Revista Colombiana de Estadística
July 2016, Volume 39, Issue 2, pp. 281 to 305
DOI: http://dx.doi.org/10.15446/rce.v39n2.55424

The Performance of Multivariate Calibration on Ratios, Means and Proportions

Sobre la calibración multivariada sobre razones, medias y proporciones

Hugo Andrés Gutiérrez Rojas[1,a], Hanwen Zhang[1,b],
Nelson Andrés Rodríguez[2,c]

[1]Facultad de Estadística, División de Ciencias Económicas y Administrativas, Universidad Santo Tomás, Bogotá, Colombia

[2]Instituto Colombiano para la Evaluación de la Educación (ICFES), Bogotá, Colombia

Abstract

In this paper, the calibration approach is revisited in order to allow new calibration weights that are subject to the restriction of multiple calibration equations on a vector of ratios, means and proportions. The classical approach is extended in such a way that the calibration equations are not based on a vector of totals, but on a vector of other nonlinear parameters. We stated some properties of the resulting estimators and carry out some empirical simulations in order to asses the performance of this approach. We found that this methodology is suitable for some practical situations like vote intention estimation, estimation of labor force, and retrospective studies. The methodology is applied in the context of the Presidential elections held in Colombia in 2014 for which we estimated the vote intention in the second round using information from an election poll, taking the results from the first round as auxiliary information.

***Key words*:** Calibration, Survey sampling, Ratio estimation, Nonlinear estimation, Monte Carlo simulation.

Resumen

En este artículo se aborda la metodología de calibración que reproduce pesos nuevos sujeto la restricción de las ecuaciones de calibración múltiple sobre un vector de razones, medias o proporciones. Se extiende la calibración clásica de tal forma que las ecuaciones de calibración no estén basados solo

[a]Professor. E-mail: hugogutierrez@usantotomas.edu.co
[b]Professor. E-mail: hanwenzhang@usantotomas.edu.co
[c]Statistic. E-mail: nrodriguez@icfes.gov.co

un vector de totales, sino un vector de parámetros no lineales. Se dan algunas propiedades de los estimadores resultantes y se llevan a cabo algunas simulaciones empíricas para verificar el desempeño de este enfoque. Encontramos que este es apropiado para algunas situaciones prácticas tales como la estimación de la intención de voto, estimación de fuerza laboral y estudios retrospectivos. La metodología es aplicada en el contexto de las elecciones presidenciales de Colombia en el 2014, donde estimamos la intención de voto en la segunda vuelta utilizando datos provenientes de una encuesta electoral tomando los resultados de la primera vuelta como información auxiliar.

***Palabras clave*:** calibración, encuestas por muestreo, estimación de razón, estimadores no lineales, simulación Monte Carlo.

1. Introduction

Consider a finite population U as a set of N units labeled as $\{1, \dots, N\}$. The size of the population U is not necessarily known. There are a vector of variables $\mathbf{x}_k = (x_{1k}, \dots, x_{Qk})'$ and a variable y_k associated with every unit k in the population. Likewise, assume that a random sample s of size n is drawn from U according to a (usually complex) sampling design $p(s)$. Let $\pi_k = Pr(k \in s)$ be the first-order inclusion probability, and $\pi_{kl} = Pr(k, l \in s)$ the second-order inclusion probability. If the purpose of the study is to unbiasedly estimate the total of y in the finite population given by $t_y = \sum_{k \in U} y_k$, then the Horvitz-Thompson estimator (HT) can be used, which is defined as:

$$\hat{t}_{y,\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k \tag{1}$$

Where $d_k = 1/\pi_k$ is known as the sampling weight. The unbiased estimator of the variance of the HT estimator is obtained by the following expression:

$$\hat{V}(\hat{t}_{y,\pi}) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. It is well known that the use of auxiliary information is important in survey sampling theory, not only in the design stage, but also in the estimation step. One of the most plausible ways to incorporate auxiliary information is by using calibration estimators, where the calibration equations only involve totals of auxiliary variables. Deville & Särndal (1992) proposed a class of linear estimators of population totals in the following form:

$$\hat{t}_{y,cal} = \sum_{k \in s} w_k y_k, \tag{2}$$

where w_k $(k \in s)$ is the calibrated weight of unit k, induced by the use of auxiliary information in the form of a vector of population totals $\mathbf{t}_\mathbf{x} = (t_{x_1}, \dots, t_{x_Q})'$. The aim of calibration weights is to satisfy the following calibration equations:

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{x} \tag{3}$$

In the classical approach, the calibration weights are defined in order to minimise a pseudo-distance Φ_s from the design weights $d_k = \frac{1}{\pi_k}$ that are subject to the calibration equation (3). In this perspective, the calibration weights are given by:

$$w_k = d_k F_k(\mathbf{x}_k' \boldsymbol{\lambda}) \tag{4}$$

For example, by minimising the chi-squared distance, $\Phi_s = \sum_s c_k(w_k - d_k)^2/d_k$,[1] the calibration weights can be expressed as

$$w_k = d_k + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},\pi})' \left(\sum_{k \in s} d_k c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} c_k d_k \mathbf{x}_k. \tag{5}$$

Särndal (2007) reviews the calibration approach and the use of auxiliary information. Note that the estimated variance of the calibration estimator (2) under the chi-squared distance is

$$\hat{V}(\hat{t}_{y,cal}) = \sum_{k \in s} \sum_{l \in s} (\Delta_{kl}/\pi_{kl})(w_k e_k)(w_l e_l), \tag{6}$$

where $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}$ and

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} w_k c_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} w_k c_k \mathbf{x}_k y_k. \tag{7}$$

Other approaches to variance estimation of calibration estimators are provided by Kim & Park (2010). Note that many weight systems may satisfy (3). For example, Estevao, Särndal & Sautory (2000) found that by taking into account a set of instrumental variables $\mathbf{z}_k$, the calibration weights become:

$$w_k = d_k + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},\pi})' \left(\sum_{k \in s} c_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} c_k \mathbf{z}_k \tag{8}$$

Observe that $dim(\mathbf{z}_k)$ must be equal to $dim(\mathbf{x}_k)$. Kott (2004) and Estevao & Särndal (2004) indicated that the optimal instrumental vector is given when $\mathbf{z}_k = \pi_k \sum_{l \in s} (\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}}) \mathbf{x}_l$. A deeper view of instrumental calibration may be found in Kott (2003), Kim & Park (2010), and Park & Kim (2014). Estevao et al. (2000) and Estevao & Särndal (2006) also considered the following set of calibration weights:

$$w_k = \mathbf{t}_{\mathbf{x}}' \left(\sum_{k \in s} c_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} c_k \mathbf{z}_k \tag{9}$$

[1] Where c_k is a constant unrelated to the sampling weights. As Deville & Särndal (1992) state, the choice $c_k = 1$ dominates many applications. However, other choices induce changes in the functional form of the calibration estimator.

In particular, when $\mathbf{z}_k = \mathbf{x}_k$, the calibration weights are given by:

$$w_k = \mathbf{t}'_{\mathbf{x}} \left(\sum_{k \in s} c_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} c_k \mathbf{x}_k \tag{10}$$

As we can see, there are many choices when calibrating on known totals, but not all of them are efficient. However, it should be mentioned that calibrating on totals is not always suitable because current totals are not frequently available. Note that this situation is more dramatic in developing countries where censuses are not carried out regularly. The nature of totals is dynamic and it changes over time, most of the times they increase year after year. However, in an official statistics context, ratios, means and proportions are more stable over time. Following the findings of Krapavickaite & Plikusas (2005), Plikusas (2006) and Lesage (2011), when a population ratio R is accurately known, it is possible to compute new calibration weights that are subject to this new benchmark constraint:

$$\frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k x_k} = R \tag{11}$$

Note that we do not necessarily know t_y or $\mathbf{t}_{\mathbf{x}}$. Moreover, with this approach we can simultaneously estimate the totals that define the ratios while maintaining their structural relation. Nevertheless, from a methodological perspective, we can extend the restriction to a vector of ratios, means and proportions. This paper deals with this issue and it can be considered a fallow an to Lesage's (2011) suggestion, who claimed that it would be interesting to determine the practical cases in which the use of complex parameters in the calibration improves the precision of the parameters of interest. He also examined calibration in terms of ratio, median and variance of auxiliary variables. Kim, Sungur & Heo (2007) proposed a calibration approach to estimate the population mean in stratified sampling by defining the calibration equation in terms of the population mean of one auxiliary variable.

In this article, we extend the ratio calibration approach to the multivariate case, that is, we propose a calibration methodology based on more than one ratio. We also present the variance estimation based on Taylor's linearization. This paper split up into the following section: after a brief introduction, Section 2 describes the estimation of total by using the calibration approach with a vector of known ratios along with some interesting properties and some specific scenarios. Section 3 reports a Monte Carlo simulation, the results of which show that, in some scenarios, the approach could be more efficient than calibrating known totals. In Section 4, the proposed methodology is applied in an electoral surveys context to estimate voting intention in a possible second round. Since the voting in the second round is influenced by the first round, we had the possibility to calibrate the sample weights using known first round ratios, which substantially improved the estimation of the second round voting intention. Section 5 concludes with a brief discussion on the use of this approach.

2. Multivariate Calibration Over Ratios

Let consider a different approach to calibration. Assume that kth population element is associated with vectors $\mathbf{x}_k$ and $\mathbf{y}_k = (y_{1k}, \dots, y_{Qk})'$. For the elements $k \in s$, we observe both $\mathbf{y}_k$ and $\mathbf{x}_k$. The population ratios $R_q = \frac{t_{y_q}}{t_{x_q}}$ $(q = 1, \dots, Q)$ are assumed to be known (even when t_{y_q} and t_{x_q} remain unknown). As such, the goal is to estimate all of the population totals $t_{y_q} = \sum_{k\in U} y_{qk}$ and $t_{x_q} = \sum_{k\in U} x_{qk}$, through the estimators $\hat{t}_{y_q} = \sum_{k\in s} w_k y_{qk}$ and $\hat{t}_{x_q} = \sum_{k\in s} w_k x_{qk}$, where new weights w_k satisfy the following constraints:

$$\frac{\sum_{k\in s} w_k y_{qk}}{\sum_{k\in s} w_k x_{qk}} = R_q, \quad \text{with } q = 1, \dots, Q$$

Equivalently, the calibration equations are defined by:

$$\hat{\mathbf{R}}_{cal} = \mathbf{R} \tag{12}$$

Where $\mathbf{R} = (R_1, \dots, R_Q)'$ and

$$\hat{\mathbf{R}}_{cal} = (\hat{R}_1, \dots, \hat{R}_Q)' = \left(\frac{\sum_{k\in s} w_k y_{1k}}{\sum_{k\in s} w_k x_{1k}}, \dots, \frac{\sum_{k\in s} w_k y_{Qk}}{\sum_{k\in s} w_k x_{Qk}} \right)'$$

The following result furthers the Lesage's (2011) idea:

Result 1. Assume that we have access to a vector of known ratios defined by:

$$\mathbf{R} = (R_1, \dots, R_Q)' \tag{13}$$

where,

$$R_q = \frac{t_{y_q}}{t_{x_q}} \quad \forall q = 1, \dots, Q. \tag{14}$$

Then, to calibrating over $\mathbf{R}$ is equivalent to calibrating over the following vector

$$\mathbf{t}_{\mathbf{z}} = (t_{z_1}, \dots, t_{z_Q})' \tag{15}$$

where,

$$t_{z_q} = \sum_{k\in U} z_{qk} = \sum_{k\in U} (y_{qk} - R_q x_{qk}) = 0 \quad \forall q = 1, \dots, Q. \tag{16}$$

That is, to calibrating over $\mathbf{R}$ is equivalent to calibrating over $\mathbf{t}_{\mathbf{z}} = \mathbf{0}$

Proof. First, notice that, from the calibration equations, we have:

$$\begin{aligned}
&\hat{\mathbf{t}}_{\mathbf{z},cal} = \mathbf{t}_{\mathbf{z}} \\
&\quad\Leftrightarrow \sum_{k\in s} w_k \mathbf{z}_k = \sum_{k\in U} \mathbf{z}_k \\
&\quad\Leftrightarrow \sum_{k\in s} w_k \left(y_{1k} - R_1 x_{1k}, \dots, y_{Qk} - R_Q x_{Qk} \right)' \\
&\qquad = \sum_{k\in U} \left(y_{1k} - R_1 x_{1k}, \dots, y_{Qk} - R_Q x_{Qk} \right)' \\
&\quad\Leftrightarrow \left(\hat{t}_{y_1,cal} - R_1 \hat{t}_{x_1,cal}, \dots, \hat{t}_{y_q,cal} - R_q \hat{t}_{x_q,cal} \right) = (0, \dots, 0)
\end{aligned}$$

In this way, for every $q = 1, \dots, Q$,

$$\begin{aligned}0 &= \hat{t}_{y_q,cal} - R_q\hat{t}_{x_q,cal}\\ \Leftrightarrow R_q &= \frac{\hat{t}_{y_q,cal}}{\hat{t}_{x_q,cal}}\\ \Leftrightarrow R_q &= \hat{R}_q = \frac{\hat{t}_{y_q,cal}}{\hat{t}_{x_q,cal}} \qquad (17)\end{aligned}$$

Then, the calibration equations $\hat{\mathbf{t}}_{\mathbf{z},cal} = \mathbf{0}$ and $\hat{\mathbf{R}}_{cal} = \mathbf{R}$ are equivalent. ■

Note that even though R_q is known, the totals t_y and t_{x_q} are not necessarily known. This condition ensures the flexibility of the approach because we can use this methodology not only to calibrate over ratios but to estimate other parameters of interest while maintaining the restriction on the calibrated weights on ratios.

Result 2. Suppose that a total of interest t_y is estimated by means of the approach given in result 2.1. As such, an asymptotically unbiased estimator of t_y is

$$\hat{t}_{y,calr} = \sum_{k\in s} w_k y_k, \qquad (18)$$

where the set of weights w_k satisfies the calibration restriction (12) over the auxiliary ratios.

Proposition 1. *For every $q = 1, \dots, Q$, the expectation of calibration estimators for totals performs according to the following expression:*

$$E(\hat{t}_{y_q,cal}) = R_q E(\hat{t}_{x_q,cal}). \qquad (19)$$

The variance of calibration estimators for totals is based on the following relation:

$$Var(\hat{t}_{y_q,cal}) = R_q^2 Var(\hat{t}_{x_q,cal}). \qquad (20)$$

The coefficient of variation of calibration estimators for totals derives from this relation:

$$CV(\hat{t}_{y_q,cal}) = CV(\hat{t}_{x_q,cal}), \qquad (21)$$

and the relative bias of calibration estimators for totals follows the expression:

$$RB(\hat{t}_{y_q,cal}) = R_q B(\hat{t}_{x_q,cal}) \qquad (22)$$

***Proof*.** The demonstration of equations (19) and (20) are straightforward. For the coefficient of variation, it should be noted that:

$$CV(\hat{t}_{y_q,cal}) = \frac{\sqrt{Var(\hat{t}_{y_q,cal})}}{\hat{t}_{y_q,cal}} = \frac{\sqrt{R_q^2 Var(\hat{t}_{x_q,cal})}}{R_q\hat{t}_{x_q,cal}} = \frac{\sqrt{Var(\hat{t}_{x_q,cal})}}{\hat{t}_{x_q,cal}} = CV(\hat{t}_{x_q,cal})$$

With respect to the relative bias

$$RB(\hat{t}_{y_q,cal}) = \frac{E(\hat{t}_{y_q,cal} - t_{y_q})}{t_{y_q}} = \frac{E(R_q\hat{t}_{x_q,cal} - R_q t_{x_q})}{R_q t_{x_q}} = R_q B(\hat{t}_{x_q,cal})$$

■

2.1. Some Particular Cases

When our parameters of interest are means or proportions, they can be estimated as particular cases of the proposed methodology. In the case of means, the corresponding calibration equation is:

$$\begin{aligned}\bar{\mathbf{y}}_{cal} &= (\bar{y}_{1,cal}, \ldots, \bar{y}_{Q,cal})' \\ &= \left(\frac{\hat{t}_{y_1,cal}}{\hat{N}}, \ldots, \frac{\hat{t}_{y_Q,cal}}{\hat{N}}\right)' \\ &= \left(\frac{\sum_{k\in s} w_k y_{1k}}{\sum_{k\in s} w_k}, \ldots, \frac{\sum_{k\in s} w_k y_{Qk}}{\sum_{k\in s} w_k}\right)' \\ &= (\bar{y}_1, \ldots, \bar{y}_Q)' = \bar{\mathbf{y}}\end{aligned}$$

That is

$$\bar{\mathbf{y}}_{cal} = \bar{\mathbf{y}} \Leftrightarrow \hat{\mathbf{t}}_{\mathbf{z},cal} = \mathbf{t}_{\mathbf{z}} \tag{23}$$

Where, for every $q = 1, \ldots, Q$,

$$z_{qk} = y_{qk} - \bar{y}_q \tag{24}$$

In the case of proportions, the corresponding calibration equation is:

$$\begin{aligned}\hat{\mathbf{P}}_{cal} &= (\hat{P}_{1,cal}, \ldots, \hat{P}_{Q,cal})' \\ &= \left(\frac{\hat{N}_{1,cal}}{\hat{N}}, \ldots, \frac{\hat{N}_{Q,cal}}{\hat{N}}\right)' \\ &= \left(\frac{\sum_{k\in s} w_k \delta_{1k}}{\sum_{k\in s} w_k}, \ldots, \frac{\sum_{k\in s} w_k \delta_{Qk}}{\sum_{k\in s} w_k}\right)' \\ &= (P_1, \ldots, P_Q)' = \mathbf{P}\end{aligned}$$

That is,

$$\hat{\mathbf{P}}_{cal} = \mathbf{P} \Leftrightarrow \hat{\mathbf{t}}_{\mathbf{z},cal} = \mathbf{t}_{\mathbf{z}} \tag{25}$$

Where, for every $q = 1, \ldots, Q$,

$$z_{qk} = \delta_{qk} - P_q \tag{26}$$

2.2. Another Perspective to Post-Stratification

Now, suppose that the population is partitioned into Q subgroups called post-strata. If ratios for those particular population subgroups are known, we should find weights satisfying the following constrains:

$$\hat{\mathbf{R}}_{cal} = (\hat{R}_{1,cal}, \ldots, \hat{R}_{Q,cal})' = (R_1, \ldots, R_Q)' = \mathbf{R} \tag{27}$$

Note that if population ratio is known, it is also possible to impose the following constraint on the calibration equations: $\hat{R}_{cal} = R_U$. Where, for every $q = 1, \ldots, Q$,

$$z_{qk} = \begin{cases} y_{qk} - R_q x_{qk} & \text{if } k \in s_h. \\ 0 & \text{Otherwise.} \end{cases} \tag{28}$$

Now, if means are known for the population subgroups, for example, in post-stratification, the proper constrains are as follows:

$$\bar{\mathbf{y}}_{cal} = (\bar{y}_{1,cal}, \ldots, \bar{y}_{Q,cal})' = (\bar{y}_1, \ldots, \bar{y}_Q)' = \bar{\mathbf{y}} \tag{29}$$

Consider that if population ratios are known, it is also possible to impose the following constraint to the calibration equations: $\bar{\mathbf{y}}_{cal} = \bar{\mathbf{y}}$. Where, for every $q = 1, \ldots, Q$,

$$z_{qk} = \begin{cases} y_{qk} - \bar{y}_q & \text{if } k \in s_h. \\ 0 & \text{Otherwise.} \end{cases} \tag{30}$$

Now, if proportions are known for population subgroups, for example, in post-stratification, the proper constrains are as follows:

$$\hat{\mathbf{P}}_{cal} = (\hat{P}_{1,cal}, \ldots, \hat{P}_{Q,cal})' = (P_1, \ldots, P_Q)' = \mathbf{P} \tag{31}$$

Note that if population proportions are known, it is also possible to impose the following constraint on the calibration equations: $\hat{P}_{cal} = \hat{P}_U$. Where, for every $q = 1, \ldots, Q$,

$$z_{qk} = \begin{cases} \delta_{qk} - P_q & \text{if } k \in s_h. \\ 0 & \text{Otherwise.} \end{cases} \tag{32}$$

2.3. Extending the Approach

Note that the calibration estimator proposed can be extended, in the sense that we can consider the situation in which we want to estimate the population totals by means of the calibration estimators $\hat{t}^*_{y_q} = \sum_{k \in s} w^*_k y_{qk}$ with $q = 1, \cdots, Q$ where the weights w^*_k satisfy the calibration equations

$$\frac{\sum_{k \in s} w^*_k \tilde{y}_{qk}}{\sum_{k \in s} w^*_k \tilde{x}_{qk}} = R^*_q, \quad \text{with } q = 1, \cdots, Q^*,$$

where $\tilde{y}_{qk}$ is any other variable, as well as $\tilde{x}_{qk}$. In a special case, these variables could represent the same characteristic of interest measured at a previous period. For example, this estimator can be useful when estimating unemployment rates for a particular period of time restricted to the calibration over the previous unemployment rate. Moreover, in runoff elections, we can calibrate using the results from the first round election in order to estimate the voting intention in the second round. As such is, Q^* may be different from Q.

Then, the aim is to find new weights w_k^* that satisfy the following calibration equations:

$$\hat{\mathbf{R}}^*_{cal} = (\hat{R}^*_{1,cal}, \cdots, \hat{R}^*_{Q^*,cal})' = (R_1^*, \cdots, R_{Q^*}^*)' = \mathbf{R}^*$$

Note that for every $q^* = 1, \cdots, Q^*$

$$z^*_{qk} = \tilde{y}_{qk} - R^*_q \tilde{x}_{qk}$$

3. Empirical Simulation

In this section some simulation experiments were carried out in order to compare the performance of the estimation of a total of interest t_y by using the calibration estimator on auxiliary ratios (CALR). This is given by (18), the classical calibration estimator on auxiliary totals (CAL), which is given by (2), and the Horvitz-Thompson (HT) estimator, which is given by (1).

A finite population of size $N = 10000$ was simulated from a superpopulation model ξ. It was supposed that the relationship between y_k and x_k can be described through a general model ξ, such as $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$. This model may adopt different forms throughout this section. The values of the vector of auxiliary information were generated from a uniform distribution and it was assumed that the ε_k values were independent and distributed as $N(0, \sigma^2)$, where σ^2 was suitably allocated to allow different values of the R-squared of the model.

In each run, random samples according to a simple random design without replacement (SI design) were drawn. We considered two sample sizes: $n = 400$ and $n = 2000$. The parameter vector $\boldsymbol{\beta}$ was estimated by (7) with $c_k = 1$. This process was repeated $M = 1000$ times. The simulation was written in the statistical software R 3.1.1. (R Development Core Team 2007). In the simulation, the performance of an estimator $\hat{t}_y$ was tracked by means of the Relative Bias (RB), Coefficient of Variation (CV) and the Relative Efficiency (RE). The RB was given by:

$$RB(\hat{t}_y) = M^{-1} \sum_{m=1}^{M} \frac{\hat{t}_{y,m} - t_y}{t_y}, \tag{33}$$

where $\hat{t}_{y,m}$ is computed in the mth simulated sample, $m = 1, \dots, M$. The CV was given by:

$$CV(\hat{t}_y) = M^{-1} \frac{\sqrt{MSE(\hat{t}_y)}}{t_y}, \tag{34}$$

where Mean Square Error (MSE) is defined as

$$MSE(\hat{t}_y) = M^{-1}\sum_{m=1}^{M}(\hat{t}_{y,m} - t_y)^2. \tag{35}$$

Finally, we computed the RE between the CALR and the CAL estimators as:

$$RE(\hat{t}_y) = \frac{MSE(\hat{t}_{y,cal})}{MSE(\hat{t}_{y,calr})} \tag{36}$$

As such, if the RE takes values higher than unity, it is concluded that the CALR estimator outperforms the CAL one. The fist two simulations dealt with super population models for the entire population. However, the remaining scenarios dealt with models involving super population groups or post-strata. In those cases, we simulated $H = 3$ with groups of the following sizes: $N_1 = 5000$, $N_2 = 2500$, and $N_3 = 2500$. In other words, population U was divided into three unequal groups U_1, U_2 and U_3.

3.1. Simple Regression Model

This first scenario deals with a single regression model:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k \tag{37}$$

We assumed that $\varepsilon_k \sim N(0, \sigma^2)$. The values of X_k were obtained from the distribution $U(10, 20)$, the values of the regression coefficients were set at $\beta_0 = 180$, $\beta_1 = -2$ and we chose convenient values for σ in order to get a predetermined R-squared.

For the CAL estimator, we assumed that the vector of auxiliary totals $\mathbf{t_x} = (N, t_x)'$ was known, and it was used in computing this estimator. Note that $t_x = \sum_{k\in U} x_k$ is the population total of the variable x. However, for the CALR estimator, it was assumed that the auxiliary ratio $R = \frac{t_y}{t_x}$ was known, and it was used when computing this estimator. Also, note that $t_y = \sum_{k\in U} y_k$ is the population total of the variable y. Tables 1 and 2 show the performance of the estimators that were considered.

Table 1: Performance of the sampling estimators for model (37) for a sample size of $n = 400$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	−5.49	0.08	18.94	18.21	7.24	6.33
0.2	−2.43	−0.64	9.28	8.07	4.65	3.00
0.4	−1.47	−0.65	6.57	4.89	3.13	2.45
0.6	−0.99	−0.53	5.37	3.27	2.19	2.25
0.8	−0.6	−0.37	4.62	2.01	1.37	2.14
0.95	−0.28	−0.18	4.18	0.92	0.64	2.08

Table 2: Performance of the sampling estimators for model (37) for a sample size of $n = 2000$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	2.06	0.88	7.80	7.64	3.10	6.06
0.2	0.91	0.62	3.74	3.38	2.02	2.82
0.4	0.55	0.41	2.61	2.05	1.35	2.32
0.6	0.37	0.28	2.12	1.37	0.93	2.16
0.8	0.23	0.17	1.83	0.84	0.58	2.09
0.95	0.10	0.08	1.67	0.39	0.27	2.05

In this scenario, we found that all of relative biases are negligible, and the lower coefficient of variation is that induced by the CALR estimator. Likewise, the relative efficiency of the CALR estimator is higher that the ones obtained in all other scenarios. When the R-squared increases, the efficiency of the CALR estimator decreases. One explication is that when the R-squared increases, the correlation between x and y also increases, and the variance of the CAL estimator gets smaller, decreasing faster than the variance of the CALR estimator. The simulations show a similar performance when the sample size increases.

3.2. Multiple Regression Model

The second scenario deals with the multivariate linear regression model:

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k, \tag{38}$$

with $\varepsilon_k \sim N(0, \sigma^2)$. The values of $X_{k,1}$, $X_{k,2}$ and $X_{k,3}$ were obtained from the distributions $U(10, 20)$, $U(100, 150)$ and $U(1, 1.8)$, respectively and the values of the regression coefficients were set at $\beta_0 = 400$, $\beta_1 = -2$, $\beta_2 = -0.8$ and $\beta_3 = 50$. We chose a convenient value for σ in order to obtain a predetermined R-squared.

For the CAL estimator, it was assumed that the vector of auxiliary totals $\mathbf{t_x} = (N, t_{x_1}, t_{x_2}, t_{x_3})'$ was known, and it was used in the computation of this estimator. Note that $t_{x_q} = \sum_{k \in U} x_{qk}$ is the population total of the variable x_q for $q = 1, 2, 3$. Moreover, for the CALR estimator it was assumed that the auxiliary vector of ratios $\mathbf{R} = \left(\frac{t_y}{t_{x_1}}, \frac{t_y}{t_{x_2}}, \frac{t_y}{t_{x_3}}\right)'$ was known, and it was used under computing this estimator. Tables 3 and 4 show the performance of the estimators that were considered.

Table 3: Performance of the sampling estimators for model (38) for a sample size of $n = 400$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	-3.09	0.25	10.81	10.49	3.76	7.79
0.2	-1.42	-0.12	5.45	4.82	2.87	2.83
0.4	-0.87	-0.27	3.86	2.95	2.10	1.96
0.6	-0.58	-0.28	3.16	1.98	1.53	1.66
0.8	-0.36	-0.23	2.73	1.21	0.99	1.50
0.95	-0.16	-0.13	2.48	0.55	0.47	1.42

Table 4: Performance of the sampling estimators for model (38) for a sample size of $n = 2000$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	1.16	0.53	4.48	4.38	1.57	7.81
0.2	0.53	0.44	2.23	2.02	1.22	2.71
0.4	0.33	0.32	1.57	1.23	0.90	1.86
0.6	0.22	0.22	1.27	0.83	0.65	1.59
0.8	0.13	0.13	1.10	0.51	0.42	1.45
0.95	0.06	0.06	1.00	0.23	0.20	1.39

In this scenario, we found that all of relative biases are negligible, and the lower coefficient of variation is that induced by the CALR estimator. In the same way, the relative efficiency of the CALR estimator is higher that those in all other scenarios. When the R-squared increases, the efficiency of the CALR estimator decreases. The simulations show a similar performance when the sample size increases.

3.3. Poststratified Calibration Over Ratios

This scenario deals with a poststratified ratio model, given by:

$$y_{kh} = \beta_h x_{kh} + \varepsilon_{kh} \qquad h = 1, 2, 3. \ - \ k = 1, \ldots, N_h. \tag{39}$$

where $\varepsilon_{kh} \sim N(0, \sigma_h^2)$ and x_{1k}, x_{2k} and x_{3k} are obtained independently from the distribution $U(1000, 2000)$. Besides, we defined $\beta_1 = 1$, $\beta_2 = 0.3$ and $\beta_3 = 0.5$. Also, $N_1 = 5000$, $N_2 = N_3 = 2500$. We chose a convenient value for σ_h^2 in order to obtain a predetermined R-squared.

For the CAL estimator, we assumed that the vector of auxiliary totals $\mathbf{t_x} = (t_x^1, t_x^2, t_x^3)'$ was known, and it was used when computing this estimator. Note that $t_x^h = \sum_{k \in U_h} x_k$ is the total of the variable x for the subpopulation U_h. Also, take into account that the population total of x over U is defined to be $t_x = \sum_{h=1}^3 t_x^h$. Moreover, for the CALR estimator, it was assumed that the auxiliary vector of means $\mathbf{R} = \left(\frac{t_y^1}{t_x^1}, \frac{t_y^2}{t_x^2}, \frac{t_y^3}{t_x^3}\right)'$ was known, and was used in the computation of this estimator. Here, $t_y^h = \sum_{k \in U_h} y_k$ is the total of the variable y for the subpopulation U_h, and the population total of y over U is $t_y = \sum_{h=1}^3 t_y^h$. Tables 5 and 6 show the performance of the estimators considered in this scenario.

In this scenario, we found that all of relative biases are negligible, the lower coefficient of variation is that induced by the CALR estimator. Also, the relative efficiency of the CALR estimator is higher than those achieved in other scenarios. When the R-squared decreases, the efficiency of the CALR estimator increases. The simulations show a similar performance when the sample size increases.

Table 5: Performance of the sampling estimators for model (39) for a sample size of $n = 400$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	5.67	10.32	99.13	98.44	23.59	17.42
0.2	2.08	10.36	49.17	45.07	23.67	3.63
0.4	0.94	10.34	35.35	28.17	23.69	1.41
0.6	0.30	10.29	29.01	18.67	23.71	0.62
0.8	−0.18	10.12	25.54	11.50	23.72	0.24
0.95	−0.60	9.02	23.90	5.35	23.71	0.05

Table 6: Performance of the sampling estimators for model (39) for a sample size of $n = 2000$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	−14.27	−0.75	42.54	41.62	9.61	18.74
0.2	−6.75	−0.74	21.49	19.38	9.65	4.04
0.4	−4.22	−0.75	15.18	11.89	9.66	1.52
0.6	−2.88	−0.76	12.39	7.94	9.66	0.68
0.8	−1.83	−0.79	10.74	4.87	9.67	0.25
0.95	−0.93	−0.98	9.88	2.24	9.67	0.05

3.4. Poststratified Calibration Over Means

This scenario deals with a poststratified mean model, given by:

$$y_{kh} = \beta_h + \varepsilon_{kh} \qquad h = 1, 2, 3. \; - \; k = 1, \dots, N_h. \tag{40}$$

with $\varepsilon_{kh} \sim N(0, \sigma^2)$ and $\beta_1 = 50$, $\beta_2 = 100$ and $\beta_3 = 150$. Also, $N_1 = 5000$, $N_2 = N_3 = 2500$. We chose a convenient value for σ in order to get a predetermined R-squared.

For the CAL estimator we assumed that the vector of auxiliary totals $\mathbf{t_x} = (N_1, N_2, N_3)'$ was known, and it was used when computing this estimator. Moreover, for the CALR estimator, it was assumed that the auxiliary vector of means $\bar{\mathbf{y}} = \left(\frac{t_y^1}{N_1}, \frac{t_y^2}{N_2}, \frac{t_y^3}{N_3}\right)'$ was known, and it was used when computing this estimator. Tables 7 and 8 show the performance of the estimators considered in this scenario.

Table 7: Performance of the sampling estimators for model (40) for a sample size of $n = 400$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	−2.45	0.97	8.80	8.60	2.01	18.31
0.2	−1.14	0.98	4.47	4.01	2.04	3.86
0.4	−0.70	0.98	3.18	2.46	2.05	1.45
0.6	−0.47	0.98	2.6	1.64	2.05	0.64
0.8	−0.29	0.99	2.26	1.01	2.06	0.24
0.95	−0.13	0.99	2.09	0.47	2.06	0.05

In this scenario, we observed that all of relative biases are negligible. We also found that the relative efficiency of the CALR estimator is higher than the one obtained when the R-square is lower than 0.4. When the R-squared decreases,

Table 8: Performance of the sampling estimators for model (40) for a sample size of $n = 2000$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	0.98	0.15	3.74	3.61	0.83	19.06
0.2	0.46	0.15	1.92	1.68	0.84	4.02
0.4	0.28	0.14	1.37	1.03	0.84	1.50
0.6	0.19	0.14	1.12	0.69	0.84	0.66
0.8	0.11	0.14	0.96	0.42	0.85	0.25
0.95	0.05	0.14	0.88	0.20	0.85	0.05

the efficiency of the CALR estimator increases. The simulations show a similar performance when the sample size increases.

3.5. Poststratified calibration over proportions

This scenario deals with a poststratified model, given by:

$$y_{kh} \sim Bernoulli(\beta_h) \qquad h = 1, 2, 3\text{and } k = 1, \ldots, N_h. \tag{41}$$

Note that in this scenario, the variable y is not continuos but discrete, taking only two values: one and zero. As such, $y_{kh} = 1$ if the element k has a certain characteristic of interest and $y_{kh} = 0$ otherwise. Besides, $N_1 = 5000$, $N_2 = 2500$ and $N_3 = 2500$. Values of β_h were chosen conveniently in order to obtain a suitable R-squared.

For the CAL estimator, we assumed that the vector of auxiliary totals $\mathbf{t_x} = (N_1, N_2, N_3)'$ was known, and it was used when computing this estimator. Moreover, for the CALR estimator, it was assumed that the auxiliary vector proportions $\mathbf{P} = \left(\frac{N^1}{N_1}, \frac{N^2}{N_2}, \frac{N^3}{N_3}\right)'$ was known, and it was used when computing this estimator. Note that $N^h = \sum_{U_h} y_{kh}$ and bear in mind that N_h is the size of the population subgroup U_h. As such, even though $\sum_{h=1}^{3} N_h = N$, $\sum_{h=1}^{3} N^h \neq N$. Tables 9 and 10 show the performance of the estimators that were considered in this scenario.

Table 9: Performance of the sampling estimators for model (41) for a sample size of $n = 400$: $10000 \times RB$ (relative bias), $1000 \times CV$ (coefficient of variation).

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	26.71	−43.72	106.29	105.06	34.72	9.16
0.2	11.15	−54.13	74.01	67.52	43.99	2.36
0.4	−15.16	−57.79	59.03	46.59	47.87	0.95
0.6	−2.27	−59.37	49.05	31.76	49.67	0.41
0.8	−0.47	172.24	43.49	19.92	145.93	0.02
0.95	−2.19	7284.13	46.82	10.47	443.52	0.0005

In this final scenario, we found that all of relative biases are negligible, and that the relative efficiency of the CALR estimator is higher that when the R-square is lower than 0.4 and when the sample size is large. When the R-squared decreases, the efficiency of the CALR estimator increases.

Table 10: Performance of the sampling estimators for model (41) for a sample size of $n = 2000$. The relative bias have been multiplied by 10000, the unit of CV is %.

R-squared	$RB(\hat{t}_{y,cal})$	$RB(\hat{t}_{y,calr})$	$CV(\hat{t}_{y,\pi})$	$CV(\hat{t}_{y,cal})$	$CV(\hat{t}_{y,calr})$	RE
0.05	−24.48	−1.61	43.53	42.50	10.01	18.02
0.2	−13.68	−2.89	30.42	27.17	13.84	3.86
0.4	6.73	−3.37	24.23	19.18	15.44	1.54
0.6	0.21	−3.39	20.18	13.42	16.18	0.69
0.8	−1.05	2.96	18.21	8.44	16.53	0.26
0.95	1.57	2541.63	19.27	4.42	423.12	0.0002

3.6. Calibration Over Any Set of Ratios

In this section we show the results of some empirical simulations when calibrating over a vector of known ratios $\mathbf{R}^*$. Generally speaking, the results obtained with a sample size of 400 are very similar to those of 1000; so we only show the results relating sample size of 400. Furthermore, the relative bias of the estimators are very small (negligible), so we just show the relative efficiency between the estimators. R_C^{HT} denotes the relative efficiency between the HT estimator and the CAL estimator, R_R^{HT} denotes the relative efficiency between the HT estimator and the CALR estimator and R_R^C denotes the relative efficiency between the CALR estimator and the CAL estimator.

3.6.1. Simple Regression Model

We first consider a simple regression model that relates the variable $\tilde{y}_k$ to $\tilde{x}_k$. As such, $\tilde{y}_k = \beta_0 + \beta_1 \tilde{x}_k + \varepsilon_k$ with $\varepsilon_k \sim N(0, \sigma^2)$: the values $\tilde{x}_k$ were simulated from the distribution $U(10, 20)$. The values of the regression coefficients were set to $\beta_0 =$ and $\beta_1 = -2$, and we chose convenient values for σ^2 in order to get a predetermined R-squared.

In order to create the variable of interest y_k, we assumed that $y_k = \gamma_0 + \gamma_1 \tilde{y}_k + \epsilon_k$, with $\epsilon_k \sim N(0, 10^2)$. We varied γ_1 to get different coefficients of correlation (ρ) between y_k and $\tilde{y}_k$. The results of this empirical study are shown in Table 11.

Table 11: Relative efficiency of the sampling estimators for the simple regression model considering a sample size of $n = 400$.

R^2	$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.6$			$\rho = 0.8$		
	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C
0.05	1.06	1.04	0.99	1.24	1.18	0.96	1.66	1.49	0.90	3.02	2.26	0.74
0.2	1.05	1.03	0.98	1.22	1.14	0.93	1.62	1.40	0.86	2.84	1.93	0.68
0.4	1.04	1.03	0.98	1.21	1.13	0.94	1.62	1.41	0.87	2.80	1.97	0.70
0.6	1.04	1.03	0.99	1.20	1.14	0.95	1.58	1.41	0.89	2.86	2.16	0.76
0.8	1.03	1.02	0.99	1.18	1.15	0.97	1.58	1.47	0.93	2.79	2.38	0.85
0.95	1.03	1.03	1.00	1.17	1.16	0.99	1.55	1.51	0.98	2.80	2.64	0.94

We can observe that the performance of the classic calibration CAL estimator improves as the correlation between y_k and $\tilde{y}_k$ increases, which is well known. With respect to the proposed CALR estimator, we can conclude that the performance of the CAL estimator is the same as the CALR estimator for lower correlation coefficients. We emphasise that the proposed CALR is useful when there are no population totals available for the variable $\tilde{y}$, which prevents the classical calibration estimator from being used. In these situations, if we know population ratios, we can use the CALR estimator. Note that the CALR estimator is always better than the Horvitz Thomson estimator and is as efficient as the classical calibration estimator when the coefficient of correlation is low.

3.6.2. Multiple Regression Model

We now consider a simple regression model that relates to the variable $\tilde{y}_k$ to x_k. As such, we consider the following model $\tilde{y}_k = \beta_0 + \beta_1\tilde{x}_{1k} + \beta_2\tilde{x}_{2k} + \beta_3\tilde{x}_{3k} + \varepsilon_k$ with $\varepsilon_k \sim N(0, \sigma^2)$. The values of $\tilde{x}_{1k}$, $\tilde{x}_{2k}$ and $\tilde{x}_{3k}$ were simulated from the distributions $U(10, 20)$, $U(100, 150)$ and $U(1, 1.8)$, respectively. The values of the regression coefficients were set to $\beta_0 = 400$, $\beta_1 = -2$, $\beta_2 = -0.8$ and $\beta_3 = 50$. Convenient values for σ^2 were proposed in order to get a predetermined R-squared.

In order to create the variable of interest y_k, we assumed that $y_k = \gamma_0 + \gamma_1\tilde{y}_k + \epsilon_k$ with $\epsilon_k \sim N(0, 10^2)$, $\gamma_0 = 100$ and we varied γ_1 to get a different coefficient of correlation between y_k and $\tilde{y}_k$. For the CAL estimator, the calibration is made over the population total $t_{\tilde{y}}$, while for the CALR estimator, the calibration is defined over population ratios $\mathbf{R} = \left(\frac{t_{\tilde{y}}}{t_{\tilde{x}_1}}, \frac{t_{\tilde{y}}}{t_{\tilde{x}_2}}, \frac{t_{\tilde{y}}}{t_{\tilde{x}_3}}\right)'$. The results of this simulation are shown in Table 12.

Table 12: Relative efficiency of the sampling estimators for the multiple regression model considering a sample size of $n = 400$.

R^2	$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.6$			$\rho = 0.8$		
	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C	R_C^{HT}	R_R^{HT}	R_R^C
0.05	1.05	1.04	0.99	1.21	1.18	0.97	1.62	1.51	0.93	2.94	2.37	0.81
0.2	1.05	1.03	0.99	1.21	1.14	0.94	1.62	1.37	0.85	2.93	1.86	0.64
0.4	1.05	1.03	0.98	1.22	1.14	0.94	1.62	1.35	0.83	2.89	1.79	0.62
0.6	1.05	1.04	0.99	1.21	1.15	0.95	1.61	1.39	0.87	2.87	1.93	0.67
0.8	1.05	1.04	0.99	1.21	1.17	0.97	1.60	1.47	0.92	2.83	2.23	0.79
0.95	1.05	1.05	1.00	1.21	1.20	0.99	1.59	1.55	0.97	2.80	2.60	0.93

Note that the results of this table are very similar to those shown in the previous simulation. The proposed CALR estimator is always better than the Horvitz-Thompson estimator, so it is a good option when there are no population totals available to perform the classical calibration estimator.

4. Estimation of Vote Intention

In a runoff election, a candidate wins in the first round if he obtains an absolute majority of the votes. If no candidate wins in the first round, then a second round must be held between the two candidates who managed to obtain the majority of the votes in the first round. The winner of that round wins the election

(Bouton & Gratton 2015). This system is used around the world for the election of presidents in Afghanistan, Argentina, Austria, Brazil, Bulgaria, Cape Verde, Chile, Colombia, Costa Rica, Croatia, The Czech Republic, Cyprus, Dominican Republic, Ecuador, Egypt, El Salvador, Finland, French, Ghana, Guatemala, India, Indonesia, Liberia, Peru, Poland, Portugal, Romania, Senegal, Serbia, Slovakia, Slovenia, Timor-Leste, Turkey, Ukraine, Uruguay and Zimbabwe.

In Latin America, as stated by Pérez-Liñán (2006), over the last two decades a majority of Latin American countries have adopted presidential runoff elections in order to strengthen the legitimacy of their elected presidents. During 2014, out of the 20 countries in Latin America, 7 had presidential elections, while 5 of them had to use the runoff elections mechanism. Table 13 shows the elections dates for the first and the second rounds in 2014 that were held in these nations, as well as the winners of these second rounds.

Table 13: Latin American presidential elections held in 2014.

Country	Date of the first round	Second round	Date of the second round	Winner
Bolivia	October 5	No		Evo Morales
Brazil	October 5	Yes	October 26	Dilma Rousseff
Colombia	May 25	Yes	June 15	Juan M. Santos
Costa Rica	February 2	Yes	April 6	Luis Guillermo Solís
El Salvador	February 2	Yes	March 9	Salvador Sánchez Cerén
Panam	May 4	No		Juan Carlos Varela
Uruguay	October 26	Yes	November 30	Tabaré Ramón Vázquez

Now, let us assume that after the first round elections, we perform a survey to a sample s of n citizens who are able to participate in the second round election. In that very survey, we ask the following estimations: for a) the vote intention in the second round; b) whether they had vote in the first round, and c) for which candidate they voted in the first round. Note that the estimates of the survey may be calibrated in order to improve the estimation of the results in the runoff by including auxiliary information from the results officially cast in the first round.

To do this, we must understand that for $k \in s$ there are four variables of interest that address the problem of vote intention. For the first round we define:

$$v_k = \begin{cases} 1 & \text{If } k\text{-th individual voted in the first round,} \\ 0 & \text{Otherwise.} \end{cases}$$

And, assuming that Q candidates (blank vote included) were contending in the first round, we define for every $q = 1, \dots, Q$,

$$x_{qk} = \begin{cases} 1 & \text{If } k\text{-th individual voted for the } q\text{-th candidate in first round,} \\ 0 & \text{Otherwise.} \end{cases}$$

For the second round, assuming that the intention of vote in the second round is going to be measured for only two candidates and a blank vote, we define the following:

$$u_k = \begin{cases} 1 & \text{If the } k\text{-th element will vote in the second round,} \\ 0 & \text{Otherwise.} \end{cases}$$

Finally, assuming that M (blank vote included) from Q candidates remain in the second round, we define for every $m = 1, \ldots, M$,

$$y_{mk} = \begin{cases} 1 & \text{If } k\text{-th individual has the intention to vote for the } m\text{-th candidate,} \\ 0 & \text{Otherwise.} \end{cases}$$

Note that, as the survey is carried out between the first and the second rounds, the vector of total votes in the first round $\mathbf{t_x} = (t_{x_1}, \ldots, t_{x_Q})'$ is already known. By defining $t_v = \sum_{k \in U} v_k$ as the amount of voters in the first round, the vector of ratios per candidate in the the first round is:

$$\mathbf{R}' = (R_1, R_2, \ldots, R_Q) = \left(\frac{\sum_{k \in U} x_{1k}}{\sum_{k \in U} v_k}, \ldots, \frac{\sum_{k \in U} x_{Qk}}{\sum_{k \in U} v_k} \right) = \left(\frac{t_{x1}}{t_v}, \ldots, \frac{t_{xQ}}{t_v} \right)$$

The approach in this paper addresses the construction of new weights w_k that areobtained by calibrating over the vector of population ratios $\mathbf{R}$. If the objective is to exactly reproduce the percentage of voters in the first round, then the calibration ratio estimators present in this paper should be used. As such, in order to create the weights w_k, it is necessary to define the following calibration equations:

$$\hat{\mathbf{R}} = \left(\frac{\sum_{k \in s} w_k x_{1k}}{\sum_{k \in s} w_k v_k}, \ldots, \frac{\sum_{k \in s} w_k x_{Qk}}{\sum_{k \in s} w_k v_k} \right) = \mathbf{R} \tag{42}$$

Applying the methodology proposed in this article, we define proper variables z_{qk}, so that for every $q = 1, \cdots, Q$, we obtain:

$$z_{qk} = \begin{cases} x_{qk} - R_q v_k & \text{If the } k\text{-th element voted for the } q\text{-th candidate in the first round} \\ 0 & \text{Otherwise} \end{cases}$$

Therefore, we also may address the estimation of the percentage of potential voters per m-th candidate in the second round by defining the following estimator:

$$\hat{R}_{m,cal} = \frac{\hat{t}_{ym,cal}}{\hat{t}_{u,cal}} = \frac{\sum_{k \in s} w_k y_{mk}}{\sum_{k \in s} w_k u_k} \tag{43}$$

Note that, in order to solve the calibration problem, if we use the chi-square distance, the estimator $\hat{t}_{yq,cal}$ adopts the following form:

$$\hat{t}_{ym,cal} = \hat{t}_{ym,\pi} + (\mathbf{t}_z - \hat{\mathbf{t}}_{z,\pi})' \hat{\mathbf{B}}_{\mathbf{yz}} = \hat{t}_{y,\pi} - \hat{\mathbf{t}}'_{z,\pi} \hat{\mathbf{B}}_{\mathbf{yz}} \tag{44}$$

where

$$\hat{\mathbf{B}}_{\mathbf{yz}} = \left(\sum_{s} \frac{\mathbf{z}_k\mathbf{z}'_k}{\pi_k}\right)^{-1} \left(\sum_{s} \frac{\mathbf{z}_k y_k}{\pi_k}\right).$$

Where $\mathbf{z}'_k = (z_{1k}, \ldots, z_{Qk})$. In the same way we can define $\hat{t}_{u,cal}$ as it follows:

$$\begin{aligned}\hat{t}_{u,cal} &= \hat{t}_{u,\pi} + (\mathbf{t}_z - \hat{\mathbf{t}}_{z,\pi})'\hat{\mathbf{B}}_{\mathbf{uz}} \\ &= \hat{t}_{u,\pi} - \hat{t}'_{\mathbf{z}.\pi}\hat{\mathbf{B}}_{\mathbf{uz}},\end{aligned}$$

where

$$\hat{\mathbf{B}}_{\mathbf{uz}} = \left(\sum_{k\in s} \frac{\mathbf{z}_k\mathbf{z}'_k}{\pi_k}\right)^{-1} \left(\sum_{k\in s} \frac{\mathbf{z}_k u_k}{\pi_k}\right),$$

and the ratio estimator takes the following form:

$$\hat{R}_{m,cal} = \frac{\hat{t}_{ym,\pi} - \hat{\mathbf{t}}'_{z,\pi}\hat{\mathbf{B}}_{\mathbf{yz}}}{\hat{t}_{u,\pi} - \hat{\mathbf{t}}'_{z,\pi}\hat{\mathbf{B}}_{\mathbf{uz}}} \tag{45}$$

4.1. Variance Estimator

We propose a variance estimator for $\hat{R}_{q,cal}$ by using a Taylor's approximation (see Särndal, Swensson & Wretman (2003) for detailed information). Then, the ratio estimator $\hat{R}_{q,cal}$ can be approximated by:

$$\hat{R}_{m,cal} \approx R_m + a_1(\hat{t}_{ym,\pi} - t_{ym}) + a_2(\hat{t}_{u,\pi} - t_u) + a'_3(\hat{\mathbf{t}}_{z,\pi} - \mathbf{t}_z) \tag{46}$$

And

$$\begin{aligned}a_1 &= \frac{\partial R_{m,cal}}{\partial \hat{t}_{ym,\pi}}\Big|_{\hat{t}_{ym,\pi}=t_{ym};\hat{t}_{u,\pi}=t_u;\hat{\mathbf{t}}_{z,\pi}=0} = \frac{1}{t_u} \\ a_2 &= \frac{\partial R_{q,cal}}{\partial \hat{t}_{u,\pi}}\Big|_{\hat{t}_{ym,\pi}=t_{ym};\hat{t}_{u,\pi}=t_u;\hat{\mathbf{t}}_{z,\pi}=0} = -\frac{t_{ym}}{t_u^2} \\ \mathbf{a}_3 &= \frac{\partial R_{q,cal}}{\partial \hat{\mathbf{t}}_{z,\pi}}\Big|_{\hat{t}_{ym,\pi}=t_{ym};\hat{t}_{u,\pi}=t_u;\hat{\mathbf{t}}_{z,\pi}=0} = \frac{t_{ym}\mathbf{B}_{\mathbf{uz}} - t_u\mathbf{B}_{\mathbf{yz}}}{t_u^2}\end{aligned}$$

Where $\mathbf{B}_{\mathbf{yz}}$ and $\mathbf{B}_{\mathbf{uz}}$ are the population counterparts of $\hat{\mathbf{B}}_{\mathbf{yz}}$ and $\hat{\mathbf{B}}_{\mathbf{uz}}$, respectively. As such, the variance estimator for $\hat{R}_{cal}$ is:

$$\begin{aligned}Var(\hat{R}_{m,cal}) &\approx Var(a_1\hat{t}_{ym,\pi} + a_2\hat{t}_{u,\pi} + \mathbf{a}'_3\hat{\mathbf{t}}_{\mathbf{z},\pi}) \\ &= Var\left(a_1\sum_{k\in s}\frac{y_{mk}}{\pi_k} + a_2\sum_{k\in s}\frac{u_k}{\pi_k} + \mathbf{a}'_3\sum_{k\in s}\frac{\mathbf{z}_k}{\pi_k}\right) \\ &= Var\left(\sum_{k\in s}\frac{1}{\pi_k}(a_1 y_{mk} + a_2 v_k + \mathbf{a}'_3\mathbf{z}_k)\right)\end{aligned}$$

Let $E_k = (a_1 y_{mk} + a_2 v_k + \mathbf{a}_3 z_{qk})$, then

$$Var(\hat{R}_{m,cal}) \approx Var\left(\sum_{k\in s}\frac{E_k}{\pi_k}\right) = \sum_{k\in U}\sum_{k\in U}\Delta_{kl}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l},$$

which can be estimated by

$$\hat{Var}(\hat{R}_{m,cal}) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l},$$

where $e_k = (\hat{a}_1 y_{mk} + \hat{a}_2 v_k + \hat{\mathbf{a}}_3'\mathbf{z}_k)$, and $\hat{a}_1 = \dfrac{1}{\hat{t}_{u,\pi}}$, $\hat{a}_2 = -\dfrac{\hat{t}_{ym,\pi}}{\hat{t}^2_{u,\pi}}$ and $\hat{\mathbf{a}_3} = \dfrac{\hat{\mathbf{B}}_{\mathbf{uz}}\hat{t}_{ym,\pi} - \hat{\mathbf{B}}_{\mathbf{yz}}\hat{t}_{u,\pi}}{\hat{t}^2_{u,\pi}}$.

4.2. Presidential Elections Held in Colombia (2014)

Presidential elections are the electoral mechanism through which citizens determine who will be the president of Colombia for a four year period (Blais, Massicotte & Dobrzynska 1997). One candidate gets elected in the first round when he or she obtains 50% of the total voters plus one (an absolute majority). If none of the candidates obtain the absolute majority, it is necessary to conduct a second round of voting: a runoff election. This will include the two candidates who obtained the most votes in the first round, as stated in Article 190 of the 1991 Colombian Constitution. Of the six presidential elections held since 1991, the second round mechanism has been used on four occasions: in 1994, 1998, 2010, and, recently, in 2014. The exceptions occured in 2002 and 2008 when the most popular politician in recent years, Álvaro Uribe Vélez, obtained on absolute majority in the first round with 53.04% and 62.35%, respectively. In Table 14, we show the results in the second rounds since 1991. We can conclude that the two candidates achieved quite similar numbers in all the second rounds, with the exception of 2010 when the candidate of the Colombian Green Party, Antanas Mockus, lost with 27.47% despite his popularity among young voters. Furthermore, the estimation of vote intention in the second round is also important because the candidates who do go on to the second round ally with those who did not. These partnerships are important as they try to get the most votes of these potential voters for they are the ones who will define the victor of the second round.

Table 14: Second round results in Colombia.

Year	Winner (Vote)	Loser (Vote)
1994	Ernesto Samper (50.57%)	Andrés Pastrana (48.45%)
1998	Andrés Pastrana (50.39%)	Horacio Serpa (46.53%)
2010	Juan M. Santos (69.13%)	Antanas Mockus (27.47%)
2014	Juan M. Santos (50.99%)	Oscar Zuluaga (44.99%)

We applied the proposed methods in this article to the results of a survey to estimate the voting intention in the second round of the presidential elections held in Colombia in 2014. This information must be conducted between the first and the second rounds by selecting a probability sample of voters made by a sampling design. Therefore inclusion probabilities must be included, which allows us to make estimations about the population of potential voters in the second round.

In this section we applied the proposed estimator to the Colombian presidential runoff election held in 2014. The first round was held on May 25th, 2014. The results of this first round are shown on Table 15.

Table 15: Results of the first round of the Colombian Presidential Elections held in 2014.

Candidate	% Votes	Total votes
Oscar Zuluaga	29.25	3759971
Juan M. Santos	25.69	3301815
Martha Ramírez	15.52	1995698
Clara López	15.23	1958414
Enrique Peñalosa	8.28	1065142
Blank Vote	5.99	770610
Total		12851650

These results indicate that if Colombia used the simple majority system, the president would have been Zuluaga and not Santos, who, in fact, is the current president of Colombia. As stated above, the candidates involved in the second round were Santos and Zuluaga. The population of interest were the voters who cast a valid vote in the first round, including votes that did not choose any candidate. This way, $N = 12.851.650$, of which 94% voted for a candidate while the other 6% did not vote for any candidate. Our goal was to estimate the number of people who planned on voting for Santos, Zuluaga or no candidate. The way to compute this estimation is by constructing new weights w_k, which are created using the voting rates for each candidate and the no vote in the first round as auxiliary information.

We also used the results of one survey carried out between the first and the second electoral rounds. This sample contains the opinion for $n = 2594$ potential voters. We present the summary information of this survey in Table 16. Note that the first round results are based on the real voting of the respondents, whereas the second round results are based on their intentions.

Table 16: Results from the survey carried out with a total of 2594 persons.

		Second round			
		Juan M. Santos	Oscar Zuluaga	Blank	Total
First round	López	172	65	64	301
	Peñalosa	47	22	28	97
	Juan M. Santos	849	23	5	877
	Ramírez	48	105	49	202
	Oscar Zuluaga	8	696	6	710
	Blank	87	86	234	407
	Total	1311	997	386	2594

In order to estimate the vote intention in the second round, and simultaneously calibrate over the known ratios of the first round, we defined the following calibration equations:

$$\begin{aligned}\hat{\mathbf{R}} &= \left(\frac{\sum_{k\in s} w_k x_{1k}}{\sum_{k\in s} w_k v_k}, \ldots, \frac{\sum_{k\in s} w_k x_{6k}}{\sum_{k\in s} w_k v_k}\right)\\ &= \mathbf{R}\\ &= (0.2925, 0.2569, 0.1552, 0.1523, 0.0828, 0.0599)\end{aligned}$$

Where the weights w_k are used to compute the proposed estimator for the total votes and the corresponding proportions given in equations (44) and (45), respectively. Additionally, it is also possible to calibrate by using the number of in the first round. That is, we computed the classic calibration estimator (CAL) using the calibration equation given by:

$$\begin{aligned}\hat{\mathbf{t}}_x &= \left(\sum_{k\in s} w_k^* x_{1k}, \ldots, \sum_{k\in s} w_k^* x_{6k}\right)\\ &= \mathbf{t}_x\\ &= (3759971, 3301815, 1995698, 1958414, 1065142, 770610)\end{aligned}$$

We computed the Horvitz-Thompson (HT) estimator, the proposed estimator (CALR) and the classic calibration estimator (CAL), and we found the new weights[2] w_k and w_k^* using the function calib from the package sampling (Tillé & Matei 2013). The dataset and the computational codes are available upon request from the main author.

Table 17 presents the results of the estimation of potential voters per candidate for the second round using the new weights. We can see that all three estimators considered Santos to be the winner of the election: this was the actual reality. However, the HT estimator gives much more percentage of a vote for no candidate than the other two estimators. The results of the CAL and CALR estimators, in this particular dataset, are similar. However, the proposed estimator in this paper does calibrate over the known ratios in the first round.

5. Discussion

In this paper, we have proposed a ratio calibration estimator considering several ratios, inducing calibration constraints. From the empirical research, we found that the proposed estimator has a smaller variance than the Horvitz & Thompson estimator and even a smaller one than the classic calibration estimator for most simulation scenarios considered in this article. Furthermore, the proposed estimator has the ability to estimate the population totals with negligible empirical bias.

[2] Note that weights w_k^* are different from w_k because they are induced by different calibration constraints.

Table 17: Estimations for the second round of the Colombian presidential elections held in 2014: estimated vote intention, proportion of estimated votes and its corresponding standard error (SE).

Candidate		CALR	CAL	HT	Real votes
Juan M. Santos	Total	5770076	6334436	4504076	7839342
	Proportion	44.89%	44.92%	45.05%	50.99%
	SE	3.99%	3.86%	3.31%	-
Oscar Zuluaga	Total	5618616	6093813	3920092	6917001
	Proportion	43.71%	43.21%	39.21%	44.99%
	SE	4.83%	5.07%	3.84%	-
Blank Vote	Total	1462958	1674619	1572545	618759
	Proportion	11.39%	11.87%	15.74%	4.02%
	SE	6.46%	6.85%	6.20%	-

We illustrated the particular usefulness of the proposed methodology in the runoff election system to estimate the vote intention in the second round. Despite the good performance of the proposed estimator, we noted that the estimated total number of voters is by far smaller than the real one, and that the estimation of vote for no candidate is too high. For future research, one way to estimate the voting intention in the second round could be by attempting to estimate the abstention percentage.

The proposed estimator can also be useful in other survey studies. For example, by taking into account the auto correlation and seasonal behaviour of macroeconomic variables, we can use the unemployment rate of a particular month of the year as auxiliary information in order to estimate the current value.

In order to keep the model-consistency and design-unbiasedness of the calibration estimators, Brewer (1999) argued that the proper choice of c_k, as in equation (5), should be $d_k - 1$. For further work, the appropriateness of these scalars should be investigated. In terms of consistency, this approach can also be used jointly, from a model-based perspective.

Further work on using this approach in the presence of non-response and frame imperfections is necessary. This methodology could also be used in surveys with multiple frames such as in the work of Elkasabi, Heeringa & Lepkowski (2015), its applicability, statistical properties and effect of misclassified domains are of great interest in further investigations.

[Received: January 2016 — Accepted: May 2016]

References

Blais, A., Massicotte, L. & Dobrzynska, A. (1997), 'Direct presidential elections: a world summary', *Electoral Studies* **16**(4), 441–455.

Bouton, L. & Gratton, G. (2015), 'Majority runoff elections: Strategic voting and duverger's hypothesis', *Theoretical Economics* **10**, 283–314.

Brewer, K. R. W. (1999), 'Cosmetic calibration with unequal probability sampling', *Survey Methodology* **25**(2), 205–212.

Deville, J.-C. & Särndal, C.-E. (1992), 'Calibration estimators in survey sampling', *Journal of the American statistical Association* **87**(418), 376–382.

Elkasabi, M. A., Heeringa, S. G. & Lepkowski, J. M. (2015), 'Joint calibration estimator for dual frame surveys', *Statistics in Transition* **16**(1), 7–36.

Estevao, V. M. & Särndal, C. E. (2004), 'Borrowing Strength Is Not the Best Technique Within a Wide Class of Design-Consistent Domain Estimators', *Journal of Official Statistics* **20**(4), 645–669.

Estevao, V. M. & Särndal, C.-E. (2006), 'Survey estimates by calibration on complex auxiliary information', *International Statistical Review* **74**(2), 127–147.

Estevao, V. M., Särndal, C.-E. & Sautory, O. (2000), 'A functional form approach to calibration', *Journal of Official Statistics* **16**, 379–399.

Kim, J. K. & Park, M. (2010), 'Calibration estimation in survey sampling', *International Statistical Review* **78**(1), 21–39.

Kim, J.-M., Sungur, E. A. & Heo, T.-Y. (2007), 'Calibration approach estimators in stratified sampling', *Statistics & Probability Letters* **77**(1), 99–103.

Kott, P. S. (2003), 'A practical use for instrumental-variable calibration', *Journal of Official Statistics* **19**(3), 265–272.

Kott, P. S. (2004), 'Comment on Demnati and Rao: Linarization variance estimators for survey data', *Survey Methodology* **30**, 27–28.

Krapavickaite, D. & Plikusas, A. (2005), 'Estimation of a ratio in the finite population', *Informatica* **16**(3), 347–364.

Lesage, E. (2011), 'The use of estimating equations to perform a calibration on complex parameters', *Survey methodology* **37**(1), 103–108.

Park, S. & Kim, J. K. (2014), 'Instrumental-variable calibration estimation in survey sampling', *Statistica Sinica* **24**, 1001–1015.

Pérez-Liñán, A. (2006), 'Evaluating presidential runoff elections', *Electoral Studies* **25**(1), 129–146.

Plikusas, A. (2006), Non-linear calibration, *in* 'Proceedings, Workshop on survey sampling', Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Särndal, C. E. (2007), 'The calibration approach in survey theory and practice', *Survey Methodology* **33**(2), 99–119.

Särndal, C.-E., Swensson, B. & Wretman, J. (2003), *Model assisted survey sampling*, Springer.

Tillé, Y. & Matei, A. (2013), *sampling: Survey Sampling.* R package version 2.6. *http://CRAN.R-project.org/package=sampling