-1}\sigma$, i.e.,

$$\mu(\gamma) = \lim_{n\to+\infty} E_{\Sigma^{-1}\sigma}[T_n(\gamma)] = \tilde{Y}$$

where $\tilde{Y}$ is the limiting value of Y as $N \to \infty$. Therefore

$$\left.\frac{\partial\mu(\gamma)}{\partial\gamma}\right|_{\gamma=\Sigma^{-1}\sigma} = \left(\left.\frac{\partial\mu(\gamma)}{\partial\gamma_1}\right|_{\gamma=\Sigma^{-1}\sigma}, \dots, \left.\frac{\partial\mu(\gamma)}{\partial\gamma_k}\right|_{\gamma=\Sigma^{-1}\sigma} \right) = (0, \dots, 0).$$

Assuming this condition, [30] showed that the limiting distribution of $T_n(\widehat{\Sigma}^{-1}\widehat{\sigma})$ is the same than the distribution of $T_n(\Sigma^{-1}\sigma)$, and hence the estimator $\widehat{Y}_{gd}$ is asymptotically unbiased and has the same asymptotic variance than $\widehat{Y}_{gD}$. This completes the proof.

3.3. Application to stratified sampling

Stratified sampling designs form an interesting and useful subclass of sampling designs. To define a stratified sampling design, on divide the population U of size N into L non overlapping subpopulations or strata U_l, having N_l units of U. In each stratum U_l we select a sample s_l by using a sampling design p_l independently of one another. For example, we consider that p_l is a simple random sampling without replacement of n_l size.

Let y_{il} be the value of the study variable y, and r_{il} the value of the randomized response for the ith population element of stratum l. If the values of auxiliary variables x_h are known for each population unit x_{hil}, in a similar way as [12] we can considered a general class of estimators in each stratum. The minimum asymptotic variance $AV_{minl} = V(\hat{Y}_l(r)) - \sigma_l' {\Sigma_l}^{-1} \sigma_l$ is achieved by the difference estimator in stratum l:

$$\widehat{Y}_{gDl} = \hat{Y}_l(r) + \sum_{h=1}^{k} d_{hl}(X_{hl} - \widehat{X}_{hl})$$

where $\hat{Y}_l(r) = N_l \sum_{i \in s_l} \frac{r_{il}}{n_l}$, $X_{hl} = \sum_{i \in U_l} x_{hil}$, $\widehat{X}_{hl} = N_l \sum_{i \in s_l} \frac{x_{hil}}{n_l}$, $d_{hl} = {\Sigma_l}^{-1} \sigma_l$ $\Sigma_l = (a_{ht})_{(k \times k)}$ with $a_{hh} = V(\widehat{X}_{hl})$, $a_{ht} = Cov(\widehat{X}_{hl}, \widehat{X}_{tl})$ and $\sigma_l = (Cov(\widehat{X}_{1l}, \hat{Y}_l(r)), \ldots, Cov(\widehat{X}_{kl}, \hat{Y}_l(r)))'$.

Thus, the separate estimator:

$$\widehat{Y}_{gD}^{st} = \sum_{l=1}^{L} \widehat{Y}_{gDl}$$

can be used for estimating the total Y. The properties of this estimator can be easily obtained by using the independence of sampling en each stratum. For example, an expression for the variance is given by:

$$V(\widehat{Y}_{gD}^{s}) = \sum_{l=1}^{L} V(\hat{Y}_l(r)) - \sigma_l' {\Sigma_l}^{-1} \sigma_l.$$

The above formulae are based on the assumption that n_l is large $\forall l$. This, however, is not always true in practice. To get over this difficulty we suggest a general class of combined estimators given by:

$$\widehat{Y}_{gc}^{(r)} = \{G(\hat{Y}_{st}(r), u_1, \ldots, u_k)\}, \qquad (6)$$

where $G(\cdot)$ is a function of $u_h = \widehat{X}_{hst}/X_h$, being $\widehat{X}_{hst} = \sum_l \widehat{X}_{hl}$ and $\hat{Y}_{st}(r) = \sum_l \hat{Y}_l(r)$.

In a similar way of section 3.1, the asymptotic variance of any estimator into the class verifies

$$AV(\widehat{Y}_{gc}^{(r)}) \geq V(\sum_{l=1}^{L} \hat{Y}_l(r)) - \sigma_{st}' {\Sigma_{st}}^{-1} \sigma_{st}$$

where $\Sigma_{st} = (a_{ht})_{(k \times k)}$ with $a_{hh} = V(\sum_l \widehat{X}_{hl})$, $a_{ht} = Cov(\sum_l \widehat{X}_{hl}, \sum_l \widehat{X}_{tl})$ and
$\sigma_{st} = (Cov(\sum_l \widehat{X}_{1l}, \sum_l \hat{Y}_l(r)), \ldots, Cov(\sum_l \widehat{X}_{kl}, \sum_l \hat{Y}_l(r)))'$.

An asymptotically optimal estimator is given by the combined difference estimator:

$$\widehat{Y}_{gcD} = \hat{Y}_{st}(r) + \left(\Theta - \widehat{\Theta}_{st}\right)' {\Sigma_{st}}^{-1} \sigma_{st}$$

being $\widehat{\Theta}_{st} = (\widehat{X}_{1st}, \ldots, \widehat{X}_{kst})'$.

The optimum sample allocation for the separate and for the combined difference estimators under a linear cost function can be obtained minimizing in n_l the above expressions given for its variances.

3.4. Other estimators in the class

For direct questioning many different estimators based on information of auxiliary variables have been proposed following different approaches. Some of them can be extended to our case of RR questioning. To the first order of approximation, some of these are equivalent to the difference estimator $\widehat{Y}_{gD}$ while others are less efficient. For space saving purpose, we do not show the plethora of estimator based on the information about parameters of auxiliary variables (see [13] for more examples of estimators in the class) .

Some example of estimators which attain the minimum variance bound of the class:

- The exponentation estimator (based on the idea of [1]), which is given by

$$\widehat{Y}_{g}^{exp} = \hat{Y}(r) \prod_{h=1}^{k} \left(\frac{X_h}{\widehat{X}_h}\right)^{\alpha_h}. \qquad (7)$$

- The exponentiation-difference estimator (based on the idea of [24]) given by

$$\widehat{Y}_{g}^{expD} = \hat{Y}(r) \prod_{h=1}^{k} \left(\frac{X_h}{\widehat{X}_h}\right)^{\alpha_h} + \sum_{h=1}^{k} b_h(X_h - \widehat{X}_h). \qquad (8)$$

Some example of estimator which are not optimum in the class.

- The exponential ratio type estimator (based on the idea of [6])

$$\widehat{Y}_g^{expR} = \hat{Y}(r)\prod_{h=1}^{k} exp\frac{X_h - \widehat{X}_h}{X_h + \widehat{X}_h} \qquad (9)$$

- The generalized regression-cum-exponential estimator (based on the idea of [25])

$$\widehat{Y}_g^{regcex} = (w_0\hat{Y}(r) + \sum_h w_h(X_h - \widehat{X}_h))exp(\frac{\sum_h(X_h - \widehat{X}_h)}{\sum_h(X_h + \widehat{X}_h)}) \qquad (10)$$

4. Simulation study

We have tested the real performance of the proposed estimators through simulation studies. The free statistical software R ([33]) was used to perform this simulation study. The library RRTCS of R ([11]) was used and, where necessary, we have developed new R-code implementing the proposed estimators.

For this purpose, we consider two studies with real and simulated populations.

The first simulation has been performed using two simulated populations used previously by [25]. The populations of size N=1000 are generated from a multivariate normal distribution (y, x_1, x_2) with the same vector of means (5 , 5, 5) and with different covariance matrices. The correlations in population 1 are $\rho_{yx_1} = 0.6844426$ and $\rho_{yx_2} = 0.6458839$, and the correlations in population 2 are $\rho_{yx_1} = 0.8659185$ and $\rho_{yx_2} = 0.8279276$.

We calculate the mean estimation of de variable of interest, dividing the proposed estimators above-named by population size.

For all populations, randomized response data were generated by using three different randomized response models. In recent years many models of randomized response have been proposed; we have included in the simulations these three models because there are some kind of kernel of RR procedures families:

- Eichhorn and Hayre model: In Eichhorn and Hayre model ([17]) each sample respondent is to report $z_i = S * y_i$ where S is a random sample from a population with known mean θ and known variance γ^2. In this model $E_R(z_i) = \theta, r_i = z_i/\theta, \phi = \gamma/\theta^2$
- Eriksson model: In Eriksson RR technique ([18]) it is assumed that the variable under study y can take any value in the known interval (a, b). M values $Q_1(= a), Q_2, ..., Q_M(= b)$ are chosen in the interval (a, b). The vector $Q = (Q_1, ..., Q_M)$ covers the range (a, b) and the value of M depends on the length of the interval. The respondent is supposed to report either the true value y_i with probability c or the Q_j value with probability $q_j(q_j > 0, \sum_j q_j = 1 - c)$ as his/her RR response. In this case $E_R(z_i) = c * y_i + \sum_j q_jQ_j$, $E_R(z_i^2) = c * y_i^2 + \sum_j q_jQ_j^2$, $r_i = (z_i - \sum_j q_jQ_j)/c$, $\phi_i = \alpha * y_i^2 + \beta * y_i + \gamma$ and $\hat{\phi}_i = \frac{\alpha r_i^2 + \beta r_i + \gamma}{1+\alpha}$ where $\alpha = \frac{1-c}{c}$, $\beta = \frac{-2\sum_j q_jQ_j}{c}$ and $\gamma = \frac{\sum_j q_jQ_j^2 - (\sum_j q_jQ_j)^2}{c^2}$
- Bar-Lev, Bobovitch and Boukai model: This model considered in ([8]) is a special case of Eriksson model. Here each of the sampled respondents is requested to rotate a spinner unobserved by the interviewer, and if the spinner stops in the shaded area, then the respondent is asked to disclose the true value y_i, otherwise, the respondent is asked to scramble their response y by multiplying it by a random variable S with known distribution. So in this method $z_i = y_i$ with probability p_1, $z_i = S * y_i$ with probability $1 - p_1$ and $r_i = \frac{z_i}{p_1 + (1-p_1) * E(S)}$.

The parameters for these models are:

- Eichhorn and Hayre $S \sim F(20, 20)$
- Eriksson $Q = (36656.0, 40200.5, 43698.0), c = 0.7, q_1 = q_2 = q_3 = 0.1$
- Bar-Lev, Bobovitch and Boukai $S \sim exp(1), p = 0.6$

For comparison purposes, the Horvitz Thompson estimator, $\hat{Y}(r)$, the difference estimator, $\widehat{Y}_{gd}$, the exponentiation estimator, $\widehat{Y}_g^{exp}$, and the exponential ratio type estimator, $\widehat{Y}_g^{expR}$, are computed.

In this context, simple random samples with different sizes ($n = 30, 50, 100, 200, 300$) have been drawn. We have tested the performance of these estimators with respect to the criteria: relative bias and mean square error through simulation studies.

$$RB = \frac{(1/T) * \sum_{i=1}^{T} |\hat{\hat{Y}} - \bar{Y}|}{\bar{Y}}; MSE = (1/T) * \sum_{i=1}^{T} (\hat{\hat{Y}} - \bar{Y})^2$$

where $\hat{\hat{Y}}$ is a given estimator and $\bar{Y}$ the population mean and T is the number of replicates, in our case 1000.

Table 1 and Table 2 present the RB and MSE statistics for population 1 and population 2 for some sample sizes. The value NAV indicates the number of auxiliary variables used in the estimation process.

Table 1. Relative bias and mean square error for some estimators and some RRT devices. Population 1 ($\rho_{yx_1} = 0.6844426$ and $\rho_{yx_2} = 0.6458839$)

		Eichhorn and Hayre		Eriksson		Bar-Lev, Bobovitch and Boukai	
	NAV	Relative bias	Mean square error	Relative bias	Mean square error	Relative bias	Mean square error
n=30							
Horvitz-Thompson	0	0.122105	0.632723	0.119747	0.56197	0.129363	0.668025
Difference	1	0.106300	0.478182	0.100275	0.405386	0.117332	0.554959
Exponentiation	1	0.106098	0.478106	0.100739	0.412917	0.117523	0.555011
ExpRatio	1	0.113011	0.546582	0.110181	0.479325	0.122619	0.600449
Difference	2	0.102131	0.441093	0.095689	0.369376	0.113429	0.521364
Exponentiation	2	0.102416	0.442217	0.096668	0.383995	0.114218	0.528864
ExpRatio	2	0.106037	0.482795	0.103034	0.419880	0.116618	0.550594
n=50							
Horvitz-Thompson	0	0.095893	0.378586	0.093214	0.340943	0.102764	0.428592
Difference	1	0.082850	0.285976	0.077843	0.238776	0.092737	0.339595
Exponentiation	1	0.083071	0.287671	0.078479	0.243594	0.093055	0.342951
ExpRatio	1	0.088917	0.327752	0.085621	0.289689	0.097201	0.382186
Difference	2	0.079662	0.269144	0.074633	0.221482	0.091103	0.332474
Exponentiation	2	0.080185	0.272712	0.075299	0.226668	0.091618	0.336719
ExpRatio	2	0.083768	0.293867	0.080110	0.254902	0.093304	0.355199
n=100							
Horvitz-Thompson	0	0.066638	0.180880	0.062108	0.155636	0.072514	0.212142
Difference	1	0.055411	0.127707	0.050513	0.102501	0.067067	0.180812
Exponentiation	1	0.055564	0.127824	0.050817	0.104094	0.067127	0.181399
ExpRatio	1	0.061385	0.153358	0.056563	0.129093	0.069137	0.193490
Difference	2	0.055044	0.125252	0.049137	0.098345	0.066055	0.177752
Exponentiation	2	0.055488	0.126739	0.049578	0.100861	0.066543	0.180119
ExpRatio	2	0.057970	0.137688	0.052878	0.112785	0.066884	0.182271
n=200							
Horvitz-Thompson	0	0.044155	0.078191	0.042876	0.076481	0.055025	0.118622
Difference	1	0.037582	0.056182	0.035919	0.052029	0.050808	0.100988
Exponentiation	1	0.037572	0.056068	0.035998	0.052412	0.050901	0.101165
ExpRatio	1	0.040683	0.066405	0.039351	0.064241	0.052819	0.109047
Difference	2	0.036508	0.053085	0.034267	0.047843	0.049808	0.097363
Exponentiation	2	0.036613	0.053426	0.034395	0.048208	0.049967	0.097604
ExpRatio	2	0.038524	0.058488	0.036468	0.055366	0.051175	0.101879
n=300							
Horvitz-Thompson	0	0.035667	0.051763	0.033776	0.046499	0.046136	0.081825
Difference	1	0.031225	0.039008	0.027283	0.031188	0.043103	0.071165
Exponentiation	1	0.031199	0.038963	0.027358	0.031354	0.043133	0.071313
ExpRatio	1	0.033400	0.044901	0.030836	0.038968	0.044535	0.076105
Difference	2	0.030198	0.036942	0.026489	0.029207	0.042907	0.069589
Exponentiation	2	0.030232	0.037094	0.026584	0.029420	0.043034	0.069915
ExpRatio	2	0.031567	0.040153	0.028728	0.033884	0.043536	0.072219

Table 2. Relative bias and mean square error for some estimators and some RRT devices. Population 2 ($\rho_{yx_1} = 0.8659185$ and $\rho_{yx_2} = 0.8279276$)

		Eichhorn and Hayre		Eriksson		Bar-Lev, Bobovitch and Boukai	
	NAV	Relative bias	Mean square error	Relative bias	Mean square error	Relative bias	Mean square error
n=30							
Horvitz-Thompson	0	0.106670	0.486278	0.084411	0.285370	0.117151	0.567523
Difference	1	0.086009	0.306308	0.060335	0.149244	0.100215	0.413828
Exponentiation	1	0.085868	0.301077	0.060513	0.151140	0.100051	0.407549
ExpRatio	1	0.095796	0.391094	0.072407	0.209705	0.108193	0.482807
Difference	2	0.081512	0.270888	0.054502	0.121669	0.094466	0.367412
Exponentiation	2	0.081203	0.263149	0.054845	0.124365	0.094024	0.356536
ExpRatio	2	0.088474	0.327594	0.063047	0.160628	0.100685	0.417619
n=50							
Horvitz-Thompson	0	0.079028	0.253276	0.065659	0.171725	0.090236	0.323928
Difference	1	0.065999	0.174221	0.046760	0.089259	0.075676	0.228148
Exponentiation	1	0.066010	0.173262	0.046973	0.090071	0.076008	0.229806
ExpRatio	1	0.071924	0.208751	0.056170	0.126811	0.083042	0.275677
Difference	2	0.061896	0.150319	0.041894	0.072563	0.072017	0.208150
Exponentiation	2	0.061813	0.148364	0.042217	0.074100	0.072274	0.209616
ExpRatio	2	0.065983	0.173529	0.048747	0.096205	0.077515	0.241732
n=100							
Horvitz-Thompson	0	0.054629	0.121016	0.046083	0.085763	0.063392	0.155941
Difference	1	0.045091	0.082673	0.033176	0.043984	0.055671	0.119216
Exponentiation	1	0.045157	0.082898	0.033443	0.045007	0.055921	0.120383
ExpRatio	1	0.049644	0.100069	0.039878	0.064161	0.059505	0.136610
Difference	2	0.041880	0.071943	0.029607	0.035960	0.052620	0.105662
Exponentiation	2	0.041917	0.072086	0.029940	0.036881	0.052676	0.106071
ExpRatio	2	0.045223	0.083422	0.034589	0.048632	0.055817	0.119154
n=200							
Horvitz-Thompson	0	0.037040	0.055113	0.028886	0.034199	0.046213	0.083300
Difference	1	0.031334	0.039189	0.021923	0.019378	0.041857	0.067203
Exponentiation	1	0.031323	0.039088	0.021899	0.019386	0.041889	0.067447
ExpRatio	1	0.033881	0.046152	0.024981	0.025738	0.043891	0.074681
Difference	2	0.028849	0.033905	0.020133	0.016159	0.040219	0.061899
Exponentiation	2	0.028735	0.033661	0.020272	0.016369	0.040227	0.062008
ExpRatio	2	0.031060	0.038938	0.022115	0.019979	0.041923	0.067677
n=300							
Horvitz-Thompson	0	0.028446	0.032859	0.022610	0.021417	0.037813	0.055494
Difference	1	0.023195	0.022167	0.017047	0.011603	0.033501	0.043550
Exponentiation	1	0.023255	0.022241	0.017101	0.011683	0.033620	0.043845
ExpRatio	1	0.025895	0.027081	0.019720	0.015993	0.035844	0.049533
Difference	2	0.021505	0.019052	0.015480	0.009754	0.032609	0.041205
Exponentiation	2	0.021502	0.019026	0.015684	0.009968	0.032655	0.041401
ExpRatio	2	0.023445	0.022483	0.017163	0.012297	0.034257	0.045268

The main conclusions derived in this study are:

- The relative absolute bias of the estimators are all within a reasonable range for the different sample sizes considered.

- More efficient estimators values are obtained if the correlations between the auxiliary variables and the principal are high.

- The values of relative bias and mean square error decrease as the sampling size increase, for all estimators and all RR techniques.

- The superiority of estimators based on auxiliary information is clear: the suggested estimators belonging to the class $\widehat{Y}_g^{(r)}$ are always more efficient than the Horvitz-Thompson estimator, whatever the adopted scrambling procedure.

- The values of relative bias and mean square error are very similar between $\widehat{Y}_{gd}$ and $\widehat{Y}_g^{exp}$. The difference in bias and MSE between these estimators is smaller as the sample size increases. This is expectable because the two estimator are asymptotically equivalents.

- Difference estimator $\widehat{Y}_{gd}$ or exponentiation estimator $\widehat{Y}_g^{exp}$ are the most efficient estimator for using one or two auxiliary variables.

- The suggested estimators $\widehat{Y}_{gd}$, and $\widehat{Y}_g^{exp}$ with two auxiliary variables perform better than the estimator with one auxiliary variable, as expected.

The second simulation study was carried out with a natural population called FAM1500 (see [19], [31]). The study considers this population of 1500 families living in an Andalusian province to investigate their income tax return. In these simulations we use as auxiliary variable, food expenses. The total for this variable is known. The sample is drawn by stratified sampling by house ownership. We select $T = 1000$ stratified samples of different samples sizes $n = 30, 50, 100, 200, 300$ with proportional allocation.

Table 3 shows the RB and MSE statistics for the FAM1500 population.

Table 3. Relative bias and mean square error for some estimators and some RRT devices in FAM1500 population

	Eichhorn and Hayre		**Eriksson**		**Bar-Lev, Bobovitch and Boukai**	
	Relative bias	Mean square error	Relative bias	Mean square error	Relative bias	Mean square error
n=30						
Horvitz-Thompson	0.062969	9970422	0.023375	1398428.1	0.075021	15303213
Difference	0.061768	9540168	0.016944	749632.8	0.073091	14721819
Exponentiation	0.061737	9521919	0.016935	749648.3	0.073146	14728094
ExpRatio	0.061755	9553421	0.018058	843063.9	0.073385	14791268
n=50						
Horvitz-Thompson	0.047696	5868140	0.017066	729256.6	0.059368	9310484
Difference	0.047041	5692488	0.012862	421624.8	0.058332	9065942
Exponentiation	0.046981	5674113	0.012844	420192.0	0.058328	9065732
ExpRatio	0.046915	5668912	0.013339	451857.9	0.058413	9070390
n=100						
Horvitz-Thompson	0.039153	3729670	0.012654	407192.7	0.044445	5153787
Difference	0.038019	3524888	0.009457	224479.4	0.044083	5021198
Exponentiation	0.038023	3526251	0.009448	224163.8	0.044082	5017795
ExpRatio	0.038218	3556232	0.009901	246333.3	0.044052	5021400
n=200						
Horvitz-Thompson	0.028094	1951266	0.008686	194422.6	0.033014	2895354
Difference	0.027709	1874658	0.007003	123289.9	0.032623	2823758
Exponentiation	0.027701	1874357	0.007000	123240.1	0.032615	2822372
ExpRatio	0.027710	1883025	0.007120	129032.5	0.032629	2827938
n=300						
Horvitz-Thompson	0.024538	1425025	0.007346	136750.26	0.028466	1981292
Difference	0.024299	1378528	0.006107	92563.45	0.028095	1946481
Exponentiation	0.024299	1378814	0.006104	92480.66	0.028102	1947923
ExpRatio	0.024296	1383988	0.006228	96837.40	0.028172	1947810

Results of this simulation are in accordance with those obtained in the previous study: for all randomized response models used, there is a decrease in the relative bias and the mean square error if we compared the Horvitz-Thompson estimator with others estimators which use auxiliary information. The gain in efficiency is relevant for the Eriksson model. The values of relative bias and mean square error for all models are very similar between $\widehat{Y}_{gd}$ and $\widehat{Y}_g^{exp}$. Nevertheless, the proposed estimators $\widehat{Y}_{gd}$ and $\widehat{Y}_g^{exp}$ dominate the other, for any choice of the sample size and the randomized technique.

5. Conclusions

Privacy protection is a crucial objective for both data collection and statistical analyses in the study of sensitive variables as tax evasion, sexual behaviours, reckless driving, indiscriminate gambling, abortion, etc. Randomized response methods can be beneficially employed for collecting and analysing information about sensitive topics. Many studies have assessed the validity of RR methods showing that they can produce more reliable answers than other conventional data collection methods but RRT estimates are affected by higher sampling variance than direct questioning estimates. The loss of efficiency represents the cost to pay for obtaining more reliable information by reducing response bias. Consequently, achieving efficient estimates which are comparable with those under direct questioning may require considerable larger sample with an obvious increasing of the survey cost. A way to reduce the sampling variance of RRT estimators is the use of auxiliary information.

This paper makes an attempt to provide a general form of estimation of a total of a sensitive variable using auxiliary information of supplementary variables. This situation is very common in the sampling practice. A lot of estimators were proposed to deal with the problem of estimating a total of a non sensitive variable when supplementary information is available. Nevertheless, in spite of different ideas followed to construct the estimators, most of them show the same efficiency. The unawareness of this aspect caused a proliferation of several types of estimators. This situation could be extended to the case of sensitive variables.

A class of estimators of a finite population total under a general randomized response model has been defined when the sample is obtained under a general sampling design. Estimators belonging to this class have been proven to be asymptotically design unbiased and their asymptotic variances has been obtained. We provide also the expression of an optimal estimator in the class, the difference estimator, that is the estimator that attains the asymptotic minimum variance bound. This estimator is studied for some elementary sampling design as simple random sampling and stratified sampling. We introduce other estimators in this class, some of them have asymptotically the same variance as the optimal difference estimator.

We have conducted a simulation study to check the performance of the proposed estimators. The results obtained from the simulation study support the theoretical background and show that, given a set of auxiliary variables, the method performs well under different scenarios in both natural and artificial populations.

In short, this paper generalizes some existing results about the use of auxiliary information in RRT ([29]) and want to contribute stopping the tentative to spread in the RRT literature new estimators by extending non-optimum estimators conceived for direct questioning surveys.

Acknowledgement

This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2015-63609-R and FPU grant program, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía).

References

1. Abu-Dayyeh WA, Ahmed MS, Ahmed RA, Muttlak HA. Some estimators of finite population mean using auxiliary information. *Applied Mathematics and Computation* 2003; 139: 287–298.
2. Al-Omari AI, Bouza CN, Herrera C. Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient. *Quality and Quantity* 2013; 47: 353–365.
3. Arcos A, Rueda M, Singh S. Generalized approach to randomized response for quantitative variables. *Quality and Quantity* 2015; 49: 1239–1256.
4. Ayachit GR. 1953 *Some aspects of large-scale sample surveys with particular reference to the ratio method of estimation*. M.Sc. Thesis, Bombay University, Bombay.
5. Arnab R. Optional randomized response techniques for complex survey designs. *Biom. J.* 2004; 46(1): 114–124.
6. Bahl S, Tuteja RK. Ratio and product type exponential estimator. *Information and Optimization Sciences* 1991; XII(I): 159–163
7. Barabesi L, Diana G, Perri PF. Design-based distribution function estimation for stigmatized populations *Metrika* 2013; 76: 919–935.
8. Bar-Lev SK, Bobovitch E, Boukai B. A note on randomized response models for quantitative data. *Metrika* 2004; 60: 255–260.
9. Bouza CN, Herrera C, Mitra PG. A review of randomized responses procedures: the qualitative variable case. *Investigación Oper.* 2010; 31(3): 240–247.
10. Chaudhuri A, Mukherjee R. 1988. *Randomized response. Theory and techniques. Statistics: Textbooks and Monographs, 85* New York: Marcel Dekker, Inc. xvi, 162 p.
11. Cobo B, Rueda M, Arcos A. 2015. *RRTCS: Randomized Response Techniques for Complex Surveys*. R package version 1.0.

12. Dalabebara M, Sahoo LN. A class of estimators in stratified sampling with two auxiliary variables. *Jour. Inti. Soc. Ag. Statistics* 1997; 50(2): 144–149.
13. Diana G, Perri PF. Estimation of finite population mean using multi-auxiliary information. *Metron* 2007; LXV(I): 99–112.
14. Diana G, Perri PF. New scrambled response models for estimating the mean of a sensitive quantitative character. *J. Appl. Stat.* 2010; 37: 1875–1890.
15. Diana G, Perri PF. A class of estimators for quantitative sensitive data. *Statistical Papers* 2011; 52(3): 633–650.
16. Diana G, Perri PF. A calibration-based approach to sensitive data: a simulation study. *Journal of Applied Statistics* 2012; 39(1): 53–65.
17. Eichhorn BH, Hayre LS. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference* 1983; 7: 306–316.
18. Eriksson SA. A new model for randomized response. *International Statistical Review* 1973; 41: 40–43.
19. Fernandez FR, Mayor JA. 1994 *Muestreo en Poblaciones Finitas: Curso Basico.* PPU, Barcelona.
20. Gjestvang CR, Singh S. A new randomized response model. *J. Royal Statist. Soc Ser. B* 2006; 68: 523–530
21. Greenberg BG, Abul-Ela AL, Simmons WR, Horvitz DG. The unrelated question RR model: Theoretical framework. *J. Amer. Statist. Assoc.* 1969; 64: 520–539.
22. Horvitz DG, Shah BV, Simmons WR. 1967. The unrelated question RR model. *In: Proc. Social Statist. Sec.* ASA, 65–72.
23. Isaki CT, Fuller WA. Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 1982; 77(377): 89–96.
24. Kadilar C, Cingi H. A new estimator using two auxiliary variables. *Applied Mathematics and Computation* 2005; 162: 901–908.
25. Koyuncu N, Gupta S, Sousa R. Exponential-Type Estimators of the Mean of a Sensitive Variable in the Presence of Nonsensitive Auxiliary Information. *Communications in Statistics - Simulation and Computation* 2014; 43 (7): 1583–1594.
26. Montanari GE. Post-sampling Efficient QR-prediction in Large-sample Surveys. *International Statistical Review* 1987; 55: 191–202.
27. Montanari GE. On regression estimation of finite population means. *Survey Methodology* 1998; 24,1: 69–77.
28. Odumade O, Singh S. An Alternative to the Bar-Lev, Bobovitch and Boukai Randomized Response Model. *Sociological Methods & Research* 2010; 39: 206–21
29. Perri PF, Diana G. Scrambled response models based on auxiliary variables. *In: Advances in Theoretical and Applied Statistics, Torelli, Nicola; Pesarin, Fortunato; Bar-Hen, Avner (Eds.),* 2013; 281-291, Spriger-Verlag, Berlin.
30. Randles RH. On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* 1982; 10: 462–474.
31. Rueda M., Arcos, A. On estimating the median from survey data using multiple auxiliary information. *Metrika* 2001; 54 (1): 59–76.
32. Saha A. A simple randomized response technique in complex surveys. *Metron* 2007; LXV: 59–66.
33. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015, URL: https://www.R-project.org/.
34. Santiago A, Bouza CN, Sautto JM, Al-Omari AI. Randomized Response Procedure for the Estimation of the Population Ratio using Ranked Set Sampling. *Journal of Mathematics and Statistics* 2016; 12 (2). DOI: 10.3844/jmssp.2016.107.114.
35. Singh S, Joarder AH, King ML. Regression analysis using scrambled responses. *Aust. J. Stat.* 1996; 38: 201–211.
36. Singh S, Kim JM. A pseudo-empirical log-likelihood estimator using scrambled responses. *Statist. Probab. Lett.* 2011; 81: 345–351.
37. Singh S, Sedory S, Arnab R. Estimation of Finite Population Variance Using Scrambled Responses in the Presence of Auxiliary Information. *Communications in Statistics - Simulation and Computation* 2015; 44(4): 1050–1065.
38. Singh , Tracy DS. Ridge regression using scrambled responses. *Metron* 1999; LVII: 147–157.
39. Srivastava SK, Jhajj HS. A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika* 1981; 68: 341–343.
40. Tracy D, Singh S. Calibration estimators in randomized response survey. *Metron* 1999; LVII: 47–68.
41. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 1965; 60(309): 63–69.