Mathematics and Computers in Simulation

New estimation techniques for ordinal sensitive variables

--Manuscript Draft--

Manuscript Number:	MATCOM-D-19-01190R1
Article Type:	SI: MACMACS 2019
Keywords:	Ordinal logistic regression; randomized response; Calibration; monte carlo simulation
Corresponding Author:	Maria del Mar Rueda University of Granada Granada, SPAIN
First Author:	Maria del Mar Rueda
Order of Authors:	Maria del Mar Rueda
	Beatriz Cobo
	Pier Francesco Perri
Abstract:	Methods to analyze multicategorical variables are extensively used in sociological, medical and educational research. Nonetheless, they have a very sparse presence in finite population sampling when sensitive topics are investigate and data are obtained by means of the randomized response technique (RRT), a survey method based on the principle that sensitive questions must not be asked directly to the respondents. The RRT is used with the aim of reducing social desirability bias, which is de ned as respondents' tendency to release personal information according to what is socially acceptable. This nonstandard data-collection approach was originally developed to deal with dichotomous responses to sensitive questions. Later, the idea has been extended to multicategory responses. Here we consider ordinal variables with more than two response categories. In particular, we first discuss the theoretical framework for estimating the frequency of the ordinal categories when data are subjected to misclassification due to the use of a particular RRT. Then we show how it is possible to improve the efficiency of the inferential process by employing auxiliary information at the estimation stage through calibration. Finally, we assess the performance of the proposed estimators in a Monte Carlo simulation study.

New estimation techniques for ordinal sensitive variables

M. Rueda[1] , B. Cobo[2] , P.F. Perri[3]

[1] *University of Granada*
[2] *Complutense University of Madrid*
[3] *University of Calabria*

Abstract

Methods to analyze multicategorical variables are extensively used in sociological, medical and educational research. Nonetheless, they have a very sparse presence in finite population sampling when sensitive topics are investigate and data are obtained by means of the randomized response technique (RRT), a survey method based on the principle that sensitive questions must not be asked directly to the respondents. The RRT is used with the aim of reducing social desirability bias, which is defined as the respondent tendency to release personal information according to what is socially acceptable. This nonstandard data-collection approach was originally developed to deal with dichotomous responses to sensitive questions. Later, the idea has been extended to multicategory responses. In this paper we consider ordinal variables with more than two response categories. In particular, we first discuss the theoretical framework for estimating the frequency of ordinal categories when data are subjected to misclassification due to the use of a particular RRT. Then, we show how it is possible to improve the efficiency of the inferential process by employing auxiliary information at the estimation stage through the calibration approach. Finally, we assess the performance of the proposed estimators in a Monte Carlo simulation study.

Keywords: Ordinal logistic regression, randomized response, calibration, Monte Carlo simulation
2008 MSC: 62D05

1. Introduction

In surveys to estimate the proportion of people bearing a stigmatizing characteristic like habitual gambling, marijuana consumption, tax evasion

and so on, people often do not respond truthfully when asked personal or embarrassing questions. Obtaining valid and reliable information depends on the cooperation of the respondents, and this depends, in turn, on the confidentiality of their responses. Empirical researches based on self-report measures, due to the so called social desirability bias, run the risk to produce measurement errors mainly ascribable to high nonresponse rates and misreporting.

To reduce these nonsampling errors, [10] developed a data-collection procedure, the randomized response (RR) technique (RRT), that allows researchers to elicit sensitive information while guaranteeing privacy to respondents. The technique decreases social desirability bias, enhances respondent cooperation and procures more reliable estimates. Although the RRT was originally developed to handle dichotomous (yes/no) responses to sensitive questions, it has been extended to polychotomous sensitive variables that include multicategory responses.

Along this line, [6] proposed a method to estimate the proportions of people classified in m mutually exclusive and exhaustive categories which are coded $1, \ldots, m$. The RR device consists of a box containing colored balls, say red and white balls. Each of the white balls is marked with one of the numbers $1, \ldots, m$. Let p_r denote the proportion of red balls in the box and p_{w_i} that of white balls marked i, $i = 1, \ldots, m$. Each survey participant is asked to select in private a ball from the box and report his/her true category of the sensitive character if the ball is red, or the number marked on the ball if it is white. Obviously, the interviewer does not know which type of colored ball has been selected and, thus, the true category of the respondents remains undisclosed and privacy protected. In order to estimate the true proportion P_i of people who possess the ith category, a simple random sample of size n is taken. Let n_i denote the number of respondents reporting the ith category; hence, an unbiased estimator of P_i is given by $\widehat{P}_i = \left(\frac{n_i}{n} - p_{wi}\right)/p_r$, for $i = 1, \ldots, m$. A generalization of the Liu-Chow method [6] is discussed in [1] by assuming a randomization device such that an interviewee belonging to the ith category reports the jth category with probability p_{ji}, $\sum_{j=1}^{m} p_{ji} = 1$. Further, [5] presented a more complex RR method for the case of two sensitive questions of interest.

In spite of the great deal of RR devices discussed in the literature, there are very few studies that address the problem of estimating sensitive behaviors from ordinal variables. The aim of this paper, thus, is to propose new estimation techniques in the RR setting when responses can be sorted.

Section 2 presents, under a general sampling designs, a RR method for estimating the frequency of ordinal categories. In Section 3, we discuss the use of auxiliary information to improve the efficiency of the estimates. Emphasis is given to the model-calibration approach. To evaluate the performance of different estimators for the prevalence of sensitive categories, we present a simulation study in Section 4 based on 2012 PISA survey data. Section 5 concludes the work with some final consideration.

2. RRT estimators under a general sampling design

Let $U = \{1, \ldots, k, \ldots, N\}$ denote a finite population of N identifiable and distinct units. Consider a sensitive variable Y with m response categories, which are labeled $1, \ldots, m$, and let y_k denote the category label for the kth population unit. Our aim is to estimate the frequency distribution of Y in the population. In so doing, let z_i be an indicator variable such that for each unit $k \in U$, $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise, $i = 1, \ldots, m$. The problem thus, is to estimate the proportion $P_i = \sum_{k \in U} z_{ki}/N$, $i = 1, \ldots, m$.

Let s be a probability sample of size n drawn from U according to a sampling design $p(\cdot)$. The considered sampling design induces first-order inclusion probabilities $\pi_k = \sum_{s \ni k} p(s)$, second-order inclusion probabilities $\pi_{kl} = \sum_{s \ni k,l} p(s)$ with $k, l \in U$, and design weights $d_k = 1/\pi_k$. Under this general framework, from sampling theory it is well known that a design-unbiased estimator of P_i is the Horvitz-Thompson (HT) estimator [4] defined as $\widehat{P}_{i,\mathrm{HT}} = \sum_{k \in s} z_{ki} d_k / N$.

Without loss of generality, we assume that sensitive data are collected on the sampled units by means of a RR method which drives the respondent to report his/her true category with probability p_T, or is forced to report the ith category with probability p_{F_i}, $p_T + \sum_{i=1}^{m} p_{F_i} = 1$ (Liu-Chow method is an example of this randomization method). Furthermore, we assume that the sampling design and the randomization stage are independent each other, and that the randomization stage is performed on each sampled units independently.

Let r_k denote the category released by kth respondent subjected to the randomization mechanism, and let us introduce the indicator variable $r_{ki} = 1$ if $r_k = i$ and $r_{ki} = 0$ otherwise. Accordingly, under the adopted RR device, let us consider the transformed randomized response $R_{ki} = (r_{ki} - p_{F_i})/p_T$.

Then, an HT-type estimator for the proportion P_i is given by

$$\widehat{P}^*_{i,\mathrm{HT}} = \frac{1}{N}\sum_{k\in s} R_{ki} d_k, \quad i = 1, \ldots, m,$$

and the following result holds.

Theorem 1. *The estimator $\widehat{P}^*_{i,\mathrm{HT}}$ is an unbiased estimator of the population proportion P_i.*

Proof. First of all, note that $\widehat{P}^*_{i,\mathrm{HT}}$ is affected by two sources of variability: one induced by the sampling design, the other by the randomization device. Accordingly, let E_d and V_d denote the expectation and variance operators with respect to the sampling design. Analogously, let E_r and V_r denote the expectation and variance operators with respect to the RR mechanism. Preliminarily observing that $E_r(r_{ki}) = p_T z_{ki} + p_{F_i}$, it is ready proved that R_{ki} is an RR-unbiased estimator of the latent value z_{ki}, i.e. $E_r(R_{ki}) = z_{ki}$. Hence, it straightforward follows that:

$$E(\widehat{P}^*_{i,\mathrm{HT}}) = E_d E_r\left(\frac{1}{N}\sum_{k\in s} R_{ki} d_k\right) = E_d\left(\frac{1}{N}\sum_{k\in s} E_r(R_{ki}) d_k\right) = E_d\left(\frac{1}{N}\sum_{k\in s} z_{ki} d_k\right).$$

Noting the the expression in the last parenthesis denotes the standard HT estimator $\widehat{P}_{i,\mathrm{HT}}$, the Theorem remains proved on the basis of previous considerations.

Theorem 2 *The variance of $\widehat{P}^*_{i,\mathrm{HT}}$ is given by*

$$V(\widehat{P}^*_{i,\mathrm{HT}}) = V(\widehat{P}_{i,\mathrm{HT}}) + \frac{1}{(Np_T)^2}\sum_{k\in U}\phi_{ki} d_k,$$

where $V(\widehat{P}_{i,\mathrm{HT}})$ is the variance of the standard HT estimator, $V(\widehat{P}_{i,HT}) = \sum_{k,l\in U}(d_k d_l \pi_{kl} - 1) z_{ki} z_{li}/N^2$ [8], and $\phi_{ki} = (p_T z_{ki} + p_{F_i})(1 - (p_T z_{ki} + p_{F_i}))$.

Proof. Using the notation introduced in the proof of Theorem 1, the variance of $\widehat{P}^*_{i,\mathrm{HT}}$ stems from the variance decomposition formula:

$$V(\widehat{P}^*_{i,\mathrm{HT}}) = V_d(E_r(\widehat{P}^*_{i,\mathrm{HT}})) + E_d(V_r(\widehat{P}^*_{i,\mathrm{HT}})).$$

Considering the single terms of the decomposition, we have:

$$V_d(E_r(\widehat{P}^*_{i,\mathrm{HT}})) = V_d\left(\frac{1}{N}\sum_{k\in s} E_r(R_{ki})d_k\right) = V_d\left(\frac{1}{N}\sum_{k\in s} z_{ki}d_k\right) = V(\widehat{P}_{i,\mathrm{HT}}).$$

Hence, the first part of $V(\widehat{P}^*_{i,\mathrm{HT}})$ is obtained. For the second addendum, let us introduce the indicator variable I_k, with $I_k = 1$ if the kth population unit is included in the sample, and, for the sake of brevity, let ϕ_{ki} denote the variance of the indicator variable r_{ki} under the adopted RR method, $V_r(r_{ki})$. Now, since $E_d(I_k) = \pi_k$ and $V_r(R_{ki}) = \phi_{ki}/p_T^2$, we have:

$$\begin{aligned} E_d(V_r(\widehat{P}^*_{i,\mathrm{HT}})) &= E_d\left(\frac{1}{N^2}\sum_{k\in s} V_r(R_{ki})d_k^2\right) = E_d\left(\frac{1}{N^2 p_T^2}\sum_{k\in U} \phi_{ki}d_k^2 I_k\right) \\ &= \frac{1}{(Np_T)^2}\sum_{k\in U}\phi_{ki}d_k^2 E_d(I_k) = \frac{1}{(Np_T)^2}\sum_{k\in U}\phi_{ki}d_k. \end{aligned}$$

3. New estimators for items with ordinal outcomes

Usually, in real surveys, additional information about auxiliary variables is available for each unit of the population that can be profitably used at the estimation stage to increase the efficiency of the inferential process. Let us assume that a set of $q > 1$ auxiliary variables is available and let $\mathbf{x}_k = (x_{1k}, \ldots, x_{q_k})'$ the vector of the observed q variables for the kth unit of the population. There are many techniques to employ the auxiliary variables. Among these, the calibration approach introduced by [3] is widely used in practice since it ensures that survey estimates are coherent with those already in the public domain, while simultaneously reducing non-coverage, nonresponse and selection biases. We therefore extend this methodology to the RR scheme considered in Section 2.

3.1. The model-calibration ordinal estimator

We consider a superpopulation ordinal logistic model by assuming that the finite population under study $\mathbf{y} = (y_1, ..., y_N)'$ is the determination of the superpopulation random variable vector $\mathbf{Y} = (Y_1, ..., Y_N)'$, that can be described by the superpopulation model:

$$P(y_k = i|\mathbf{x}_k) = E_\xi(z_{ki}|\mathbf{x}_k) = \begin{cases} \frac{\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_i+\beta\mathbf{x}_k)}, & i = 1 \\ \frac{\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)} - \frac{\exp(\alpha_{i-1}+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_{i-1}+\boldsymbol{\beta}\mathbf{x}_k)}, & i = 2, \ldots, m \end{cases},$$

where E_ξ denotes the expectation with respect to the superpopulation model. In the RR setting, the model can be rewritten as:

$$\begin{aligned} P(r_k = i|\mathbf{x}_k) &= E_\xi(r_{ki}|\mathbf{x}_k) = \mu_i(\mathbf{x}_k, \boldsymbol{\theta}) \\ &= \begin{cases} \frac{\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)} p_T + p_{F_i}, & i = 1 \\ \left(\frac{\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_i+\boldsymbol{\beta}\mathbf{x}_k)} - \frac{\exp(\alpha_{i-1}+\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\alpha_{i-1}+\boldsymbol{\beta}\mathbf{x}_k)} \right) p_T + p_{F_i}, & i = 2, \ldots, m \end{cases} . \end{aligned}$$

Pseudo-maximum likelihood (ML) estimates, say $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_m, \widehat{\boldsymbol{\beta}})$, for the parameter $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_m, \boldsymbol{\beta})$ are obtained by maximizing the pseudo log-likelihood function:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{k \in s} d_k r_{ki} \log \mu_i(\mathbf{x}_k, \boldsymbol{\theta}).$$

Using pseudo-ML estimates, the probabilities of the categories are estimated as:

$$\widehat{p}_{ki} = \begin{cases} \frac{\exp(\widehat{\alpha}_i+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)}{1+\exp(\widehat{\alpha}_i+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)} p_T + p_{F_i}, & i = 1 \\ \left(\frac{\exp(\widehat{\alpha}_i+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)}{1+\exp(\widehat{\alpha}_i+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)} - \frac{\exp(\widehat{\alpha}_{i-1}+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)}{1+\exp(\widehat{\alpha}_{i-1}+\widehat{\boldsymbol{\beta}}\mathbf{x}_k)} \right) p_T + p_{F_i}, & i = 2, \ldots, m \end{cases} \quad (1)$$

The calibration approach allows us to account for auxiliary information and modify, as little as possible, the design weights d_k to define new sample weights w_k, and obtain more efficient estimates of the proportion P_i. By using the idea of model-calibration proposed by [11], and the probabilities calculated in (1), a new calibration estimator for P_i can be defined as:

$$\widehat{P}_{i,\mathrm{MC}} = \frac{1}{N} \sum_{k \in s} w_k R_{ki}, \quad i = 1, \ldots, m,$$

where the sample weights w_k are such that:

$$\min \sum_{k \in s} G(w_k, d_k) \qquad \text{s.t.} \qquad \sum_{k \in s} w_k \widehat{p}_{ki} = \sum_{k \in U} \widehat{p}_{ki}, \quad (2)$$

and $G(w, d)$ is a distance measure satisfying the usual conditions required in the calibration approach given in [3]. Different calibration estimators can be obtained by using different distance measure. In many instances, numerical methods are required to solve the the minimization problem in (2). However, it is well known that, the choice of the chi-square distance function

$G(w_k, d_k) = (w_k - d_k)^2/2d_k$ yields to an analytic solution of the constrained minimization problem.

Theorem 3. *Under the chi-square distance function, the model-calibration estimator is given by:*

$$\widehat{P}_{i,\mathrm{MC}} = \frac{1}{N}\sum_{k\in s} d_k R_{ki} + \frac{1}{N}\left(\sum_{k\in U}\widehat{p}_{ki} - \sum_{k\in s} d_k \widehat{p}_{ki}\right)\widehat{B}_i, \tag{3}$$

where $\widehat{B}_i = (\sum_{k\in s} d_k \widehat{p}_{ki}^2)^{-1}(\sum_{k\in s} d_k \widehat{p}_{ki} R_{ki})$.

Proof. The estimator can be considered a particular case of the model-calibration estimator given in [7], Section 4.5, changing y_k with R_{ki} and deleting the constrain $\sum_{k\in s} w_k = 1$. Thus, adapting expression (4.4) in [7] we obtain expression (3).

We now discuss and prove the asymptotic unbiasedness of the proposed calibration estimator. In so doing, we consider the usual asymptotic framework in survey sampling where the finite population U and the sampling design $p(\cdot)$ are embedded into a sequence of populations and designs indexed by N, $\{U_N, p_N\}$, with $N \to \infty$. We assume therefore, that n tends to infinity as $N \to \infty$, with $n/N < t < 1$. In order to prove the property of the the estimator, we make the following technical assumptions:

- ***Assumption 1***. Let $\boldsymbol{\theta}_N$ denote the population parameter for U_N. Assume that $\boldsymbol{\theta} = \lim_{N\to\infty} \boldsymbol{\theta}_N$ exists and that the pseudo-ML estimator is $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_N + O_p(n^{-1/2})$.
- ***Assumption 2***. Let $B_{Ni} = \sum_{k\in U_N}(\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N))^{-2}\sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)R_{ki}$. Assume that $B_i = \lim_{N\to\infty} B_{Ni}$ exists for $i = 1, \ldots, m$, and the sampling design is such that $\widehat{B}_i = B_{Ni} + o_p(1)$ with $\widehat{B}_i$ defined in Theorem 3.

Within this framework, the following result holds.

Theorem 4. *Under Assumptions 1 and 2, the estimator* $\widehat{P}_{i,\mathrm{MC}}$ *is asymptotically unbiased for* P_i.

Proof. By applying the Taylor series expansion at $\boldsymbol{\theta}_N$ we have:

$$\widehat{p}_{ki} = \mu_i(\mathbf{x}_k, \widehat{\boldsymbol{\theta}}) = \mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) + O_p(n^{-1/2}),$$

and

$$\frac{1}{N}\sum_{k\in U_N}\widehat{p}_{ki} - \frac{1}{N}\sum_{k\in s} d_k\widehat{p}_{ki} = \frac{1}{N}\sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) - \frac{1}{N}\sum_{k\in s} d_k\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) + O_p(n^{-1/2}).$$

The proposed estimator can be expressed as:

$$\begin{aligned}\widehat{P}_{i,\mathrm{MC}} &= \frac{1}{N}\sum_{k\in s} d_k R_{ki} + \frac{1}{N}\left(\sum_{k\in U_N}\widehat{p}_{ki} - \sum_{k\in s} d_k\widehat{p}_{ki}\right) B_{Ni} + \\ &+ \frac{1}{N}\left(\sum_{k\in U_N}\widehat{p}_{ki} - \sum_{k\in s} d_k\widehat{p}_{ki}\right)\left(\widehat{B}_i - B_{Ni}\right).\end{aligned}$$

Thus, since $\widehat{B}_i = B_{Ni} + o_p(1)$, we have:

$$\widehat{P}_{i,\mathrm{MC}} = \frac{1}{N}\sum_{k\in s} d_k R_{ki} + \frac{1}{N}\left(\sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) - \sum_{k\in s} d_k\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)\right) B_{Ni} + o_p(n^{-1/2}).$$

Taking into account the two sources of variability induced by the sampling design and the randomization device, we have:

$$\begin{aligned}\lim_{N,n\to\infty} E(\widehat{P}_{i,\mathrm{MC}}) &= \lim_{N,n\to\infty} E_d E_r(\widehat{P}_{i,\mathrm{MC}}) = E_d E_r\left(\frac{1}{N}\sum_{k\in s} d_k R_{ki}\right) \\ &+ \frac{B_i}{N}\left(\sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) - E_d\left(\sum_{k\in s} d_k\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)\right)\right).\end{aligned}$$

From Theorem 1, we have $E_d E_r\left(\frac{1}{N}\sum_{k\in s} d_k R_{ki}\right) = E(\widehat{P}^*_{i,\mathrm{HT}}) = P_i$, and observing that $E_d\left(\sum_{k\in s} d_k\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)\right) = \sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)$, it is ready proved that $\lim_{N,n\to\infty} E(\widehat{P}_{i,\mathrm{MC}}) = P_i$

We remark that from calibration theory [3], it is well known that all other calibration estimators that use different distance functions are asymptotically equivalent to the estimator $\widehat{P}^*_{i,\mathrm{MC}}$, under certain regularity conditions concerning the shape of the distance function. Thus, the choice of the distance measure $G(w_k, d_k)$ has only a modest impact on important properties of the estimators like the variance.

3.2. The difference ordinal estimator

We can formulate an alternative estimator of the proportion P_i, for $i = 1, \ldots, m$, that uses the predicted values $\widehat{p}_{ki}$. Preliminary, let P_i be reformulated as:

$$P_i = \frac{1}{N}\left(\sum_{k\in U}\widehat{p}_{ki} + \sum_{k\in U}(z_{ki} - \widehat{p}_{ki})\right).$$

Note that the total $\sum_{k\in U}\widehat{p}_{ki}$ is known, whereas the total of the differences $\sum_{U}(z_{ki} - \widehat{p}_{ki})$ is unknown (since z_{ki} is unknown) although it can be unbiasedly estimated through the an HT-type estimator. Thus, we define a new estimator of P_i as:

$$\widehat{P}_{i,\mathrm{DIF}} = \frac{1}{N}\left(\sum_{k\in U}\widehat{p}_{ki} + \sum_{k\in s} d_k(R_{ki} - \widehat{p}_{ki})\right).$$

The estimator includes two components: a sum of predicted values for the population, and an adjustment term $\sum_{k\in s} d_k(R_{ki} - \widehat{p}_{ki})$. The motivation underlying the construction of $\widehat{P}_{i,\mathrm{DIF}}$ is the possibility to achieve highly accurate estimates through a fitted model which produces small residuals $R_{ki} - \widehat{p}_{ki}$.
The estimator can be rewritten as:

$$\widehat{P}_{i,\mathrm{DIF}} = \frac{1}{N}\sum_{k\in s} d_k R_{ki} + \frac{1}{N}\left(\sum_{k\in U}\widehat{p}_{ki} - \sum_{k\in s} d_k\widehat{p}_{ki}\right),$$

and takes the form of a difference-type estimator [8] adapted to RR data. It is worthwhile highlighting the similarity of the estimator $\widehat{P}_{i,\mathrm{DIF}}$ with $\hat{P}_{i,\mathrm{MC}}$: they have the same expression except for the value of $\widehat{B}_i$ which is equal to 1 in $\widehat{P}_{i,\mathrm{DIF}}$.
Within this framework, the following result holds.

Theorem 5. *Under Assumption 1, the estimator $\widehat{P}_{i,\mathrm{DIF}}$ is asymptotically unbiased for P_i.*

Proof. From the Assumption 1, we have:

$$\frac{1}{N}\sum_{k\in U_N}\widehat{p}_{ki} - \frac{1}{N}\sum_{k\in U_N}\mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) = O_p(n^{-1/2})$$

and

$$\frac{1}{N}\sum_{k\in s} d_k \widehat{p}_{ki} - \frac{1}{N}\sum_{k\in s} d_k \mu_i(\mathbf{x}_k, \boldsymbol{\theta})_N = O_p(n^{-1/2}).$$

Accordingly, $\widehat{P}_{i,\mathrm{DIF}}$ can be written as:

$$\widehat{P}_{i,\mathrm{DIF}} = \frac{1}{N}\left(\sum_{k\in s} d_k R_{ki} - \sum_{k\in s} d_k \mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N)\right) + \frac{1}{N}\sum_{k\in U_N} \mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) + O_p(n^{-1/2})$$

and, analogously to the proof of Theorem 1, it follows that:

$$\begin{aligned}\lim_{N,n\to\infty} E(\widehat{P}_{i,\mathrm{DIF}}) &= \lim_{N,n\to\infty} E_d E_r(\widehat{P}_{i,\mathrm{DIF}}) \\ &= \frac{E_d E_r}{N}\left(\sum_{k\in s} d_k R_{ki} - \sum_{k\in s} d_k \mu_i(\mathbf{x}_k, \boldsymbol{\theta}_N) + \sum_{k\in U_N} \mu_i(\mathbf{x}, \boldsymbol{\theta}_N)\right) = P_i,\end{aligned}$$

which proves the asymptotic unbiasedness of $\widehat{P}_{i,\mathrm{DIF}}$.

4. Monte Carlo simulation

To assess the performance of the estimators proposed in Section 3, a Monte Carlo (MC) simulation study has been designed using real data from the 2012 PISA survey (Programme for International Student Assessment) and assuming as target population the 15-year-old Spanish students ($N = 15499$) who participated in the survey. From the questionnaire, we chose the question "*How strongly do you agree with the statement: I enjoy reading about mathematics?*" to identify the variable of interest (Y) with four ordinal response categories: 1 = *"strongly agree"*, 2 = *"agree"*, 3 = *"disagree"* and 4 = *"strongly disagree"*. The population percentages obtained for these categories were 0.03, 0.149, 0.413 and 0.408, respectively.

We considered two scenarios: auxiliary variables with low and high correlation, with the aim of investigating whether the performance of estimators improves by increasing the correlation with the study variable. In the scenario characterized by low correlation, we employed three auxiliary variables, say X_1, X_2 and X_3, which correspond to the three following questions included in the questionnaire: *"Making an effort in maths is worth because it will help me in the work that I want to do later on"; " Learning maths is worthwhile for me because it will improve my career"; " I will learn many*

things in maths that will help me get a job". The three variables admit the same four ordinal response categories of the study variable and show a correlation with it around 0.3; specifically $\rho_{YX_1} = 0.39, \rho_{YX_2} = 0.33, \rho_{YX_3} = 0.35$. For the second scenario, we simulated three variables by perturbing the above said auxiliary variables in order to produce higher correlation with the study variable. For the synthetic variables X_1, X_2 and X_3 we have now: $\rho_{YX_1} = 0.69, \rho_{YX_2} = 0.78$ and $\rho_{YX_3} = 0.85$.

Under this setting, we considered a single-stage stratified cluster design with unequal selection probabilities. In particular, the PISA students were assumed as the target population that we grouped in five strata according to the location of schools (villages, small towns, towns, cities and large cities). Within each stratum, a sample of schools was selected with probabilities proportional to the number of enrolled students according to the Midzuno sampling design [9]. Three sample sizes for schools were considered in the study, $n = 25, 50, 100$. For each student in the selected schools, we simulated the RR model discussed in Section 2 for $m = 4$ response categories by assuming that student is asked to release the true response category with $p_T = 0.6$ or he/she is forced to declare to posses the ith category with probability $p_{F_i} = 0.1$, $i = 1, \ldots, 4$. The randomization of the category was achieved by means of a random number generated from a Uniform distribution in the interval (0,1).

To estimate the proportion P_i of students with the ith category, we employed the estimators $\widehat{P}^*_{i,\mathrm{HT}}$, $\widehat{P}_{i,\mathrm{MC}}$ and $\widehat{P}_{i,\mathrm{DIF}}$ discussed in Section 3. For each category, and each estimator $\widehat{P}_i = \widehat{P}^*_{i,\mathrm{HT}}, \widehat{P}_{i,\mathrm{MC}}, \widehat{P}_{i,\mathrm{DIF}}$, we computed the percentage bias, $\mathrm{B}(\widehat{P}_i) = 100 \times E_{MC}(\widehat{P}_i - P_i)$, and the percentage mean squared error, $\mathrm{MSE}(\widehat{P}_i) = 100 \times E_{MC}(\widehat{P}_i - P_i)^2$ on the basis of 1000 simulation runs. Hence, the percentage relative efficiency, $\mathrm{RE}(\widehat{P}_i) = 100 \times \mathrm{MSE}(\widehat{P}_i)/\mathrm{MSE}(\widehat{P}^*_{i,\mathrm{HT}})$, was computed to compare the performance of the estimators. The free statistical software R [2] was used to perform the simulation study.

The results for bias, mean squared error and relative efficiency are shown in Table 1 for different sample sizes and for each of the four response categories, under the two correlation scenarios. Note that the sample size in Table 1 refers to the number of schools (primary sampling unit) selected. All the students enrolled in the school represent the elementary units that are surveyed and whose size surveyed cannot be determined in advance since it depends on the size of the schools selected at the first sampling stage.

Overall, we observe that the use of the auxiliary information through the estimators $\widehat{P}_{i,\mathrm{MC}}$ and $\widehat{P}_{i,\mathrm{DIF}}$ may yield more efficient estimates than the HT estimator. The model calibration estimator $\widehat{P}_{i,\mathrm{MC}}$ outperforms $\widehat{P}^{*}_{i,\mathrm{HT}}$ in all the situations and the efficiency gain is particularly high for categories 3 and 4. On the contrary, the difference estimator $\widehat{P}_{i,\mathrm{DIF}}$ may be less efficient than $\widehat{P}^{*}_{i,\mathrm{HT}}$ in some cases. A feature which is common to both $\widehat{P}_{i,\mathrm{MC}}$ and $\widehat{P}_{i,\mathrm{DIF}}$ is that their performance improves when the correlation between the auxiliary variables and the study variable increases. Finally, we observe that all the considered estimators are consistent since their MSE sharply decreases when the sample size of the primary sampling units (schools) increases passing from $n = 25$ to $n = 100$.

5. Conclusions

Nowadays, nonstandard survey techniques that reduce social desirability bias by increasing respondent cooperation in sensitive research are attracting more and more the interest of statisticians, survey practitioners and researchers involved in social, medical and behavioral sciences. Although many RR methods have been proposed in the literature since the sixties to perturb the responses, improving the efficiency of the estimators of sensitive characteristics remains still an open problem that needs methodological advances since privacy protection and efficiency of the estimates are opposite aspects. In general, methods that offer a high degree of privacy protection lead to estimators with high variance too; on the contrary, methods which produce very efficient estimators tend to jeopardize the respondent privacy. Finding a trade-off between privacy and efficiency is a matter of concern in real studies about sensitive topics.

In this paper, we have tried to tackle the challenge of improving the efficiency of the estimates for the frequency of ordinal sensitive categories without infringing privacy protection. In our proposal, we have investigated under a RR setting, the model-calibration estimator and the difference estimator. Both the estimators are based on the use of auxiliary information available in advance for each unit of the population under study.

The asymptotic unbiasedness of the estimators has been theoretically proved and their performance assessed in a simulation study based on the 2012 PISA survey. The simulation results point out the good performance of the model-calibration estimator which, therefore, becomes eligible for future research. In particular,the model-calibration approach could be used in a

generalization of the Liu-Chow RR method [6] according to the respondent belonging to the ith category deliberately missclassifies his/her response and reports the jth category with a transition probability p_{ji}.

Acknowledgements

This work is partially supported by Ministerio de Economía y Competitividad of Spain (grant MTM2015-63609-R).

References

[1] A. Chaudhuri, R. Mukerjee, Randomized Response: Theory and Techniques, Marcel Dekker, New York, 1988.

[2] Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015, https://www.Rproject. org/.

[3] J.C. Deville, C.E. Särndal, Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87, 418 (1992) 376–382.

[4] D.G. Horvitz, D.J. Thompson, A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc. 47, 260 (1952) 663–685.

[5] J.M. Kim, W.D. Warde, Some New Results on the Multinomial Randomized Response Model, Commun. Stat. - Theory Methods 34, 4 (2005) 847–856.

[6] P.T. Liu, L.P. Chow, A New Discrete Quantitative Randomized Response Model, J. Amer. Statist. Assoc. 71, 353 (1976) 72–73.

[7] C.E. Särndal, The calibration approach in survey theory and practice, Surv. Methodology 33, 2 (2007) 99–119.

[8] C.E. Särndal, B. Swenson, J. Wretman, Model Assisted Survey Sampling, Springer-Verlag, New York, 1992.

[9] P.V. Sukhatme, B.V.Sukhatme, S. Sukhatme, C. Asok, Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, Iowa, 1984.

[10] S.L. Warner, Randomized response: A survey technique for eliminating evasive answer bias, J. Amer. Statist. Assoc. 60, 309 (1965) 63–69.

[11] C. Wu, R.R. Sitter, A model-calibration approach to using complete auxiliary information from survey data, J. Amer. Statist. Assoc. 96, 453 (2001) 185–193.

Estimator		Categories 1	2	3	4
		$n = 25$			
$\widehat{P}^*_{i,\mathrm{HT}}$	B	0.275	1.422	3.538	-5.140
	MSE	0.094	0.163	0.609	0.604
	RE	100	100	100	100
		Low correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.259	1.428	3.531	-5.218
	MSE	0.093	0.147	0.358	0.462
	RE	98.936	90.184	58.785	76.490
$\widehat{P}_{i,\mathrm{MC}}$	B	0.255	1.492	3.483	-5.230
	MSE	0.085	0.142	0.317	0.440
	RE	90.426	87.117	52.053	72.848
		High correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.128	1.053	3.595	-4.776
	MSE	0.087	0.147	0.332	0.416
	RE	92.553	90.184	54.516	68.874
$\widehat{P}_{i,\mathrm{MC}}$	B	0.189	1.057	3.694	-4.940
	MSE	0.074	0.106	0.248	0.340
	RE	78.723	65.031	40.723	56.291
		$n = 50$			
$\widehat{P}^*_{i,\mathrm{HT}}$	B	0.242	1.418	3.685	-4.831
	MSE	0.045	0.088	0.358	0.406
	RE	100	100	100	100
		Low correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.181	1.334	3.509	-5.024
	MSE	0.048	0.079	0.231	0.350
	RE	106.667	89.773	64.525	86.207
$\widehat{P}_{i,\mathrm{MC}}$	B	0.255	1.492	3.483	-5.230
	MSE	0.043	0.076	0.207	0.337
	RE	95.556	86.364	57.821	83.005
		High correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.126	1.228	3.546	-4.900
	MSE	0.045	0.074	0.216	0.332
	RE	100	84.091	60.335	81.773
$\widehat{P}_{i,\mathrm{MC}}$	B	0.164	1.243	3.550	-4.956
	MSE	0.039	0.061	0.179	0.295
	RE	86.667	69.318	50.000	72.660
		$n = 100$			
$\widehat{P}^*_{i,\mathrm{HT}}$	B	0.209	1.230	3.395	-4.943
	MSE	0.022	0.052	0.209	0.326
	RE	100	100	100	100
		Low correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.226	1.253	3.436	-4.914
	MSE	0.024	0.048	0.170	0.291
	RE	109.091	92.308	81.340	89.264
$\widehat{P}_{i,\mathrm{MC}}$	B	0.217	1.263	3.441	-4.921
	MSE	0.022	0.047	0.165	0.287
	RE	100	90.385	78.947	88.037
		High correlation			
$\widehat{P}_{i,\mathrm{DIF}}$	B	0.234	1.360	3.300	-4.893
	MSE	0.022	0.048	0.160	0.288
	RE	100.000	92.308	76.555	88.344
$\widehat{P}_{i,\mathrm{MC}}$	B	0.234	1.277	3.391	-4.903
	MSE	0.020	0.039	0.145	0.264
	RE	90.909	75.000	69.378	80.982

Table 1: Bias, mean squared error and relative efficiency (in percentage) of the estimators for P_i.

Mathematics and Computers in Simulation

New estimation techniques for ordinal sensitive variables

MATCOM-D-19-01190

Replies to Comments by Reviewer 1

Dear Referee,

We are very grateful for your carefully reading, comments and suggestions regarding our work. We have revised the paper following your advice and you can easily find the changes by looking for the blue text in Sections

A point-by-point reply to the issues you had raised follows.

Comment 1.*My first comment regards the two proposed estimators. While the model calibrated ordinal estimator is well presented in Section 3.1, the discussion of the difference ordinal estimator is limited to few lines after estimator's formula, referring the reader to the cited literature. Furthermore, Authors state that the estimator "is closely related to the model-calibrated estimator, but is simpler to handle mathematically". Is this the only reason that motivate Authors to propose this estimator as an alternative? A wider theoretical discussion about this estimator may also help in understanding the performance of the estimator in the Monte Carlo simulation study presented in Section 4.*

Response.

We changed Section 3.2. In particular, we explained how to obtain the difference estimator and we added Theorem 5 to prove its asymptotic unbiasedness. Similarly, we have integrated Section 3.1 to prove the asymptotic unbiasedness of the model-calibration estimator.

Comment 2.*My second comment regards the final discussion and concluding remarks that are missing in the paper. Authors should consider to include a Conclusion section*

that summarizes novelties, main results, limitations of the proposal (if any) and possibilities of further research

Response.

We have included a final section with remarks and possible ideas for further research.

Comment 3.*I have few minor comments:*

- *Throughout the paper, citations are not uniform: most often they are in square brackets, but sometimes they are in parentheses and square brackets (see page 2 line 30 and page 8 line 15). Also citation at page 6 line 2 after proof should be standardize according to the journal guidelines;*

  Response: The citations have been changed according to the rules of the journal
- *Page 5 line 11: the term "of" need to be included after the word vector*;

  Response: done
- *Page 5 line 12: use a dot after the word population;*

  Response: done

New estimation techniques for ordinal sensitive variables

MATCOM-D-19-01190

Replies to Comments by Reviewer 2

Dear Referee,

We are very grateful for your carefully reading, comments and suggestions regarding our work. We have revised the paper following your advice and you can easily find the changes by looking for the blue text through the paper.

A point-by-point reply to the issues you had raised follows.

Comment 1.*Please check references carefully for accuracy.*

Response.

You are right. We mentioned the wrong article. As per your suggestion, we have now included:

Liu, P. T., and Chow, L. P. (1976). A new discrete quantitative RR model. *Journal of the American Statistical Association,* 71, 72-3.

Comment 2.*In Section 2, the design weights* $d_k = 1/\pi_k$ *are used as the probability sampling design based on first-order inclusion probabilities* π_k*Then, the second-order inclusion probabilities* π_{kl}*should be addressed with some details in this article. On Page 5, line 5, the equation of* $E_d\left(V_r\left(\hat{P}^*_{i,HT}\right)\right)$*should be* $\frac{1}{(Np_T)^2}\sum_{k\in U}\phi_{ki}$*. Moreover, the authors should define what*ϕ_{ki} *is.*

Response.

We are sorry, but it is not clear to us which details are requested for π_{kl}. The first-order inclusion probabilitiesπ_k are just used to define the basic design weightsd_k, $k \in U$,which, together withπ_{kl} are used in Theorem 2 to express the variance of the

Horvitz-Thompson estimator, say $V\left(\hat{P}_{i,HT}^{*}\right)$.In any case, in Section 2 we have explicitly given the definition of the inclusion probabilities in order to improve the clearness.
Please, note that, through the paper, theoretical results are obtained and discussed under a generic sampling design and, consequently, it is not possible the give the exact (or approximate) expressions for $\pi_k$and π_{kl} which, as it is now, can be defined only after a sampling design is specified.

We have corrected the expression for $E_d\left(V_r\left(\hat{P}_{i,HT}^{*}\right)\right)$. Although the meaning of ϕ_{ki} was defined in the first version of the manuscript in the proof of Theorem 1, we realize now, after your comment, that, perhaps, it was the wrong place. Therefore, in the revisited version of the paper,we have definedϕ_{ki} directly in Theorem 2 and its proof.

Comment 3. *In Section 3, page 6, line 4, the log-likelihood function should be the sum of i=1 to m and $k \in s$. And it is unclear what calibration weights w_k and distance function $G(w_{k,}d_k)$ are.*

Response.

Thanks for your comment. We have corrected the log-likelihood function and included in Section 3, after eqn. (1), more information about the calibration approach, $\mathrm{w_k}$and $\mathrm{G(w_{k,}d_k)}$.

Comment 4.

In Section 4, the authors should explain how they generated the auxiliary variables with low and high correlation, and what the performance of (0.39, 0.33, 0.85) is. In Table 1, the authors write that for n=25, the percentage bias of $\hat{P}_{i,HT}^{*}$ is 0,275, but it should be 0.275. And if $\mathrm{MSE}(\hat{P}_{i,DIF}^{*}) = 0.093$ *and*$\mathrm{MSE}(\hat{P}_{i,DIF}^{*}) = 0.094$*, then the*$\mathrm{RE}(\hat{P}_{i,DIF}^{*})$ *should be 0.093/0.094=0.9893. It is necessary to check the details in Table 1. Finally, it will be very interesting if the authors can clearly stress*

mathematically that the best performance is achieved for $\hat{P}^*_{i,MC}$ *rather than* $\hat{P}^*_{i,DIF}$, *and* $\hat{P}^*_{i,MC}$ *rather than* $\hat{P}^*_{i,HT}$.

Response.

In Section 4, we have explained how the low and high correlated variables are generated. In particular, in the case of low correlation, we have included three auxiliary variables from the original 2012 PISA database. High correlated variables have been produced by perturbing the low correlation variable.

We have checked and corrected all the values in the table. The differences were due to the decimals used. We changed the values using three decimals in order to avoid reader confusion.

We are sorry but we do not have a rigorous proof of the different performance of the estimators and, at the moment, we are not able to produce it. Deriving theoretical comparisons may be certainly interesting but perhaps out the aim of the present paper which, given the aims of *Mathematics and Computers in Simulation,* has been deliberately oriented to a simulation study. More important thing, analytical and theoretical considerations will require a large amount of study which final outcome is, however, uncertain. In fact, first we must explore the performance of the competitive estimators in the standard sampling framework without the RR stage, than we must compare them in the RR setting. It may happen that the uniform superiority of $\hat{P}^*_{i,MC}$ cannot be ascertained in all the situations.Certainly, your suggestion may be object of future research. We thank you for this incentive.

To conclude to point, note, please, that we have included in the paper Theorem 4 and 5 for a wider theoretical discussion about $\hat{P}^*_{i,MC}$ and $\hat{P}^*_{i,DIF}$ as requested by another referee.