Revista Colombiana de Estadística
January 2015, Volume 38, Issue 1, pp. 267 to 293
DOI: http://dx.doi.org/10.15446/rce.v38n1.48814

Estimating Population Proportions by Means of Calibration Estimators

Estimación de proporciones poblacionales mediante estimadores de calibración

Sergio Martínez[1,a], Antonio Arcos[2,b], Helena Martínez[1,c], Sarjinder Singh[3,d]

[1]Math Department, University of Almería, Almería, España

[2]Department of Statistics and Operational Research, University of Granada, Granada, España

[3]Department of Mathematics, Texas A&M University-Kingsville, Kingsville Texas, United States

Abstract

This paper considers the problem of estimating the population proportion of a categorical variable using the calibration framework. Different situations are explored according to the level of auxiliary information available and the theoretical properties are investigated. A new class of estimator based upon the proposed calibration estimators is also defined, and the optimal estimator in the class, in the sense of minimal variance, is derived. Finally, an estimator of the population proportion, under new calibration conditions, is defined. Simulation studies are considered to evaluate the performance of the proposed calibration estimators via the empirical relative bias and the empirical relative efficiency, and favourable results are achieved.

***Key words*:** Auxiliary Information, Calibration, Estimators, Finite Population, Sampling Design.

Resumen

El artículo considera el problema de la estimación de la proporción poblacional de una variable categórica usando como marco de trabajo la calibración. Se exploran diferentes situaciones de acuerdo con la información auxiliar disponible y se investigan las propiedades teóricas. Una nueva clase de estimadores basada en los estimadores de calibración propuestos también

[a]Professor. E-mail: spuertas@ual.es
[b]Professor. E-mail: arcos@ugr.es
[c]Ph.D. Research Assistant. E-mail: hmartinez@ual.es
[d]Associate Professor. E-mail: sarjinder.singh@tamuk.edu

es definida y el estimador óptimo en la clase, en el sentido de varianza mínima, es obtenido. Finalmente, un estimador de la proporción poblacional, bajo nuevas condiciones de calibración es también propuesto. Estudios de simulación para evaluar el comportamiento de los estimadores calibrados propuestos a través del sesgo relativo empírico y de la eficiencia relativa empírica son incluidos, obteniéndose resultados satisfactorios.

***Palabras clave*:** calibración, diseño muestral, estimadores, información auxiliar, población finita.

1. Introduction

In the presence of auxiliary information, various approaches may be used to improve the precision of estimators at the estimation stage. The book of Singh (2003) contains several examples, including ratio, difference or calibration estimators, following the methodology proposed by Deville & Särndal (1992) and Särndal (2007), or regression estimators, as the papers of Arnab, Shangodoyin & Singh (2010) and Singh, Singh & Kozak (2008) show. These techniques are generally more efficient than other methods not using auxiliary information. Usually social surveys are focused on categorical variables as sex, race, potential voters, etc.

Efficient insertion of available auxiliary information would improve the precision the estimations for the proportion of a categorical variable of interest. Conceptually, it is difficult to justify using a regression estimator for estimating proportions. Duchesne (2003) considered estimators of a proportion under different sampling schemes and presented an estimator which used the logistic regression estimator. The model calibration technique proposed by Wu & Sitter (2001) can be also used to estimate a proportion by using a logistic regression model. Based on logistic models, these estimators efficiently facilitate good modeling of survey data assuming that unit-specific auxiliary data in the population U are available. In this case it is assumed that the values of auxiliary variables are known for the entire finite population (referred to as complete auxiliary information) but the values of main variable are known only if the unit is selected in the sample.

It is very common for population data associated with auxiliary variables to be obtained from census results, administrative files, etc., and these sources often provide different parameters for these auxiliary variables. For example, position measures (mean, median and other moments) are normally provided, but there is no access to data for each individual. In the present study, it is assumed that the only datum known is the proportion of individuals presenting one or more characteristics related to the study variable.

Under this assumption, Rueda, Muñoz, Arcos, Álvarez & Martínez (2011) defined an estimator and various confidence intervals for a proportion using the ratio method. The results of their simulation studies show that ratio estimators are more efficient than traditional estimators. Confidence intervals outperform alternative methods, especially in terms of interval width.

Calibration techniques were first employed by Deville & Särndal (1992) to estimate the total population, but this approach is also applicable to the estimation more complex of parameters than the total population. Relevant papers estimating population variances are Singh (2001), Singh, Horn, Chowdhury & Yu (1999) and Farrell & Singh (2005). The estimation of finite population distribution functions is studied in papers by Harms & Duchesne (2006) or Rueda, Martínez, Martínez & Arcos (2007) and the estimation of quantiles in Rueda, Martínez-Puertas, Martínez-Puertas & Arcos (2007). In Section 2 we review a proportion estimator using the calibration technique. Section 3 describes alternative methods for deriving the calibration estimator for the proposed parameter. In Section 4 we extend these methods to the multiple case. A simulation study is performed in Section 5 and our conclusions are reported in Section 7.

2. Calibration Estimators for the Proportion

2.1. Definition of the Calibration Estimator

Assume a sample s with size n from a finite population $U = \{1, 2, \ldots, N\}$ with size N, selected by a specific sampling design d, with inclusion probabilities π_k and π_{kl} assumed to be strictly positive. Let A be an attribute of study in the population U, defining $A_k = 1$ when a unit k of the population U has the attribute A and $A_k = 0$ otherwise. The population proportion of attribute A in the population U is given by:

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

To estimate (1), the usual design-weighted Horvitz-Thompson estimator is:

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

where $d_k = 1/\pi_k$.

If we consider an auxiliary attribute B in which the value B_k is known for every unit k in the sample s and P_B is also known, the above estimator cannot incorporate the information provided by the attribute B, in estimating the population proportion of A. One way of incorporating auxiliary information in the parameter estimation is via replacing the weights d_k by new weights ω_k, using calibration techniques.

Calibration is a highly desirable property for survey weights, as Särndal (2007) argues, for the following reasons:

- it provides a systematic way of taking auxiliary information into account;
- it is a means of obtaining consistent estimates, with known aggregates;

- it is used by statistical agencies for estimating different finite population parameters. Several national statistical agencies have developed software designed to compute calibration weights based on auxiliary information available in population registers and other sources. Such agencies include CLAN (Statistics Sweden) and BASCULA (Central Bureau of Statistics, The Netherlands).

Following Deville & Särndal (1992) to obtain a calibration estimator for the attribute A based on the attribute B, we calculate the weights ω_k minimizing the chi-square distance

$$\chi^2 = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{3}$$

subject to the condition

$$P_B = \frac{1}{N} \sum_{k \in U} B_k = \frac{1}{N} \sum_{k \in s} \omega_k B_k \tag{4}$$

where q_k are known positive constants unrelated to d_k and $0 < P_B < 1$.

By minimizing (3) under (4), the new weights ω_k are given by:

$$\omega_k = d_k + \frac{\lambda d_k q_k B_k}{N} \tag{5}$$

where λ is the following Lagrange multiplier

$$\lambda = \frac{N^2 (P_B - \widehat{P}_{BH})}{\sum_{k \in s} d_k q_k B_k}$$

and $\widehat{P}_{BH}$ is the usual Horvitz-Thompson estimator for the attribute B.

With the calibration weights (5), assuming $\sum_{k \in s} d_k q_k B_k \neq 0$, the resulting estimator is:

$$\widehat{P}_{AW} = \frac{1}{N} \sum_{k \in s} \omega_k A_k = \widehat{P}_{AH} + \frac{(P_B - \widehat{P}_{BH})}{\sum_{k \in s} d_k q_k B_k} \cdot \sum_{k \in s} d_k q_k B_k A_k \tag{6}$$

By (4), when the estimator is applied to estimate the population proportion of B, it coincides with P_B.

2.2. Properties of the Calibration Estimator

Following Deville & Särndal (1992), it can be shown that the estimator $\widehat{P}_{AW}$ is an asymptotically unbiased estimator for P_A and its asymptotic variance is given by

$$AV(\widehat{P}_{AW}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} (d_k E_k)(d_l E_l) \tag{7}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$; $E_k = A_k - D \cdot B_k$ and $D = \dfrac{\sum_{k\in U} q_k B_k A_k}{\sum_{k\in U} q_k B_k}$

An estimator for this variance is

$$\widehat{V}(\widehat{P}_{AW}) = \sum_{k\in U}\sum_{l\in U}\frac{\Delta_{kl}}{\pi_{kl}}(d_k e_k)(d_l e_l) \tag{8}$$

with $e_k = A_k - B_k \cdot \widehat{D}$ and $\widehat{D} = \dfrac{\sum_{k\in s} d_k q_k B_k A_k}{\sum_{k\in s} d_k q_k B_k}$

Example 1. Under SRSWOR and $q_k = 1$ for all $k \in U$ the estimator $\widehat{P}_{AW}$ is:

$$\widehat{P}_{AW} = \widehat{p}_A + (P_B - \widehat{p}_B)\cdot\frac{\widehat{p}_{AB}}{\widehat{p}_B}$$

where

$$\widehat{p}_A = \frac{1}{n}\sum_{k\in s} A_k; \quad \widehat{p}_B = \frac{1}{n}\sum_{k\in s} B_k \quad \text{and} \quad \widehat{p}_{AB} = \frac{1}{n}\sum_{k\in s} A_k B_k$$

and the asymptotic variance is

$$\begin{aligned} AV(\widehat{P}_{AW}) &= V(\widehat{P}_{AVW}) = V(\widehat{p}_A) + D^2 V(\widehat{p}_B) - 2DCov(\widehat{p}_A, \widehat{p}_B) \\ &= \frac{(1-f)}{n}\frac{N}{N-1}\left[P_A Q_A + \left(\frac{P_{AB}}{P_B}\right)^2 \right. \\ &\left. \times P_B Q_B - 2\left(\frac{P_{AB}}{P_B}\right)(P_{AB} - P_A P_B)\right] \end{aligned} \tag{9}$$

where $Q_A = 1 - P_A$; $Q_B = 1 - P_B$, $P_{AB} = \dfrac{1}{N}\sum_{k\in U} A_k B_k$ and $f = \dfrac{n}{N}$. This variance can be estimated by

$$\widehat{V}(\widehat{P}_{AW}) = \frac{1-f}{n-1}\left[\widehat{p}_A\widehat{q}_A + \left(\frac{\widehat{p}_{AB}}{\widehat{p}_B}\right)^2 \cdot \widehat{p}_B\widehat{q}_B - 2\left(\frac{\widehat{p}_{AB}}{\widehat{p}_B}\right)(\widehat{p}_{AB} - \widehat{p}_A\widehat{p}_B)\right] \tag{10}$$

with $\widehat{q}_A = \dfrac{1}{n}\sum_{k\in s}(1 - A_k)$ and $\widehat{q}_B = \dfrac{1}{n}\sum_{k\in s}(1 - B_k)$

3. Alternative Calibration Estimators

The usual estimator under SRSWOR, $\widehat{p}_A$ has the following shift invariance property

$$\widehat{p}_A = 1 - \widehat{q}_A \tag{11}$$

Hence $\widehat{p}_A$ has the same performance in the estimation of P_A as the performance of $\widehat{q}_A$ in the estimation of Q_A. In general, this property is not satisfied by $\widehat{P}_{AW}$. It is easy to see that this property is fulfilled if

$$1 = \frac{1}{N}\sum_{k\in s}\omega_k \tag{12}$$

Thus, we have two ways of obtaining an estimator with the above property:

(i) By considering a calibration estimator $\widehat{Q}_{AW}$ for Q_A based on Q_B, and determining when the estimator $\widehat{P}_{AW}$ has a smaller variance than the estimator $\widehat{Q}_{AW}$ in order to define a new estimator based on these two.

(ii) By considering a calibration estimator for P_A based on P_B and Q_B because if we derive a calibration estimator that provides perfect estimates for P_B and Q_B, then:

$$1 = P_B + Q_B = \frac{1}{N}\sum_{k\in s}\omega_k B_k + \frac{1}{N}\sum_{k\in s}\omega_k(1-B_k) = \frac{1}{N}\sum_{k\in s}\omega_k$$

3.1. An Estimator Based on the Complementary: The $\widehat{P}_{AT}$ Estimator

The first alternative is developed only under SRSWOR, minimizing (3) subject to

$$Q_B = 1 - P_B = \frac{1}{N}\sum_{k\in U}\omega_k(1-B_k) \tag{13}$$

The resulting estimator, assuming $\widehat{q}_B \neq 0$, can be expressed by

$$\widehat{Q}_{AW} = \widehat{q}_A + (Q_B - \widehat{q}_B)\cdot\frac{\widehat{q}_{AB}}{\widehat{q}_B} \tag{14}$$

with

$$\widehat{q}_{AB} = \frac{1}{n}\sum_{k\in s}(1-B_k)(1-A_k)$$

In the same way as with the estimator $\widehat{P}_{AW}$ in Example 1, the asymptotic variance of $\widehat{Q}_{AW}$ is given by:

$$\begin{aligned} AV(\widehat{Q}_{AW}) = \frac{(1-f)}{n}\frac{N}{N-1}\Bigg[P_AQ_A + \left(\frac{Q_{AB}}{Q_B}\right)^2\cdot P_BQ_B \\ - 2\left(\frac{Q_{AB}}{Q_B}\right)(Q_{AB} - Q_AQ_B)\Bigg] \end{aligned} \tag{15}$$

where $Q_{AB} = \dfrac{1}{N}\displaystyle\sum_{k\in U}(1-B_k)(1-A_k)$.

An estimator $\widehat{V}(\widehat{Q}_{AW})$ for (15) can be easily defined by

$$\widehat{V}(\widehat{Q}_{AW}) = \frac{1-f}{n-1}\left[\widehat{p}_A\widehat{q}_A + \left(\frac{\widehat{q}_{AB}}{\widehat{q}_B}\right)^2 \cdot \widehat{p}_B\widehat{q}_B - 2\left(\frac{\widehat{q}_{AB}}{\widehat{q}_B}\right)(\widehat{q}_{AB} - \widehat{q}_A\widehat{q}_B)\right] \tag{16}$$

Let us now compare the asymptotic variance of $\widehat{P}_{AW}$ with the following estimator of P_A, $\widehat{P}_{AQ} = 1 - \widehat{q}_{AB}$. We have (see Appendix A) $AV(\widehat{P}_{AW}) < AV(\widehat{P}_{AQ})$ when

$$\frac{P_{AB}}{P_B} < \frac{Q_{AB}}{Q_B}. \tag{17}$$

Hence, asymptotically, a more efficient estimator for the population proportion P_A is

$$\widehat{P}_{AT} = \begin{cases} \widehat{P}_{AW} & \text{if } \dfrac{\widehat{p}_{AB}}{\widehat{p}_B} < \dfrac{\widehat{q}_{AB}}{\widehat{q}_B} \text{ or } \widehat{q}_B = 0 \\ \\ \widehat{P}_{AQ} & \text{if } \dfrac{\widehat{p}_{AB}}{\widehat{p}_B} \geq \dfrac{\widehat{q}_{AB}}{\widehat{q}_B} \text{ or } \widehat{p}_B = 0 \end{cases}$$

Note that the asymptotic variance of $\widehat{P}_{AT}$ is

$$AV(\widehat{P}_{AT}) = \begin{cases} AV(\widehat{P}_{AW}) & \text{if } \dfrac{P_{AB}}{P_B} < \dfrac{Q_{AB}}{Q_B} \\ \\ AV(\widehat{P}_{AQ}) & \text{otherwise} \end{cases} \tag{18}$$

To estimate the asymptotic variance of $\widehat{P}_{AT}$ the estimator can be defined by (10) when the following condition is satisfied

$$\frac{\widehat{p}_{AB}}{\widehat{p}_B} < \frac{\widehat{q}_{AB}}{\widehat{q}_B} \quad \text{or} \quad \widehat{q}_B = 0 \tag{19}$$

and it can be defined by (16) if

$$\frac{\widehat{p}_{AB}}{\widehat{p}_B} \geq \frac{\widehat{q}_{AB}}{\widehat{q}_B} \quad \text{or} \quad \widehat{p}_B = 0. \tag{20}$$

When in condition (20) the equality is satisfied, we have $\widehat{P}_{AW} = \widehat{P}_{AQ}$, which implies that $AV(\widehat{P}_{AW}) = AV(\widehat{P}_{AQ})$ and $\widehat{V}(\widehat{P}_{AW}) = \widehat{V}(\widehat{P}_{AQ})$.

The reason for defining $\widehat{P}_{AT}$ is to obtain an estimator with the property $\widehat{P}_{AT} = 1 - \widehat{Q}_{AT}$. Accordingly, the estimator $\widehat{Q}_{AT}$ is defined by

$$\widehat{Q}_{AT} = \begin{cases} \widehat{Q}_{AW} & \text{if } \dfrac{\widehat{q}_{AB}}{\widehat{q}_B} < \dfrac{\widehat{p}_{AB}}{\widehat{p}_B} \text{ or } \widehat{p}_B = 0 \\ \\ 1 - \widehat{P}_{AW} & \text{f } \dfrac{\widehat{q}_{AB}}{\widehat{q}_B} \geq \dfrac{\widehat{p}_{AB}}{\widehat{p}_B} \text{ or } \widehat{q}_B = 0 \end{cases}$$

and

$$1-\widehat{Q}_{AT}=\begin{cases}1-\widehat{Q}_{AW} & \text{if } \dfrac{\widehat{q}_{AB}}{\widehat{q}_B}<\dfrac{\widehat{p}_{AB}}{\widehat{p}_B} \text{ or } \widehat{p}_B=0\\[2ex] \widehat{P}_{AW} & \text{if } \dfrac{\widehat{q}_{AB}}{\widehat{q}_B}\geq\dfrac{\widehat{p}_{AB}}{\widehat{p}_B} \text{ or } \widehat{q}_B=0\end{cases}$$

Since $\widehat{P}_{AW}=1-\widehat{Q}_{AW}$ when $\dfrac{\widehat{q}_{AB}}{\widehat{q}_B}=\dfrac{\widehat{p}_{AB}}{\widehat{p}_B}$, we have $1-\widehat{Q}_{AT}=\widehat{P}_{AT}$.

Another important question relative to the estimator $\widehat{P}_{AT}$ is: for low proportions, estimators such as $\widehat{P}_{AW}$ cannot be calculated when $\widehat{p}_B=0$. The estimator $\widehat{P}_{AT}$ does not present this problem, because if $\widehat{p}_B=0$ the estimator $\widehat{P}_{AT}=1-\widehat{Q}_{AW}$ and since $\widehat{q}_B=1$ we have

$$\widehat{P}_{AT}=1-\widehat{Q}_{AW}=1-\widehat{q}_A+(Q_B-\widehat{q}_B)\cdot\widehat{q}_{AB}$$

Then the estimator $\widehat{P}_{AT}$ can be obtained for low proportions.

Finally, the estimator $\widehat{P}_{AT}$ has the following drawback: by (9); (15) and (18), its asymptotic behaviour is worse than that of the usual estimator when

$$P_{AB}-P_AP_B=Q_{AB}-Q_AQ_B<0 \tag{21}$$

Thus, if condition (21) occurs, we propose to use the attribute $\bar{B}=B^c$ because

$$P_{A\bar{B}}-P_AP_{\bar{B}}=\frac{1}{N}\sum_{k\in U}A_k(1-B_k)-P_A(1-P_B)=$$

$$=P_A-P_{AB}-P_A+P_AP_B=-P_{AB}+P_AP_B>0$$

Therefore, with attribute $\bar{B}$, the above problem, when (21) occurs, is solved.

3.2. An Optimal Estimator: The $\widehat{P}_{A\alpha}$ Estimator

Another way to improve the asymptotic behaviour is as follows: let us define

$$\widehat{P}_{A\alpha}=\alpha\cdot\widehat{P}_{AW}+(1-\alpha)(1-\widehat{Q}_{AW}) \tag{22}$$

and take the value α that minimizes the variance.

The minimum variance of $\widehat{P}_{A\alpha}$ is given by (see Appendix B)

$$AV(\widehat{P}_{A\alpha})=G\cdot\left(P_AQ_A-\frac{\phi^2}{P_BQ_B}\right)$$

with

$$G=\frac{1-f}{n}\left(\frac{N}{N-1}\right)\quad\text{and}\quad\phi=P_{AB}-P_AP_B=Q_{AB}-Q_AQ_B$$

and this is achieved where

$$\alpha = \frac{(P_{AB} - P_A P_B) - P_B Q_{AB}}{P_B Q_B \left(\dfrac{P_{AB}}{P_B} - \dfrac{Q_{AB}}{Q_B} \right)} = \frac{P_{AB} - P_B}{P_B \left(\dfrac{P_{AB}}{P_B} - \dfrac{Q_{AB}}{Q_B} \right)}. \tag{23}$$

The estimator $\widehat{P}_{A\alpha}$ has the desirable property

$$\widehat{P}_{A\alpha} = 1 - \widehat{Q}_{A\alpha}$$

because if we define

$$\widehat{Q}_{A\alpha} = \beta \cdot \widehat{Q}_{AW} + (1 - \beta)(1 - \widehat{P}_{AW})$$

then

$$1 - \widehat{Q}_{A\alpha} = (1 - \beta) \cdot \widehat{P}_{AW} + \beta \cdot (1 - \widehat{Q}_{AW}) = \alpha \cdot \widehat{P}_{AW} + (1 - \alpha)(1 - \widehat{Q}_{AW})$$

with $\alpha = 1 - \beta$. Since $AV(1 - \widehat{Q}_{A\alpha}) = AV(\widehat{Q}_{A\alpha})$, we find that minimizing the variance of $\widehat{Q}_{A\alpha}$ with respect to β is equal to minimizing the variance $1 - \widehat{Q}_{A\alpha}$ with respect to α, and consequently α is again given by (44) and $\widehat{P}_{A\alpha} = 1 - \widehat{Q}_{A\alpha}$.

The $\widehat{P}_{A\alpha}$ estimator presents the following disadvantages:

- It cannot be calculated if $\widehat{p}_B = 0$ or $\widehat{q}_B = 0$
- The optimum value α given by (44) depends on theoretical variances and covariances, which are generally unknown.

With respect to the second question, the value α can be easily estimated when the sample is drawn by

$$\tilde{\alpha} = \frac{\widehat{p}_{AB} - \widehat{p}_B}{\widehat{p}_B \left(\dfrac{\widehat{p}_{AB}}{\widehat{p}_B} - \dfrac{\widehat{q}_{AB}}{\widehat{q}_B} \right)}$$

Therefore, we obtain the following estimator

$$\tilde{P}_{A\alpha} = \tilde{\alpha} \widehat{P}_{AW} + (1 - \tilde{\alpha})(1 - \widehat{Q}_{AW})$$

The estimator $\tilde{P}_{A\alpha}$ also has the desired property (11).

3.3. An estimator that Calibrate in P_B and Q_B at the Same Time: The $\widehat{P}_{AR}$ Estimator

When using the population size and the population proportion of B, the auxiliary information is the same as in the case of a post-stratified estimator. The second way (ii) to obtain the property (11), based on an estimator for P_A calibration with P_B and Q_B, can be developed in any sampling design. To do so, we must minimize the distance (3) under the conditions

$$\begin{cases} P_B = \dfrac{1}{N}\displaystyle\sum_{k\in s}\omega_k B_k \\ \\ Q_B = \dfrac{1}{N}\displaystyle\sum_{k\in s}\omega_k (1-B_k) \end{cases} \tag{24}$$

The calibration weights with this new calibration process are

$$\omega_k = d_k + \frac{N(P_B - \widehat{P}_{BH})}{\displaystyle\sum_{k\in s} d_k q_k B_k} d_k q_k B_k + \frac{N(Q_B - \widehat{Q}_{BH})}{\displaystyle\sum_{k\in s} d_k q_k (1-B_k)} d_k q_k (1-B_k) \tag{25}$$

and the resulting estimator is

$$\widehat{P}_{AR} = \widehat{P}_{AH} + (P_B - \widehat{P}_{BH}) \cdot \widehat{B}_1 + (Q_B - \widehat{Q}_{BH}) \cdot \widehat{B}_2 \tag{26}$$

where

$$\widehat{B}_1 = \frac{\displaystyle\sum_{k\in s} d_k q_k B_k A_k}{\displaystyle\sum_{k\in s} d_k q_k B_k} \quad \text{and} \quad \widehat{B}_2 = \frac{\displaystyle\sum_{k\in s} d_k q_k (1-B_k) A_k}{\displaystyle\sum_{k\in s} d_k q_k (1-B_k)}$$

Therefore, when

$$\begin{cases} \displaystyle\sum_{k\in s} d_k q_k B_k = 0 \\ \text{or} \\ \displaystyle\sum_{k\in s} d_k q_k (1-B_k) = 0 \end{cases} \tag{27}$$

the estimator (26) cannot be obtained. This problem is solved as follows: since the two conditions are mutually exclusive, if

$$\sum_{k\in s} d_k q_k B_k = 0$$

we can calibrate the estimator using only the attribute $\bar{B}$. On the other hand, if

$$\sum_{k\in s} d_k q_k (1-B_k) = 0$$

the estimator $\widehat{P}_{AR}$ can be developed only with the attribute B. Thus, the estimator $\widehat{P}_{AR}$ is well defined.

To prevent this article from becoming excessively long, in the same way as with the estimator $\widehat{P}_{AW}$, the asymptotic variance of (26) is given by

$$AV(\widehat{P}_{AR}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k U_k)(d_l U_l) \tag{28}$$

with $U_k = A_k - B_1 \cdot B_K - B_2 \cdot (1 - B_k)$ and

$$B_1 = \frac{\sum_{k \in U} q_k B_k A_k}{\sum_{k \in U} q_k B_k} \quad ; \quad B_2 = \frac{\sum_{k \in U} q_k (1 - B_k) A_k}{\sum_{k \in U} q_k (1 - B_k)}$$

(28) is determined using the following estimator

$$\widehat{V}(\widehat{P}_{AR}) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} (d_k u_k)(d_l u_l) \tag{29}$$

where $u_k = A_k - \widehat{B}_1 \cdot B_k - \widehat{B}_2 \cdot (1 - B_k)$

Example 2. Under SRSWOR and $q_k = 1$ for all $k \in U$, the estimator $\widehat{P}_{AR}$ can be expressed by

$$\widehat{P}_{AR} = \widehat{p}_A + (P_B - \widehat{p}_B) \frac{\widehat{p}_{AB}}{\widehat{p}_B} + (Q_B - \widehat{q}_B) \frac{\widehat{p}_{A\bar{B}}}{q_B} \tag{30}$$

with

$$\widehat{p}_{A\bar{B}} = \frac{1}{n} \sum_{k \in s} A_k (1 - B_k).$$

In the same way as before with the estimator $\widehat{P}_{AW}$, the asymptotic variance of $\widehat{P}_{AR}$ is (see Appendix C).

$$\begin{aligned} AV(\widehat{P}_{AR}) &= \frac{(1-f)}{n} \left(\frac{N}{N-1} \right) \left[P_A Q_A + \frac{(P_{AB} - P_A P_B)^2}{P_B Q_B} - 2 \frac{(P_{AB} - P_A P_B)^2}{P_B Q_B} \right] \\ &= G \cdot \left[P_A Q_A - \frac{(P_{AB} - P_A P_B)^2}{P_B Q_B} \right] = G \cdot \left(P_A Q_A - \frac{\phi^2}{P_B Q_B} \right) \end{aligned} \tag{31}$$

To estimate (45) the following estimator is defined

$$\widehat{V}(\widehat{P}_{AR}) = \frac{(1-f)}{n-1} \left[\widehat{p}_A \widehat{q}_A - \frac{\widehat{\phi}^2}{\widehat{p}_B \widehat{q}_b} \right] \tag{32}$$

with $\widehat{\phi} = \widehat{p}_{AB} - \widehat{p}_A \widehat{p}_B$.

Thus, the estimator $\widehat{P}_{AR}$ has the same asymptotic variance as the estimator $\widehat{P}_{A\alpha}$. Then, by (45), under SRSWOR the estimator $\widehat{P}_{AR}$ is always more efficient than the estimators $\widehat{p}_A$ and $\widehat{P}_{AT}$.

Note that the proposed estimator is essentially a post-stratified estimator and, in this sense, is not new. Here, we look at it from a different point of view. In a practical situation, the estimator $\widehat{P}_{AR}$ is preferred the estimator $\widehat{P}_{A\alpha}$, since $\widehat{P}_{AR}$ does not need estimation of any unknown population quantity. The case of estimator $\widehat{P}_{A\alpha}$ requires estimating value α, which is generally unknown.

4. Extension to Multivariate Auxiliary Information

The previous section considered only an auxiliary attribute B; let us now assume that the study attribute A is related to J auxiliary attributes $B_1, \dots B_J$.

To develop the usual way of incorporating the information provided by J attributes in the estimation of $\widehat{P}_A$ with calibration techniques, we consider a new weight ω_k subject to the following conditions

$$P_j = \frac{1}{N}\sum_{k\in U} B_{jk} = \frac{1}{N}\sum_{k\in s}\omega_k B_{jk} \qquad j = 1, \dots, J. \tag{33}$$

Next, we denote

$$P^{'} = (P_1, \dots, P_J); \quad (\widehat{P}_H)^{'} = (\widehat{P}_{1H}, \dots, \widehat{P}_{JH}) \text{ and } \quad (B_k)^{'} = (B_{1k}, \dots, B_{Jk})$$

where

$$\widehat{P}_{jH} = \frac{1}{N}\sum_{k\in s} d_k B_{jk} \qquad j = 1, \dots, J.$$

By T we denote the following matrix

$$T = \sum_{k\in s} d_k q_k B_k (B_k)'.$$

With the minimization of (3) under the P conditions given by (33), the new weights obtained are:

$$\omega_k = d_k + d_k q_k N (P - P_H)^{'} T^{-1} B_k. \tag{34}$$

The calibration estimator based on (34) is given by:

$$\widehat{P}_{AWM} = \widehat{P}_{AH} + (P - P_H)^{'}\widehat{H} \tag{35}$$

with $\widehat{H} = T^{-1}\sum_{k\in s} d_k q_k B_k A_k$.

Note that the weights (34) and the estimator $\widehat{P}_{AWM}$ cannot be obtained if the matrix T is singular.

Following Rueda *et al.* (2007a), the asymptotic variance of the estimator $\widehat{P}_{AWM}$ is

$$AV(\widehat{P}_{AWM}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}(d_k Z_k)(d_l Z_l) \tag{36}$$

with $Z_k = A_k - (B_k)'H$ where

$$H = \Big(\sum_{k\in U} q_k B_k (B_k)^{'}\Big)^{-1}\Big(\sum_{k\in U} q_k B_k A_k\Big).$$

The asymptotic variance (36) can be estimated by

$$\widehat{V}(\widehat{P}_{AWM}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}(d_k z_k)(d_l z_l) \tag{37}$$

with $z_k = A_k - (B_k)^{'}\widehat{H}$

Example 3. Under SRSWOR and $q_k = 1$, when only the two auxiliary attributes B and C are considered, the matrix T can be expressed by

$$T = \begin{pmatrix} N\widehat{p}_B & N\widehat{p}_{BC} \\ N\widehat{p}_{BC} & N\widehat{p}_C \end{pmatrix}.$$

Consequently, $|T| = N^2\Big(\widehat{p}_B \cdot \widehat{p}_A - (\widehat{p}_{BC})^2\Big)$ and

$$T^{-1} = \frac{\begin{pmatrix} N\widehat{p}_C & -N\widehat{p}_{BC} \\ -N\widehat{p}_{BC} & N\widehat{p}_B \end{pmatrix}}{|T|} = \frac{\begin{pmatrix} \widehat{p}_C & -\widehat{p}_{BC} \\ -\widehat{p}_{BC} & \widehat{p}_B \end{pmatrix}}{N(\widehat{p}_B \cdot \widehat{p}_C - (\widehat{p}_{BC})^2)}.$$

Next, we have

$$\sum_{k\in s} d_k q_k B_k A_k = \begin{pmatrix} N \cdot \widehat{p}_{AB} \\ N \cdot \widehat{p}_{AC} \end{pmatrix}$$

and

$$\widehat{H} = T^{-1} \cdot \sum_{k\in s} d_k q_k B_k A_k = \frac{\begin{pmatrix} \widehat{p}_C\widehat{p}_{AB} - \widehat{p}_{AB} \cdot \widehat{p}_{BC} \\ \widehat{p}_B\widehat{p}_{AC} - \widehat{p}_{AB} \cdot \widehat{p}_{BC} \end{pmatrix}}{(\widehat{p}_B \cdot \widehat{p}_C - (\widehat{p}_{BC})^2)}.$$

Thus, the estimator $\widehat{P}_{AWM}$, under SRSWOR with two auxiliary attributes, is

$$\begin{aligned} \widehat{P}_{AWM} = \widehat{p}_A &+ (P_B - \widehat{p}_B)\left[\frac{\widehat{p}_C\widehat{p}_{AB} - \widehat{p}_{AC}\widehat{p}_{BC}}{(\widehat{p}_B\widehat{p}_C - (\widehat{p}_{BC})^2)}\right] \\ &+ (P_C - \widehat{p}_C) \cdot \left[\frac{\widehat{p}_B\widehat{p}_{AC} - \widehat{p}_{AB}\widehat{p}_{BC}}{(\widehat{p}_B \cdot \widehat{p}_C - (\widehat{p}_{BC})^2)}\right] \end{aligned} \tag{38}$$

Now, if we denote

$$A_1 = \frac{P_C P_{AB} - P_{AC}P_{BC}}{P_B P_C - (P_{BC})^2} \quad \text{and} \quad A_2 = \frac{P_B P_{AC} - P_{AB}P_{BC}}{P_B P_C - (P_{BC})^2}$$

The asymptotic variance of the estimator $\widehat{P}_{AWM}$ is

$$\begin{aligned} AV(\widehat{P}_{AWM}) &= \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\Big[P_AQ_A + A_1^2 \cdot P_BQ_B + A_2^2 \cdot P_CQ_C \\ &\quad + 2A_1A_2(P_{BC} - P_BP_C) - 2A_1(P_{AB} - P_AP_B) \\ &\quad - 2A_2(P_{AC} - P_AP_C)\Big] = \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\Big[P_AQ_A \\ &\quad + A_1^2 \cdot P_B + A_2^2 \cdot P_C + 2A_1A_2P_{BC} \\ &\quad - (A_1P_B + A_2P_C)^2 - 2A_1(P_{AB} - P_AP_B) \\ &\quad - 2A_2(P_{AC} - P_AP_C)\Big]. \end{aligned} \tag{39}$$

To determine (39) we use the following estimator:

$$\begin{aligned} \widehat{V}(\widehat{P}_{AWM}) &= \frac{(1-f)}{n-1}\Big[\widehat{p}_A\widehat{q}_A + (\widehat{A}_1)^2 \cdot \widehat{p}_B + (\widehat{A}_2)^2 \cdot \widehat{p}_C + 2(\widehat{A}_1)(\widehat{A}_2)\widehat{p}_{BC} \\ &\quad - ((\widehat{A}_1)\widehat{p}_B + \widehat{A}_2\widehat{p}_C)^2 - 2(\widehat{A}_1)(\widehat{p}_{AB} - \widehat{p}_A\widehat{p}_B) - 2(\widehat{A}_2)(\widehat{p}_{AC} - \widehat{p}_A\widehat{p}_C)\Big]. \end{aligned} \tag{40}$$

5. Simulation study

A limited study was carried out to investigate the design-based finite sample performance of the proposed estimators in comparison with that of conventional estimators.

5.1. Simulated data

The estimators were evaluated using 15 simulated populations with a population size $N = 1000$. These populations were generated as a random sample of 1000 units from a Bernoulli distribution with parameter $P_A = \{0.5, 0.75, 0.9\}$, and the attributes of interest were thus achieved with the aforementioned population proportions. Auxiliary attributes were also generated, using the same distribution, but a given proportion of values were randomly changed so that Cramer's V coefficient between the attribute of interest and the auxiliary attribute took the values 0.5, 0.6, 0.7, 0.8 and 0.9.

For each simulation, 1000 samples with sizes n= 50, 75, 100 and 125, were selected under SRSWOR to compare the estimators:

(1) the Horvitz-Thompson estimator $\widehat{P}_{AH}$

(2) the ratio estimator $\widehat{P}_{Aratio}$ (see Rueda *et al.*, 2011),

(3) $\widehat{P}_{AW}$ estimator (W-calibrated),

(4) $\widehat{P}_{AT}$ estimator (T-calibrated),

(5) $\tilde{P}_{A\alpha}$ estimator with α estimated (α-calibrated),

(6) $\widehat{P}_{AR}$ estimator (R-calibrated),

(7) the multivariate ratio estimator $\widehat{P}_{AratioM}$ (see Rueda *et al.*, 2011) (Multiple ratio), and

(8) the multivariate calibration estimator $\widehat{P}_{AWM}$ (Multiple calibrated).

in terms of relative bias (RB) and relative efficiency with respect to the ratio estimator (RE), where

$$RB = \frac{E[\widetilde{p}_A] - P_A}{P_A}, \qquad RE = \frac{MSE[\widetilde{p}_A]}{MSE[\widehat{P}_{Aratio}]},$$

$\widetilde{p}_A$ is a given estimator and $E[\cdot]$ and $MSE[\cdot]$ denote, respectively, the empirical mean and the mean square error. Values of RE less than 1 indicate that $\widetilde{p}_A$ is more efficient than $\widehat{P}_{Aratio}$.

The results derived from this simulation study gave values of RB within a reasonable range. All the calibration estimators produced absolute relative bias values of less than 1% except in case P_A=0.9 and ϕ=0.9. Univariate ratio estimator produced the highest bias values, especially for small sample sizes.

Figures 1, 2 and 3 show the values of RE for the various populations.

These figures show:

- The ratio estimator performs poorly when there is little association between the variables. When $\phi = 0.5$ this estimator is worse than the Horvitz-Thompson estimator. Even when $\phi = 0.6$ as is sometimes the case ($P_A = 0.75$ and $P_A = 0.9$) the ratio estimator has a large MSE. In populations with a large ϕ this problem does not arise.

- With large ϕ values, all the estimators that use auxiliary information produce good results: for $\phi \geq 0.7$ all calibration and ratio estimators are better than the Horvitz-Thompson estimator. It is also seen that as ϕ increases, all the estimators achieve greater precision, which is particularly marked for very high proportions.

- Of all the calibration estimators, the first one proposed $\widehat{P}_{AW}$ has the lowest degree of efficiency. Although it performs better than the Horvitz-Thompson estimator on most occasions (except when P_A=0.9, ϕ=0.5 and 0.6) the others produce a smaller MSE.

- The $\widehat{P}_{AT}$, $\tilde{P}_{A\alpha}$ and $\widehat{P}_{AR}$ estimators perform very well in all cases. For high proportions (P_A=0.75 and 0.9) the efficiency of the estimators is fairly similar; only in the case of P_A=0.5 and small values of ϕ is there a noticeable difference between them, in terms of efficiency. In these cases, the best results are achieved by the $\widehat{P}_{AR}$ estimators that calibrate in P_B and Q_B at the same time.

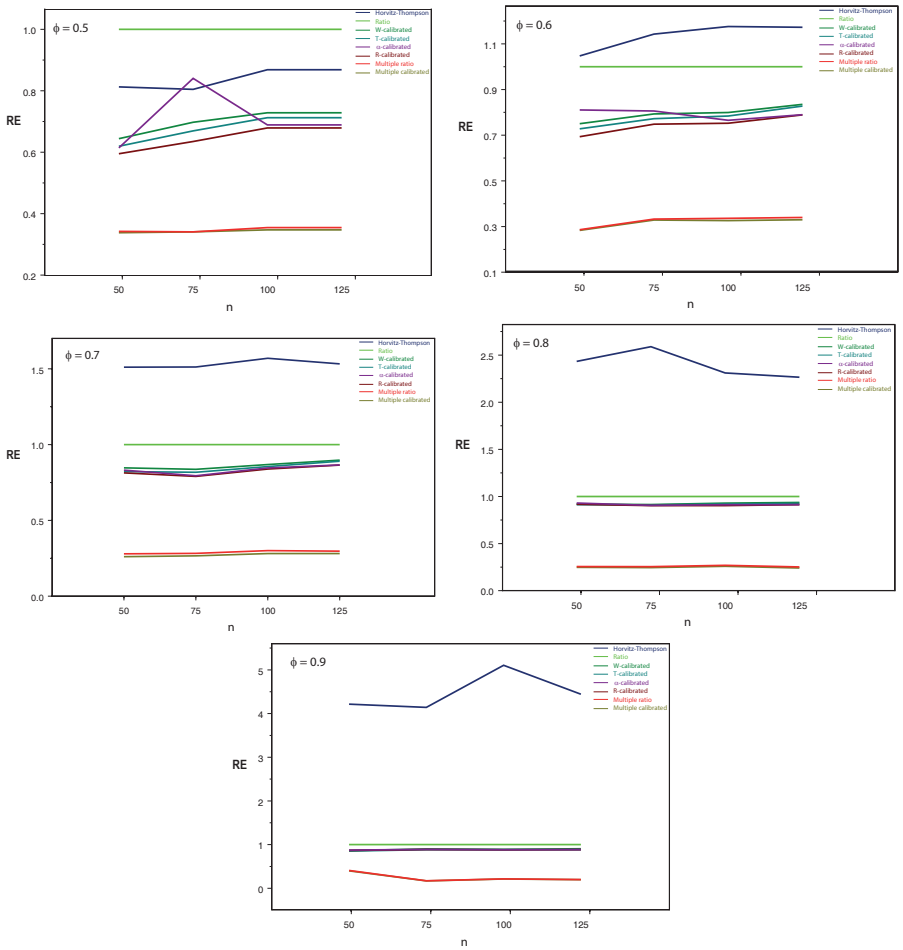

FIGURE 1: Empirical relative efficiency (RE) of the different estimators for the simulated populations when $P_A = 0.5$ and $\phi = 0.5, 0.6, 0.7, 0.8, 0.9$.

- The sample size does not produce a clear effect on the behaviour of the estimators; in some cases, as the sample size increases, the efficiency of the estimators increases, while in others, it decreases (as when $\phi = 0.9$ and P_A=0.9).
- Ratio and calibration estimators using two auxiliary variables always have a lower RE than those using a single auxiliary variable. For P_A=0.5 and $0.5 \leq \phi \leq 0.7$ the multiple calibration estimator is slightly more efficient than the multiple ratio estimator. For P_A=0.75 and 0.9 both estimators have similar levels of efficiency.

Ratio estimation is usually known to work well when the variables (auxiliary and of interest) are positively correlated. In this case it is applied to 0-1 variables, so that a positive association is expected for the method to work (higher frequen-

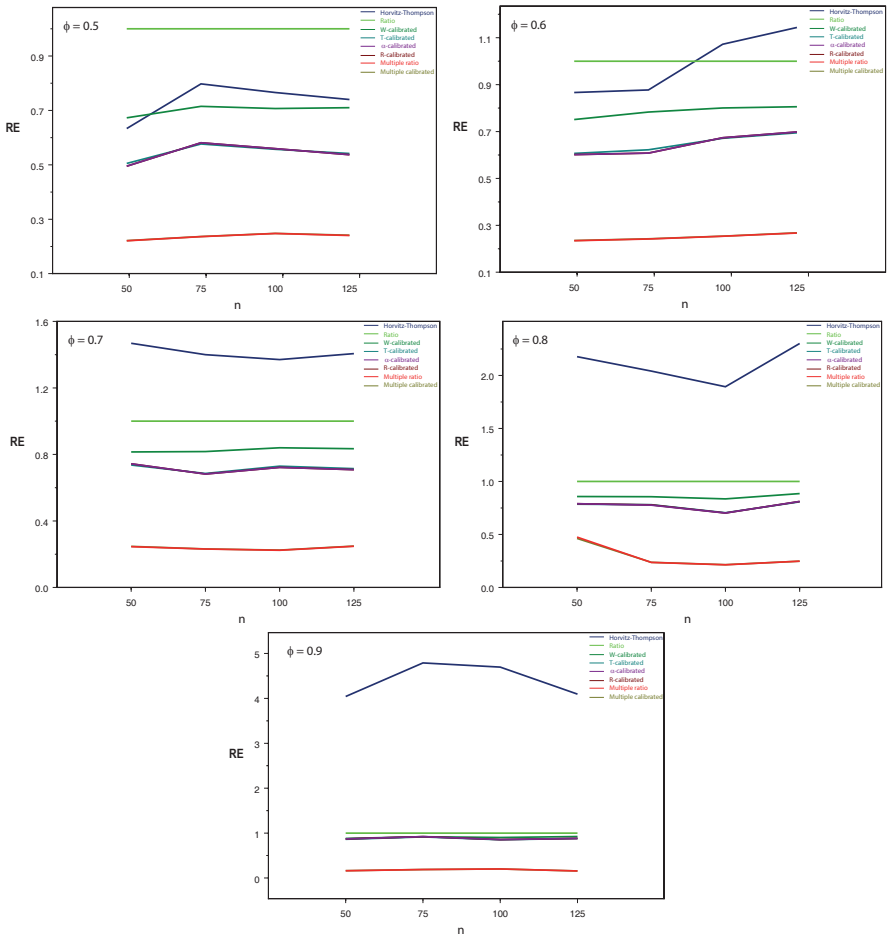

FIGURE 2: Empirical relative efficiency (RE) of the different estimators for the simulated populations when P_A=0.75 and $\phi = 0.5, 0.6, 0.7, 0.8, 0.9$.

cies for the A=1; B=1 (A=0; B=0) cases instead of the 0;1 (1;0) cases). From this study we can conclude that the association between the variables is the most important factor influencing the behaviour of ratio and also of calibration estimators. As expected, as ϕ increases the MSE of the calibrated estimators decreases. Even for moderate values of ϕ the calibration estimators improve considerably, in terms of efficiency, on the Horvitz-Thompson estimator. The behaviour of calibration estimators is similar for small proportions, whereas when the proportion approaches 1 there are larger differences among the proposed calibration estimators. Hovewer, it is not an easy task to quantify how much association is needed for a good improvement in terms of efficiency, or when too small that it becomes harmful to introduce extra variables in the calibration constraints.

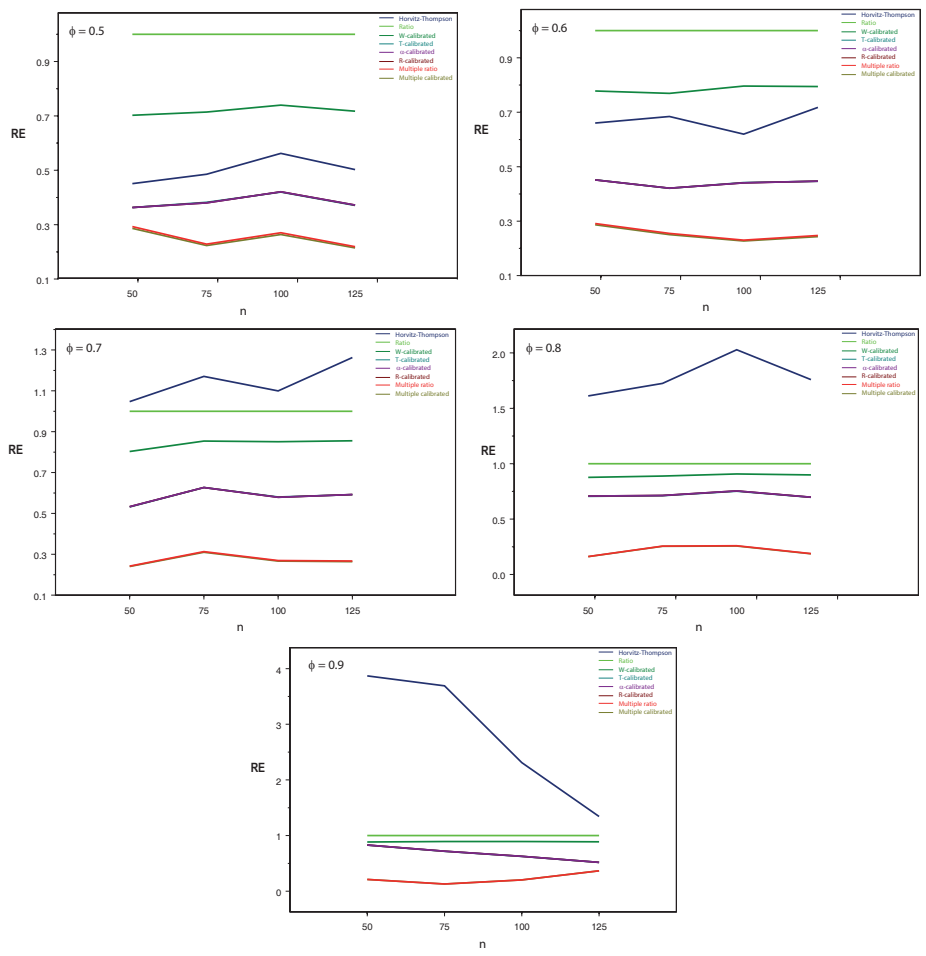

FIGURE 3: Empirical relative efficiency (RE) of the different estimators for the simulated populations when P_A=0.9 and $\phi = 0.5, 0.6, 0.7, 0.8, 0.9$.

5.2. Real Data

In this section we apply some proposed estimators to data obtained in a survey on perceptions of immigration in a certain region in Spain. A sample of size $n = 1919$ was selected from a population with size $N = 4982920$, using stratified random sampling.

Among topics of interest in the survey was estimating the percentage of citizens who believe that the authorities should make immigration more difficult by imposing stricter conditions. The auxiliary variable available is the respondent's gender. This variable was observed in the sample and the totals are known for each province (stratum).

Three main variables are included in this study, related to "goodness of immigration" and "amount of immigration". The main variables are the answers to the following questions:

- in general, do you think that for Andalusia, immigration is . . . ? C1-Very bad, C2 Bad, C3 Neither good nor bad, C4 Good, C5 Very good,
- and in relation to the number of immigrants currently living in Andalusia, do you think there are . . . ? C1-Too many, C2-A reasonable number, C3-Too few.

In this simulation study, we use the sample as population and we draw stratified random samples of size $n = 240$ with proportional allocation (eight stratum). Relative efficiency with respect to the ratio estimator is computed, as in the previous case, for compared estimators over 1000 simulation runs. We computed this relative efficiency for each category of the main variable (5 categories in the first case, and 3 in the second case) and the average over categories is also computed. At the same time, confidence intervals based on a normal distribution and using proposed estimated variances are computed for each proportion. Table 1 shows the average length of the 1000 simulation runs for each category and the average over categories. In a similar way, the empirical coverage of the confidence estimation is computed.

Tables 1 and 2 show that, from an efficiency standpoint $\widehat{P}_{A\alpha}$ is best. Looking the average length of confidence intervals for the proportion in each category, and the average over categories, the best estimator is $\widehat{P}_{AR}$, but the optimal $\widehat{P}_{A\alpha}$ has very similar results. However, the empirical coverage of confidence intervals is closer to the nominal level when the optimal estimator is used.

6. Application

IESA, the Institute for Advanced Social Studies conducted a survey between January 14th and February 13th, 2011 on the perception of culture in the Spanish region of Andalusia (Barometer of Culture of Andalusia - BACU). It is based on a sample drawn from a landline phone frame (N = 5,064,304).

From this frame a stratified random sample without replacement of dimension $n = 641$ was selected, where strata were made up by eight geographical regions. Strata population sizes are $N_h = (274128, 919124, 463008, 502450, 237183, 441936, 856392, 1370083)$ and the corresponding strata sample sizes are $n_h = (53, 99, 66, 62, 38, 49, 131, 143)$.

Among several topics of interest in the survey, is the interest to estimate perception of their culture in relation to European citizens. An auxiliary variable available is gender which totals are known for each strata. From Table 3 we observe that the $\widehat{P}_{A\alpha}$ estimator produces the best confidence intervals.

Table 1: Relative efficiency with respect to the ratio estimator; length and empirical coverage of 95% confidence level estimation of proportions. Main variable: "goodness of immigration". Stratified random sampling.

Estimator	c1	c2	c3	c4	c5	avg
			Relative Efficiency			
$\widehat{P}_{AW}$	0.915	0.607	0.921	0.786	0.986	0.843
$\widehat{P}_{AT}$	0.917	0.590	0.923	0.795	1.003	0.846
$\widehat{P}_{A\alpha}$	0.911	0.490	0.901	0.711	0.986	0.800
$\widehat{P}_{AR}$	0.945	0.650	0.894	0.742	1.021	0.850
			Length			
$\widehat{P}_{AW}$	0.082	0.131	0.053	0.112	0.037	0.083
$\widehat{P}_{AT}$	0.081	0.128	0.053	0.112	0.037	0.082
$\widehat{P}_{A\alpha}$	0.080	0.118	0.053	0.108	0.037	0.079
$\widehat{P}_{AR}$	0.078	0.115	0.051	0.105	0.036	0.077
			Coverage			
$\widehat{P}_{AW}$	0.935	0.950	0.941	0.946	0.921	0.938
$\widehat{P}_{AT}$	0.940	0.947	0.935	0.948	0.913	0.936
$\widehat{P}_{A\alpha}$	0.936	0.952	0.938	0.948	0.920	0.939
$\widehat{P}_{AR}$	0.923	0.938	0.937	0.931	0.900	0.925

Table 2: Relative efficiency with respect to the ratio estimator; length and empirical coverage of 95% confidence level estimation of proportions. Main variable: "amount of immigration". Stratified random sampling.

Estimator	c1	c2	c3	avg
		Relative Efficiency		
$\widehat{P}_{AW}$	0.644	0.628	0.972	0.748
$\widehat{P}_{AT}$	0.638	0.614	0.981	0.744
$\widehat{P}_{A\alpha}$	0.547	0.522	0.974	0.681
$\widehat{P}_{AR}$	0.566	0.830	0.988	0.795
		Length		
$\widehat{P}_{AW}$	0.125	0.132	0.052	0.103
$\widehat{P}_{AT}$	0.123	0.128	0.052	0.101
$\widehat{P}_{A\alpha}$	0.116	0.118	0.052	0.095
$\widehat{P}_{AR}$	0.113	0.116	0.050	0.093
		Coverage		
$\widehat{P}_{AW}$	0.952	0.945	0.930	0.942
$\widehat{P}_{AT}$	0.949	0.938	0.924	0.937
$\widehat{P}_{A\alpha}$	0.953	0.947	0.928	0.943
$\widehat{P}_{AR}$	0.940	0.925	0.909	0.924

TABLE 3: Estimated proportion ($\hat{P}$), lower bound (LB), upper bound (UB) and length (L) of a 95% confidence interval under stratified random sampling.

Do you think that in Andalusia the cultural level, compared to the European Union, is...?				
Estimator	PROP	LB	UB	LEN
	Much lower			
$\widehat{P}_{AW}$	12.899	11.342	14.456	3.114
$\widehat{P}_{AT}$	12.899	11.342	14.456	3.114
$\widehat{P}_{AR}$	12.913	11.380	14.447	3.067
$\widehat{P}_{A\alpha}$	13.235	11.726	14.744	3.018
	lower			
$\widehat{P}_{AW}$	46.126	43.563	48.689	5.126
$\widehat{P}_{AT}$	46.126	43.563	48.689	5.126
$\widehat{P}_{AR}$	46.182	43.881	48.483	4.602
$\widehat{P}_{A\alpha}$	45.241	43.030	47.452	4.422
	Equal			
$\widehat{P}_{AW}$	5.085	4.074	6.095	2.021
$\widehat{P}_{AT}$	5.085	4.074	6.095	2.021
$\widehat{P}_{AR}$	5.092	4.092	6.093	2.001
$\widehat{P}_{A\alpha}$	5.268	4.274	6.262	1.988
	Higher			
$\widehat{P}_{AW}$	29.838	27.649	32.027	4.378
$\widehat{P}_{AT}$	29.838	27.649	32.027	4.378
$\widehat{P}_{AR}$	29.867	27.756	31.978	4.222
$\widehat{P}_{A\alpha}$	29.899	27.864	31.934	4.070
	Much higher			
$\widehat{P}_{AW}$	2.372	1.702	3.042	1.340
$\widehat{P}_{AT}$	2.372	1.702	3.042	1.340
$\widehat{P}_{AR}$	2.374	1.704	3.043	1.339
$\widehat{P}_{A\alpha}$	2.603	1.936	3.270	1.334

7. Conclusions

In practice, it is important to make the best possible use of available auxiliary information so as to obtain the most efficient estimator possible.

When a proportion can be estimated in the case of complete auxiliary information (i.e., when auxiliary information is available at the population level for each unit) it is possible to consider estimators that use the logistic regression model (Duchesne, 2003, Wu and Sitter, 2001), as an improvement on the simple estimator. When there is merely a rearrangement of the population proportion of an attribute with respect to the study variable, then traditional indirect methods

such as the ratio by Rueda *et al.* (2011) or the calibration studied in this work can be applied.

We have studied four calibration estimators for the proportion which are simple to calculate from standard calibration packages and can give rise to considerable increases in the precision achieved, as illustrated by the theoretical results reported here and by the simulation performed. The proposed estimators $\widehat{P}_{AW}$ and $\widehat{P}_{AR}$ can be obtained from any arbitrary sampling design, whereas $\widehat{P}_{A\alpha}$ and $\widehat{P}_{AT}$ estimators are defined under SRSWOR. However, the extension to a stratified random sampling is straightforward.

Confidence intervals based on the estimated variances of the studied calibration estimators is also investigated through a limited simulation study, under a more realistic survey (stratified random sampling) using real data. $\widehat{P}_{A\alpha}$ and $\widehat{P}_{AR}$ have good properties in confidence estimation, and in some sense (a balance between length and coverage) the optimal estimator $\widehat{P}_{A\alpha}$ provides the best results.

Acknowledgements

The authors thank the Editor and the two anonymous referees of this journal for their helpful comments on an earlier version of this paper. This study was partially supported by Ministerio de Educación y Ciencia (grant MTM2012-35650, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía).

[Received: October 2013 — Accepted: November 2014]

References

Arnab, R., Shangodoyin, D. K. & Singh, S. (2010), 'Variance estimation of a generalized regression predictor', *Journal of the Indian Society of Agricultural Statistics* **64**(2), 273–288.

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

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

Farrell, P. & Singh, S. (2005), 'Model-assisted higher-order calibration of estimators of variance', *Australian & New Zealand Journal of Statistics* **47**(3), 375–383.

Harms, T. & Duchesne, P. (2006), 'On calibration estimation for quantiles', *Survey Methodology* **32**, 37–52.

Rueda, M., Martínez-Puertas, S., Martínez-Puertas, H. & Arcos, A. (2007), 'Calibration methods for estimating quantiles', *Metrika* **66**(3), 355–371.

Rueda, M., Martínez, S., Martínez, H. & Arcos, A. (2007), 'Estimation of the distribution function with calibration methods', *Journal of Statistical Planning and Inference* **137**(2), 435–448.

Rueda, M., Muñoz, J. F., Arcos, A., Álvarez, E. & Martínez, S. (2011), 'Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies', *Journal of Biopharmaceutical Statistics* **21**(3), 526–54.

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

Singh, H. P., Singh, S. & Kozak, M. (2008), 'A family of estimators of finite-population distribution functions using auxiliary information', *Acta Applicandae Mathematicae* **104**(2), 115–130.

Singh, S. (2001), 'Generalized calibration approach for estimating variance in survey sampling', *Annals of the Institute of Statistical Mathematics* **53**(2), 404–417.

Singh, S. (2003), *Advanced Sampling Theory with Applications: How Michael 'selected' Amy*, Kluwer Academic Publishers.

Singh, S., Horn, S., Chowdhury, S. & Yu, F. (1999), 'Calibration of the estimator of variance', *Australian and New Zealand Journal of Statistics* **41**, 199–212.

Wu, C. & Sitter, Y. R. (2001), 'A model-calibration approach to using complete auxiliary information from survey data', *Journals - American Statistical Association* **96**, 185–193.

Appendix A. Comparison between $AV(\widehat{P}_{AW})$ and $AV(\widehat{P}_{AQ})$

Because $AV(\widehat{P}_{AQ}) = AV(\widehat{q}_{AB})$ we have $AV(\widehat{P}_{AW}) < AV(\widehat{P}_{AQ})$ when

$$\left[P_AQ_A + \left(\frac{P_{AB}}{P_B}\right)^2 P_BQ_B - 2\left(\frac{P_{AB}}{P_B}\right)(P_{AB} - P_AP_B)\right] < \left[P_AQ_A + \left(\frac{Q_{AB}}{Q_B}\right)^2 P_BQ_B - 2\left(\frac{Q_{AB}}{Q_B}\right)(Q_{AB} - Q_AQ_B)\right]$$

or equivalently

$$\frac{P_{AB}^2}{P_B}Q_B - 2\left(\frac{P_{AB}}{P_B}\right)(P_{AB} - P_AP_B) < \frac{Q_{AB}^2}{Q_B}P_B - 2\left(\frac{Q_{AB}}{Q_B}\right)(Q_{AB} - Q_AQ_B).$$

Now, we have

$$Q_{AB} - Q_A Q_B = \frac{1}{N}\sum_{k\in U}(1-A_k)(1-B_k) - (1-P_A)(1-P_B) = P_{AB} - P_A P_B.$$

Then $AV(\widehat{P}_{AW}) < AV(\widehat{P}_{AQ})$ when

$$\left(\frac{P_{AB}}{P_B}\right)^2 P_B Q_B - \left(\frac{Q_{AB}}{Q_B}\right)^2 P_B Q_B - 2\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)(P_{AB} - P_A P_B) < 0$$

therefore

$$\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)\left[\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)P_B Q_B - 2(P_{AB} - P_A P_B)\right] = K_1 K_2 < 0.$$

Since

$$K_2 = P_B - P_{AB} + P_B(P_A - P_B) = P_{\bar{A}B} Q_B + P_B P_{A\bar{B}} \geq 0$$

where

$$P_{\bar{A}B} = \frac{1}{N}\sum_{k\in U}(1-A_k)B_k \quad \text{and} \quad P_{A\bar{B}} = \frac{1}{N}\sum_{k\in U} A_k(1-B_k),$$

we deduce that $AV(\widehat{P}_{AW}) < AV(\widehat{P}_{AQ})$ when $k_1 < 0$, that is

$$\frac{P_{AB}}{P_B} < \frac{Q_{AB}}{Q_B}. \tag{41}$$

Appendix B. Obtaining the minimum variance of $\widehat{P}_{A\alpha}$

If we denote $V_1 = AV(\widehat{P}_{AW})$; $V_2 = AV(\widehat{Q}_{AW})$ and $C = Cov(\widehat{P}_{AW}, \widehat{Q}_{AW})$ the minimum variance of $\widehat{P}_{A\alpha}$ is

$$\frac{V_2 \times V_1 - C^2}{V_1 + V_2 + 2C} \tag{42}$$

and this is achieved when

$$\alpha = \frac{(V_2 + C)}{(V_1 + V_2 + 2C)} \tag{43}$$

It is easy to see that

$$C = Cov(\widehat{P}_{AW}, \widehat{Q}_{AW}) = \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\Bigg[- P_A Q_A - \left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)P_B Q_B + \left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)(P_{AB} - P_A P_B)\Bigg]$$

Therefore, we have:

$$\begin{aligned} V_2 + C &= \frac{1-f}{n}\left(\frac{N}{N-1}\right)\Bigg[(P_{AB} - P_A P_B)\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right) \\ &\quad + P_B Q_B\left(\left(\frac{Q_{AB}}{Q_B}\right)^2 - \left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)\right)\Bigg] \\ &= \frac{1-f}{n}\left(\frac{N}{N-1}\right)\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)\Bigg[(P_{AB} - P_A P_B) - P_B Q_{AB}\Bigg] \end{aligned}$$

Similarly

$$\begin{aligned} V_1 + V_2 + 2C &= \frac{1-f}{n}\left(\frac{N}{N-1}\right) \\ &\quad \times P_B Q_B\left[\left(\frac{P_{AB}}{P_B}\right)^2 + \left(\frac{Q_{AB}}{Q_B}\right)^2 - 2\left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)\right] \\ &= \frac{1-f}{n}\left(\frac{N}{N-1}\right) P_B Q_B\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 \end{aligned}$$

By substituting the values $V_2 + C$ and $V_1 + V_2 - 2C$ in (43), the value of α is found to be:

$$\alpha = \frac{(P_{AB} - P_A P_B) - P_B Q_{AB}}{P_B Q_B\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)} = \frac{P_{AB} - P_B}{P_B\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)}. \tag{44}$$

$$\begin{aligned} V_1 V_2 &= G^2\Bigg[\left(P_A Q_A\right)^2 + \left[\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 + 2\left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)\right] P_A Q_A P_B Q_B \\ &\quad + \left(\frac{P_{AB}}{P_B}\right)^2\left(\frac{Q_{AB}}{Q_B}\right)^2 \left(P_B Q_B\right)^2 + 4\left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)\phi^2 \\ &\quad - 2\phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\left(P_A Q_A + P_{AB} Q_{AB}\right)\Bigg] \\ &= G^2\Bigg[\left(P_A Q_A - \phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\right)^2 - \phi^2\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 \\ &\quad + \left(\frac{P_{AB}}{P_B}\right)^2\left(\frac{Q_{AB}}{Q_B}\right)^2 \left(P_B Q_B\right)^2 + \Bigg[\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 \\ &\quad + 2\left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)\Bigg] P_A Q_A P_B Q_B - 2\phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\left(P_{AB} Q_{AB}\right)\Bigg] \end{aligned}$$

$$
\begin{aligned}
&= G^2\Bigg[\left(P_AQ_A - \phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\right)^2 \\
&+ P_BQ_B\left(P_AQ_A - \frac{\phi^2}{P_BQ_B}\right)\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 \\
&+ 2(P_{AB}Q_{AB})\left(P_AQ_A - \phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\right) \\
&+ \left(\frac{P_{AB}}{P_B}\right)^2\left(\frac{Q_{AB}}{Q_B}\right)^2(P_BQ_B)^2\Bigg] \\
&= G^2P_BQ_B\left(P_AQ_A - \frac{\phi^2}{P_BQ_B}\right)\left(\frac{P_{AB}}{P_B} - \frac{Q_{AB}}{Q_B}\right)^2 + G^2 \times K
\end{aligned}
$$

where

$$
\begin{aligned}
K = &\left(P_AQ_A - \phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\right)^2 + 2(P_{AB}Q_{AB})\left(P_AQ_A - \phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right)\right) \\
&+ \left(\frac{P_{AB}}{P_B}\right)^2\left(\frac{Q_{AB}}{Q_B}\right)^2(P_BQ_B)^2.
\end{aligned}
$$

On the other hand

$$
C^2 = G^2\left[\phi\left(\frac{P_{AB}}{P_B} + \frac{Q_{AB}}{Q_B}\right) - P_AQ_A - \left(\frac{P_{AB}}{P_B}\right)\left(\frac{Q_{AB}}{Q_B}\right)P_BQ_B\right]^2 = -G^2 \times K
$$

Thus, by substituting the values C^2; $V_1 \times V_2$ and $V_1 + V_2 + 2C$ in (42), we have:

$$
AV(\widehat{P}_{A\alpha}) = G\left(P_AQ_A - \frac{\phi^2}{P_BQ_B}\right)
$$

Appendix C. Obtaining $AV(\widehat{P}_{AR})$

$$
\begin{aligned}
AV(\widehat{P}_{AR}) = \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\Bigg[&P_AQ_A + \left(\frac{P_{AB}}{P_B} - \frac{P_{A\bar{B}}}{Q_B}\right)^2 P_BQ_B \\
&- 2\left(\frac{P_{AB}}{P_B}\right)(P_{AB} - P_AP_B) - 2\left(\frac{P_{A\bar{B}}}{Q_B}\right)(P_{A\bar{B}} - P_AQ_B)\Bigg]
\end{aligned}
$$

with

$$
P_{A\bar{B}} = \frac{1}{N}\sum_{k\in U} A_k(1 - B_k).
$$

Now, taking into account that

$$P_{A\bar{B}} - P_A Q_B = P_A - P_{AB} - P_A + P_A P_B = P_A P_B - P_{AB}$$

the asymptotic variance is

$$AV(\widehat{P}_{AR}) = \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\left[P_A Q_A + \left(\frac{P_{AB}}{P_B} - \frac{P_{A\bar{B}}}{Q_B}\right)^2 P_B Q_B + 2(P_{AB} - P_A P_B)\left(\frac{P_{A\bar{B}}}{Q_B} - \frac{P_{AB}}{P_B}\right)\right].$$

Since

$$\frac{P_{A\bar{B}}}{Q_B} - \frac{P_{AB}}{P_B} = \frac{P_{A\bar{B}} P_B + P_{AB} P_B - P_{AB}}{P_B Q_B} = \frac{P_A P_B - P_{AB}}{P_B Q_B}$$

we have

$$\begin{aligned} AV(\widehat{P}_{AR}) &= \frac{(1-f)}{n}\left(\frac{N}{N-1}\right)\left[P_A Q_A + \frac{(P_{AB} - P_A P_B)^2}{P_B Q_B} - 2\frac{(P_{AB} - P_A P_B)^2}{P_B Q_B}\right] \\ &= G\left[P_A Q_A - \frac{(P_{AB} - P_A P_B)^2}{P_B Q_B}\right] = G\left(P_A Q_A - \frac{\phi^2}{P_B Q_B}\right) \end{aligned} \tag{45}$$