Estimating the proportion of a categorical variable with probit regression

Sergio Martínez[1], Antonio Arcos[2], María del Mar Rueda[2], Helena Martínez[1]
1 Department of Mathematics. University of Almería, Spain
2 Department of Statistics and Operational Research. University of Granada, Spain

This paper discusses the estimation of a population proportion, using the auxiliary information available, which is incorporated into the estimation procedure by a probit model fit. Three probit regression estimators are considered, using model-based and model-assisted approaches. The theoretical properties of the proposed estimators are derived and discussed. Monte Carlo experiments were carried out for simulated data and for real data taken from a database of confirmed dengue cases in Mexico. The probit estimates gives valuable results in comparison to alternative estimators. Finally, the proposed methodology is applied to data obtained from an immigration survey.

Keywords: *Auxiliary information, Calibration estimator, Probit Regression, Finite population, Sampling design.*

1. Introduction

Economic status, education levels and public health conditions are estimated through surveys conducted by national organisations. Marketing analysts estimate people's preferences by means of consumer panels. Private agencies determine the populations opinions on issues such as the economy, school budgets or changes in legislation. Polls are conducted to ascertain voting intentions. To estimate proportions and percentages in all these cases,we use statistical methods of inference in survey sampling.

Most methods for estimating a population proportion and forming confidence intervals are based on the assumption of a simple random sample drawn from a large population. However, this scenario is not always present in practice, i.e., many surveys assume a finite population with samples extracted from complex sampling designs. For example, the National Health and Nutrition Examination Survey (NHANES) carried out by the National Center for Health Statistics in the USA, the National Assessment of Educational Progress (NAEP) and the OECD Programme for International Student Assessment (PISA), all use complex sampling designs (including stratification and multistage sampling). In this situation, the use of estimation methods involving sampling weights can provide better estimates than the customary approaches. Sampling weights are needed to make valid inferences about the populations from which they were drawn. Appropriate sampling weights are computed to obtain unbiased estimates of

population characteristics.

In many populations, particularly ones that have been previously sampled or surveyed, a frame of units is available, together with auxiliary data on each unit. In other cases, a full frame of all units is not directly present but can be constructed by sampling in stages and assembling a partial frame at each stage. In both single and multi-stage sampling, auxiliary data can be used to construct efficient estimators of totals. The growing availability of information from census data, administrative registers and previous surveys provides a wide range of variables concerning the population of interest, which can legitimately be employed as auxiliary information. An obvious example of the availability of auxiliary information is that of election polls: the data recorded in previous elections, electoral section levels, respondents age, sex or past vote recall are all elements of auxiliary information that can be used to improve the quality of sampling estimates (e.g. [25], [33], [32].)

In the presence of auxiliary information, various design-based approaches may be used to improve the precision of estimators at the estimation stage (see [41]), including ratio, difference and post-stratification methods

These techniques are generally more efficient than other methods which do not make use of auxiliary information. However, the above-mentioned techniques were originally proposed to estimate means and totals of quantitative variables, and while their extension to the estimation of proportions is possible, it requires further investigation. For example, the analyst should be aware of the risks that may arise when confidence intervals are constructed for a population proportion, since limits outside $[0, 1]$ could be achieved. This situation does not occur with respect to means associated with quantitative variables.

The efficient insertion of the auxiliary information available would improve the precision of the estimations for the proportion of a categorical variable of interest, but other methods are more appropriate. In fact, the post-stratification techniques and ratio estimators currently used in the polling industry in order to reduce deviations do not show enough capacity to mend the biases introduced when collecting data ([30], [31]). A real example was used in [17], namely a series of CBS/New York Times national polls from the 1988 election campaign, to show how post-stratification and other weighting techniques using auxiliary information really improve the proportion estimation by correcting known differences between the sample and the population. This study estimated the proportion of voters who supported the Republican candidate, by means of weighting schemes incorporating as auxiliary variables the number of adults and the number of telephone lines in the sampled household, the region of the country, and the respondent's sex, ethnicity, age and education level. All of these variables have an important effect on levels of nonresponse.

If the expectation of the response variable can be assumed to depend linearly on the auxiliary variables, as can be the case for continuous response variables, it is advisable to use the generalised regression ([39]) or the calibration estimator ([15, 38]). However, a linear model is not the best choice for binary response variables. For such variables we introduce a class of estimators based on a probit model describing the joint distribution of the class indicators. As is well known in the sociological literature ([44]) the probit or logit specifications are usually

preferable to the linear model because the former take account of the ceiling and floor effects on the dependent variable whereas the linear model does not. Logit and probit regression models for dichotomous data has been extensively used in sociological and educational studies ([1], [6], [28], [14]...) although its use for parameter estimation in finite populations sampling is very sparse ([22]).

This paper is organised as follows. Section 2 presents the notation used and briefly reviews the methods proposed in the literature to estimate a population proportion using auxiliary information. Then, Section 3 illustrates the proposed class of estimators using the probit model, first addressing the design based approach and then considering the model based approach. The theoretical properties of the proposed estimators are also investigated. Section 4 reports the results obtained from an extensive simulation study run on a set of simulated and real finite populations in which the performance of the proposed class of estimators is investigated for finite size samples. Section 5 shows an application of the proposed methods for data from a real survey and, finally, Section 6 presents the conclusions drawn.

2. Estimation of a proportion under a general sampling design

Consider the scenario of a finite population $U = \{1, \ldots, N\}$ containing N units. Let $A_1, \ldots, A_N$ denote the values of an attribute of interest A, where $A_k = 1$ if the kth unit possesses the attribute A and $A_k = 0$ otherwise. We also assume that the sample s is selected according to a specified sampling design with inclusion probabilities π_k and π_{kl} is assumed to be strictly positive. The aim is to estimate the population proportion of individuals that possess the attribute A, i.e. $P_A = N^{-1} \sum_{k=1}^{N} A_k$.

The customary design for an unbiased estimator of P_A is the Horvitz-Thompson estimator given by

$$\widehat{P}_{AHT} = \frac{1}{N} \sum_{k \in s} \frac{A_k}{\pi_k} = \frac{1}{N} \sum_{k \in s} A_k d_k. \tag{1}$$

The values $d_k = 1/\pi_k$ are known as design weights and are commonly thought of as the number of population units represented by unit k in the sample. The weights permit valid inferences to be drawn between the samples and the respective populations from which they were drawn and, most importantly, ensure that the results of the assessments are fully representative of the target populations.

Note that estimator (1) makes no use of the auxiliary information at the estimation stage.

In a sample survey, auxiliary information is often used at the estimation stage to increase the precision of estimators of totals or means ([41, 3, 4]), variance ([42, 40]), covariance ([34]) distribution functions ([43]), quantile ([36]), etc. by using ratio, regression, difference and calibration estimators.

When the interest is the estimation of a proportion, it is also common to have auxiliary information related to the attribute of interest.

Let B denote an auxiliary attribute associated with A and values given by $B_1, \ldots, B_N$. Ratio type estimators are known methods involving auxiliary information that possess various desirable properties including an important gain in efficiency. The ratio estimator for a proportion is defined by

$$\widehat{P}_r = \widehat{R} P_B, \tag{2}$$

where $\widehat{R} = \widehat{P}_{AHT}/\widehat{P}_{BHT}$, $\widehat{P}_{BHT} = \sum_{k \in s} \frac{B_k}{\pi_k}$ and $P_B = N^{-1} \sum_{k=1}^{N} B_k$. We assume that the population proportion of individuals that possess the attribute B, P_B, is known from a census or is estimated without error, but we do not have access to each individual datum.

In [37] it was proposed that regression type estimators could be used to estimate P_A. The regression type estimator of P_A is

$$\widehat{P}_{reg} = \widehat{P}_{AHT} + b(P_B - \widehat{P}_{BHT}), \tag{3}$$

where b is defined minimising the variance of the estimator.

A more restrictive situation is to assume that unit-specific auxiliary data for every unit are available. This is known as complete auxiliary information. If the auxiliaries are known for every unit in the population, this implies that a sampling frame has been constructed that lists every unit in the survey universe. In a universe of elementary-level schools, auxiliaries could include the number of students and teachers, the location of the school (urban, suburban or rural) and the total budget in a recent year; in election polls, auxiliaries could include age, sex or past vote recall.

In the following analysis, we use the usual notation in survey sampling. Let y be the study variable (which in this case corresponds to the presence or absence of an attribute and takes one of two values, 1 or 0). Let us assume the existence of a vector $\mathbf{x} = (x_1, x_2, \ldots x_P)^{'}$ of auxiliary information, such that for every population unit k the value $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Pk})$ is known. We also assume that the variables included in the vector $\mathbf{x}$ can be either numeric or binary attributes of the same type as the study variable y.

In many populations, some of the most useful auxiliaries are qualitative rather than quantitative. For example, in surveys of persons, demographic variables like age group, race-ethnicity, and gender are useful predictors of response variables. Quantitative x can also be used in combination with qualitative ones.

Beyond reducing the variance and compensating for nonresponse or undercoverage, calibration on control totals is widely used in practice for its highly desirable feature of producing estimates that are consistent with external sources. For example, the population may have 100,000 urban residents and 30,000 rural residents. Because of nonresponse, a simple random sample may obtain 1100 urban residents and 200 rural residents, and thus urban residents are overrepresented in the sample. If the urban and rural areas have different means (proportions), an unweighted estimator is biased for the purpose of estimating the population mean (proportion). Calibration is desirable so that demographic counts and other quantities will be consistent across surveys, and consistent with the census ([24]).

The calibration method ([15]) considers the estimation of P_A by an estimator given by

$$\widehat{P}_{CAL} = \frac{1}{N}\sum_{k\in s}\omega_k A_k$$

where the calibration weights ω_k are chosen to minimise an average distance Φ_s from the basic design weights $d_k = 1/\pi_k$, subject to a set of calibration constraints. Thus, the calibration technique replaces the basic weights $d_k = \frac{1}{\pi_k}$ from the Horvitz-Thompson with a new system of weights ω_k minimising the chi-square distance (in practice, this is the most commonly used distance)

$$\chi = \sum_{k\in s}\frac{(\omega_k - d_k)^2}{d_k q_k} \tag{4}$$

subject to the condition

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k = \frac{1}{N}\sum_{k\in s}\omega_k\mathbf{x}_k \tag{5}$$

where q_k are known positive constants unrelated to d_k.

Calibration estimators have some desirable properties. First, weights satisfying (5) provide sample estimates for the totals of the auxiliary variables in $\mathbf{x}_k$ that exactly match the known population totals for these variables. If the population totals of the auxiliary variables were published before the survey results are produced, then using calibration estimators for the survey guarantees that the survey estimates are coherent with those already in the public domain. The second desirable property is simplicity, namely the fact that given the weights ω_k calibration estimates are linear in y. This means that each survey record can carry a single weight to estimate all survey variables. The third property of calibration estimators is their flexibility to incorporate auxiliary information, to include continuous, discrete or both types of benchmark variables at the same time. If the auxiliary totals represent counts of the numbers of population units in certain classes of categorical (discrete) variables, then the values of the corresponding x variables are simply indicators of the units that are members of the corresponding classes.Calibration estimators also yield some degree of integration, in the sense that some widely used estimators are special cases, for example, ratio, regression and poststratification estimators.

It is an implicit assumption in the calibration that the study variable y and $\mathbf{x}$ are linearly related ([45]); however, since the vector $\mathbf{x}$ includes numeric and binary attributes and the characteristic of study y is also a binary attribute, the use of a linear model is difficult to justify.

In the binary case [22] considered a logistic regression model and defined the logistic generalised regression estimator (LGREG) given by

$$\widehat{P}_{LGREG} = \frac{1}{N}\left(\sum_{k\in U} pl_k + \sum_{k\in s}\frac{A_k - pl_k}{\pi_k}\right) \tag{6}$$

where

$$pl_k = \frac{exp(x_k\widehat{\beta})}{1 + exp(x_k\widehat{\beta})} \tag{7}$$

and where $\widehat{\beta}$ is the BLUP estimator of the β parameter of the logistic regression. [16] provided some codes to compute the LGREG estimator and a Monte Carlo study to empirically investigate the accuracy of the confidence intervals when HT and LGREG estimators are used.

3. Estimation of a proportion by using a probit model

In the following we assume the existence of a vector $\mathbf{x} = (x_1, x_2, \dots x_P)^{'}$ of auxiliary information, such that for every population unit k the value $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{Pk})$ is known. We also assume that the variables included in the vector $\mathbf{x}$ can be either numeric or binary attributes of the same type as the study attribute A. For this purpose, we consider a sample $s = \{1, 2, \dots, n\}$ selected according to a specified sampling design with inclusion probabilities π_k and π_{kl} that are assumed to be strictly positive. The value A_k is only available for the sample units.

Under this scenario, we can consider the prediction theory for sampling surveys. The general prediction theory is based on superpopulation models, in which it is assumed that the population under study $\mathbf{y} = (y_1, ..., y_N)'$ constitutes a body of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)'$ containing a super-population model ξ. The value of the variable of interest, associated with the k-th unit of the population, is comprised of a deterministic element $\mu(\mathbf{x}_k)$ (known) and a random element, $Y_k = \mu(\mathbf{x}_k) + e_k,\ k = 1, ..., N$. The random vector $e = (e_1, ..., e_N)$ is assumed to have a zero mean and a positive definite covariance matrix which is diagonal (Y_k are mutually independent).

A superpopulation model is a way of formalising the relationship between a target variable and auxiliary data. For example, in a survey of hospitals, the number of patients discharged in a particular calendar quarter may be related to the number of beds in the hospital and the type of hospital (e.g., general medical and surgical, rehabilitation, children's hospital, military, etc.). Superpopulation models have been used in previous sociological studies: thus, [11] used the superpopulation approach to estimate the average costumer satisfaction from a probability sample, and [30] used superpopulation models to reduce nonresponse bias in electoral pools.

Traditionally, parametric methods utilise regression models to incorporate auxiliary information: $E_\xi(Y_k) = \mu_k = \beta \mathbf{x}_k$ (E_ξ denotes the expected value with respect to the model). The selection of a linear model is fully justified for a continuous response variable.

As is well known in sociological literature ([44]), for binary measurements a linear model might be unrealistic, and ordinarily the logistic or the probit model would be preferred. For example, we might want to estimate the proportion of students with learning disabilities. For such 0-1 Y variables, a logistic or other type of nonlinear model usually produces a better fit than a linear model. This type of estimator can be prediction-unbiased if a linear model holds, but can be seriously biased if, for example, the correct underlying model is logistic. Another problem with using a linear model for a binary variable in the presence of auxiliaries is that the predicted value for a given unit need not be confined to

[0,1], as a probability should be.

In typical settings, the data do not differentiate between probit and logit conditional link functions. Probit models are obtained by discretising a latent normal distribution, a process that has been used extensively in biometrics and econometrics (see [5, 26] and [20]). Indeed, if we consider a random variable z_k following a normal regression model, we only observe the variable $A_k = 1$ when $z_k > 0$). The difference between logit and probit models is to be found on this. For example, if Y were whether a child was born to a woman in a given year, the logit model would express the effects of X on the log of the odds of a birth versus a non-birth. In the probit model, a unit change in X produces a "b" unit change in the cumulative normal probability, or Z score, that Y will fall in a particular category. For example, the probit model would express the effect of a unit change in X on the cumulative normal probability that a woman will give birth within a year.

Probit models have attractive properties compared to logit models ([9]), although the literature on probit models in survey sampling is not very extensive([10]).

We assume that the relationship between the attribute A (the main variable y) and the auxiliary vector $\mathbf{x}$ can be described by the probit model:

$$\mu_k = E_\xi(Y_k) = P(A_k = 1) = F(\beta^{'} \cdot \mathbf{x}_k), \quad (k = 1, ..., N) \tag{8}$$

where F is the normal-standard distribution function and β is a parameter vector.

We now define a new estimator for a proportion, using the probit regression model. To do so, we consider the estimation of the superpopulation parameter β by the units of the sample s. We estimate β by maximising the π-weighted likelihood ([18], [27]). The sample estimator $\widehat{\beta}$ of β is the solution to the following equation

$$\sum_{k \in s} d_k \cdot \frac{f(\beta^{'} \cdot \mathbf{x}_k)}{F(\beta^{'} \cdot \mathbf{x}_k)(1 - F(\beta^{'} \cdot \mathbf{x}_k))} (A_k - F(\beta^{'} \cdot \mathbf{x}_k)) \cdot \mathbf{x_k} = 0 \tag{9}$$

where the function f denotes the normal-standard density function. The value of $\widehat{\beta}$ is thus obtained either by the standard Newton-Raphson method or by Fisher's scoring method. Under certain regularity conditions (similar to those used by [7]), it can be shown that $\widehat{\beta} = \beta + O(n^{-1/2})$ ([46]).

With the estimation $\widehat{\beta}$ of β, we consider the following auxiliary variable

$$p_k = \widehat{P}[A_k = 1] = F(\widehat{\beta}^{'} \cdot \mathbf{x}_k) \tag{10}$$

Since the vector $\mathbf{x}_k$ is known for all units of the population U, the values p_k are available $\forall k \in U$ and we propose the use of the values p_k to obtain new estimators for P_A. Statistical inference procedures can be considered in which stochastic elements are introduced through the randomisation aspect (the fixed population approach) or the stochastic model (the superpopulation approach).

It is in the area of sample surveys where the debate between randomisation-based and model-based inference is most sharply drawn (eg. [23]). The theoretical underpinnings of the two approaches remain strikingly different. We will therefore consider both points of view and propose several estimators based on each approach.

The predictive probit estimator

The first estimator considered in this section is a model based estimator. The proposed estimator is based on the following idea: the population proportion to be estimated is given by

$$P_A = \frac{1}{N}\sum_U A_k = \frac{1}{N}\left(\sum_{k\in s} A_k + \sum_{k\in U-s} A_k\right),$$

where U denotes the set of N units in the population. If a sample s of n units is selected, we can observe the sample total, $\sum_s A_k$. The total for the remainder of the population, $U-s$, is equal to $\sum_{U-s} A_k$, which is unknown and must be estimated using the predicted values p_k instead of the true values.

Thus we define the probit predictive estimator:

$$\widehat{P}_{PP} = \frac{1}{N}\left(\sum_{k\in s} A_k + \sum_{k\in U-s} p_k\right) \tag{11}$$

Theorem 1. Assume the working model used to construct the estimators has the general structure (8); then, the predictive estimator is an asymptotically model unbiased estimator for the proportion P_A.

Proof.

Consider the difference:

$$(\widehat{P}_{PP} - P_A) = \frac{1}{N}(\sum_{k\in s} A_k + \sum_{k\in U-s} p_k - \sum_{k\in s} A_k - \sum_{k\in U-s} A_k) =$$

$$\frac{1}{N}\sum_{k\in\overline{s}}(p_k - A_k) = \frac{1}{N}(\sum_{k\in U-s}(p_k-\mu_k) + (\mu_k - A_k))$$

Thus

$$E_\xi(\widehat{P}_{PP} - P_A) = \frac{1}{N}\sum_{j\in U-s} E_\xi(p_k-\mu_k) + \frac{1}{N}\sum_{j\in U-s} E_\xi(\mu_k - A_k) =$$

$$\frac{1}{N}\sum_{j\in U-s} E_\xi(p_k-\mu_k) \simeq 0$$

because $E_\xi(p_k) \simeq \mu_k, \forall k \in U$.

This estimator is similar to the model-based estimator proposed by [21] but replacing the censored regression or the tobit model with the probit model.

The model-assisted probit estimator

The second case we consider is a model-assisted estimator. Now we write the population proportion as

$$P_A = \frac{1}{N}\left(\sum_{k\in U} p_k + \sum_{k\in U}(A_k - p_k)\right).$$

The proxy total $\sum_U p_k$ is known (as the auxiliary vector $\mathbf{x}_k$ for all units in the population) whereas the total of the differences $\sum_U(A_k - p_k)$ is unknown (since A_k are unknown), but we can use the Horvitz-Thompson estimator to form an unbiased estimator of the unknown term.

Thus, we define the model-assisted probit estimator given by:

$$\widehat{P}_{MAP} = \frac{1}{N}\left(\sum_{k\in U} p_k + \sum_{k\in s} d_k(A_k - p_k)\right) \tag{12}$$

where p_k are the values given by (10).

The estimator $\widehat{P}_{MAP}$ has two parts: a sum of estimated expectations for the population and an adjustment term $\sum_{k\in s} d_k(A_k - p_k)$. The obvious motivation for this construction is the prospect of achieving a highly accurate estimate $\widehat{P}_{MAP}$ through a close fitting assisting model that leaves small residuals $A_k - p_k$.

This estimator is similar to the LGREG estimator but changes the pl_k-values to p_k values.

Theorem 2. The estimator $\widehat{P}_{MAP}$ is approximately design unbiased for P_A with the approximate design variance:

$$AV_d(\widehat{P}_{MAP}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k E_k)(d_l E_l) \tag{13}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ and $E_k = (A_k - F(\beta^{'}\cdot\mathbf{x}_k))$ are the population fit residuals.

Proof.

Denote by E_d and V_d the expected value and the variance with respect to the design.

$$E_d(\widehat{P}_{MAP}) = E_d(\frac{1}{N}(\sum_{k\in U} p_k + \sum_{k\in s} d_k(A_k - p_k))) =$$

$$\frac{1}{N}(\sum_{k\in U} p_k + E_d(\sum_{k\in s} d_k A_k) - E_d(\sum_{k\in s} d_k p_k)) = \frac{1}{N}\sum_{k\in U} p_k + \frac{1}{N}\sum_{k\in U} A_k - \frac{1}{N}\sum_{k\in U} p_k = P_A.$$

On the other hand:

$$V_d(\widehat{P}_{MAP}) = V_d(\sum_{k\in s} d_k(A_k - p_k)) \simeq V_d(\sum_{k\in s} d_k E_k) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k E_k)(d_l E_l)$$

from the properties of the Horvitz-Thompson estimator.

From this expression we can obtain the variance estimator:

$$\widehat{V}(\widehat{P}_{MAP}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k e_k)(d_l e_l) \tag{14}$$

with $e_k = A_k - p_k$ the sample fit residuals.

This estimator can be seen as a particular case of the generalised difference estimator ([13], pag 24), in which we consider $(p_1, ..., p_N)$ instead of the constant vector.

The calibrated probit estimator

Using the idea of model-calibration given in [45], we now define a new model-calibration estimator for P_A using a probit model.

The idea underlying our calibration estimator is that we wish to find new weights ω_k that are similar to the design weights d_k(so as to preserve the unbiased property of the Horvitz-Thompson estimator) and that give perfect estimations for the predicted values, that is $\frac{1}{N}\sum_{k\in s}\omega_k p_k = \frac{1}{N}\sum_{k\in U} p_k$. This condition implies that the calibration estimator for our proportion, $\frac{1}{N}\sum_{k\in s}\omega_k A_k$ will be close to P_A if the predicted values p_k are close to the A_k values.

To obtain this calibration estimator, we consider the minimisation of (4) subject to the following conditions:

$$\begin{aligned}\frac{1}{N}\sum_{k\in s}\omega_k p_k &= \bar{P} = \frac{1}{N}\sum_{k\in U} p_k \\ \frac{1}{N}\sum_{k\in s}\omega_k &= 1.\end{aligned}$$

By taking the Euclidean (or χ^2-statistic) type of distance function, an analytic solution to this problem can be obtained. Denoting by

$$C_1 = \sum_{k\in s} d_k q_k \;\; ; \;\; C_2 = \sum_{k\in s} d_k q_k p_k \;\; ; \;\; C_3 = \sum_{k\in s} d_k q_k p_k^2 \;\text{ and }\; \bar{P}_{HT} = \frac{1}{N}\sum_{k\in s} d_k p_k$$

the new calibration weights are:

$$\begin{aligned}\omega_k &= d_k + d_k q_k p_k N \frac{\left[(\bar{P}-\bar{P}_{HT})\cdot C_1 - \left(1-\frac{1}{N}\cdot\sum_{k\in s} d_k\right)\cdot C_2\right]}{[C_3\cdot C_1 - C_2^2]} + \\ &+ d_k q_k N \frac{\left[\left(1-\frac{1}{N}\cdot\sum_{k\in s} d_k\right)\cdot C_3 - (\bar{P}-\bar{P}_{HT})\cdot C_2\right]}{[C_3\cdot C_1 - C_2^2]} \qquad (15)\end{aligned}$$

and the calibrated estimator for P_A is:

$$\begin{aligned}\widehat{P}_{CP} &= \widehat{P}_{AHT} + (\bar{P}-\bar{P}_{HT})\frac{\left[C_1\cdot\sum_{k\in s} d_k q_k p_k A_k - C_2\cdot\sum_{k\in s} d_k q_k A_k\right]}{[C_3\cdot C_1 - C_2^2]} \\ &+ \left(1-\frac{1}{N}\cdot\sum_{k\in s} d_k\right)\frac{\left[C_3\cdot\sum_{k\in s} d_k q_k A_k - C_2\cdot\sum_{k\in s} d_k q_k p_k A_k\right]}{[C_3\cdot C_1 - C_2^2]} \\ &= \widehat{P}_{AHT} + \left(1-\frac{1}{N}\cdot\sum_{k\in s} d_k\right)\cdot\widehat{D}_1 + (\bar{P}-\bar{P}_{HT})\cdot\widehat{D}_2 \qquad (16)\end{aligned}$$

Theorem 3. If the basic design weights satisfy the condition that the Horvitz-Thompson estimator is asymptotically normally distributed, the estimator $\widehat{P}_{CP}$ is an asymptotically unbiased estimator for P_A and its asymptotic variance is given by

$$AV(\widehat{P}_{CP}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k E_k)(d_l E_l) \qquad (17)$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$; $E_k = A_k - D_1 - D_2\cdot p_k$ and

$$D_1 = \frac{\left[C_U\cdot\sum_{k\in U} q_k A_k - B_U\cdot\sum_{k\in U} q_k p_k A_k\right]}{[C_U\cdot A_U - B_U^2]}$$

$$D_2 = \frac{\left[A_U\cdot\sum_{k\in U} q_k p_k A_k - B_U\cdot\sum_{k\in U} q_k A_k\right]}{[C_U\cdot A_U - B_U^2]}$$

$$A_U = \sum_{k\in U} q_k \quad ; \quad B_U = \sum_{k\in U} q_k p_k \quad \text{and} \quad C_U = \sum_{k\in U} q_k p_k^2$$

Proof.

$\widehat{\beta} = \beta + O(n^{-1/2})$, the function $F(\widehat{\beta}'\cdot\mathbf{x}_k)$ verifies condition ii) and the design weights verify condition iii) of Theorem 1 of [45]. Thus we can apply this Theorem and prove that $\widehat{P}_{CP} = \widehat{P}_{AHT} + O(n^{-1/2})$ and obtain an asymptotically

design-unbiased estimator for P_A.

In order to obtain the variance of this estimator, we apply the formula (3.1) of the variance to the calibration estimator given in [15], and substitute $E_k = y_k - \beta^{'} \cdot \mathbf{x}_k$ by the new residuals $E_k = A_k - D_1 - D_2 \cdot p_k$.

The asymptotic variance given by (17) can be estimated by

$$\widehat{V}(\widehat{P}_{CP}) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} (d_k e_k)(d_l e_l) \tag{18}$$

with $e_k = A_k - \widehat{D}_1 - \widehat{D}_2 \cdot p_k$.

4. Simulation studies

An extensive simulation study was conducted to analyse the performance of the proposed estimators for surveys from finite populations. Our simulations are programmed in `R`, with some new code developed to compute the estimators to be compared.

To evaluate the performance of the proposed estimators for different scenarios, after the simulations have been performed, it is necessary to consider the criteria for evaluating the performance of the results obtained,using each of the statistical approaches being studied.The comparison of the simulated results with the true values used to simulate the data provides a measure of the performance and precision of the simulation process. Performance measures often include an assessment of bias and accuracy. In general, the expectation of the simulated estimates is the main aspect of interest and hence the average of the estimates for all simulations is used to calculate accuracy measures, such as the bias. Methods that result in an unbiased estimate with large variability or in a biased estimate with little variability may be considered of little practical use. Let us now consider the most commonly used performance measures.

Assessment of bias. The bias is the deviation in an estimate from the true quantity, and can indicate the performance of the methods being assessed. One assessment of bias is the difference between the average estimate and the true value. Another approach is to calculate the bias as a percentage of the true value, providing the true value is not equal to zero. The bias as a percentage can be more informative than the former approach. A standardised bias exceeding 40 per cent in either direction has a noticeable adverse impact on rates of efficiency and error.

Assessment of accuracy. The mean squared error (MSE) provides a useful measure of overall accuracy, as it incorporates measures of both bias and variability.

The performance of each proportion estimator was measured and compared in terms of relative bias (RB) and relative efficiency (RE). The simulated values of RB and RE for a particular proportion estimator T were computed as

$$\text{RB} = B^{-1} \sum_{b=1}^{B} (T^b - P)/P, \quad \text{RE} = MSE(\widehat{P}_{HT})/MSE(T^b)$$

where P is the true value for the estimate of interest, $MSE(T^b) = B^{-1}\sum_{b=1}^{B}(T^b - P)^2$, $MSE(\widehat{P}_{HT}) = B^{-1}\sum_{b=1}^{B}(\widehat{P}_{HT}^b - P)^2$, and T^b and $\widehat{P}_{HT}^b$ are the values of T and $\widehat{P}_{HT}$ from the bth simulation, respectively.

RE is the relative efficiency of each estimator with respect to the Horvitz-Thompson (HT) estimator. In either case RE > 100 means the estimator is preferred (the estimator is inefficient relative to the Horvitz-Thompson (HT) estimator). The gain in efficiency, relative to the Horvitz-Thompson (HT) estimator was computed as: GE = RE -100.

The following estimators were compared: Horvitz-Thompson (HT), calibration (CAL), regression (REG) , calibration probit (CP), model assisted probit regression (MAP), predictive probit (PP) and logistic regression (LGREG)

The first simulation studies conducted were based on five simulated populations of $N = 10000$ units, as previously used by [29] and by [2] (see these papers for a more detailed description). These populations covered a wide range of scenarios, including small and large Cramer V coefficients between the attribute of interest and the auxiliary attribute.

For example, for a population proportion of the attribute A as $P = 0.1$, a random sample of 10000 units from a Bernoulli distribution with parameter $P = 0.1$ is obtained. Then,the attribute B whose Cramer's V coefficient with A is 0.9 can be obtained by changing just a few values in A. The progressive change of more values in A can allow found B that verify that the Cramer's V coefficient with A is 0.8, 0.7, 0.6 and 0.5, respectively.

For each of the 5 populations, $B = 10000$ samples were selected to compare the various estimators in terms of relative bias (RB) and relative efficiency (RE).

Two survey designs were performed: simple random sampling without replacement and unequal probability sampling with Midzuno's method of sizes n =50, 75, 100 and 150. Samples in Midzuno sampling are selected with inclusion probabilities proportional to variable $z_k = N(300, 200)$, for $k = 1, \ldots, N$ unrelated with the variable of interest.

Tables 1 and 2 give the values of RB, RE and GE in percentages for the binomial populations under the two survey designs considered.

The results derived from this simulation study gave values for RB within a reasonable range. All the estimators considered produced absolute relative bias values of less than 0.5%.

Incorporating no auxiliary information, HT estimators usually have a larger MSE than calibration, probit and logit regression estimators. With large Cramer's V coefficient (ϕ) values, all the estimators that use auxiliary information produce good results. It can also be seen that as ϕ increases, all the estimators achieve greater precision, which is particularly marked for very high correlations.

Of all the estimates that use auxiliary information, the calibration estimator has the lowest degree of efficiency. Although it performs better than the HT

estimator on most occasions (expect when ϕ=0.5), the others produce a smaller MSE, and then a larger gain in efficiency.

The regression (REG), calibration probit (CP), model-assisted probit regression (MAP), predictive probit (PP) and logistic regression (LGREG) estimators perform very well in all cases. For high correlations, the efficiency of the estimators is similar. All the proposed probit estimators behave well, but two stand out: the calibration probit and the model-assisted probit regression estimators have similar levels of efficiency and achieve the best results for all sample sizes, all populations and all survey designs.

The sample size produces a clear effect on the behaviour of the estimators: as the sample size increases, so does the efficiency of the estimators.

The second population used in this simulation study was a real database of size $N = 10850$ referring to dengue cases confirmed by the Rio State Public Health laboratory in Guerrero State, Mexico (2006). The data reflect the records of patients diagnosed with dengue and the characteristics associated with symptoms of a typical case of dengue.

The main variable is the type of dengue (classical, Y=1, or haemorrhagic, Y=0). For the selection of the auxiliary variable, we took into consideration the relationship between the main variable and the different auxiliary variables, from a probit regression model. Thus, we selected the variables that best classified the patients (if the patient had headache, x_1, abdominal pain x_2 or diarrhoea x_3). We use the variable body temperature to assign the probability in the Midzuno sampling.

In this population, the results are slightly different. The relative biases remain negligible (less than 0.5 %) but the efficiency is different:

- The calibration estimator performs poorly. Even when n =1000 the calibration estimator is worse than the HT estimator.
- The remaining estimates are more efficient than the HT estimator but the gain in efficiency is not as great as in the previous example.
- The proposed probit estimators often work better than the other estimators. The calibration probit and the model-assisted probit regression estimators have similar levels of efficiency, but the predictive estimator is the most efficient for all sample sizes and for two designs.

In summary, we conclude that the question of the associations between the variables is the most important factor influencing the behaviour of the proposed probit estimators. Even for moderate correlation values, the proposed estimators are more efficient than the HT estimator. Of the three estimators proposed, the calibration probit and the model-assisted probit regression present the same levels of efficiency, and none are outstanding in this respect.

5. Application in an immigration survey

In this section we apply the proposed estimators to data corresponding to a survey on perceptions of immigration in a certain region in Spain. A sample

of size $n = 1919$ was selected from a population with size $N = 4982920$, using stratified random sampling.

Among the topics of interest in the survey is the question of estimating the percentage of citizens who believe that the authorities should make immigration more difficult by imposing stricter conditions. The auxiliary variables available are the respondent's sex and age (in four categories). Both variables were observed in the sample and their totals are known for each province (stratum).

We compared estimates of the mean of this binary variable of interest without using any auxiliary information and also using the auxiliary information provided by age and sex (see Table 4). The confidence intervals (and their length) based on Jackknife variance estimation were also determined. It should be noted that the proportion to be estimated is very small. The probit point estimates are at least half a percentage point below the estimates provided by the HT estimator (and those of the CAL estimator). Half a percentage point in this population corresponds to about 25000 people. Moreover, the length of the confidence intervals is around 2 percentage points. Reducing this length by a quarter percentage point, as is achieved with the PP estimator, may be of considerable significance.

Table 4 shows that the inclusion of auxiliary information provides estimates with shorter confidence intervals except in the case of the usual calibration estimator (as was to be expected, because the calibration technique assumes that the response variable depends linearly on the auxiliary variables). This is particularly true when probit regression is used. By including all available auxiliary information by means of a probit model, we obtained the best empirical performance and an estimate that is coherent with that provided by the other models.

6. Conclusions

The estimation of proportions is an important subject with many practical applications. The Horvitz-Thompson estimators for the mean or total can be improved by using the general regression estimator, because this estimator can incorporate auxiliary information. In estimating a proportion, we might also wish to incorporate auxiliary information and it is more natural to motivate the use of logistic and probit models for a discrete variable.

In this paper, we show how this could be done, using various probit regression estimators, based on a probit model. We propose three estimators: $\widehat{P}_{PP}$, $\widehat{P}_{MAP}$ and $\widehat{P}_{CP}$. The first is a model-based estimator and the other two are design based.

In the simulation section, we use several populations to show that smaller variances might be obtained with the probit regression estimator than with the HT estimator or with other classical regression estimators. In our empirical study, the probit estimators gave accurate estimations under various sampling plans and are competitive with the logistic estimator. In conclusion, when estimating a proportion, the use of auxiliary information may provide large

gains in efficiency, while the choice of an appropriate model may enable smaller variances to be achieved. We also conclude that both approaches to using probit models seem plausible, and that both model-based and model-assisted philosophies of statistical analysis can be adopted, according to the practical conditions in question.

References

[1] Allison, P.D. 1999. Comparing Logit and Probit Coefficients Across Groups. Sociological Methods & Research 28:186-208.

[2] Álvarez, E., Arcos, A., González, S., Muñoz, J.F., Rueda, M. Estimating population proportions in the presence of missing data. Journal of Computational and Applied Mathematics 237, 470-476 (2013).

[3] Arnab, R. , Singh, S.: A note on variance estimation for the generalized regression predictor. Australian and New Zealand Journal of Statistics. 47(2), 231-234 (2005).

[4] Arnab, R., Shangodoyin, D.K., Singh, S.: Variance estimation of a generalized regression predictor. Journal of the Indian Society of Agricultural Statistics. 64(2), 273-288 (2010).

[5] Ashford, J., Sowden, R.: Multi-variate probit analysis. Biometrics. 26, 535-546 (1970).

[6] Breen, R., Bernt, K. and Holms, A.: Total, Direct, and Indirect Effects in Logit and Probit Models. Sociological Methods & Research 42(2) 164-191 (2013).

[7] Binder, D. A.: On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review. 51, 279-292 (1983).

[8] Breidt, F.J., Opsomer, J.D.: Local polynomial regression estimators in survey sampling. The Annals of Statistics. 28, 1026-1053 (2000).

[9] Caffo B., Griswold, M.: A user-friendly introduction to link.probit-normal models. The American Statistician. 60(2), 139-145 (2006).

[10] Carnes, N.: Probit.survey: Survey-Weighted Probit Regression for Dichotomous Dependent Variables. In Imai, K., King, G., Lau, O. (eds.) Zelig: Everyoneâs Statistical Software. (2008)

[11] Cassel,C.M.(2005) Measuring costumer satisfaction on a national level using a superpopulation model. Total quality management 11(7) 909–915 (2000).

[12] Cassel, C.M., Särndal, C.E., Wretman, J.H.: Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. Biometrika. 63, 615-620 (1976).

[13] Cassel, C.M., Särndal, C.E., Wretman, J.H.: Foundations of inference in survey sampling. Wiley, New York (1977).

[14] Cramer, J. S. 2003. Logit Models. From Economics and Other Fields. Cambridge, England: Cambridge University Press.

[15] Deville, J.C., Särndal, C.E.: Calibration Estimators in Survey Sampling. Journal of the American Statistical Association. 87, 376-382 (1992).

[16] Duchesne, P.: Estimation of a proportion with survey data. Journal of Statistics Education [Online],11(3) (www.amstat.org/publications/jse/v11n3/duchesne.pdf) (2003).

[17] Gelman, A.: Struggles with Survey Weighting and Regression Modeling. Statistical Science 22(2), 153164 (2007).

[18] Godambe, V. P., Thompson, M. E.: Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review. 54, 127–138 (1986).

[19] Harms, T., Duchesne, P.: On calibration estimation for quantiles. Survey Methodology. 32, 37-52 (2006).

[20] Hausman, J., Wise, D.: A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogenous preferences. Econometrica. 46(2), 403-426 (1978)

[21] Krapavickaité, D.: Estimation of a Finite Population Total for a Censored Regression Model for a Study Variable. Acta Appl Math. 96, 339â347 (2007).

[22] Lehtonen, R., Veijanen, A.: Logistic generalized regression estimators. Survey Methodology. 24, 51-55 (1998).

[23] Little, R.J.: To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association. 99(446), 546-556 (2004).

[24] Lohr, S.L.: Comment: Struggles with Survey Weighting and Regression Modeling. Statistical Science 22(2), 175-178 2007.

[25] Martin, E.A., Traugott, M.W. and Kennedy, C.: A Review and Proposal for a New Measure of Poll Accuracy. Public Opinion Quarterly, 69, 342-369 (2005).

[26] McFadden, D.: A method of simulated moments for estimation of discrete repsonse models without numerical integration. Econometrica. 57(2), 995-1026 (1989).

[27] Molina, E.A., Skinner, C.J.: Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. Computational Statistics & Data Analysis. 13, 395–405 (1992).

[28] Morgan, S.L.: Models of College Entry and the Challenges of Estimating Primary and Secondary Effects. Sociological Methods and Research 41: 17-56 (2012).

[29] Muñoz, J. F., Arcos, A.,Álvarez-Verdejo, E., Rueda, M., Martínez-Puertas, S.: Estimators and confidence intervals for the proportion using auxiliary information with applications to the estimation of prevalences. Journal Of Biopharmaceutical Statistics 1, 1–32 (2011).

[30] Pavía, J.M., Larranz, B.: Nonresponse Bias and Superpopulation Models in Electoral Polls. Reis 137, 237–264 (2012).

[31] Pavía, J.M.: Improving predictive accuracy of exit polls International Journal of Forecasting 26, 68-81 (2010).

[32] Pavía, J.M., Larraz,B., Montero, J.M.: Election Forecasts Using Spatio-Temporal Models, Journal of the American Statistical Association, 103, 1050-1059 (2008).

[33] Pavía-Miralles, J.M.: Forecasts from Non-Random Samples: The Election Night Case. Journal of the American Statistical Association 100: 1113-1122 (2005).

[34] Plikusas, A., Pumputis, D.: Calibrated Estimators of the Population Covariance. Acta Appl Math. 97, 177–187 (2007).

[35] Rueda, M., Martínez, S., Martínez, H., Arcos, A.: Estimation of the distribution function with calibration methods. Journal of Statistical Planning and Inference. 137, 435-448 (2007).

[36] Rueda, M., Martínez, S., Martínez, H., Arcos, A.: Calibration methods for estimating quantiles. Metrika. 66, 355-371 (2007).

[37] Rueda,M., Muñoz, J.F., Arcos, A., Álvarez, E.: Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies. Journal of Biopharmaceutical Statistics. 21(3), 526â554 (2011).

[38] Särndal, C.E.: The calibration approach in survey theory and practice. Survey Methodology. 33, 99-119, (2007).

[39] Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (1992).

[40] Singh, S.: Generalized calibration approach for estimating variance in survey sampling. Annals of the Institute of Statistical Mathematics. 53, 404-417 (2001).

[41] Singh, S.: Advanced sampling theory with applications: How Michael "selected" Amy. Kluwer Academic Publisher, The Netherlands (2003).

[42] Singh, S., Horn, S., Chowdhury, S., Yu, F.: Calibration of the estimator of variance. New Zealand J Stat. 40, 199-212 (1999).

[43] Singh, H.P, Singh, S., Kozak, M.: A Family of Estimators of Finite-Population Distribution Function Using Auxiliary Information. Acta Appl Math. 104, 115-130 (2008).

[44] Winship, C. and Mare, R. D.: Regression Models with Ordinal Variables. American Sociological Review 49 (4), 512-525 (1984).

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

[46] Wu, C.: The Effective Use of Complete Auxiliary Information From Survey Data. Unpublished doctoral dissertation. Simon Fraser University, Canada. (1999)

Table 1: RB %, RE % and GE % respect to Horvitz-Thompson estimator, for several sample sizes of the estimators compared: Horvitz-Thompson (HT), calibration (CAL), regression (REG) , calibration probit (CP), model assisted probit regression (MAP), predictive probit (PP) and logistic regression (LGREG). SRSWOR from the BINOMIAL populations.

		RB	RE	GE	RB	RE	GE	RB	RE	GE
	ϕ	$n = 50$			$n = 100$			$n = 150$		
HT	0.5	0.00	100.00	0.00	0.06	100.00	0.00	0.10	100.00	0.00
CAL		0.05	92.56	-3.44	0.11	93.57	-2.43	0.21	93.15	-6.85
REG		-0.02	103.08	3.08	0.05	106.74	6.74	0.13	108.21	8.21
CP		-0.02	103.13	3.13	0.05	106.80	6.80	0.13	108.21	8.21
MAP		-0.02	103.09	3.09	0.05	106.79	6.79	0.13	108.21	8.21
PP		0.23	103.04	3.04	0.17	106.77	6.77	0.20	108.12	8.12
LGREG		-0.02	103.07	3.07	0.05	106.74	6.74	0.13	108.20	8.20
HT	0.6	-0.25	100.00	0.00	-0.18	100.00	0.00	-0.16	100.00	0.00
CAL		-0.19	101.38	1.38	-0.19	103.00	3.00	-0.29	103.74	3.74
REG		-0.25	121.27	21.27	-0.21	128.94	28.94	-0.14	129.45	29.45
CP		-0.26	121.65	21.65	-0.21	129.08	29.08	-0.14	129.53	29.53
MAP		-0.26	121.62	21.62	-0.21	129.05	29.05	-0.14	129.55	29.55
PP		-0.04	121.39	21.39	-0.14	128.75	8.75	-0.11	129.38	29.38
LGREG		-0.25	121.27	21.27	-0.20	128.92	28.92	-0.14	129.46	29.46
HT	0.7	0.35	100.00	0.00	0.15	100.00	0.00	0.24	100.00	0.00
CAL		0.37	120.03	20.03	0.20	119.03	19.03	0.20	122.61	22.61
REG		0.28	146.92	46.92	0.11	153.50	53.50	0.21	159.78	59.78
CP		0.27	147.81	47.81	0.10	153.82	53.82	0.21	159.85	59.85
MAP		0.27	147.69	47.69	0.10	153.86	53.86	0.21	159.79	59.79
PP		0.35	145.76	45.76	0.11	151.68	51.68	0.20	157.90	57.90
LGREG		0.29	146.88	46.88	0.11	153.54	53.54	0.21	159.74	59.74
HT	0.8	0.10	100.00	0.00	-0.07	100.00	0.00	0.06	100.00	0.00
CAL		-0.13	154.26	54.26	-0.17	153.93	53.93	-0.10	160.89	60.89
REG		0.12	187.30	87.30	0.05	199.84	99.84	0.02	212.29	112.29
CP		0.08	189.01	89.01	0.02	201.38	101.38	0.00	212.65	112.65
MAP		0.09	188.84	88.84	0.02	201.40	101.40	0.01	212.75	112.75
PP		0.29	184.23	84.23	0.19	196.53	96.53	0.17	207.74	107.74
LGREG		0.13	187.30	87.30	0.05	199.91	99.91	0.03	212.41	112.41
HT	0.9	0.17	100.00	0.00	0.18	100.00	0.00	0.17	100.00	0.00
CAL		-0.02	275.73	175.73	0.04	275.57	175.57	-0.04	276.90	176.90
REG		-0.39	308.61	208.61	-0.23	340.37	240.37	-0.22	379.22	279.22
CP		-0.37	309.40	209.40	-0.19	344.47	244.47	-0.18	381.35	281.35
MAP		-0.37	309.44	209.44	-0.19	344.26	244.26	-0.18	380.92	280.92
PP		-0.45	300.47	200.47	-0.32	328.52	228.52	-0.34	360.51	260.51
LGREG		-0.39	308.84	208.84	-0.23	340.44	240.44	-0.22	379.04	279.04

Table 2: RB %, RE % and GE respect to Horvitz-Thompson estimator, for several sample sizes of the estimators compared: Horvitz-Thompson (HT), calibration (CAL), regression (REG) , calibration probit (CP), model assisted probit regression (MAP), predictive probit (PP) and logistic regression (LGREG). Midzuno sampling schemes from BINOMIAL populations.

		RB	RE	GE	RB	RE	GE	RB	RE	GE
	ϕ	$n = 50$			$n = 100$			$n = 150$		
HT	0.5	0.06	100.00	0.00	0.11	0.00	100.00	-0.17	100.00	0.00
CAL		0.02	92.83	-7.17	0.04	94.58	-5.42	-0.12	94.80	-5.20
REG		-0.01	103.65	3.65	0.02	108.55	8.55	-0.13	110.16	10.16
CP		-0.01	103.74	3.74	0.03	108.61	8.61	-0.13	110.17	10.17
MAP		-0.01	103.67	3.67	0.03	108.62	8.62	-0.13	110.16	10.16
PP		-0.01	103.43	3.43	0.02	108.51	8.51	-0.13	110.03	10.03
LGREG		-0.02	103.63	3.63	0.02	108.55	8.55	-0.13	110.15	10.15
HT	0.6	0.02	100.00	0.00	0.07	100.00	0.00	0.10	100.00	0.00
CAL		-0.03	101.69	1.69	0.03	103.52	3.52	0.14	104.33	4.33
REG		-0.07	123.66	23.66	0.02	130.09	30.09	0.14	135.12	35.12
CP		-0.08	124.14	24.14	0.01	130.28	30.28	0.14	135.16	35.16
MAP		-0.08	124.07	24.07	0.01	130.30	30.30	0.14	135.18	35.18
PP		-0.14	123.05	23.05	-0.03	129.61	29.61	0.09	134.42	34.42
LGREG		-0.07	123.64	23.64	0.02	130.11	30.11	0.14	135.14	35.14
HT	0.7	-0.24	100.00	0.00	-0.10	100.00	0.00	0.08	100.00	0.00
CAL		-0.08	120.72	20.72	-0.14	118.69	18.69	0.05	119.94	19.94
REG		-0.24	144.03	44.03	-0.14	154.83	54.83	0.04	158.67	58.67
CP		-0.25	145.30	45.30	-0.14	155.15	55.15	0.04	158.60	58.60
MAP		-0.25	145.32	45.32	-0.14	155.15	55.15	0.04	158.63	58.63
PP		-0.31	143.09	43.09	-0.21	152.47	52.47	-0.03	155.61	55.61
LGREG		-0.25	144.13	44.13	-0.14	154.85	54.85	0.04	158.71	58.71
HT	0.8	0.00	100.00	0.00	0.03	100.00	0.00	-0.16	100.00	0.00
CAL		-0.02	159.35	59.35	0.02	161.20	61.20	-0.14	160.47	60.47
REG		0.23	185.89	85.89	0.13	213.98	113.98	-0.02	218.74	118.74
CP		0.17	187.62	87.62	0.08	215.55	115.55	-0.05	218.88	118.88
MAP		0.18	187.68	87.68	0.09	215.60	115.60	-0.05	218.89	118.89
PP		0.21	182.68	82.68	0.13	208.93	108.93	-0.01	211.42	111.42
LGREG		0.23	186.09	86.09	0.13	214.07	114.07	-0.02	218.80	118.80
HT	0.9	0.11	100.00	0.00	-0.01	100.00	0.00	0.13	100.00	0.00
CAL		0.12	282.33	182.33	-0.04	279.27	179.27	-0.03	280.96	180.96
REG		-0.31	298.34	198.34	-0.35	349.31	249.31	-0.22	386.95	286.95
CP		-0.30	300.45	200.45	-0.29	353.25	253.25	-0.18	389.10	289.10
MAP		-0.30	300.79	200.79	-0.30	353.42	253.42	-0.18	389.02	289.02
PP		-0.38	291.62	191.62	-0.44	331.28	231.28	-0.32	362.28	262.28
LGREG		-0.31	298.80	198.80	-0.35	349.71	249.71	-0.22	387.14	287.14

Table 3: RB %, RE % and GE respect to Horvitz-Thompson estimator, for several sample sizes of the estimators compared: Horvitz-Thompson (HT), calibration (CAL), regression (REG) , calibration probit (CP), model assisted probit regression (MAP), predictive probit (PP) and logistic regression (LGREG). SRSWOR and Midzuno sampling schemes from DENGUE population.

	Simple random sampling								
	RB	RE	GE	RB	RE	GE	RB	RE	GE
	n=500			n=600			n=700		
HT	0.11	100.00	0.00	0.11	100.00	0.00	0.06	100.00	0.00
CAL	0.10	99.19	-0.81	0.11	99.57	-0.43	0.07	99.20	-0.80
REG	0.11	105.91	5.91	0.04	105.72	5.72	0.06	104.97	4.97
CP	0.11	105.98	5.98	0.04	105.82	5.82	0.06	105.02	5.02
MAP	0.11	105.95	5.95	0.04	105.79	5.79	0.06	104.94	4.94
PP	0.09	106.19	6.19	0.02	106.02	6.02	0.04	105.15	5.15
LGREG	0.12	105.86	5.86	0.05	105.69	5.69	0.07	104.89	4.89
	n=800			n=900			n=1000		
HT	0.10	100.00	0.00	-0.13	100.00	0.00	-0.01	100.00	0.00
CAL	0.10	99.44	-0.56	-0.14	99.49	-0.51	0.00	99.49	-0.51
REG	0.10	106.52	6.52	-0.14	106.28	6.28	-0.05	105.41	5.41
CP	0.10	106.60	6.60	-0.14	106.34	6.34	-0.05	105.48	5.48
MAP	0.11	106.61	6.61	-0.14	106.34	6.34	-0.05	105.41	5.41
PP	0.08	106.82	6.82	-0.16	106.54	6.54	-0.07	105.62	5.62
LGREG	0.11	106.52	6.52	-0.14	106.27	6.27	-0.05	105.34	5.34
	Midzuno								
	n=500			n=600			n=700		
HT	0.31	100.00	0.00	-0.03	100.00	0.00	-0.05	100.00	0.00
CAL	0.31	99.27	-0.73	-0.05	99.18	-0.82	-0.06	99.72	-0.28
REG	0.27	106.44	6.44	-0.08	105.00	5.00	-0.04	106.45	6.45
CP	0.26	106.56	6.56	-0.08	105.04	5.04	-0.04	106.55	6.55
MAP	0.26	106.58	6.58	-0.08	105.02	5.02	-0.03	106.49	6.49
PP	0.12	107.03	7.03	-0.20	105.43	5.43	-0.14	106.85	6.85
LGREG	0.27	106.44	6.44	-0.07	104.98	4.98	-0.03	106.39	6.39
	n=800			n=900			n=1000		
HT	-0.27	100.00	0.00	-0.03	100.00	0.00	0.02	100.00	0.00
CAL	-0.27	99.48	-0.52	-0.03	99.63	-0.37	0.03	99.63	-0.37
REG	-0.28	106.05	6.05	-0.07	106.05	6.05	0.05	105.28	5.28
CP	-0.28	106.14	6.14	-0.08	106.11	6.11	0.04	105.37	5.37
MAP	-0.28	106.09	6.09	-0.08	106.06	6.06	0.05	105.34	5.34
PP	-0.37	106.40	6.40	-0.16	106.35	6.35	-0.03	105.64	5.64
LGREG	-0.28	106.01	6.01	-0.07	106.00	6.00	0.05	105.24	5.24

Table 4: Estimated proportion ($\hat{P}$), lower bound (LB), upper bound (UB) and length (LEN) of a 95% confidence interval for alternative estimators in the Inmigration Survey.

	$\hat{P}$	LB	UB	LEN
HT	5.088	4.086	6.090	2.004
CAL	5.175	4.147	6.204	2.057
REG	4.353	3.414	5.293	1.878
CP	4.353	3.448	5.257	1.809
MAP	4.364	3.424	5.303	1.878
PP	4.539	3.668	5.409	1.742
LGREG	4.363	3.365	5.362	1.997