REVSTAT – Statistical Journal
Volume 16, Number 3, July 2018, 323–348

MODEL-ASSISTED AND MODEL-CALIBRATED ESTIMATION FOR CLASS FREQUENCIES WITH ORDINAL OUTCOMES

Authors: Maria del Mar Rueda
– Department of Statistics and Operational Research,
University of Granada, Spain
mrueda@ugr.es

Antonio Arcos
– Department of Statistics and Operational Research,
University of Granada, Spain
arcos@ugr.es

David Molina
– Department of Statistics and Operational Research,
University of Granada, Spain
dmolinam@ugr.es

Manuel Trujillo
– Institute of Social Studies of Andalusia (IESA),
Spanish National Research Council (CSIC), Spain
mtrujillo@iesa.csic.es

Received: February 2016 Revised: September 2016 Accepted: September 2016

Abstract:

- This paper considers new techniques for complex surveys in the case of estimation of proportions when the variable of interest has ordinal outcomes. Ordinal model-assisted and ordinal model-calibrated estimators are introduced for class frequencies in a population, taking two different approaches. Theoretical properties and numerical methods are investigated. Simulation studies using data from a real macro survey are considered to evaluate the performance of the proposed estimators. The empirical coverage and the length of confidence intervals are computed using several techniques in variance estimation. We also use data from an opinion survey to show the behavior of the proposed estimators in real applications.

Key-Words:

- *complex surveys; model calibration; ordinal data; weighted least squares; weighted maximum likelihood.*

AMS Subject Classification:

- 62D05.

1. INTRODUCTION

Questions with categorical outcomes are quite common in surveys, especially in health, marketing, public opinion and official surveys. In the simplest case, questions have only two possible responses, which are often used to represent the "success" or the "failure" of an experiment (such as the occurrence or nonoccurrence of an event or the presence or absence of a characteristic). These items can be modelled statistically using binary logit regression ([1]; [25]).

In more complex situations, items have three or more possible options and respondents must select one of them. When analyzing a polytomous variable it is necessary to determine whether its categories can be ordered according to an intrinsic characteristic of the categories themselves. If so, the number of outcomes attributable to each category is modelled by an ordinal distribution, which gives rise to an ordinal logit model. Otherwise, we should use a multinomial logit model, which is based on the multinomial probability distribution.

Most studies related to binary, multinomial or ordinal logit regression are based on the assumption of a simple random sample drawn from a large population. However, this scenario is not always present in practice: many surveys assume a finite population with samples extracted from complex sampling designs. For example, the Educational Longitudinal Study developed by the National Center for Education Statistics, the Post Enumeration Survey conducted by the Portuguese Statistical Office ([6]) and the Programme for International Student Assessment (PISA) study conducted by the Organisation for Economic Co-operation and Development, all applied complex sampling survey designs. These designs have in common the use of strata, clusters and unequal probabilities of selection in data collection. In this respect, it has been shown that ignoring weights, clusters and strata can lead to biased parameter estimates and erroneous standard errors in ordinal logistic regression analysis [24].

[23] used binary and multinomial logistic regressions in the context of survey sampling. In this context, [38] used a logistic regression model to obtain a calibration estimator for the finite population distribution function under a general sampling design, while [22] developed point and variance estimators for the total of finite population characteristics from a clustered sample assisted by a logistic regression model.

Ordinal regression models have been used extensively in sociological, medical and educational research but have a very sparse presence in parameter estimation in finite population sampling, which motivated this work. Therefore, the objective of this paper is to introduce new ordinal model-assisted estimators and ordinal model-calibrated estimators for the proportions of the categories of a response variable with ordinal outcomes.

This article proceeds as follows: Section 2 reviews the estimation methods that have been suggested to determine the proportion of categories of an ordinal response variable in finite population sampling. In section 3, the use of ordinal regression models in survey sampling is introduced. In section 4 we propose several estimators for the proportions of categories: the first of these is based on the procedure used by [23] for the case of a nominal variable, and the second is defined using calibration techniques ([8]; [31]). A brief discussion is then offered of numerical methods for parameter estimation. The main theoretical properties of the proposed estimators are studied in section 5. Variance estimation is addressed in section 6. Section 7 describes how the performance of the proposed estimators is measured through simulation experiments. Section 8 presents the results obtained from the different estimation strategies with respect to an opinion survey dataset. Finally, section 9 summarizes the conclusions drawn.

2. ESTIMATORS FOR CLASS RELATIVE FREQUENCIES OF A DISCRETE RESPONSE VARIABLE UNDER A GENERAL SAMPLING DESIGN

Let U denote a finite population with N units, $U = \{1, ..., k, ..., N\}$. Assume that data are collected from respondents who provide a single choice from a list of alternatives coded $1, 2, ..., i, ..., m$. Consider a discrete m-valued survey variable Y and denote the value observed for the kth individual of the population as y_k. Our aim is to estimate the frequency distribution of Y in the population U. To do so, we define a class of indicators z_i $(i = 1, ..., m)$ such that for each unit $k \in U$ $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. The problem thus, is to estimate the proportions $P_i = 1/N \sum_{k \in U} z_{ki}$, $i = 1, 2, ..., m$.

Let s be a probability sample of size n drawn from population U using a sampling design p_d. The sampling design considered induces first-order inclusion probabilities π_k, second-order inclusion probabilities π_{kl} and design weights $d_k = 1/\pi_k$, for $k, l = 1, ..., N$.

The customary design unbiased estimator of P_i is given by

$$\widehat{P}_{\mathrm{HT}i} = \frac{1}{N} \sum_{k \in s} \frac{z_{ki}}{\pi_k} = \frac{1}{N} \sum_{k \in s} z_{ki} d_k, \tag{2.1}$$

where the subindex HT refers to the Horvitz–Thompson estimator [20]. The design weights d_k are commonly thought of as the number of population units represented by unit k in the sample. [10] discussed the estimation of proportions using Bernoulli sampling and stratified designs.

In sample surveys, the use of auxiliary variables has been widely discussed by survey practitioners since this approach can increase the efficiency of the estimates in different contexts (see e.g. [9]). Thus, it is common practice to use auxiliary information on a character x related to the main variable y. A variety of approaches are available to construct more efficient estimators including design-based and model-based methods (see e.g. [35]; [32]).

Let us now consider a general situation where the auxiliary variable can be either numeric or binary. Let x_k be the value of the study variable x for the kth population element, available for all of U. For the sample s, the values of the two variables (y_k, x_k), $k \in s$, are observed. Under this scenario, we can consider the use of superpopulation models for sampling surveys. A superpopulation model is a way of formalising the relationship between a target variable and the auxiliary data. In previous research, superpopulation models have been used in sociological and electoral studies. For example, [5] used the superpopulation approach to estimate average customer satisfaction and [29] used superpopulation models to analyze electoral polls. Traditionally, linear regression models have been used to incorporate auxiliary information but (as is well known in sociological literature, see e.g. [36]) for qualitative variables a linear model might be unrealistic.

A first procedure is to consider the superpopulation multinomial logistic model given in [23]: we assume that the population under study $\boldsymbol{y} = (y_1, ..., y_N)^\top$ constitutes a body of superpopulation random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$, containing a superpopulation model, ξ, such that

$$\mu_i(x_k) = P(Y_k = i|x_k) = \mathrm{E}_\xi(Z_{ki}|x_k) = \frac{\exp(\alpha_i + \beta_i x_k)}{\sum_{j=1}^m \exp(\alpha_j + \beta_j x_k)},$$

$i = 1, ..., m,\ \ k = 1, ..., N$ (E_ξ denotes the expected value with respect to the model) and assume that Y_k are conditionally independent given x_k.

Usually, population parameters α_i and β_i involved in the model ξ are unknown and should be estimated from the sample. Considering $\hat{\alpha}_i$ and $\hat{\beta}_i$ as the maximum likelihood estimations of α_i and β_i, we can define an estimator for probabilities for each category as follows:

$$p_{ki}^M = \hat{\mu}_i(x_k) = \frac{\exp(\hat{\alpha}_i + \hat{\beta}_i x_k)}{\sum_{j=1}^m \exp(\hat{\alpha}_j + \hat{\beta}_j x_k)}, \quad i = 1, ..., m, \ \ k = 1, ..., N.$$

[23] used the values p_{ki}^M as auxiliary information to define an estimator of class frequencies for nominal response variables. This estimator is in the form

$$\hat{F}_{\mathrm{LV}i} = \sum_{k \in U} p_{ki}^M + \sum_{k \in s} d_k (z_{ki} - p_{ki}^M), \quad i = 1, ..., m, \tag{2.2}$$

where the subindex LV refers to the Lehtonen and Veijanen estimator. An estimator of class proportions can be obtained simply by dividing in (2.2) by population size, N, which is assumed to be known, as follows:

$$\hat{P}_{\mathrm{LV}i} = \frac{1}{N}\hat{F}_{\mathrm{LV}i} = \frac{1}{N}\left(\sum_{k\in U} p_{ki}^M + \sum_{k\in s} d_k(z_{ki} - p_{ki}^M)\right), \quad i = 1, ..., m. \tag{2.3}$$

The sum $\sum_{k\in U} p_{ki}^M$ implies that auxiliary information is known for every element in the population. However, when categorical variables (such as gender or the professional status of the individual) or quantitative categorized variables (such as the age of the individual, grouped in classes) are used as auxiliary information in a survey, we may not have a complete list of individuals. Nevertheless, the proposed estimators can still be computed since the population information needed can be found in the databases of national statistical agencies.

3. THE USE OF ORDINAL REGRESSION MODELS IN SURVEY SAMPLING

Let us now assume that the m possible values of Y can be sorted, such that $1 < \cdots < m$. A disadvantage of using multinomial models for ordinal data is that information about the ordering is discarded. Ordinal regression provides a better fit and hence more accurate results. Within ordinal regression models, the most popular is the cumulative logit model, which assumes a linear model for the logit of cumulative probabilities for categories of Y. Given a particular point, the cumulative probability can be defined as the probability that Y falls at or below that point. For the ith category, its cumulative probability can be expressed as

$$P(Y \le i) = \mu_1 + \cdots + \mu_i, \quad i = 1, ..., m,$$

with $\mu_i = P(Y = i)$. Logit transformations of the cumulative probabilities are, for $i = 1, ..., m-1$,

$$\mathrm{logit}(P(Y \le i)) = \log\left(\frac{P(Y \le i)}{1 - P(Y \le i)}\right) = \log\left(\frac{P(Y \le i)}{P(Y > i)}\right) = \log\left(\frac{\mu_1 + \cdots + \mu_i}{\mu_{i+1} + \cdots + \mu_m}\right).$$

Note that no logit transformation can be defined for the mth category since, in this case, $P(Y \le m) = 1$, and so $1 - P(Y \le m) = 1 - 1 = 0$ and therefore the denominator would be cancelled out. An important property that is usually assumed to be satisfied is that of proportional odds, according to which the effects of the predictors are the same across categories. This implies that β parameters associated with the independent variables are fixed and independent of the category in question. Let us consider

$$P(Y \le i | X = x_k) = \frac{\exp(\alpha_i + \beta x_k)}{1 + \exp(\alpha_i + \beta x_k)}, \quad i = 1, ..., m-1, \;\; k = 1, ..., N.$$

The cumulative probability for the last category, $P(Y \le m | X = x_k)$, is always equal to 1. The probability for each category can, then, be calculated as the difference of the cumulative probabilities.

Thus, we propose a superpopulation model ξ with random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$ such that

$$\mu_1(x_k) = \mathrm{E}_\xi(Z_{k1}|x_k) = \frac{\exp(\alpha_1 + \beta x_k)}{1 + \exp(\alpha_1 + \beta x_k)},$$
$$\mu_i(x_k) = \mathrm{E}_\xi(Z_{ki}|x_k) = \frac{\exp(\alpha_i + \beta x_k)}{1+\exp(\alpha_i+\beta x_k)} - \frac{\exp(\alpha_{i-1} + \beta x_k)}{1+\exp(\alpha_{i-1}+\beta x_k)}, \quad i = 2, ..., m-1,$$
$$\mu_m(x_k) = \mathrm{E}_\xi(Z_{km}|x_k) = 1 - \frac{\exp(\alpha_{m-1} + \beta x_k)}{1 + \exp(\alpha_{m-1} + \beta x_k)}.$$

To define a new estimator for a proportion, using this regression model, we estimate the superpopulation parameter $\theta = (\alpha_1, ..., \alpha_{m-1}, \beta)$ from the units of sample s. After calculating the optimal estimators of the m parameters involved in the model, we can define estimators for individual probabilities

(3.1)

$$p_{k1} = \hat{\mu}_1(x_k) = \frac{\exp(\hat{\alpha}_1 + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_1 + \hat{\beta} x_k)},$$
$$p_{ki} = \hat{\mu}_i(x_k) = \frac{\exp(\hat{\alpha}_i + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_i + \hat{\beta} x_k)} - \frac{\exp(\hat{\alpha}_{i-1} + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_{i-1} + \hat{\beta} x_k)}, \quad i = 2, ..., m-1,$$
$$p_{km} = \hat{\mu}_m(x_k) = 1 - \frac{\exp(\hat{\alpha}_{m-1} + \hat{\beta} x_k)}{1 + \exp(\hat{\alpha}_{m-1} + \hat{\beta} x_k)}.$$

Now, we consider the question of estimating the model parameters. Two general approaches can be adopted to find the optimal estimations of these parameters: (1) by minimizing the sum of the squared distances between the observed and the predicted values (i.e., least squares estimation); or (2) by maximizing the likelihood function (i.e., maximum likelihood estimation or ML estimation).

Weighted least squares method. One way to estimate the parameters of the ordinal logistic regression model is that of least squares. However, in our case, instead of using ordinary least squares, weighted least squares (WLS) must be used. The main difference between the two is that in WLS each observation is weighted using its corresponding survey weight (see e.g. [37]). In this context, WLS involves minimizing, with respect to the residual standard squared error, the weighted distance between the observed outcome (or a function of the observed outcome) and non-linear estimates. In the present case, the function to minimize is

$$S = \sum_{i=1,...,m} \sum_{k \in s} d_k r_{ki}^2,$$

with $r_{ki} = \log\left(P(Y \le i)/(1 - P(Y \le i))\right) - \alpha_i - \beta x_k$. This typically requires a numerical procedure, such as the Gauss-Newton method with the Levenberg-Marquardt adjustment (see [19]), which uses derivatives or estimates of derivatives to select the optimal fit. In an iterative fitting process for WLS, assuming ordinal data, at some settings of explanatory variables, the estimated mean may fall below the lowest score or above the highest one and then the fit fails (see [1]).

Maximum likelihood method. Ordered regression models are usually implemented using ML. For ML estimation, the ordinal likelihood function must be numerically maximized to find the parameter values below which the observed data were most likely produced. In theory, these estimates might have the properties of asymptotic efficiency and invariance under parameterization, which makes ML estimation [28] an attractive option in general.

The Nelder Mead simplex [27] is a popular and powerful direct search procedure for likelihood-based optimization. The attraction of this method is that it does not use any derivatives and does not assume that the objective function being optimized has continuous derivatives. In cases such as the present, we expect continuity in the first derivatives and so the latter advantage is not so important. However, this method may be much less efficient or even highly unstable, compared to derivative-based ML estimation methods when sample sizes are as large as the datasets commonly found in complex survey designs.

Let us now examine the logistic likelihood function for modelling ordinal outcomes. As the available data are limited to the sample s, the likelihood function is defined as:

$$L(\boldsymbol{\theta}) = \prod_{i=1,\ldots,m} \prod_{k\in s} \mu_i(x_k)^{z_{ki} d_k}.$$

The pseudolikelihood ([17]; [32]), which is more convenient for use in optimization procedures is given by

$$\log\left(L(\boldsymbol{\theta})\right) = \sum_{i=1,\ldots,m} \sum_{k\in s} d_k z_{ki} \log\left(\mu_i(x_k)\right).$$

ML estimates are obtained by solving a system of m nonlinear equations. Traditionally, two alternatives can be used to address the solution of these equations numerically: Fisher scoring or Newton-Raphson algorithms. Since the results obtained by either method are nearly the same, the decision as to which one to use is trivial (see e.g. [18]).

Various statistical packages can be used to compute the ML estimates of an ordinal logistic model, such as SAS (PROC SURVEYLOGISTIC) or library ordinal for R, but all of them use the Newton-Raphson algorithm to solve the weighted ML equations. The SAS SURVEYLOGISTIC procedure also implements the Fisher scoring algorithm.

4. PROPOSED ESTIMATORS FOR ITEMS WITH ORDINAL OUTCOMES

The estimated individual probabilities (3.1) may be used to define new estimators. We consider a model-assisted approach and a model-calibrated approach to define the following ordinal estimators:

The model-assisted ordinal estimator. Using the idea of the generalized difference predictor given in [5], we define an estimator for proportions of the ordered categories of the response variable as follows

$$\hat{P}_{\mathrm{MA}i} = \frac{1}{N}\left(\sum_{k\in U} p_{ki} + \sum_{k\in s} d_k(z_{ki} - p_{ki})\right), \quad i = 1, ..., m, \tag{4.1}$$

where the subindex MA stands for Model-Assisted.

This estimator is similar to the $\hat{P}_{\mathrm{LV}i}$ estimator proposed by [23] but changes the p_{ki}^{M} values to p_{ki} values.

The model-calibrated ordinal estimator. A new calibration estimator, let us say P_{MC} (the subindex MC stands for Model-Calibrated), can be defined using the probabilities calculated in (3.1). This estimator is in the form

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} w_k z_{ki}, \quad i = 1, ..., m, \tag{4.2}$$

where, in this case, the weights w_k minimize $G(w_k, d_k)$, and where $G(\cdot,\cdot)$ is a particular distance measure, subject to

$$\sum_{k\in s} w_k p_{ki} = \sum_{k\in U} p_{ki}. \tag{4.3}$$

This is an extension of the model calibration approach proposed by [39]. The distance measure that is usually considered is the chi square

$$\chi = \sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k}, \tag{4.4}$$

where the q_k's are known positive weights unrelated to d_k. Following [31], section 4.2, and using p_{ki} as an auxiliary variable with a known total $\sum_{k\in U} p_{ki}$, by minimizing (4.4) subject to (4.3) we obtain new weights w_k. By substituting these weights in (4.2) we obtain the following analytic expression for the chi-square calibration estimator:

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right)\hat{B}_i,$$

where $\hat{B}_i = (\sum_{k\in s} d_k p_{ki}^2)^{-1}(\sum_{k\in s} d_k p_{ki} z_{ki})$.

From calibration theory (see [8]), it is well known that all other calibration estimators that use different distance functions are asymptotically equivalent to the chi-square calibration estimator, under additional regularity conditions concerning the shape of the distance function.

So far, we have considered only one auxiliary variable when defining the estimators. These estimators can be easily extended to the general case of p auxiliary variables $\boldsymbol{x} = (x_1, ..., x_p)^\top$ observed for each individual in the population U.

5. PROPERTIES OF THE PROPOSED ESTIMATORS

The most significant properties of the proposed estimators $\hat{P}_{\text{MA}i}$ and $\hat{P}_{\text{MC}i}$ are summarized in this section. To illustrate the asymptotic properties of the proposed classes of estimators, we consider the asymptotic framework of [21], in which the finite population U and the sampling design $p_d(\cdot)$ are embedded into a sequence of populations and designs indexed by N, $\{U_N, p_{d_N}\}$, with $N \to \infty$. We assume therefore, that n tends to infinity as $N \to \infty$. We further assume that $N > 0$. The subscript N may be discarded for ease of notation, although all limiting processes are understood as $N \to \infty$. We denote by E_p the expected value with respect to the sampling design.

The following assumptions are imposed for the sampling design p_d and for the variables:

i) Let $\boldsymbol{\theta_U}$ be the census level parameter estimate obtained by maximizing the likelihood $L(\boldsymbol{\theta})$. Assume that $\boldsymbol{\theta} = \lim_{N\to\infty}\boldsymbol{\theta_U}$ exists and that the pseudomaximum likelihood estimator is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta_U} + O_p(n^{-1/2})$. [1]

ii) For any study variable h the sampling designs are such that the Horvitz–Thompson estimator for $\overline{h}_N = N^{-1} \sum\limits_{k\in U} h_k$ is asymptotically normal distributed.

iii) Let $B_{iU} = \sum\limits_{k\in U} (\mu_i(x_k)^2)^{-1} \sum\limits_{k\in U} \mu_i(x_k) z_{ki}$. Assume that $B_i = \lim\limits_{N\to\infty} B_{Ui}$ exists, and the sampling design is such that B_i are consistently estimated by $\hat{B}_i$ for $i = 1, ..., m$.

Theorem 5.1. *Under conditions i) and ii) the estimator $\hat{P}_{\text{MA}i}$ is approximately design unbiased for P_i, asymptotically normal distributed and the asymptotic design variance is given by*

$$AV_p(\hat{P}_{\text{MA}i}) = \frac{1}{N^2} \sum_{k\in U} \sum_{l\in U} \Delta_{kl} (d_k c_{ki})(d_l c_{li}),$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$; $c_{ki} = z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})$.

[1] This is true under certain regularity conditions given by [3].

Proof: The proof for unbiasedness is very similar to the one presented in [32], page 223, for the difference estimator.

Let us consider the parametric vector $\boldsymbol{t} = (t_1, t_2, ..., t_m)$ and the function $\mu_i(x_k, t_1, t_2, ..., t_m) = \exp(t_i + t_m x_k)/(1 + \exp(t_i + t_m x_k)) - \exp(t_{i-1} + t_m x_k)/(1 + \exp(t_{i-1} + t_m x_k))$. This function has partial derivatives, for each x_k, $\partial\mu_i(x_k, t)/\partial t_j$, which are continuous in $\boldsymbol{t}$ and

$$\frac{\partial\mu_i(x_k,\boldsymbol{t})}{\partial t_i}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq 1,$$

$$\frac{\partial\mu_i(x_k,\boldsymbol{t})}{\partial t_{i-1}}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq 1,$$

$$\frac{\partial\mu_i(x_k,\boldsymbol{t})}{\partial t_m}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} \leq x_k, \text{ and}$$

$$\frac{\partial\mu_i(x_k,\boldsymbol{t})}{\partial t_j}|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)} = 0 \text{ for } j \neq i, i-1, m.$$

Thus, by applying the Taylor series expansion at $\boldsymbol{t} = \boldsymbol{\theta_U}$

$$p_{ki} = \mu_i(x_k, \hat{\boldsymbol{\theta}}) = \mu_i(x_k, \boldsymbol{\theta_U}) + \sum_{j=1,...,m} \partial\mu_i(x_k, \boldsymbol{t})/\partial t_j|_{\boldsymbol{t}=(x_k,\alpha_1,...,\beta)}(\hat{\theta}_j - \theta_{Uj}).$$

Under condition i)

$$p_{ki} = \mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}),$$

and then

$$\frac{1}{N}\sum_{k\in U} p_{ki} - \frac{1}{N}\sum_{k\in U} \mu_i(x_k, \boldsymbol{\theta_U}) = O_p(n^{-1/2}), \text{ and}$$

$$\frac{1}{N}\sum_{k\in s} d_k p_{ki} - \frac{1}{N}\sum_{k\in s} d_k \mu_i(x_k, \boldsymbol{\theta_U}) = O_p(n^{-1/2}).$$

Thus

$$\hat{P}_{\mathrm{MA}i} = \frac{1}{N}\left(\sum_{k\in s} d_k z_{ki} - \sum_{k\in s} d_k \mu_i(x_k, \boldsymbol{\theta_U})\right) + \frac{1}{N}\sum_{k\in U} \mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}),$$

and the asymptotic design variance of $\hat{P}_{\mathrm{MA}i}$ is the same as that the Horvitz–Thompson estimator $\hat{C}_{\mathrm{HT}i} = 1/N \sum_{k\in s} d_k(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U}))$.

Condition ii) ensures that estimator $\hat{C}_{\mathrm{HT}i}$ is asymptotically normal distributed and, therefore, estimator $\hat{P}_{\mathrm{MA}i}$ is also asymptotically normal distributed. □

Theorem 5.2. *Under conditions i), ii) and iii) the calibration estimator $\hat{P}_{\mathrm{MC}i}$ is approximately design unbiased for P_i, asymptotically normal distributed and the asymptotic design variance is given by*

$$AV_p(\hat{P}_{\mathrm{MC}i}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k e_{ki})(d_l e_{li}),$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ and $e_{ki} = z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})B_{iU}$.

Proof:

$$\begin{aligned}\hat{P}_{\mathrm{MC}i} &= \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right) B_{iU}\\ &\quad + \frac{1}{N}\left(\sum_{k\in U} p_{ki} - \sum_{k\in s} d_k p_{ki}\right)\left(\hat{B}_i - B_{iU}\right).\end{aligned}$$

Under condition iii) $\hat{B}_i - B_{iU} = o(1)$; under conditions i) and ii)

$$\frac{1}{N}\sum_{k\in U} p_{ki} - \frac{1}{N}\sum_{k\in s} d_k p_{ki} = \frac{1}{N}\sum_{k\in U}\mu_i(x_k, \boldsymbol{\theta_U}) - \frac{1}{N}\sum_{k\in s} d_k\mu_i(x_k, \boldsymbol{\theta_U}) + O_p(n^{-1/2}).$$

Thus

$$\hat{P}_{\mathrm{MC}i} = \frac{1}{N}\sum_{k\in s} d_k z_{ki} + \frac{1}{N}\left(\sum_{k\in U}\mu_i(x_k, \boldsymbol{\theta_U}) - \sum_{k\in s} d_k\mu_i(x_k, \boldsymbol{\theta_U})\right) B_{iU} + o_p(n^{-1/2}),$$

and consequently

$$\mathrm{E}_p(\hat{P}_{\mathrm{MC}i}) \to \mathrm{E}_p\left(\frac{1}{N}\sum_{k\in s} d_k z_{ki}\right) = P_i,$$

and

$$\mathrm{V}_p(\hat{P}_{\mathrm{MC}i}) \to \mathrm{V}_p\left(\frac{1}{N}\sum_{k\in s} d_k\left(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U})\right) B_{iU}\right).$$

Under condition ii), estimator $(1/N)\sum_{k\in s} d_k(z_{ki} - \mu_i(x_k, \boldsymbol{\theta_U}))B_{iU}$ is asymptotically normal distributed, and therefore we conclude that estimator $\hat{P}_{\mathrm{MC}i}$ is also asymptotically normal distributed.

□

6. ESTIMATION FOR THE VARIANCE OF ORDINAL ESTIMATORS

The next theorem gives analytic expressions for the estimators of the design variances $\mathrm{V}_p(\hat{P}_{\mathrm{MA}i})$ and $\mathrm{V}_p(\hat{P}_{\mathrm{MC}i})$, obtained using the linearization method.

Theorem 6.1. *Under conditions i), ii) and iii) and assuming that all second order probabilities are non null,*

$$\widehat{\mathrm{V}}(\hat{P}_{\mathrm{MA}i}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li}) \quad \textit{(Lin)} \tag{6.1}$$

is approximately design unbiased for $V_p(\hat{P}_{\mathrm{MA}i})$ *and*

$$\widehat{\mathrm{V}}(\hat{P}_{\mathrm{MC}i}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{e}_{ki})(d_l\tilde{e}_{li}) \quad \textit{(Lin)} \tag{6.2}$$

is approximately design unbiased for $V_p(\hat{P}_{\mathrm{MC}i})$ *where* $\tilde{c}_{ki} = z_{ki} - p_{ki}$ *and* $\tilde{e}_{ki} = z_{ki} - p_{ki}\widehat{B}_i$.

Proof: We denote by I_k the sample membership indicator of element k. Thus, for each $i = 1, ..., m$:

$$\mathrm{E}_p(\widehat{\mathrm{V}}(\hat{P}_{\mathrm{MA}i})) = \frac{1}{N^2}\mathrm{E}_p\sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li})I_k(s)I_l(s) =$$

$$= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_{kl}}(d_k\tilde{c}_{ki})(d_l\tilde{c}_{li})\pi_{kl} \to \mathrm{V}_p(\hat{P}_{\mathrm{MA}i}),$$

using the theorem 5.1. From the theorem 5.2, the estimator of the design variance $\mathrm{V}_p(\hat{P}_{\mathrm{MC}i})$ can be derived. □

These variance estimators require knowledge of second-order inclusion probabilities, which are often impossible to compute or unavailable to data analysts for complex sampling designs. A simple alternative is to use with-replacement variance estimators (see [32], page 99). For the $\hat{P}_{\mathrm{MA}i}$ estimator, the with-replacement variance estimator is

$$\hat{\mathrm{v}}_{\text{W-R}}(\hat{P}_{\mathrm{MA}i}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{\tilde{c}_{ki}}{pr_k} - \frac{1}{n}\sum_{j\in s}\frac{\tilde{c}_{ji}}{pr_j}\right)^2 \quad \text{(W-R)},$$

where $pr_k = \pi_k/n$ when we have a simple random sampling without replacement design. For other sampling designs, the relationship between pr_k and π_k is $\pi_k = 1-(1-pr_k)^n$ according to expression (2.9.5), page 51 in [32].

The with-replacement variance estimator for $\hat{P}_{\mathrm{MC}i}$ is obtained in a similar way:

$$\hat{\mathrm{v}}_{\text{W-R}}(\hat{P}_{\mathrm{MC}i}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{k\in s}\left(\frac{\tilde{e}_{ki}}{pr_k} - \frac{1}{n}\sum_{j\in s}\frac{\tilde{e}_{ji}}{pr_j}\right)^2 \quad \text{(W-R)}.$$

These with-replacement variance estimators are not without bias. An expression for the bias can be obtained using the theory of sampling with probability proportional to size (see [11] or [32]).

Alternative variance estimators can be obtained using implicit differentiation [3] or replicated sampling methods (see [37] for a detailed description of these techniques in finite population sampling). The replicated methods estimate the variance of a parameter by generating replicated subsamples and examining the variability of the subsample estimates. The replicated methods, also referred to as resampling methods, include balanced repeated replication (BRR), jackknife repeated replication (JRR) [34] and the bootstrap method [12]. This article focuses on jackknife techniques due to their simplicity and because they are implemented in general purpose software packages, such as R (see for example the packages sampling [33], samplingVarEst [14] and samplingEstimates [13]).

For a non stratified design, the jackknife estimator of the variance for any of the model-assisted estimators, $\hat{P}_{\mathrm{MA}i}$ is given by

$$\hat{\mathrm{v}}_{\mathrm{J}}(\hat{P}_{\mathrm{MA}i}) = \frac{n-1}{n}\sum_{j\in s}(\hat{P}_{\mathrm{MA}i}(j) - \overline{P}_{\mathrm{MA}i})^2 \quad \text{(Tukey)}, \tag{6.3}$$

where $\hat{P}_{\mathrm{MA}i}(j)$ is the value of the estimator $\hat{P}_{\mathrm{MA}i}$ after dropping unit j from s and where $\overline{P}_{\mathrm{MA}i}$ is the mean of values $\hat{P}_{\mathrm{MA}i}(j)$.

The jackknife estimator may present an important bias when designs without replacement are used in finite populations. In such a case, an approximated finite-population correction could be incorporated into the estimation in order to achieve unbiasedness. A modified jackknife estimator of variance, $\hat{\mathrm{v}}^*_{\mathrm{J}}(\hat{P}_i)$, can be calculated by replacing $\hat{P}_{\mathrm{MA}i}(j)$ in (6.3) with $\hat{P}^*_{\mathrm{MA}i}(j) = \hat{P}_{\mathrm{MA}i} + \sqrt{1-\overline{\pi}}(\hat{P}_{\mathrm{MA}i}(j) - \overline{P}_{\mathrm{MA}i})$, where $\overline{\pi} = \sum_{k\in s}\pi_k/n$.

Using the idea of the unequal probability jackknife variance estimator given by [4], we can obtain a new estimator $\hat{\mathrm{v}}_{\mathrm{JC}}(\overline{P}_{\mathrm{MA}i})$ by replacing $\tilde{c}_{ki}$ in (6.1) with $\tilde{cm}_{ki} = 1 - \widetilde{d}_k(\widetilde{C}_{\mathrm{HT}i} - \widetilde{C}_{\mathrm{HT}i}(k))$ where $\widetilde{C}_{\mathrm{HT}i} = 1/N\sum_{k\in s} d_k\tilde{c}_{ki}$, $\widetilde{C}_{\mathrm{HT}i}(k)$ is the Horvitz–Thompson estimator dropping the unit k of the sample and $\widetilde{d}_k = d_k/\sum\limits_{l\in s} d_l$. The design consistency of this type of variance estimator was highlighted in [2].

More recently, [15] formulated a new design-consistent variance estimator for the population mean. Based on this idea, we can obtain a new variance estimator $\hat{\mathrm{v}}_{\mathrm{JEB}}(\overline{P}_{\mathrm{MA}i})$ by replacing $\tilde{c}_{ki}$ in (6.1) with $\widetilde{ce}_{ki} = d_k^{\alpha_k}(\widetilde{C}_{\mathrm{HT}i} - \widetilde{C}_{\mathrm{HT}i}(k))$. The authors propose d the use of $\alpha_k = 1\ \forall k \in s$.

Similarly, we define jackknife variance estimators for the ordinal calibration estimator.

7. MONTE CARLO SIMULATION EXPERIMENTS

To determine the behaviour of the estimators when they are applied to real data obtained through complex sampling designs, we consider data from the 2012 PISA survey. This is a macro-surveying procedure that is conducted every three years to collect information about 15-year-old students in each of the 65 countries participating. The main aim of the survey is to determine how well students are prepared to meet the challenges of the future. To do so, their performance and attitudes are measured in three key areas: mathematics, reading and science.

The 2012 PISA survey was focused on mathematics in particular, and so the students were asked to indicate their degree of agreement with various statements related to mathematics. The population considered for our study was composed of $N = 15,499$ 15-year-old Spanish students who responded to the survey, and who attended $C = 838$ different schools. We chose the question “How strongly do you agree with the statement: I enjoy reading about mathematics?” as the main variable, where the possible options were $\mathtt{1}$ = strongly agree, $\mathtt{2}$ = agree, $\mathtt{3}$ = disagree and $\mathtt{4}$ = strongly disagree. The population percentages obtained for these categories were 0.03, 0.149, 0.413 and 0.408, respectively. We then considered the degree of agreement (expressed as Strongly agree, agree, disagree and strongly disagree) with to the following sentences: “Making an effort in mathematics is worth it because it will help me in the work that I want to do later on”, “Learning mathematics is worthwhile for me because it will improve my career” and “I will learn many things in mathematics that will help me get a job” as auxiliary variables.

With these data as population, we used a stratified design, selecting a sample of schools with probabilities proportional to their size within each stratum. Then, the values of all the students at the selected schools were observed. The population was divided into five different strata depending on the type of location of each school: villages (fewer than 3,000 people), small towns (3,000 to 15,000 people), towns (15,000 to 100,000 people), cities (100,000 to 1,000,000 people) and large cities (over 1,000,000 people). The number of schools (C_h) and students (N_h) by stratum is detailed in Table 1.

Table 1: Strata population data.

	Villages	Small towns	Towns	Cities	Large cities	Total
C_h	48	239	254	269	28	838
N_h	831	4,312	4,795	5,046	515	15,499

Two sample sizes for schools ($c = 25$ and $c = 50$) are included in the study. A sample of schools using a Midzuno sampling scheme was drawn from each stratum considering probabilities proportional to the school size (taken as the number of students enrolled in the school).

The free statistical software R ([30]) was used to perform this simulation study. The library ordinal of R ([7]) was used, where necessary, to estimate the parameters of the ordinal model. We have developed new R-code implementing the proposed estimators. The R libraries samplingVarEst ([14]) and samplingEstimates ([13]) were used to estimate the variance of the estimators according to the different methods discussed. For each estimator, we computed the percent relative bias $\mathrm{RB\%} = \mathrm{E}_{\mathrm{MC}}(\hat{P} - P)/P * 100\%$ and the percent relative mean squared error $\mathrm{RMSE\%} = \mathrm{E}_{\mathrm{MC}}[(\hat{P} - P)^2]/P^2 * 100\%$ for each category of the main variable Y based on 1,000 simulation runs. We used RMSE% to calculate the percent relative efficiency gain with respect to the HT estimator for the three remaining estimators. The minimum, maximum and mean percent over the categories are also calculated (in absolute values for the relative bias).

The results for relative bias and relative efficiency based on 1,000 simulated samples are shown in Table 2. Additionally, the mean number of students finally observed in each scenario, $\bar{n}$, is included for informative purposes.

Table 2: Relative bias (in % and Italics) and Relative efficiency (with respect to the HT estimator) of the estimators. Auxiliary variables: "Making an effort in...", "Learning mathematics is...", "I will learn many...".

Estimator	1	2	3	4	min	max	mean
	$c = 25$ $(\bar{n} = 482.88)$						
HT	*0.35*	*−0.02*	*−0.34*	*−0.25*	*0.02*	*0.35*	*0.24*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
LV	*0.74*	*0.30*	*−0.10*	*−0.06*	*0.06*	*0.74*	*0.30*
	111.37	141.94	329.21	318.49	111.37	329.21	225.25
MA	*0.73*	*0.43*	*−0.08*	*−0.13*	*0.08*	*0.73*	*0.34*
	111.71	143.76	372.79	367.04	111.71	372.79	248.82
MC	*0.78*	*0.48*	*−0.11*	*−0.12*	*0.11*	*0.78*	*0.37*
	110.65	144.05	381.46	374.35	110.65	381.46	252.62
	$c = 50$ $(\bar{n} = 966.43)$						
HT	*−0.40*	*−0.37*	*−0.58*	*−0.45*	*0.37*	*0.58*	*0.45*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
LV	*−0.02*	*0.09*	*−0.11*	*0.07*	*0.02*	*0.11*	*0.07*
	111.92	131.05	333.83	313.76	111.92	333.83	222.64
MA	*0.03*	*0.17*	*−0.09*	*0.03*	*0.03*	*0.17*	*0.08*
	112.31	131.56	389.68	381.50	112.31	389.68	253.76
MC	*−0.02*	*0.19*	*−0.09*	*0.03*	*0.02*	*0.19*	*0.08*
	112.83	132.42	397.13	389.41	112.83	397.13	257.94

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

Relative bias is below 1% in all cases and can be considered negligible. Both model-assisted and model-calibrated estimators show good performance in terms of efficiency, with the first of these showing slightly better results. Whatever the estimator, the most accurate estimations are achieved in categories 3 and 4, those with the largest population sizes.

The efficiency of these estimators is greater than that of the HT estimator in all cases, and is especially high for categories 3 and 4. As the sample size increases, so does the relative efficiency of the ordinal estimators, with values close to 400% in categories 3 and 4 for $c = 50$.

An alternative model was then fitted for the same variable response, taking the student's gender and the educational level of the father and mother as auxiliary variables. With these covariates it was not possible to obtain a good model fit, since they achieved a very low association with the main variable in the population. Indeed, the Akaike Information Criterion (AIC) in this case was noticeably higher than the value obtained for the previous model fit.

Table 3 shows the results of relative bias and relative efficiency of the estimators for these variables.

Table 3: Relative bias (in % and Italics) and relative efficiency (with respect to the HT estimator) of the estimators. Auxiliary variables: Sex of the student, educational level of of the father and that of the mother.

Estimator	1	2	3	4	min	max	mean
	$c = 25 \quad (\bar{n} = 460.03)$						
HT	*0.56*	*−0.08*	*−0.41*	*−0.13*	*0.08*	*0.56*	*0.29*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
LV	*1.01*	*0.37*	*−0.24*	*0.04*	*0.04*	*1.01*	*0.41*
	108.00	126.32	329.57	293.39	108.00	329.57	214.32
MA	*0.79*	*0.27*	*−0.22*	*0.06*	*0.06*	*0.79*	*0.33*
	110.00	128.24	333.88	288.83	110.00	333.88	215.23
MC	*0.68*	*0.31*	*−0.24*	*0.08*	*0.08*	*0.68*	*0.33*
	110.11	127.97	335.45	291.91	110.11	335.45	216.36
	$c = 50 \quad (\bar{n} = 920.96)$						
HT	*−0.36*	*−0.38*	*−0.61*	*−0.40*	*0.36*	*0.61*	*0.44*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
LV	*0.53*	*0.36*	*−0.20*	*0.03*	*0.03*	*0.53*	*0.28*
	107.71	113.25	340.53	284.22	107.71	340.53	211.42
MA	*0.13*	*0.19*	*−0.15*	*0.07*	*0.07*	*0.19*	*0.14*
	109.11	114.43	344.48	278.70	109.11	344.48	211.68
MC	*0.06*	*0.24*	*−0.17*	*0.08*	*0.06*	*0.24*	*0.14*
	109.68	113.39	345.99	280.24	109.68	345.99	212.32

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

The results for this population presented a similar pattern: both model-assisted and model-calibrated ordinal estimators achieved very good performance, but in this case, the differences between the estimators are not significant. Efficiency gains with respect to the HT estimator are smaller in this scenario than in the previous one. However, once again, the largest efficiency gains are obtained in categories 3 and 4.

Alternative scenarios were also considered, and these yielded similar results. Specifically, even in the case in which tests of the proportional odds assumption provided evidence of the non-proportional odds context, the efficiency results were comparable.

In a similar way, we computed confidence intervals using different methods to estimate the variance of the estimators. Tables 4 and 5 show the relative length (length / parameter) in % and the empirical coverage of the confidence intervals, in the first case for a good model fit and in the second for a bad one.

Table 4: Relative length in % (LEN) and empirical coverage (COV) of confidence intervals for the estimators using different estimators for variance. Auxiliary variables: ”Making an effort in...”, ”Learning mathematics is...”, ”I will learn many...”. Nominal level 95%.

Estimator		LV		MA		MC		LV		MA		MC	
		LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV
		$c = 25$						$c = 50$					
Lin	1	115.00	89.8	115.83	90.3	115.75	90.7	84.69	92.0	84.97	91.8	84.95	92.0
	2	49.82	92.8	49.22	93.9	49.23	93.2	35.42	92.6	35.03	92.5	35.02	92.7
	3	25.45	91.4	23.96	91.5	23.94	91.8	18.09	92.2	17.13	92.5	17.12	93.1
	4	26.68	91.7	24.22	93.0	24.22	93.2	18.99	93.0	17.35	91.8	17.36	92.0
	mean	54.24	91.4	53.31	92.0	53.29	92.2	39.30	92.4	38.62	92.2	38.61	92.5
W-R	1	111.24	89.8	111.97	90.4	111.74	90.4	81.96	92.3	82.20	92.2	82.12	92.1
	2	46.77	91.9	46.36	92.0	46.32	91.9	33.21	90.8	32.92	91.1	32.91	90.8
	3	24.57	92.7	24.05	93.6	24.05	94.1	17.38	92.7	16.98	94.0	16.98	94.2
	4	23.98	91.1	22.98	92.6	22.98	93.1	16.94	90.6	16.23	91.2	16.23	91.7
	mean	51.64	91.4	51.34	91.2	51.27	92.4	37.37	91.6	37.08	92.1	37.06	92.2
EB	1	114.31	89.4	115.35	90.6	115.17	90.6	84.74	91.7	85.11	91.6	85.10	91.6
	2	49.98	93.3	48.94	93.2	48.91	93.3	35.71	92.7	34.99	92.7	34.98	92.8
	3	24.90	91.1	23.81	91.4	23.81	91.9	17.92	91.8	17.13	93.1	17.13	93.6
	4	26.13	92.0	24.10	92.9	24.10	93.1	18.80	91.7	17.36	92.1	17.37	92.1
	mean	53.83	91.5	53.05	92.0	53.00	92.2	39.29	92.0	38.65	92.4	38.65	92.5
CBS	1	114.30	89.4	115.34	90.6	115.16	90.6	84.73	91.7	85.11	91.6	85.10	91.6
	2	49.98	93.3	48.94	93.2	48.90	93.3	35.71	92.7	34.99	92.7	34.98	92.8
	3	24.90	91.1	23.81	91.4	23.81	91.9	17.92	91.8	17.13	93.1	17.13	93.6
	4	26.13	92.0	24.10	92.9	24.10	93.1	18.80	91.7	17.36	92.1	17.37	92.1
	mean	53.83	91.5	53.05	92.0	52.99	92.2	39.29	92.0	38.65	92.4	38.64	92.5
Tukey	1	110.94	90.0	111.68	90.3	111.44	90.3	82.08	92.3	82.32	92.4	82.25	92.2
	2	46.74	92.1	46.30	92.4	46.26	92.6	33.31	91.7	32.99	91.3	32.98	90.9
	3	24.47	92.8	23.97	93.6	23.98	94.4	17.40	92.5	17.01	93.9	17.01	94.4
	4	23.88	90.7	22.91	92.4	22.90	92.7	16.95	90.6	16.25	91.7	16.25	92.3
	mean	51.51	91.4	51.22	92.2	51.14	92.5	37.43	91.8	37.14	92.3	37.12	92.5

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

Table 5: Relative length in % (LEN) and empirical coverage (COV) of confidence intervals for compared estimators using different estimators for variance. Auxiliary variables: Sex of the student, educational level of father and educational level of mother. Nominal level 95%.

Estimator		LV		MA		MC		LV		MA		MC	
		LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV
		$c = 25$						$c = 50$					
Lin	1	123.09	90.1	122.30	89.1	122.31	88.6	90.16	91.7	89.86	91.4	89.87	91.9
	2	53.27	93.3	53.04	92.9	53.03	93.2	37.78	93.1	37.60	92.8	37.62	93.0
	3	25.45	90.6	25.43	90.7	25.43	91.4	18.13	91.8	18.13	92.3	18.13	92.1
	4	27.84	91.6	27.84	92.6	27.84	92.3	20.00	92.5	20.05	92.6	20.05	92.3
	mean	57.41	91.4	57.15	91.3	57.15	91.3	41.52	92.3	41.41	92.3	41.42	92.3
W-R	1	119.22	89.2	118.28	89.6	118.36	89.9	87.81	91.6	87.14	92.1	87.18	91.9
	2	49.39	91.8	48.94	92.2	48.91	91.5	34.96	91.6	34.60	91.8	34.59	91.8
	3	25.18	92.9	25.26	93.0	25.26	92.5	17.80	92.6	17.83	93.2	17.83	93.1
	4	25.52	91.4	25.34	91.4	25.34	91.0	18.07	90.7	17.95	90.6	17.95	90.4
	mean	54.82	91.3	54.45	91.5	54.47	91.2	39.66	91.6	39.38	91.9	39.38	91.8
EB	1	122.34	89.5	121.73	89.0	121.82	88.9	90.16	91.9	89.97	91.9	90.02	92.1
	2	52.43	92.9	52.81	93.7	52.78	93.2	37.40	92.6	37.64	92.9	37.62	93.1
	3	25.19	90.9	25.25	91.0	25.25	91.1	18.09	91.6	18.13	92.3	18.13	92.5
	4	27.62	91.6	27.70	92.2	27.69	92.6	19.95	92.4	20.06	92.4	20.06	92.8
	mean	56.89	91.2	56.87	91.4	56.89	91.4	41.40	92.1	41.45	92.4	41.45	92.6
CBS	1	122.33	89.5	121.73	89.0	121.81	88.9	90.15	91.9	89.97	91.9	90.01	92.1
	2	52.43	92.9	52.81	93.7	52.78	93.2	37.40	92.6	37.63	92.9	37.62	93.1
	3	25.19	90.9	25.25	91.0	25.25	91.1	18.09	91.6	18.13	92.3	18.13	92.5
	4	27.61	91.6	27.69	92.2	27.69	92.6	19.95	92.4	20.06	92.4	20.06	92.8
	mean	56.89	91.2	56.87	91.4	56.88	91.4	41.40	92.1	41.45	92.4	41.45	92.6
Tukey	1	118.85	89.5	117.96	88.5	118.04	88.5	87.93	91.9	87.27	92.3	87.31	92.1
	2	49.28	92.6	48.87	92.7	48.84	91.8	35.01	91.7	34.69	91.6	34.67	92.1
	3	25.08	92.8	25.17	93.1	25.17	92.9	17.82	92.2	17.86	93.0	17.85	93.0
	4	25.43	91.6	25.26	91.8	25.25	91.4	18.09	91.2	17.98	90.9	17.98	91.0
	mean	54.66	91.6	54.31	91.5	54.33	91.1	39.72	91.7	39.45	91.9	39.45	92.0

1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree

It is no easy matter to compare all the estimators and all the variance estimation techniques over all the categories. However, the tables obtained show that the lengths of the EB (Escobar-Berger, [15]) and CBS (Campbell, [4]; Berger and Skinner, [2]) intervals are practically the same, and also that the intervals with the LV estimator have longer lengths, while those with the MC estimators have shorter ones, for both sample sizes. Obviously, the length of the confidence intervals decreases as the sample size increases.

The coverage is below the nominal value in every case. The MC estimator obtains the closest coverage to the nominal level, but with small differences with respect to the other estimators.

8. APPLICATION TO AN OPINION SURVEY

In this section, the ordinal regression approach is illustrated using a real survey, deriving the proposed estimates and comparing these to alternative ones.

This population-based survey was conducted by the Institute of Social Studies of Andalusia, a public scientific research institute specialising in the social sciences. Its aim is to reflect the opinions of the population of Andalusia, a region in Southern Spain, with regard to various aspects of policies. Taking into account the time and budget available, 1,890 interviews were performed by qualified interviewers, specially trained in survey techniques. The interviews were carried out by the Statistics and Surveys sections of the institute using Computer Assisted Telephone Interviewing data input techniques. A stratified random sampling design with eight strata, each one corresponding to a municipality in the region, was considered. In each stratum, a simple random sampling without replacement design was considered. The design weights were modified to adjust for coverage and non-response bias. The two main variables included in this study, related to "education" and "housing", are the answers to the following questions:

- *Do you think that education issues have improved, remain the same or have worsened in recent years?*
- *Do you think that housing issues have improved, remain the same or have worsened in recent years?*

each one with three possible response categories. As in the simulation study, R software and the library ordinal were used to analyze the data. Together with the ordinal model-assisted MA (4.1) and the ordinal model-calibrated MC (4.2) estimators, the HT estimator (2.1) and the LV estimator (2.3) were computed for comparison purposes. As auxiliary information we took into account the sex of the respondents, a categorical variable with two possible outcomes, and their age, categorized into four age ranges. The population information for the auxiliary variables needed to compute the LV estimator and the two proposed estimators is shown in Table 6.

Table 6: Population information for the auxiliary variables (Sex and Age).

Sex	*Age*			
	18–29	30–44	45–59	≥ 60
Male	411,501	699,378	636,061	578,775
Female	460,834	649,434	615,057	731,410

Table 7 shows the point and 95% confidence interval estimation of proportions of each category of the main variables.

Table 7: Point (PROP) and 95% confidence level estimation (lower bound, LB, upper bound, UB, and length, LEN) of percentages. Auxiliary variables: Sex and Age.

Estimator	*In recent years, education issues...*				*In recent years, housing issues...*			
	PROP	LB	UB	LEN	PROP	LB	UB	LEN
	... have improved				*... have improved*			
HT	3.88	2.61	5.16	2.55	7.15	5.74	8.55	2.81
LV	4.49	3.11	5.87	2.76	7.65	6.10	9.20	3.10
MA	3.92	2.67	5.18	2.51	7.08	5.69	8.47	2.78
MC	3.94	2.68	5.19	2.51	7.07	5.68	8.47	2.79
	... remain the same				*... remain the same*			
HT	17.69	15.44	19.94	4.50	9.42	7.80	11.04	3.24
LV	18.12	15.87	20.37	4.50	9.87	8.12	11.63	3.51
MA	17.84	15.64	20.03	4.39	9.35	7.74	10.96	3.22
MC	17.82	15.63	20.02	4.39	9.36	7.76	10.97	3.21
	... have worsened				*... have worsened*			
HT	78.41	74.59	82.24	7.65	83.42	79.52	87.32	7.80
LV	77.38	74.76	79.99	5.23	82.47	80.00	84.93	4.93
MA	78.22	75.82	80.63	4.81	83.56	81.52	85.59	4.07
MC	78.23	75.83	80.63	4.80	83.55	81.52	85.59	4.07

Whatever the category of either of the two main variables, the lengths of the confidence intervals of the proposed estimators are shorter than that of the corresponding confidence interval associated with the LV estimator, which uses the same amount of auxiliary information. In part, these differences are due to the better fit of the ordinal logistic model than the multinomial logistic model in both cases. Indeed, for the two response variables, the AIC is larger for the multinomial model than for the ordinal model. To highlight these discrepancies, we computed the relative length reduction of the confidence intervals of the proposed estimators with respect to the corresponding confidence intervals of the LV estimator. The results are shown in Table 8.

The length reductions are significant in all categories of the response variables (6.5% on average for the first variable and 12% for the second).

Tables 9 and 10 show the point estimation for the proposed estimators, classified by sex and age. In the first of these respects, it is noticeable that more men than women believe that education and housing issues have improved or remain the same, while the women are slightly more pessimistic.

The general perception that these issues have worsened is common to all age groups, with the highest such proportion being found among respondents aged 45-59 years.

Table 8: Relative length reduction in % of the 95% confidence intervals of the proposed estimators with respect to the LV estimator.

In recent years, education issues...				
Estimator	REDUCTION			
	... have improved	*... remain the same*	*... have worsened*	MEAN
MA	9.326725	2.359543	8.013819	6.566695
MC	9.349046	2.371333	8.014386	6.578255

In recent years, housing issues...				
Estimator	REDUCTION			
	... have improved	*... remain the same*	*... have worsened*	MEAN
MA	10.12919	8.518985	17.41836	12.022178
MC	10.10099	8.540746	17.42298	12.021572

Table 9: Point estimation of percentages by sex.

Estimator	*In recent years, education issues...*			*In recent years, housing issues...*		
	ALL	MEN	WOMEN	ALL	MEN	WOMEN
	... have improved			*... have improved*		
HT	3.88	4.65	3.15	7.15	10.00	4.39
LV	4.49	4.96	4.05	7.65	10.33	5.11
MA	3.92	4.69	3.21	7.08	9.97	4.34
MC	3.94	4.69	3.16	7.07	9.99	4.38
	... remain the same			*... remain the same*		
HT	17.69	20.71	14.80	9.42	11.59	7.33
LV	18.12	21.27	15.15	9.87	11.97	7.89
MA	17.84	20.83	15.01	9.35	11.49	7.32
MC	17.82	20.72	14.84	9.36	11.58	7.32
	... have worsened			*... have worsened*		
HT	78.41	74.63	82.05	83.42	78.40	88.27
LV	77.38	73.76	80.80	82.47	77.69	86.99
MA	78.22	74.47	81.78	83.56	78.52	88.32
MC	78.23	74.58	81.99	83.55	78.42	88.28

Table 10: Point estimation of percentages by age groups.

Estimator	*In last years, education issues...*					*In last years, housing issues...*				
	ALL	18–29	30–44	45–59	≥60	ALL	18–29	30–44	45–59	≥60
	... have improved					*... have improved*				
HT	3.88	3.99	4.24	1.92	5.61	7.15	8.66	8.34	4.94	6.85
LV	4.49	4.20	4.15	1.87	7.54	7.65	8.90	8.24	4.86	8.87
MA	3.92	4.01	4.26	1.84	5.51	7.08	8.52	8.50	4.99	6.66
MC	3.94	4.09	4.22	1.89	5.66	7.07	8.59	8.29	4.92	6.76
	... remain the same					*... remain the same*				
HT	17.69	16.75	17.02	17.06	20.07	9.42	10.16	11.47	8.90	7.07
LV	18.12	16.87	16.97	16.99	21.23	9.87	10.44	11.33	8.74	9.07
MA	17.84	16.31	17.17	17.35	20.02	9.35	9.68	11.61	9.13	7.01
MC	17.82	16.84	17.08	17.22	20.16	9.36	10.11	11.43	8.86	7.02
	... have worsened					*... have worsened*				
HT	78.41	79.25	78.73	81.01	74.31	83.42	81.17	80.17	86.14	86.07
LV	77.38	78.92	78.87	81.13	71.21	82.47	80.65	80.42	86.38	82.04
MA	78.22	79.67	78.56	80.79	74.46	83.56	81.78	79.88	85.87	86.32
MC	79.06	78.69	80.87	80.78	74.16	83.55	81.29	80.27	86.21	86.20

9. CONCLUSIONS

Data collected from surveys are often organized into discrete categories. Analyzing variables with ordinal outcomes, obtained from a complex survey, often requires specialised techniques. To improve the accuracy of estimation procedures, a survey statistician often makes use of the auxiliary data available from administrative registers and other sources.

In this paper, we present estimation techniques applied to the results of complex surveys when the variable of interest has ordinal outcomes, and describe the joint distribution of the class indicators by an ordinal model. Ordinal model-assisted estimators and ordinal model-calibrated estimators are introduced for class frequencies, using two different approaches to estimation.

We show that the proposed estimators are asymptotically normal distributed and we derive expressions for their asymptotic variances. Resampling techniques are obtained when joint inclusion probabilities are unavailable to data analysts.

We used the weighted ML estimation procedure to obtain the estimators for the model parameters because in the iterative fitting process for WLS, assuming ordinal data, at some settings of explanatory variables the estimated mean may fall below the lowest score or above the highest one and then the fit will fail [1]. When numerical maximization for the pseudolikelihood is feasible, good estimates may be obtained in certain cases by WLS. This approach is usable when working with discrete predictors [23].

We also include a limited simulation study with a real population, finding that the ordinal logistic formulation yields better results than the classical estimators that implicitly assume individual linear models for the variables.

The effective use of auxiliary information from survey data depends on the population quantities to be estimated and on the actual relation between the response variable and the covariates. The simulation results obtained show that these estimators are robust against misspecified models.

Ordinal model-assisted and model-calibrated estimators also have some drawbacks: they require a sampling frame, complete with all the explanatory variables used in the assisting model, for all units in the population. This situation frequently arises, for example, when categorical variables (such as gender or the professional status of the individual) or quantitative categorized variables (such as the age of the individual, grouped into classes) are used as auxiliary information in a survey. In this context, although we do not have a complete list of individuals, the proposed estimators can still be computed because the necessary population information can be found in the databases published by national statistical agencies and in business registers and trade association lists. This is the case in our application of the estimators to data from the survey on opinions and attitudes, as discussed in Section 8.

ACKNOWLEDGMENTS

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R and FPU grant programme, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grants SEJ2954 and HUM1413 Junta de Andalucía, Spain).

REFERENCES

[1] Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, second ed. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[2] Berger, Y.G. and Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling, *Journal of the Royal Statistical Society B*, **67**, 79–89.

[3] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review/Revue Internationale de Statistique*, **51**, 279–292.

[4] Campbell, C. (1980). A different view of finite population estimation, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 319–324.

[5] Cassel, C.M.; Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika*, **63**, 615–620.

[6] Coelho, P.S. and Casimiro, F. (2008). Post Enumeration Survey of the 2001 portuguese population and housing censuses, *REVSTAT – Statistical Journal*, **6**(3), 231–252.

[7] Christensen, R.H.B. ordinal - Regression Models for Ordinal Data. R package version 2015, 6–28.

[8] Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

[9] Diana, G. and Perri, P.F. (2013). A class of estimators in two-phase sampling with subsampling the non-respondents, *Applied Mathematics and Computation*, **219**(19), 10033–10043.

[10] Duchesne, P. (2003). Estimation of a proportion with survey data, *Journal of Statistics Education*, **11**(3), 1–24.

[11] Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities, *Journal of the Royal Statistical Society*, **15**, 262–269.

[12] Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

[13] Escobar, E.L. samplingestimates: Sampling estimates. R package version 0.1-3.

[14] Escobar, E.L. and Barrios, E. Samplingvarest: Sampling variance estimation. R package version 0.9-9.

[15] Escobar, E.L. and Berger, Y.G. (2013). A jackknife variance estimator for self-weighted two-stage samples, *Statistica Sinica*, **23**, 595–613.

[16] Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion), *Statistical Science*, **22**(2), 153–164.

[17] Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation, *International Statistical Review/Revue Internationale de Statistique*, **54**(2), 127–138.

[18] Hardin, J.W. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*, College Station, Texas: Stata Press.

[19] Henri, C. and Gavin, P. (2016). The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems, *Department of Civil and Environmental Engineering. Duke University*, `http://people.duke.edu/~hpgavin/ce281/lm.pdf`

[20] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**(377), 663–685.

[21] Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**(377), 89–96.

[22] Kennel, T.L. and Valliant, R. (2010). Logistic generalized regression (LGREG) estimator in cluster samples, *Proceedings of the Section on Survey Research Methods*, 4756–4770.

[23] Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators, *Survey Methodology*, **24**(1), 51–55.

[24] Liu, X. and Koirala, H. (2013). Fitting proportional odds models to educational data with complex sampling designs in ordinal logistic regression, *Journal of Modern Applied Statistical Methods*, **12**(1), 235–248.

[25] Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variable*, Advanced Quantitative Techniques in the Social Sciences, Sage Publications, Thousand Oaks, CA.

[26] Mert, Y. (2015). Generalized least squares and weighted least squares estimation methods for distributional parameters, *REVSTAT – Statistical Journal*, **13**(3), 263–282.

[27] Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization, *The Computer Journal*, **7**, 308–313.

[28] Nordberg, L. (1989). Generalized linear modeling of sample survey data, *Journal of Official Statistics*, **5**, 223–239.

[29] Pavía, J.M. and Larraz, B. (2012). Nonresponse bias and superpopulation models in electoral polls, *Reis*, **137**, 121–150.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015, `https://www.R-project.org/`.

[31] Särndal, C.E. (2007). The calibration approach in survey theory and practice, *Survey Methodology*, **33**, 99–119.

[32] Särndal, C.E.; Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

[33] Tillé, Y. and Matei, A. sampling: Survey sampling. r package version 2.6.

[34] Tukey, J.W. (1958). Bias and confidence in not-quite large samples, *The Annals of Mathematical Statistics*, **29**(2), 614.

[35] Valliant, R.; Dorfman, A. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley, New York.

[36] Winship, C. and Mare, R.D. (1984). Regression models with ordinal variables, *American Sociological Review [Internet]*.

[37] Wolter, K.M. (2007). *Introduction to Variance Estimation*, second ed., Springer.

[38] Wu, C. (2003). Optimal calibration estimators in survey sampling, *Biometrika*, **90**(4), 937–951.

[39] Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, **96**(453), 185–193.