This is an Accepted Manuscript of an article published by Taylor & Francis in the International Journal of Computer Mathematics on 25th January 2018 available at: https://doi.org/10.1080/00207160.2018.1425798
Please cite as: Martínez, S., Illescas, M., Martínez, H., & Arcos, A. (2020). Calibration estimator for Head Count Index. *International Journal of Computer Mathematics*, 97(1-2), 51-62.

ACCEPTED MANUSCRIPT

Calibration estimator for Head Count Index

Sergio Martínez[a*], María Illescas[b], Helena Martínez[a] and Antonio Arcos[c]

[a] Department of Mathematics, University of Almería, Almería, Spain; [b] Department of Economics and Business, University of Almería, Almería, Spain; [c] Department of Statistics and Operational Research, University of Granada, Granada, Spain

ARTICLE HISTORY
Compiled December 19, 2023

ABSTRACT
This paper considers the problem of estimating a poverty measure, the Head Count Index, using the auxiliary information available, which is incorporated into the estimation procedure by calibration techniques. The proposed method does not directly use the auxiliary information provided by auxiliary variables related to the variable of interest in the calibration process, but the auxiliary information, after a transformation, is incorporated by calibration techniques applied to the distribution function of the study variable. Monte Carlo experiments were carried out for simulated data and for real data taken from the Spanish living conditions survey to explore the performance of the new estimation methods of the Head Count Index.

KEYWORDS
Auxiliary information; calibration estimator; poverty index; survey sampling; Monte Carlo simulation

AMS CLASSIFICATION
62D05

1. Introduction

The estimation of a proportion in finite populations is a interesting topic in many areas such as medical and pharmaceutical statistics, marketing research, sociological studies and has important applications in the field of economics. Indeed, the analysis of poverty and social exclusion measures is a topic of increased interest to society. For governments is of high interest the estimation of poverty, inequality and life condition indicators and many social indicators related to the measurement of poverty are based upon binary variables or require the use of proportions to obtain such indicators. Among these poverty measures, we can find the Head Count Index that is widely used by institutions to elaborate their reports on poverty. The Head Count Index (HCI) can be calculated as the proportion of persons (or households) with an equivalised disposable income below the 60% of the national median equivalised income. In the literature, numerous references discuss about the HCI and related poverty indicators. For instance, some references are [2, 12–15]. The real HCI is unknown in practice, but it is estimated by using survey data, therefore estimation methods for proportions are

*Corresponding author. Email: spuertas@ual.es

required, since the HCI can be expressed as a proportion. Usually the method for estimating the HCI is by using direct estimators without using auxiliary information, but official surveys on income and living conditions generally contain additional variables related to the variable of interest and the efficient insertion of the auxiliary information available would improve the precision of the estimations for the proportion of a categorical variable of interest. These additional variables includes numeric and binary attributes and the HCI can have stronger relationship with auxiliary quantitative variables. In the presence of auxiliary information, there exist several procedures to obtain more efficient estimators for the proportion of a categorical variable of interest, maybe some of them ([9, 18]) assume that the auxiliary information is given by binary variables and consequently the auxiliary quantitative variables can not include at the estimation stage. In the case that the auxiliary information available includes both categorical and numerical attributes, we can use the logistic generalised regression estimator, proposed by [6] but has the problem of estimating the parameter associated to the logistic model.
In this paper, we consider the problem of estimating the population proportion of a categorical variable using the calibration framework. Calibration techniques were first employed by [3] to estimate the total population, but this approach is also applicable to the estimation of parameters more complex than the total population. [5, 16, 17] use different ways to implement the calibration approach in the estimation of the distribution function and the quantiles. The use of calibration techniques in the estimation of population proportion of a categorical variable is not new. In [9] the authors proposed estimation procedures for a proportion and based on calibration framework but as we discussed previously, the estimator obtained cannot be applied for the estimation of the HCI, since they assume that the auxiliary information is exclusively given by binary variables. Another calibration alternative when the auxiliary information includes both categorical and numerical attributes is given in [8] where it was proposed a calibration estimator based on probit regression. In this paper, we consider the incorporation of the auxiliary information with calibration techniques applied to the distribution function of the study variable under simple random sampling. The article is arranged as follows. In Section 2, the HCI and indirect estimation methods are introduced. Section 3 gives a alternative calibration estimator for HCI based on the estimation of the distribution function. In Section 4, we derive optimum estimators in the sense of minimum variance when the sample is selected under simple random sampling without replacement (SRSWR). Finally, in Section 5, simulation studies are carried out to analyze the performance of estimator proposed in this paper. Simulation studies are based upon real survey data and simulated finite populations. The real data is obtained from the Spanish living conditions survey. Section 6 gives some concluding remarks.

2. The Head Count Index and indirect estimation of population proportion

Let $U = \{1, 2, \ldots, N\}$ be a finite population consisting of N different elements. Let $s = \{1, 2, \ldots, n\}$ be the set of the units included in a sample, selected according to a specified sampling design with inclusion probabilities π_k and π_{kl} assumed to be strictly positive. We assume that y is the quantitative variable used to obtain the HCI and L is the poverty line used to classify the population into poor and nonpoor, that is, an individual (or households) is considered as poor if its income or expenditure y is

less than the poverty line L. Thus, the real HCI can be defined as the population proportion of the attribute A in the population U:

$$P_A = \frac{1}{N}\sum_{k\in U} A_k. \tag{1}$$

where $A_k = 1$ if the unit k is classified as poor ($y_k \leq L$) and $A_k = 0$ otherwise. The value A_k is only available for the sample units. We assume that the poverty line L is established by the corresponding authority, i.e., L is fixed at some official quantity. For instance, Eurostat fixes the relative poverty line in the 60% of the median of the equivalised net income.

To estimate P_A, the usual design-weighted Horvitz-Thompson estimator is:

$$\widehat{P}_{AHT} = \frac{1}{N}\sum_{k\in s} d_k A_k \tag{2}$$

where $d_k = 1/\pi_k$. Most official surveys on income and living conditions contain auxiliary variables related to the variable of interest, these auxiliary variables can be quantitative variables or qualitative attributes. Thus, we 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 numeric variables, ordinal variables, multinomial variables or binary attributes of the same type as the study attribute A. In the case of ordinal variables or multinomial variables, these variables are not directly included in the definition of the auxiliary vector $\mathbf{x}$, but the different categories (except one of them) are incorporated as dummy variables in the definition of the auxiliary vector $\mathbf{x}$.
The Horvitz-Thompson estimator $\widehat{P}_{AHT}$ is an unbiased estimator for P_A but does not use the auxiliary information provided by the vector $\mathbf{x}$. The incorporation of auxiliary information in estimating the population proportion P_A is not new and has been treated in many works. If the auxiliary vector $\mathbf{x}$ only includes binary attributes, we can use the estimation methods proposed by [18]. In the case that the vector $\mathbf{x}$ includes both categorical and numerical attributes, we can use the logistic generalised regression estimator, proposed by [6]. This estimator is given by:

$$\widehat{P}_{LGREG} = \frac{1}{N}\left(\sum_{k\in U} pl_k + \sum_{k\in s}\frac{A_k - pl_k}{\pi_k}\right) \tag{3}$$

where $pl_k = exp(x_k\widehat{\beta})/(1+exp(x_k\widehat{\beta}))$ and $\widehat{\beta}$ is the BLUP estimator of the β parameter of the logistic regression. [4] 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.

One way of incorporating auxiliary information provided by $\mathbf{x}$ in the estimation of P_A is via replacing the weights d_k of the estimators $\widehat{P}_{AHT}$ by new weights ω_k, using calibration techniques. Following [3], to obtain a calibration estimator for the attribute

A, we calculate the weights ω_k minimizing the chi-square distance

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

subject to a set of calibration constraints, where q_k are known positive constants unrelated to d_k. Thus, if $\mathbf{x}$ only includes binary attributes, we can estimate P_A through calibration techniques proposed by [9]. Calibration techniques have also recently been used in the estimation of P_A when the vector of auxiliary variables $\mathbf{x}$ contains both binary and numerical attributes. Thus, in [8], it was proposed a calibration estimator $\widehat{P}_{CP}$ based on probit regression, where the calibrated weights ω_k are obtained by minimizing (4) subject to the following conditions:

$$\frac{1}{N} \sum_{k \in s} \omega_k p_k = \bar{P} = \frac{1}{N} \sum_{k \in U} p_k, \tag{5a}$$

$$\frac{1}{N} \sum_{k \in s} \omega_k = 1, \tag{5b}$$

with $p_k = \widehat{P}[A_k = 1] = F(\widehat{\beta}' \cdot \mathbf{x}_k)$, where F is the normal-standard distribution function and $\widehat{\beta}$ is the π-weighted likelihood estimator of the β parameter of the probit regression ([9]).

In the next section we consider the estimation of the population proportion by estimating the distribution function $F_A(t)$ of the attribute of study A.

3. Calibration estimation of population proportion by estimating distribution function

In this section, we describe alternative calibration estimation methods for the problem of estimating P_A, based on auxiliary vector $\mathbf{x}$ that includes numeric and binary attributes. This calibration methods uses the value of the vector $\mathbf{x}$ in the definition of a new indirect estimator for P_A through the estimation of the distribution function $F_A(t)$ of the attribute of study A. For it, we consider $A_k = 1$ if the kth unit possesses the attribute A and $A_k = 0$ otherwise. The distribution function $F_A(t)$ of the variable A_k is given by:

$$F_A(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - A_k)$$

where

$$\Delta(t - A_k) = \begin{cases} 0 & \text{if } t < A_k \\ 1 & \text{if } t \geq A_k \end{cases}$$

Now, the variable A_k only takes two values in the population U, consequently the distribution function $F_A(t)$ is given by:

$$F_A(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 - P_A & \text{if } 0 \leq t < 1 \\ 1 & \text{if } 1 \leq t \end{cases}$$

Since the aim is to estimate the population proportion P_A by estimating the distribution function, we will consider the complementary attribute $\bar{A}$ of the attribute A, this is $\bar{A}_k = 1 - A_k$. Thus, the distribution function associated with the complementary attribute is given by:

$$F_{\bar{A}}(t) = \begin{cases} 0 & \text{if } t < 0 \\ P_A & \text{if } 0 \leq t < 1 \\ 1 & \text{if } 1 \leq t \end{cases}$$

since $P_A = F_{\bar{A}}(0)$, the estimate of the population proportion P_A can be obtained by the methods of estimating the distribution function. The usual estimator of distribution function is the Horvitz-Thompson estimator given by

$$\widehat{F}_{\bar{A}HT}(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - \bar{A}_k) \tag{6}$$

From (2) and (6) it is easy to see that $\widehat{P}_{AHT} = \widehat{F}_{\bar{A}HT}(0)$. Thus, we will obtain new indirect estimators of P_A through calibration techniques applied to $F_{\bar{A}}(t)$ at the point $t = 0$.
Recently, the calibration approach have been employed for the estimation of the distribution function and quantiles in different ways ([1, 5, 16, 17]). Following [16], we consider the definition of a pseudo-variable $g_k = \widehat{\beta}' \mathbf{x}_k$ for $k = 1, 2, \ldots, N$ with

$$\widehat{\beta}' = \left(\sum_{k \in s} d_k q_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} d_k q_k \mathbf{x}_k \bar{A}_k$$

where q_k are known positive constants unrelated to d_k. With the pseudo-variable g, we consider the estimation of $P_A = F_{\bar{A}}(0)$ with the calibration estimator obtained with the minimization of (4) subject to the following conditions:

$$\frac{1}{N} = \sum_{k \in s} \omega_k \Delta(\mathbf{t_g} - g_k) = F_g(\mathbf{t_g}) \tag{7}$$

with $\mathbf{t_g} = (t_1, \ldots, t_P)'$ is a vector chosen arbitrarily, where $t_1 < t_2 < \ldots < t_P$. Assuming that the inverse of symmetric matrix

$$T = \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k) \Delta(\mathbf{t_g} - g_k)'$$

exists, the resulting estimator ([16]) is given by

$$\widehat{P}_{AC} = \widehat{F}_{\bar{A}C}(0) = \widehat{P}_{AHT} + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)' \cdot \widehat{D} \tag{8}$$

where

$$\widehat{D} = T^{-1} \cdot \sum_{k \in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(0 - \bar{A}_k)$$

and $\widehat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz-Thompson estimator of $F_g(\mathbf{t_g})$.

The asymptotic variance of $\widehat{F}_{\bar{A}C}(0)$ ([16]) is given by:

$$AV(\widehat{F}_{\bar{A}C}(0)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl}(d_k E_k)(d_l E_l) \tag{9}$$

where $E_k = \Delta(0 - \bar{A}_k) - \Delta(\mathbf{t_g} - g_k) \cdot D$, with

$$D = \left(\sum_{k \in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)' \right)^{-1} \cdot \left(\sum_{k \in U} \Delta(\mathbf{t_g} - g_k)\Delta(0 - \bar{A}_k) \right)$$

4. Determining optimal calibration estimators

The precision of $\widehat{F}_{\bar{A}C}(0)$ changes with the selection of $\mathbf{t_g}$. In [7, 10], the authors studied, for a fixed P, the problem of selection the optimal vector $\mathbf{t_g}$ under simple random sampling and $q_k = 1$ for all $k \in U$, that gives the best estimation of $F_y(t)$ with the calibration estimator $\widehat{F}_{yc}(t)$ developed in [16], that is, the problem of determining an auxiliary vector $\mathbf{t_g} = (t_1, \ldots, t_P)'$, with $t_1 < t_2 < \ldots < t_P$ that minimizes the variance of the estimator $\widehat{F}_{yc}(t)$ given a point t for which we want to estimate $F_y(t)$. Moreover, in [11], the problem of the optimal dimension P of the auxiliary vector $\mathbf{t}_P$ and the optimal vector of this dimension is studied for the calibrated estimator of [16]. Following [10], the minimization of the asymptotic variance (9) under simple random sampling, is equivalent to minimizing the following function:

$$G(t_1, t_2, \ldots, t_P) = 2NP_A \cdot k_P - \sum_{j=1}^{P} \frac{(k_j - k_{j-1})^2}{(F_g(t_j) - F_g(t_{j-1}))} - k_P^2 \tag{10}$$

where

$$k_i = \sum_{k \in U} \Delta(0 - \bar{A}_k)\Delta(t_i - g_k) \quad i = 1, 2 \ldots, P \text{ and } k_0 = 0$$

and t_0 is a value such that $F_g(t_0) = 0$.

If we consider the auxiliary vector $\mathbf{t}_P = t_1$ (dimension $P = 1$), the value of t_1 at which the calibration estimator $\widehat{P}_{AC}$ is optimum ([7]) is given by

$$t_{opt} = \arg \min_{a_k \in A_0} G(a_k)$$

where

$$A_0 = \{g_k : k \in U,\, \bar{A}_k = 0\} = \{g_k : k \in U,\, A_k = 1\} = \{a_1, a_2, \ldots, a_M\} \tag{11}$$

with $a_1 < a_2 < \ldots < a_M$.

The optimal value of t_1, (t_{opt}) depends on some unknown values, so we go to replace the optimal vector t_{opt} by sample-based estimates. For it, we consider the following set based on the sample s

$$A_{0_s} = \{g_k : k \in s,\, \bar{A}_k = 0\} = \{g_k : k \in s,\, A_k = 1\} = \{a_{1_s}, a_{2_s}, \ldots, a_{m_s}\} \tag{12}$$

and the global minimum of the function $\widehat{G}(t_1)$ (the usual estimation of $G(t_1)$), is at one point of A_{0_s} ([7]). Thus, we can define a new calibration estimator $\widehat{P}_{AC1}$ based on the auxiliary point $\widehat{t}_{opt}$ that minimizes the function $\widehat{G}(t_1)$. The asymptotic behaviour of the estimator $\widehat{P}_{AC1}$ is the same as the estimator based on optimum point t_{opt} ([7]). Thus the asymptotic variance of $\widehat{P}_{AC1}$ is given by (9) with $\mathbf{t_g} = t_{opt}$.

On the other hand, if the dimension of the auxiliary vector is $P > 1$, the global minimum of the function $G(\mathbf{t}_g)$ ([10]) is a vector $\mathbf{t_{GP}} = (t_1, t_2, \ldots, t_P)$, with $t_1 < t_2 < \ldots < t_P$ and $t_i \in A_0$ or $t_i \in B_0$ for $i = 1, 2, \ldots, P$, where

$$B_0 = \{b_1, b_2, \ldots, b_M\} \tag{13}$$

with

$$b_1 = \max_{l \in U_1}\{g_l\} \text{ where } U_1 = \{l \in U : g_l < a_1\}$$

$$b_h = \max_{l \in U_h}\{g_l\} \text{ where } U_h = \{l \in U : a_{h-1} \leq g_l < a_h\}, \quad h = 2, 3, \ldots, M$$

It is clear that $b_1 < b_2 < \ldots < b_M$. Since the sets A_0 and B_0 are finite, finding the global minimum is computationally simple. For some h in $1, 2, \ldots, M$ the corresponding point b_h may not exist, but in this case, the minimization problem is simpler than the current case ([10]). Again, the optimal auxiliary vector $\mathbf{t_{GP}}$ depends on some unknown values, therefore we will replace the optimal vector with sample-based estimates. For it, we consider the usual estimation of the function G denoted by $\widehat{G}(\mathbf{t}_g)$, the sample-based set A_{0_s} and the following set:

$$B_{0_s} = \{b_{1_s}, b_{2_s}, \ldots, b_{m_s}\} \tag{14}$$

with

$$b_{1_s} = \max_{l \in U_{1_s}} \{g_l\} \text{ where } U_{1_s} = \{l \in s : g_l < a_{1_s}\}$$

$$b_{h_s} = \max_{l \in U_{h_s}} \{g_l\} \text{ where } U_{h_s} = \{l \in s : a_{(h-1)_s} \leq g_l < a_{h_s}\}, \quad h = 2, 3, \ldots, m$$

The potential points for the global minimum of $\widehat{G}(\mathbf{t}_g)$ are $\widehat{\mathbf{t}}_{\mathbf{GP}} = (\widehat{t}_1, \widehat{t}_2, \ldots, \widehat{t}_P)$ with $\widehat{t}_i \in A_{0_s}$ or $\widehat{t}_i \in B_{0_s}$ ([10]). The calibration estimator $\widehat{P}_{ACP}$ based on $\widehat{\mathbf{t}}_{\mathbf{GP}}$ has the same asymptotic behaviour that the estimator based on $\mathbf{t}_{\mathbf{GP}}$ and the asymptotic variance is given by (9) with $\mathbf{t}_{\mathbf{g}} = \mathbf{t}_{\mathbf{GP}}$.

Following [11], the optimal dimension P of the auxiliary vector $\mathbf{t}_g$ is $2M$ if b_1 exists and for all $i = 1, \ldots, M-1$, $b_{i+1} \neq a_i$. The optimal vector in this case is

$$\mathbf{t}_{\mathbf{OPT}} = (b_1, a_1, b_2, a_2, \ldots, b_M, a_M) \tag{15}$$

If for some $i_1, i_2, \ldots i_R \in \{0, 1, \ldots, M-1\}$; $a_{i_1} = b_{i_1+1}$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$ the optimal dimension is $P = 2M - R$ and the optimal auxiliary vector $\mathbf{t}_{\mathbf{OP}} = (t_{O1}, \ldots, t_{O(2M-R)})$ is given by:

$$\mathbf{t}_{\mathbf{OP}} = (b_1, a_1, \ldots, b_{i_1}, a_{i_1}, a_{i_1+1}, b_{i_1+2}, \ldots, b_{i_h}, a_{i_h}, a_{i_h+1}, b_{i_h+2}, \ldots b_M, a_M) \tag{16}$$

The optimal auxiliary vector $\mathbf{t}_{\mathbf{OP}}$ depends on some unknown values, thus a calibration estimator based on this vector cannot be calculated. Furthermore, although the vector $\mathbf{t}_{\mathbf{OP}}$ be known, we could have incompatible restrictions in (7) when a sample s is selected. Thus, we replace $\mathbf{t}_{\mathbf{OP}}$ by a estimated vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ based on the set A_{0_s} and B_{0_s}. If b_{1_s} exists and for all $i = 1, \ldots, m_s - 1$, $b_{(i+1)_s} \neq a_{i_s}$ the estimated vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ is given by

$$\widehat{\mathbf{t}}_{\mathbf{OP}} = (b_{1_s}, a_{1_s}, \ldots, b_{m_s}, a_{m_s}) \tag{17}$$

If we defined $a_{0_s} = \min_{k \in U} g_k$ and for some $i_1, i_2, \ldots i_r \in \{0_s, 1_s, \ldots, (m-1)_s\}$; $a_{i_1} = b_{i_1+1}$ with $r \leq m_s$ and $i_h \neq i_j$ if $h \neq j$ the estimated vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$ is given by

$$\widehat{\mathbf{t}}_{\mathbf{OP}} = (b_{1_s}, a_{1_s}, \ldots, b_{i_1}, a_{i_1}, a_{i_1+1}, b_{i_1+2}, \ldots, b_{i_h}, a_{i_h}, a_{i_h+1}, b_{i_h+2}, \ldots b_{m_s}, a_{m_s}) \tag{18}$$

Now, we can define a new calibration estimator $\widehat{P}_{ACOPT}$ based on the auxiliary vector $\widehat{\mathbf{t}}_{\mathbf{OP}}$.

The Horvitz-Thompson estimator $\widehat{P}_{AHT}$, under SRSWOR, has the following shift invariance property $\widehat{P}_{AHT} = 1 - \widehat{Q}_{AHT}$, where $\widehat{Q}_{AHT}$ is the Horvitz-Thompson estimator for $Q_A = 1 - P_A$. Thus, $\widehat{P}_{AHT}$ has the same performance in the estimation of P_A as the performance of $\widehat{Q}_{AHT}$ in the estimation of Q_A. In general, this property is not satisfied by the calibration estimators considered $\widehat{P}_{AC1}$, $\widehat{P}_{ACP}$ and $\widehat{P}_{ACOPT}$. It is

easy to see that this property is fulfilled by a calibration estimator if

$$1 = \frac{1}{N}\sum_{k\in U}\omega_k \tag{19}$$

A way to obtain the condition (19) consists in the incorporation of the value $g_{max} = \max_{k\in U} g_k$ in the auxiliary optimum vectors. Thus, we can define a calibration estimator $\widehat{P}_{AQ1}$ based on the auxiliary vector $(\widehat{t}_{opt}, g_{max})$, a calibration estimator $\widehat{P}_{AQP}$ based on $(\widehat{\mathbf{t}}_{GP}, g_{max})$ and a calibration estimator $\widehat{P}_{AQOPT}$ based on $(\widehat{\mathbf{t}}_{OP}, g_{max})$. Nothing guarantees that we can use this vector in the calibration constraints given by (7), when selecting a sample s, since we could have incompatible restrictions. For example, suppose that, for the selected sample s, we have

$$\frac{1}{N}\sum_{k\in s}\omega_k\Delta(\widehat{t}_P - g_k) = \frac{1}{N}\sum_{k\in s}\omega_k$$

we have two incompatible restrictions. In this case, we consider the calibration constraints given by (7) without $g_{max} = \max_{k\in U} g_k$.

It is easy to see that the incorporation of the value g_{max} in the calibration conditions (7) does not produces a negative effect in the asymptotic variance. For it, from equation (10) we have

$$G(\widehat{\mathbf{t}}_{\mathbf{GP}}) - G((\widehat{\mathbf{t}}_{GP}, g_{max}) = (NP_A - k_P)^2 - \frac{(NP_A - k_P)^2}{(1 - F_g(t_P))} \leq 0$$

Another way to incorporate shift invariance property, consists in the minimization of the function (10) when the auxiliary vector considered is (t_1, g_{max}). Following [7] the optimal auxiliary vector is $\mathbf{t}_{\mathbf{GMAX}} = (t_1, g_{max})$ where $t_1 \in A_0$ or $t_1 \in B_0$. Similarly to the previous cases, we can define a new calibration estimator $\widehat{P}_{ACMAX}$ based on $\widehat{\mathbf{t}}_{\mathbf{GMAX}}$, a sample-based estimation.

5. Numerical comparison

In this sections, we present the results of a Monte Carlo simulation study where we compare the precision of the proposed calibration estimators: $\widehat{P}_{AC1}$, $\widehat{P}_{ACP}$, $\widehat{P}_{ACOPT}$, $\widehat{P}_{AQ1}$, $\widehat{P}_{AQP}$, $\widehat{P}_{AQOPT}$ and $\widehat{P}_{ACMAX}$ with the Horvitz-Thompson estimator $\widehat{P}_{AHT}$; the multivariated ratio estimator $\widehat{P}_{AMratio}$ (see [18]); calibration estimators $\widehat{P}_{AR}$ and the multivariate calibration estimator $\widehat{P}_{AWM}$ (see [9]); the logistic generalised regression estimator $\widehat{P}_{LGREG}$ ([6] and calibration estimator $\widehat{P}_{CP}$ based on probit regression (see [8]). The estimators $\widehat{P}_{ACP}$ and $\widehat{P}_{AQP}$ are based on auxiliary vector with dimension $P = 2$. Our simulations are programmed in `R`, with some new code developed to compute the estimators to be compared. 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 $\hat{T}$ were computed as

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

where $MSE(\hat{T}) = B^{-1}\sum_{b=1}^{B}(\hat{T}^b - P)^2$, $MSE(\widehat{P}_{AHT}) = B^{-1}\sum_{b=1}^{B}(\widehat{P}^b_{AHT} - P)^2$, and $\hat{T}^b$ and $\widehat{P}^b_{AHT}$ are the values of $\hat{T}$ and $\widehat{P}_{AHT}$ from the bth simulation, respectively.

To investigate the efficiency of the proposed estimators under a variety of situations, we conside different stages. First, we will consider the estimation of a population proportion in simulated populations and secondly we will use the proposed estimators in the estimation of the Head Cont Index. For the estimation of population proportion, we consider 5 populations generated as a random sample of 10000 units from a Bernoulli distribution with parameter $P = 0.9$, and the attributes of interest were thus achieved with the aforementioned population proportion. Auxiliary attributes were also generated using the same distribution, but a given proportion of values was randomly changed so that Cramer's V coefficient between the attribute of interest and the auxiliary attribute would range from 0.5 to 0.9. For each of the 5 populations, $B = 10000$ samples of sizes n =150, 250, 350 and 450 were selected, under simple random sampling, to compare the considered estimators in terms of relative bias (RB) and relative efficiency (RE). Table 1 give the values of RB and RE in percentages for the binomial populations with Cramer's V coefficient range from 0.5 to 0.7. and Table 2 give the values of RB and RE for the binomial populations with Crame's V coefficient range from 0.8 to 0.9

[Table 1 about here.]

[Table 2 about here.]

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%. The estimators $\widehat{P}_{AHT}$, $\widehat{P}_{AMratio}$, $\widehat{P}_{AC1}$ and $\widehat{P}_{AQ1}$ has the same variance and have a larger variance than the other estimators considered. With large Cramer's V coefficient (ρ) values, the proposed estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ produce good results. It can also be seen that as ρ 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 estimators $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ has the highest degree of efficiency for small values of ρ while for the large values, the estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$ and $\widehat{P}_{ACMAX}$ present a greater efficiency. In all cases, the calibration estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ perform better than the estimators $\widehat{P}_{AR}$, $\widehat{P}_{AWM}$, $\widehat{P}_{LGREG}$ and $\widehat{P}_{CP}$.
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.

For the estimation of Head Count Index, we consider real data taken from the 2008 Spanish living conditions survey carried out by the Instituto Nacional de Estadística (INE) of Spain. For our simulation study, we considered the survey data collected as a population with size $N = 12990$, from which samples are selected. The

poverty threshold is calculated each year, using the distribution of the equivalised net income for the previous year. Following the criteria recommended by Eurostat, this threshold is set at 60% of the median of the equivalised net income. The value of the population HCI is 0.1968. We considered the variable x_1 ="Returns and additional revenue from adjustments in taxes" and the attribute x_2 ="Home with own car" (1 for home with own car, 0 otherwise) as the auxiliary variables. The correlation coefficient between the main variable with x_1 is -0.12 and the correlation coefficient between the main variable with x_2 is -0.09. Again, $B = 10000$ samples of sizes n =500, 600, 700 and 800 were selected, under simple random sampling, to compare the relative bias (RB) and relative efficiency (RE) of the considered estimators. Table 3 give the values of RB and RE in percentages for the real population.

[Table 3 about here.]

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

- The calibration estimator $\widehat{P}_{AR}$, $\widehat{P}_{AC1}$ and $\widehat{P}_{AQ1}$ performs poorly and has worse efficiency than the estimator $\widehat{P}_{AHT}$.
- The remaining estimates are more efficient than the estimator $\widehat{P}_{AHT}$ but the gain in efficiency is not as great as in the previous example.
- The proposed calibration estimator $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ often work better than the other estimators. The estimator $\widehat{P}_{ACMAX}$ is the most efficient for all sample sizes, except for the sample size $n = 600$ where the estimators $\widehat{P}_{ACP}$ and $\widehat{P}_{AQP}$ present the best results.

6. Concluding remarks

In recent years, the use of calibration technique has attracted significant attention in survey methodology and applications. Calibration techniques are used for improving more efficient estimators for a finite population by using the incorporation of available auxiliary population information. This paper presents a new calibration technique for estimating proportions in finite populations, based on a vector of auxiliary information that includes both quantitative and qualitative variables. For it, the proposed calibration technique considers the incorporation of the auxiliary information with calibration techniques applied to the distribution function of the study variable under simple random sampling. The estimation of a proportion in finite populations is a interesting topic in many areas and has important applications in the field of economics and in recent years there has been growing interest in the need to precise indicators of poverty, inequality and living conditions. Many of these indicators, such as the Head Count Index, are based on population proportions of binary variables.
The HCI is a poverty indicator commonly used in comparisons of poverty across countries, but is unknown in practice and therefore it is necessary to estimate its value. The calibration technique proposed in this paper allows obtaining HCI estimates incorporating qualitative and quantitative auxiliary information.

Simulation studies are conducted to evaluate the performance of the proposed calibration technique in terms of various empirical measures under different scenarios. First, we evaluated the estimation of a population proportion in simulated populations

where we compared the precision of the proposed calibration estimators to several existing indirect estimators. We observed that the proposed estimators have a good performance in terms of relative biases and we observed that the proposed estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ perform better than the others indirect estimators considered in the simulation study.
The simulation study also compares the calibration technique proposed in the estimation of Head Count Index. For it, we consider real data taken from the 2008 Spanish living conditions survey. In the results of this study, we can also observe that the proposed estimators present good results in terms of relative bias and calibration estimator $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ often work better than the other estimators but the gain in efficiency is not as great as in the case of the estimation of a population proportion in simulated populations.

In summary, the simulation studies indicate that the proposed calibration estimators can be an alternative estimation method for the problem of estimating population proportion and therefore can be a calibration estimation method for the estimation of HCI.

References

[1] A. Arcos, S. Martínez, M. Rueda and H. Martínez, *Distribution function estimates from dual frame context*, J. Comput. Appl. Math. 318 (2017), pp. 242–252.

[2] E. Crettaz and C. Suter, *The impact of adaptive preferences on subjective indicators: An analysis of poverty indicators*, Social Indicators Research. 114 (2013), pp. 139–152.

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

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

[5] T. Harms and P. Duchesne, *On calibration estimation for quantiles*, Survey Methodology. 32 (2006), pp. 37–52.

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

[7] S. Martínez, M. Rueda, A. Arcos and H. Martínez, *Optimum calibration points estimating distribution functions*, J. Comput. Appl. Math. 233(9) (2010), pp. 2265–2277.

[8] S. Martínez, M. Rueda, A. Arcos and H. Martínez, *Estimating the proportion of a categorical variable with probit regression*, Sociological Methods and Research. In Press.

[9] S. Martínez, A. Arcos, H. Martínez and S. Singh, *Estimating population proportions by means of calibration estimators*, Revista Colombiana de Estadistica. 38(1) (2015), pp. 267–293.

[10] S. Martínez, M. Rueda, H. Martínez and A. Arcos, *Determining P optimum calibration points to construct calibration estimators of the distribution function*, J. Comput. Appl. Math. 275 (2015), pp. 281–293.

[11] S. Martínez, M. Rueda, H. Martínez and A. Arcos, *Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function*, J. Comput. Appl. Math. 318 (2017), pp. 444–459.

[12] M. Medeiros, *The rich and the poor: The construction of an affluence line from the poverty line*, Social Indicators Research. 78 (2006), pp. 1–18.

[13] D. Morales, M. Rueda and D. Esteban, *Model-assisted estimation of small area poverty measures: An application within the Valencia Region in Spain*, Social Indicators Research. https://doi.org/10.1007/s11205-017-1678-1.

[14] J. F. Múñoz, E. Álvarez-Verdejo, R. M. García-Fernández and L. J. Barroso, *Efficient es-*

timation of the Headcount Index, Social Indicators Research. 123 (2015), pp. 713–732.

[15] J. Navicke, O. Rastrigina and H. Sutherland, *Nowcasting indicators of poverty risk in the European Union: A microsimulation approach*, Social Indicators Research. 119(1) (2014), pp. 101–119.

[16] M. Rueda, S. Martínez, H. Martínez and A. Arcos, *Estimation of the distribution function with calibration methods*, J. Statist. Plann. Inference. 137 (2007), pp. 435–448.

[17] M. Rueda, S. Martínez, H. Martínez and A. Arcos, *Calibration methods for estimating quantiles*, Metrika. 66 (2007), pp. 355–371.

[18] M. Rueda, J. F. Muñoz, A. Arcos, E. Álvarez and S. Martínez, *Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies*, Journal of biopharmaceutical statistics. 21(3) (2011), pp. 526–554.

Table 1. RB % and RE % for several sample sizes of the estimators compared. SRSWOR from the BINOMIAL populations. Cramer's V coefficient range from 0.5 to 0.7

	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%
	$\rho = 0.5$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00
$\widehat{P}_{AMratio}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00
$\widehat{P}_{AR}$	0.031	109.21	-0.017	108.44	0.004	108.16	-0.037	109.90
$\widehat{P}_{AWM}$	-0.002	112.48	-0.038	111.70	-0.008	112.08	-0.036	113.86
$\widehat{P}_{LGREG}$	0.006	113.63	-0.034	112.99	-0.006	113.47	-0.035	115.42
$\widehat{P}_{CP}$	0.002	113.29	-0.037	112.60	-0.007	113.13	-0.036	115.09
$\widehat{P}_{AC1}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00
$\widehat{P}_{AQ1}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00
$\widehat{P}_{ACP}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52
$\widehat{P}_{AQP}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52
$\widehat{P}_{ACOPT}$	0.040	115.21	-0.027	115.07	0.003	115.79	-0.029	118.14
$\widehat{P}_{AQOPT}$	0.040	115.21	-0.027	115.07	0.003	115.79	-0.029	118.14
$\widehat{P}_{ACMAX}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52
	$\rho = 0.6$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	0.055	100.00	0.010	100.00	-0.016	100.00	-0.020	100.00
$\widehat{P}_{AMratio}$	0.055	100.00	0.010	100.00	-0.016	100.00	-0.020	100.00
$\widehat{P}_{AR}$	0.040	109.78	-0.007	116.88	-0.009	117.98	-0.009	116.28
$\widehat{P}_{AWM}$	0.010	112.12	-0.035	127.92	-0.023	130.18	-0.020	126.98
$\widehat{P}_{LGREG}$	0.018	113.39	-0.030	131.35	-0.021	134.15	-0.018	130.46
$\widehat{P}_{CP}$	0.014	112.99	-0.037	130.66	-0.025	133.37	-0.021	129.62
$\widehat{P}_{AC1}$	0.055	100.00	0.010	100.00	-0.016	100.00	-0.020	100.00
$\widehat{P}_{AQ1}$	0.055	100.00	0.010	100.00	-0.016	100.00	-0.020	100.00
$\widehat{P}_{ACP}$	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
$\widehat{P}_{AQP}$	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
$\widehat{P}_{ACOPT}$	0.046	115.85	-0.002	135.30	-0.007	138.81	-0.010	133.41
$\widehat{P}_{AQOPT}$	0.046	115.85	-0.002	135.30	-0.007	138.81	-0.010	133.41
$\widehat{P}_{ACMAX}$	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
	$\rho = 0.7$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00
$\widehat{P}_{AMratio}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00
$\widehat{P}_{AR}$	0.011	127.74	-0.003	130.28	0.012	133.02	0.002	131.23
$\widehat{P}_{AWM}$	-0.130	140.46	-0.077	144.91	-0.029	148.35	-0.028	146.13
$\widehat{P}_{LGREG}$	-0.120	150.79	-0.070	156.15	-0.025	160.36	-0.025	158.11
$\widehat{P}_{CP}$	-0.150	148.42	-0.092	153.53	-0.038	157.63	-0.034	155.26
$\widehat{P}_{AC1}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00
$\widehat{P}_{AQ1}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00
$\widehat{P}_{ACP}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07
$\widehat{P}_{AQP}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07
$\widehat{P}_{ACOPT}$	0.004	173.86	-0.006	179.09	0.014	181.82	0.001	179.94
$\widehat{P}_{AQOPT}$	0.004	173.86	-0.006	179.09	0.014	181.82	0.001	179.94
$\widehat{P}_{ACMAX}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07

Table 2. RB % and RE % for several sample sizes of the estimators compared. SRSWOR from the BINOMIAL populations. Cramer's V coefficient range from 0.8 to 0.9

	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%
	$\rho = 0.8$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{AMratio}$	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{AR}$	0.016	165.15	-0.032	167.61	0.003	162.67	0.005	166.53
$\widehat{P}_{AWM}$	-0.145	187.40	-0.095	193.44	-0.048	189.55	-0.034	190.52
$\widehat{P}_{LGREG}$	-0.130	200.75	-0.081	209.88	-0.038	206.88	-0.028	208.68
$\widehat{P}_{CP}$	-0.160	197.77	-0.098	205.53	-0.050	202.48	-0.036	203.90
$\widehat{P}_{AC1}$	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{AQ1}$	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{ACP}$	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
$\widehat{P}_{AQP}$	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
$\widehat{P}_{ACOPT}$	0.002	218.32	-0.011	225.62	0.005	221.95	-0.004	223.73
$\widehat{P}_{AQOPT}$	0.002	218.32	-0.011	225.62	0.005	221.95	-0.004	223.73
$\widehat{P}_{ACMAX}$	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
	$\rho = 0.9$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00
$\widehat{P}_{AMratio}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00
$\widehat{P}_{AR}$	-0.015	289.45	0.006	287.74	-0.001	286.77	-0.003	294.96
$\widehat{P}_{AWM}$	-0.181	331.63	-0.123	344.48	-0.080	345.35	-0.057	356.19
$\widehat{P}_{LGREG}$	-0.191	360.15	-0.127	377.50	-0.079	389.81	-0.053	400.60
$\widehat{P}_{CP}$	-0.209	349.52	-0.143	364.93	-0.091	372.18	-0.062	382.58
$\widehat{P}_{AC1}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00
$\widehat{P}_{AQ1}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00
$\widehat{P}_{ACP}$	-0.115	411.46	-0.050	437.04	-0.014	470.50	-0.006	469.58
$\widehat{P}_{AQP}$	-0.115	411.46	-0.050	437.04	-0.014	470.50	-0.006	469.58
$\widehat{P}_{ACOPT}$	-0.050	428.22	-0.020	445.22	-0.009	461.61	-0.006	463.43
$\widehat{P}_{AQOPT}$	-0.050	428.59	-0.020	445.22	-0.009	461.61	-0.006	463.43
$\widehat{P}_{ACMAX}$	-0.115	411.80	-0.050	437.04	-0.014	470.50	-0.006	469.58

Table 3. RB % and RE % for several sample sizes of the estimators compared under SRSWOR from the 2008 Spanish living conditions survey population.

	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%
Estimator	$n = 500$		$n = 600$		$n = 700$		$n = 800$	
$\widehat{P}_{AHT}$	-0.066	100.00	-0.333	100.00	-0.256	100.00	0.044	100.00
$\widehat{P}_{AMratio}$	-0.066	100.00	-0.333	100.00	-0.258	100.00	0.044	100.00
$\widehat{P}_{AR}$	-0.064	94.80	-0.363	90.14	-0.286	96.98	0.027	96.65
$\widehat{P}_{AWM}$	-0.184	104.88	-0.227	102.49	-0.306	102.10	0.017	102.03
$\widehat{P}_{LGREG}$	-0.151	105.04	-0.190	102.50	-0.276	103.01	0.037	102.73
$\widehat{P}_{CP}$	-0.155	104.96	-0.221	102.47	-0.274	102.71	0.042	102.57
$\widehat{P}_{AC1}$	-0.199	99.08	-0.288	100.23	-0.342	99.53	-0.041	99.66
$\widehat{P}_{AQ1}$	-0.107	99.199	-0.287	100.85	-0.273	99.68	0.027	99.68
$\widehat{P}_{ACP}$	-0.648	104.31	-0.337	112.15	-0.532	103.16	-0.053	102.95
$\widehat{P}_{AQP}$	-0.638	104.39	-0.380	112.90	-0.522	103.22	-0.053	103.04
$\widehat{P}_{ACOPT}$	-0.167	105.05	0.388	102.45	-0.157	103.06	0.057	102.55
$\widehat{P}_{AQOPT}$	0.169	105.04	0.788	102.23	0.156	103.23	0.058	102.77
$\widehat{P}_{ACMAX}$	-0.238	105.70	-0.443	103.63	-0.238	104.32	0.050	103.60