Journal of Statistics and Management Systems

ISSN: 0972-0510 (Print) 2169-0014 (Online) Journal homepage: http://www.tandfonline.com/loi/tsms20

Calibrated Bayesian shrinkage of finite population totals in survey sampling

Andrés Gutiérrez, Hanwen Zhang, Cristian Tellez & Stalyn Guerrero

To cite this article: Andrés Gutiérrez, Hanwen Zhang, Cristian Tellez & Stalyn Guerrero (2018) Calibrated Bayesian shrinkage of finite population totals in survey sampling, Journal of Statistics and Management Systems, 21:2, 225-249, DOI: 10.1080/09720510.2017.1367477

To link to this article: https://doi.org/10.1080/09720510.2017.1367477

Published online: 14 Mar 2018.

Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 21 (2018), No. 2, pp. 225–249
DOI : 10.1080/09720510.2017.1367477

Taylor & Francis
Taylor & Francis Group

Calibrated Bayesian shrinkage of finite population totals in survey sampling

Andrés Gutiérrez *
Hanwen Zhang †
Cristian Tellez §
Faculty of Statistics
Universidad Santo Tomás
Tomas
Colombia

Stalyn Guerrero ‡
Colombian Institute for the Assessment of Education
Colombia

Abstract

In this article a Bayesian methodology (parametric and nonparametric) is proposed in order to estimate, by means of calibration, the population total from a sample with unequal inclusion probabilities. By means of some simulation studies, it was empirically determined that, through an adequate choice of an prior distribution with the proposed methodology, unbiased estimators are obtained. In addition to this, with an appropriate sample size, a smaller variance and confidence intervals with higher levels and minor length were obtained in comparison with the intervals induced by classic estimators (Horvitz-Thompson and calibration estimators). Finally, the implementation of the methodology to estimate the income is exemplified using real data from a labor force survey in Colombia.

Subject Classification: *62D05, 62F15, 62G07, 62F40.*

Keywords: *Density estimation, Bayesian inference, bootstrap, Jackknife, sampling theory, sample surveys.*

*E-mail: hugogutierrez@usantotomas.edu.co (Corresponding Author)
†E-mail: hanwengutierrez@usantotomas.edu.co
§E-mail: cristiantellez@usantotomas.edu.co
‡E-mail: sguerrero@contratistia.icfes.gov.co

1. Introduction

A parameter of interest in social research is the population total. In probability sampling there are several alternatives to estimate this parameter (Horvitz and Thompson, 1952, Deville and Särndal, 1992). In this article we will focus on proposing a Bayesian methodology based on calibration estimation. Furthermore, since the total parameter is social research is almost always positive, from the Bayesian perspective, this parameter is a random variable that can be modeled using distributions with positive supports such as the Gamma, the log-normal or the Truncated normal distribution, among other distributions.

In the specialized literature little is found about the integration between the estimation theory based on probability sampling and the Bayesian theory; one of the first approaches between these two theories was made by Ericson (1969), in which he created a Bayesian model with random variables that were interchangeable in finite populations sampling. Subsequently, Binder (1982) introduced the nonparametric Bayesian models to estimate some parameters in finite population sampling. Furthermore, Ericson (1988), from a Bayesian point of view, reviewed some results that were obtained by using the Bayesian statistics to face inference and sampling design problems in finite populations. A few years later, Meeden (1999) used the non-informative Polya posterior distribution in several one-stage sampling procedures, , spreading this kind of reasoning to the two-stage cluster sampling.

On the other hand, Hamner et al. (2001) focus on the necessary assumptions for the robustness of their statistical methods for predicting the population total using mixed models. Sedransk (2008) discusses several applications where the Bayesian statistics is helpful to estimate parameters in finite populations. Following this, Aitkin (2008) extends the Bayesian bootstrap analysis, applying it to studies with a unique population, where the regression model is fitted using numeric variables, obtained from a stratified cluster sampling. Lazar et al. (2008) state that the Polya posterior distribution has developed as a non-informative Bayesian approach in probability sampling. This is an appropriate method when little or no prior information about the target population is available. This paper shows that it is possible to enlarge this procedure to incorporate some kind of prior information which is partially obtained from auxiliary variables.

In a recent paper about the usefulness of the Bayesian bootstrap in the inference of finite populations, Carota (2009) provides reasonable answers

to the problems that are not solved in Aitkin (2008). One of these problems is the choice of the population parameter, for this parameter cannot be solved using the Bayesian bootstrap, because it is based on a multinomial probability and these parameters don't have any restrictions. Chen et al. (2010) propose a Bayesian Penalized Spline Predictive (BPSP) estimator for a finite population proportion in an unequal probability sampling. Later, Little et al. (2011) describe the Calibrated Bayesian (CB) approach for models with missing data, where the problem of such missing data is approached from a Bayesian perspective. Then, Piñerez et al. (2014) published an article where they propose a Bayesian methodology to estimate a proportion in probability sampling.

This article presents a proposal to estimate, by using the calibration methodology, the population total taking into account any available prior information about the parameter. This estimator will be defined in both a parametric and a nonparametric way. In the first case the assumption that the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and the calibration estimator (Deville and Särndal, 1992) are asymptotically normal will be used. In the second case the selected technique is the bootstrap method, used to approximate the likelihood of the data.

1.1 *Calibration Estimators*

A calibration estimator is a linear estimator that has the ability to replicate (with no errors) the population total of one or more auxiliary variables in any sampling design; although the term Calibration was recently coined, some authors agree they have been using it since a long time before this process was given this name (Deville and Särndal, 1992). Let's suppose we have access to a vector of auxiliary information, $\boldsymbol{x}_k = (x_{1k}, x_{2k}, \ldots, x_{pk})'$, of p auxiliary variables and which is known for the individuals selected in the sample; due to administrative records or other reliable sources, the population total of the auxiliary information vector $\boldsymbol{t}_x = \sum_{k \in U} \boldsymbol{x}_k$ is known. The purpose of the study is to estimate the total of the characteristic of interest $\boldsymbol{t}_y = \sum_{k \in U} \boldsymbol{y}_k$ using the information given by $\boldsymbol{x}_k$ with $k \in s$ in the estimation stage. Although the Horvitz-Thompson estimator is unbiased, it requires that estimates meet the following constraint, which is given by

$$\boldsymbol{t}_x = \sum_{k \in s} w_k \boldsymbol{x}_k \tag{1}$$

and is known as calibration equation. The key idea is to find these w_k weights as close as it is possible to the inverse of the inclusion probability

of the k-th element $d_k = \frac{1}{\pi_k}$. Then the calibration estimator for the total of the characteristic of interest is defined as

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

The difficulty in using the calibration estimators is to build new w_k weights. This disadvantage can be regarded as an optimization problem where it is necessary to minimize the pseudo-distance[1] $G(w_k/d_k)$ between w_k y d_k in the sample, which is defined by

$$\sum_{k \in s} d_k G\left(w_k / d_k\right)$$

and also to comply with restriction (1). In addition, $G(w_k/d_k)$ must be strictly convex and nonnegative; $G(1) = 0$, i. e., the distance between equal weights is zero; $G'(1) = 0$, when the weights are equal, the function must have a critical point; $G''(1) = 0$, that critical point must correspond to the minimizer. The optimization problem can be solved using the Lagrange multiplier, for a detailed solution see Deville and Särndal (1992) and Deville et al. (1993), where it is shown that there are several types of distance that can be used in a calibration estimator construction.

After a brief introduction, Section 2 introduces the most relevant results of the proposed methodology; Section 3 presents the simulation study where the properties of the proposed estimators (unbiasedness and minor variance) are empirically shown; Section 4 illustrates the proposed methodology through the use of data from a labor force survey in Colombia (DANE, 2014).

2. Bayesian Inference for a Total

Consider a finite population of size N, denoted as, $U = \{u_1, u_2, ..., u_k, ..., u_N\}$ where each unit u_k with $k = 1, 2, ..., N$ is associated to observation y_k, which takes positive values. A random sample s is selected from U, according to a sampling design $p(\cdot)$. In the sample, values of y are observed for all selected items. The interest here is to estimate the posterior probability distribution for the parameter t_y, defined as $t_y = \Sigma_U y_k$, using the sample values and the inclusion probabilities arising from the $p(\cdot)$ design.

[1] $G(w_k/d_k)$ is a pseudo-distance since it does not necessarily have to comply with the symmetry property.

This Bayesian methodology considers that the point estimator of the parameter of interest is function of the distribution function from which random sample s comes, which has been chosen with a sampling design $p(\cdot)$. The calibration estimator is defined as:

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

Suppose the conditional probability distribution $P(\hat{t}_y \mid t_y)$ exists[2]; this, in turn, is the likelihood conditioned by t_y. Let $P(t_y)$ be the prior density function of parameter t_y. According to Bayes theorem we have that:

$$P(t_y \mid \hat{t}_y) \propto P(\hat{t}_y \mid t_y) P(t_y) \tag{2}$$

where $P(t_y \mid \hat{t}_y)$ is the posterior distribution of t_y, given the observations in the sample. As for the prior distribution for t_y, there is a variety of distributions that can be considered as informative and non-informative, such as Uniform, Normal, Gamma distributions or any distribution that has positive supports. With respect to the distributional assumption for $\hat{t}_y$, conditioned on parameter t_y, one must consider two cases: in the first one, none distributional assumption will be assumed, since such assumptions are not made in sampling theory, thus, they are said to be free distribution assumptions (nonparametric approach); in the second case, a normal distribution will be assumed (parametric approach) because it has been shown that the Horvitz-Thompson estimator is asymptotically normal distributed (Berger, 1998).

2.1 *Nonparametric Bayesian Inference for the Total*

According to Shao and Tu (1995) the Bayesian bootstrap method avoids the assumption of a parametric form of the distribution that produces the data. Let assume that one wants to estimate the total of the variable y, i.e. t_y, and there is access to prior information about t_y summarized in $P(t_y)$. Also suppose that $\boldsymbol{y}_s = (y_1, y_2, ..., y_n)$ returns the values of the variable of interest for every unit in the probability sample - chosen through the design $p(\cdot)$ with inclusion probabilities $\pi_k > 0$, $\forall k \in U$. Note that under this approach it is likely to assume that $\boldsymbol{y}_s$ comes from an unknown density. Suppose that the conditional probability distribution $P(\hat{t}_y \mid t_y)$, exists, then it is possible to approximate $P(\hat{t}_y \mid t_y)$ using a density estimator $\hat{P}(\hat{t}_y \mid t_y)$ and then find an estimator of the posterior distribution like:

---

[2] Note that if $\boldsymbol{y} = \{y_1, ..., \boldsymbol{y}_N\}$ has a prior distribution, then $\hat{t}_y$ and t_y will have a joint distribution, and that conditional distribution will exist.

$$P(t_y \mid \hat{t}_y) \propto P(t_y)\hat{P}(\hat{t}_y \mid t_y) \tag{3}$$

where $\hat{P}(\hat{t}_y \mid t_y)$ is a bootstrap estimate of the likelihood for a function of $\boldsymbol{y}_s$, proportional to a $\hat{P}(\hat{t}_y \mid t_y)$. Next, a sequence of the steps necessary to determine a $\hat{P}(\hat{t}_y \mid t_y)$ is presented:

1. By using the sample $\boldsymbol{y}_s$ an artificial population U^* that resembles the population of interest is built by replicating all of the items as in Särndal et al. (2003, section 11.6)
2. Select a bootstrap sample from U^* with an identical design to the one used to select the original sample s from U. Repeat the selection independently B times, and for each bootstrap sample $s_b^*(b=1,2,\ldots,B)$, calculate $\hat{t}^*_{y,cal,b}$ defined as

   $$\hat{t}^*_{y,cal,b} = \sum_{k\in s^*} w_k^* y_{kb}^*.$$

   Where w_k^* are the calibration weights obtained for each element in the bootstrap sample and y_{kb}^* is the k-th element of the b-th bootstrap sample.
3. With the above mentioned estimators $\hat{t}^*_{y,cal,1},\ldots,\hat{t}^*_{y,cal,B}$ calculate the kernel density estimator defined as:

   $$f_B(u) = \frac{1}{Bh_B}\sum_{b=1}^{B} K\left(\frac{u-\left(\hat{t}^*_{y,cal,b}-\hat{t}_{y,cal}\right)}{h_B}\right) \tag{4}$$

Where the function K is the core function (kernel) and, in general, is a continuous, unimodal and symmetrical around zero density function. Parameter h_B is known as a smoothing parameter (Wand and Jones, 1994; Hardle, 1990). Hollander (1999) shows the most used Kernel densities. Replacing $u=\hat{t}_{y,cal}-t_y$ in the previous equation, given t_y, an estimate of the sampling density of $\hat{t}_{y,cal}$ is obtained. By assessing $u=\hat{t}-t_y$ it turns out to be a function of t_y to be used as a likelihood.

$$\hat{P}\left(\hat{t}_y \mid t_y\right) = \frac{1}{Bh_B}\sum_{b=1}^{B} K\left(\frac{2\hat{t}_{y,cal}-t_y-\hat{t}^*_{y,cal,b}}{h_B}\right) \tag{5}$$

4. Then the resulting posterior distribution $P(t_y \mid \hat{t}_y)$ is proportional to $P(t_y)\hat{P}(\hat{t}_y \mid t_y)$ and the normalization constant can be found through numerical integration.

Note that this algorithm works well under regularity conditions that ensures that (4) will be a good estimate of the density of $\hat{t}_{y,cal}-t_y$. This way,

it is possible to construct a Bayesian estimator of the posterior distribution of t_y:

$$P\left(t_y \mid \hat{t}_{y,cal}\right) = c(\boldsymbol{y}_s) \times P(t_y) \times \hat{P}\left(\hat{t}_y \mid t_y\right)$$

where $c(y)$ can be obtained by numerical integration as in

$$c(\boldsymbol{y}_s) = \frac{1}{\int P(t_y) \times \hat{P}\left(\hat{t}_y \mid t_y\right) dt_y}$$

For obtaining posterior distribution samples, numerical methods can be used to generate the values of this distribution. In this case, a random sample[3] $t_y^1, t_y^2, ..., t_y^m$ is produced through the posterior distribution $P(t_y \mid \hat{t}_{y,cal})$, by means of generating a grid of values from a suitable distribution with support in $\mathbb{R}$; then, we have to evaluate each generated value t_y^i in the posterior distribution as $P(t_y^i \mid \hat{t}_{y,cal})$, with $i = 1, 2, ..., m$, thus obtaining the selection probability for each value. Finally, the required sample is obtained by taking a sample with selection probability $P(t_y^i \mid \hat{t}_{y,cal})$ for $i = 1, 2, ..., m$. Functions that are commonly used in order to achieve the estimation of the posterior distributions are the quadratic loss function, the absolute error loss function and the step function (Box 1973).

2.2 *Parametric Bayesian Inference for the Total*

In this section, we will assume that the calibration estimator of the total is distributed asymptotically normal; i.e.,

$$\hat{t}_{y,cal} \sim N(t_y, V)$$

with an priori distribution for the given parameters as it follows,

$$t_y \sim N(\mu, \tau^2) \text{ and } V \sim \text{inv} - \text{gamma}\left(\frac{n_0}{2}, \frac{n_0 v_0}{2}\right)$$

Where n_0 y v_0 refer to the sample size and the variance of the calibration estimator for a previous survey. An interesting deviance of this approach is to consider that the variance of the calibration estimator indeed depend on its sample size. If we do expect that the sample size and variance are similar to those of a previous survey, then v_0 could be considered as the

[3] With this sample, the aim is to estimate the parameter t_y f which considers an estimation error that must be minimized. To achieve this there must be a function relating the estimate of the parameter t_y with its true value.

variance of the calibration estimator (which has the same variance and sample size as in the previous survey) while n_0 could be interpreted as a shape parameter for the prior inverse gamma distribution. This way we can construct informative prior distribution.

Assuming the independence of t_y and V in the prior distribution, the joint posterior distribution of t_y y given $\hat{t}_{y,cal}$ is:

$$P\left(t_y, V \mid \hat{t}_{y,cal}\right) \propto P\left(\hat{t}_{y,cal} \mid t_y, V\right) P(t_y, V)$$
$$= V^{-\left(\frac{1+n_0}{2}\right)-1} \exp\left\{-\frac{1}{2V}\left[n_0 v_0 + \left(\hat{t}_{y,cal} - t_y\right)^2\right] - \frac{1}{2\tau^2}(t_y - \mu)^2\right\} \quad (6)$$

Note that the joint posterior distribution has no structural known form and therefore is not possible to use the analytical integration method for obtaining an integration constant (Migon and Gamerman, 1999). Then, conditional densities of t_y y V are found in the following way:

$$t_y \mid V, \quad \hat{t}_{y,cal} \sim N\left(\mu_n, \tau_n^2\right)$$

with,

$$\mu_n = \frac{\frac{\hat{t}_{y,cal}}{V} + \frac{\mu}{\tau^2}}{\frac{1}{V} + \frac{1}{\tau^2}}$$
$$\tau_n^2 = \left(\frac{1}{V} + \frac{1}{\tau^2}\right)^{-1}$$

and,

$$V \mid t_y, \hat{t}_{y,cal} \sim \text{inv-gamma}\left(\frac{n_0+1}{2}, \frac{n_0 v_0 + \left(\hat{t}_{y,cal} - t_y\right)^2}{2}\right)$$

Thus, the posterior distribution for the total is found by iterating the above conditional densities. In particular, it is possible to use the Markov chain Monte Carlo Method (MCMC) methodology and the Gibbs sampler (Gilks 2005). This approach can vary since it is possible to assume a relationship of dependency between the likelihood parameters, for example, assuming that

$$V \sim \text{inv}-\text{gamma}\left(\frac{n_0}{2}, \frac{n_0 v_0}{2}\right) \text{ and } t_y \mid V \sim N\left(\mu, \frac{V}{c_0}\right)$$

for an appropriate c_0. Then, it is possible to show that the marginal distribution of t_y follows a Student's t-distribution with $n_0 + 1$ degrees of freedom (see appendix 1.).

3. Empirical Study

The simulation study aims to evaluate the performance of the proposed methodology and compare it with the traditional procedure carried out in the estimation of a total in probability sampling. Simulation design simulation is described below.

3.1 *Design of the Simulation*

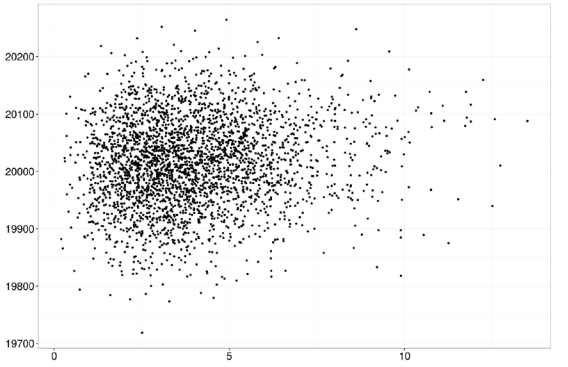
The procedure consists of simulating an artificial population (X, Y) of $N = 3000$, where X is the auxiliary variable to be used in calibration estimators. The values of this variable are obtained from a *Gamma*(α, k), distribution, where the shape parameter $\alpha = 4$ and the scale parameter $k = 1$. Furthermore, the probability of inclusion in the population are defined as $\frac{n}{N}$. The variable of interest Y is generated from the linear model $y_i = 20000 + 4x_i + \varepsilon_i$ for, where $\varepsilon_i \sim N(0, \sigma)$ (See Figure 1). In each simulation scenario a finite population is generated, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ have linear correlation coefficients for 0.1, 0.5 and 0.9, which are achieved by setting different values of σ. Then in each population simple random samples with replacement sizes $n = 50$,400 and 1000 are drawn. Based on the selected sample, the population total $\hat{t}_y$ is estimated with the Horvitz-Thompson estimator, calibration and the proposed methodologies. This process is repeated 1000 times, estimates $\hat{t}_j (j = 1, ..., 1000)$ are obtained and the relative bias B is calculated, which is defined as

$$B = \frac{1}{1000 \times t_y} \sum_{j=1}^{1000} (\hat{t}_j - t_y)$$

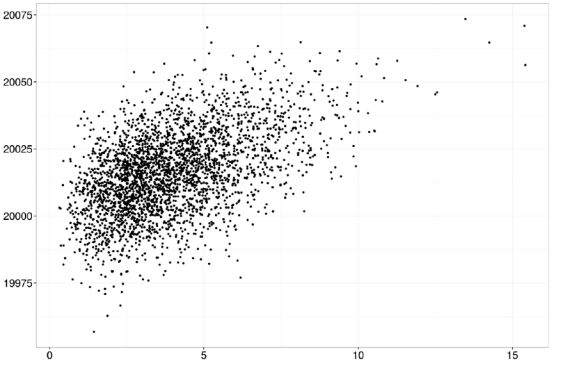
Also, the relative length of the confidence or the credible interval is calculated, which is defined as

$$\text{Length} = \frac{1}{1000} \sum_{i=1}^{1000} \left(\frac{L_{upper_i} - L_{lower_i}}{2 \times \hat{t}_{y_i}} \right)$$

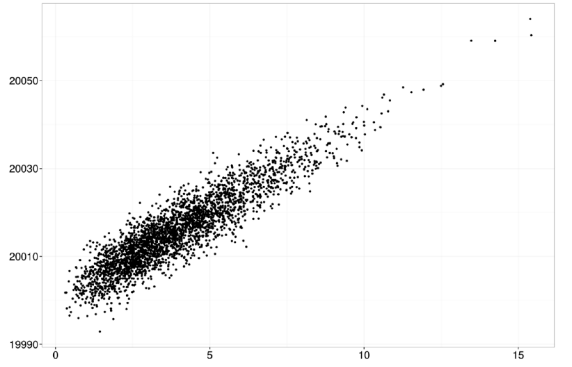
Where, L_{upper_i} and L_{lower_i} are the upper and lower limits of the corresponding credible (or confidence) intervals computed for the estimator $\hat{t}_{yi}$. The coefficient of variation (CV) and coverage rate are calculated as well. In the simulation study, for both parametric and nonparametric approach, the normal distribution $N(t_y, \sigma)$ with $\sigma = 1000$



(a) $\rho = 0.1$



(b) $\rho = 0.5$



(c) $\rho = 0.9$

Figure 1

Different correlation between the variable of interest and the auxiliary variable generates different simulated scenarios for the population of interest.

is used as non-informative prior distribution and the same distribution with $\sigma = 50$ is used as informative prior distribution. Additionally, in the nonparametric case, the uniform distribution $U(0, 10^{10})$ is considered as non-informative prior and the distribution $Gamma(\alpha, \beta)$ with $\alpha\beta = t_y$ and $\alpha\beta^2 = \sigma$ as informative prior distribution.

Moreover, as the value of $t_y \in \mathbb{R}^+$ is known, the confidence interval at 95% of the Horvitz-Thompson estimator, increasing its ends by half of its own length, was used to define the informative prior distributions. Then, with a 95% probability, the values of informative priors distributions will have their values within the following interval:

$$\left(L_{lower} - \frac{L_{upper} - L_{lower}}{2}, \quad L_{upper} + \frac{L_{upper} - L_{lower}}{2} \right)$$

with $L_{lower} = \hat{t}_{y\pi} - Z_{(1-\alpha/2)}\sqrt{Var(\hat{t}_{y\pi})}$ and $L_{lower} = \hat{t}_{y\pi} + Z_{(1-\alpha/2)}\sqrt{Var(\hat{t}_{y\pi})}$ where $\hat{t}_{y\pi}$ and $Var(\hat{t}_{y\pi})$ represent the total estimated and the variance of the HT estimator. Figures 2 and 3 show the prior functions (informative and non-informative) used in the different samples in the case where the correlation between X and Y variables is equal to 0.9, likewise, posterior and likelihood

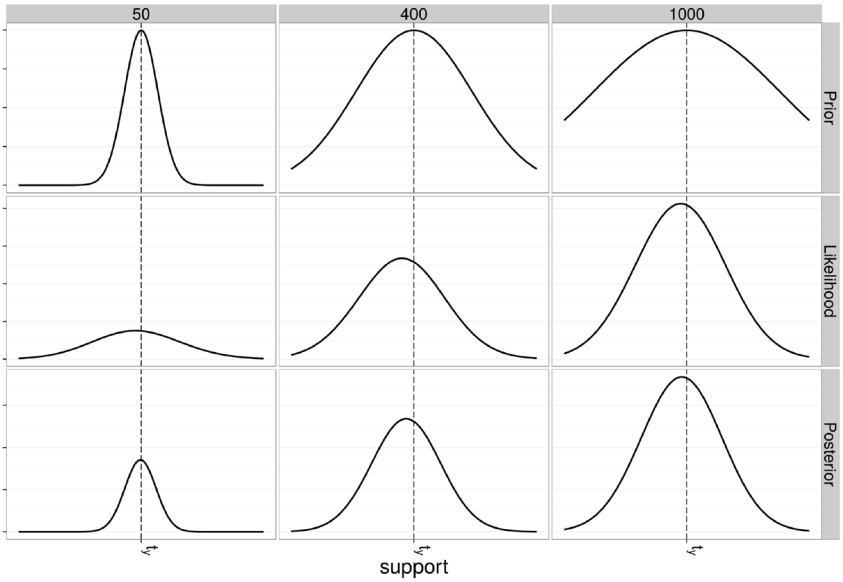


Figure 2

Non-informative prior, likelihood and posterior distributions when X and Y correlation is equal to 0.5 for random samples of size n = 50, 400 and 2000.

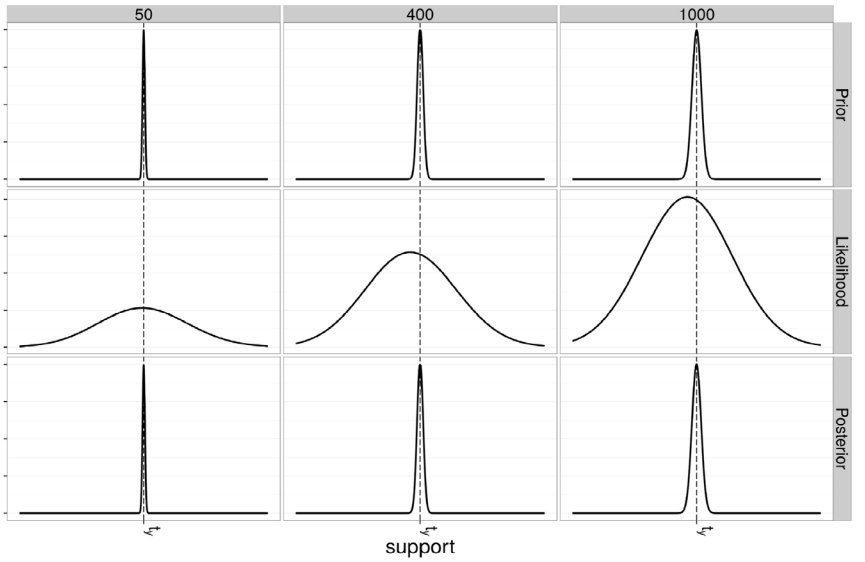


Figure 3

Informative prior, likelihood and posterior distributions when X and Y correlation is equal to 0.5 for random samples of size n = 50, 400, 2000.

functions can be observed. Note that when n grows posterior distribution concentrates in increasingly short intervals, which, in turn, leads to better estimates of t_y.

3.2 *Simulation Results*

Table 1 shows the coverage, relative length, relative bias and coefficient of variation for the Horvitz Thompson estimator, the Bootstrap calibrated Bayesian nonparametric estimator (BBCNP), and the Bayesian calibrated parametric estimator (BCP) in a population with an average linear structure and normal priors. It could evinced that the estimations obtained from the BCNP methodology with normal prior distribution yield larger values of relative length and CV. However, it is important to note that with large sample size and high correlation between X and Y, these values decrease significantly. On the other hand, when the Gamma distribution is used as prior distribution (see Table 2), the results obtained improve substantially.

Regarding the results of Calibrated parametric Bayesian methodology (BCP), this approach yield best results with both informative and non informative prior distribution, outstripping other estimators presented in this paper. It is also valuable to emphasize the good performance of

Table 1

Coverage (%), Relative Length (×1000), Relative Bias (×1000) and Coefficient of variation (×1000), of the H.T., Calibration BBCNP and BCP estimators in a population with an average linear structure and normal priors.

			$\rho = 0,1$			$\rho = 0,5$			$\rho = 0,9$		
	Estimator	$P(t_y)$	$n = 50$	$n = 400$	$n = 1000$	$n = 50$	$n = 400$	$n = 1000$	$n = 50$	$n = 400$	$n = 1000$
Coverage	HT	-----	94.6	95.1	94.1	94.2	96.0	95.6	95.4	94.6	96.0
	Calibration	-----	93.2	94.8	95.4	93.2	94.8	95.4	93.2	94.8	95.4
	BBCNP	$N(t_y, 10^3)$	99.9	99.6	99.9	99.9	99.8	99.9	99.6	99.9	99.9
		$N(t_y, 50)$	99.9	99.8	99.9	99.9	99.9	99.9	99.6	99.9	99.9
	BCP	$N(t_y, 10^3)$	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
		$N(t_y, 50)$	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
Relative Length	HT	-----	1.047	0.350	0.194	0.219	0.073	0.040	0.118	0.039	0.022
	Calibration	-----	1.028	0.347	0.193	0.188	0.063	0.035	0.049	0.017	0.009
	BBCNP	$N(t_y, 10^3)$	129.100	68.100	45.399	122.264	60.790	40.117	87.469	30.908	19.580
		$N(t_y, 50)$	8.210	8.166	8.092	8.218	8.156	8.056	8.210	7.953	7.567
	BCP	$N(t_y, 10^3)$	0.032	0.032	0.032	0.032	0.030	0.029	0.030	0.025	0.023
		$N(t_y, 50)$	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Relative Bias	HT	-----	0.000	0.001	0.001	0.002	-0.001	-0.001	0.000	-0.002	0.000
	Calibration	-----	0.022	0.008	-0.004	0.004	0.001	-0.001	0.001	0.000	0.000
	BBCNP	$N(t_y, 10^3)$	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000
		$N(t_y, 50)$	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000
	BCP	$N(t_y, 10^3)$	-0.006	0.004	0.004	0.005	0.001	-0.001	0.001	0.000	0.000
		$N(t_y, 50)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CV	HT	-----	0.534	0.179	0.099	0.112	0.037	0.021	0.060	0.020	0.011
	Calibration	-----	0.525	0.177	0.098	0.096	0.032	0.018	0.025	0.008	0.005
	BBCNP	$N(t_y, 10^3)$	66.022	34.826	23.231	62.480	31.093	20.523	44.771	15.779	10.016
		$N(t_y, 50)$	4.203	4.176	4.143	4.203	4.175	4.123	4.185	4.062	3.877
	BCP	$N(t_y, 10^3)$	0.017	0.016	0.016	0.016	0.015	0.015	0.015	0.013	0.012
		$N(t_y, 50)$	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.003	0.003

Table 2

Coverage (%), Relative Length (×1000), Relative Bias (×1000) and Coefficient of variation (×1000), of the BCNP estimators in a population with an average linear structure and priors Gamma and Uniform.

		ρ = 0, 1			ρ = 0, 5			ρ = 0, 9		
	$P(t_y)$	$n = 50$	$n = 400$	$n = 1000$	$n = 50$	$n = 400$	$n = 1000$	$n = 50$	$n = 400$	$n = 1000$
Coverage	$U(0, 10^{10})$	99.9	99.6	99.9	99.9	99.8	99.9	98.8	99.8	99.9
	Gamma (α, β)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
Relative Length	$U(0, 10^{10})$	211.300	74.840	47.222	184.765	65.417	41.389	89.607	31.712	20.211
	Gamma (α, β)	1.163	1.163	1.163	1.163	1.161	1.162	1.164	1.161	1.159
Relative Bias	$U(0, 10^{10})$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Gamma (α, β)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CV	$U(0, 10^{10})$	107.966	38.299	24.165	94.530	33.461	21.174	45.796	16.226	10.319
	Gamma (α, β)	0.595	0.595	0.595	0.595	0.594	0.594	0.594	0.594	0.593

the classic calibration estimator. It is well known that these estimators are better than the Horvitz-Thompson estimator when highly correlated auxiliary variable is available.

An important result of this paper is that regardless the sample size, the proposed estimators always yield better coverage rate better larger than the nominal value, with smaller interval length than the traditional estimators. Finally, if the two estimators proposed in this article are compared, it is possible to say that with the same priori distribution (normal distribution) the parametric estimator, in general terms, outperforms the nonparametric estimator.

4. Real Application: Estimating the Average Income in Colombia

To illustrate the proposed methodology, we used the data for the year 2014 from *La Gran Encuesta Integrada de Hogares* (GEIH) (Official Household Survey), which is carried out monthly in Colombia since 2009 by the National Administrative Department of Statistics (DANE). This survey addresses different aspects, including the socio-economic one, which contains information on the employment conditions of people (if they work, what they do at work, how they work, how much they earn, if they have health and social security or if they are looking for a job). In addition to the general characteristics of the population such as sex, age, marital status and educational level, the survey asks about their sources of income and their expenses (what they buy, how often they buy and where they buy).

The GEIH gives to the country and its citizens information at national, regional, urban and rural levels and for each of the capitals of the departments of the country. The results of the survey, as well as the data bases consolidated during the data collection process are available for the general public at the official website of the National Administrative Department of Statistics - DANE (http://www.dane.gov.co/). The database that was selected contains information of employed people living in municipal seats. The objective is to estimate the average net income of the last month of the year in the households of municipal seats by way of professional fees, businesses, profession or real estate by using as an auxiliary variable labor-based earnings of people who were working in December 2014. In this month, 26257 households were observed and the variables showed a correlation of 0.987 (see Figure 4). The survey questions that refer to these revenues are identified inside the base as P6750 and Inglabo.

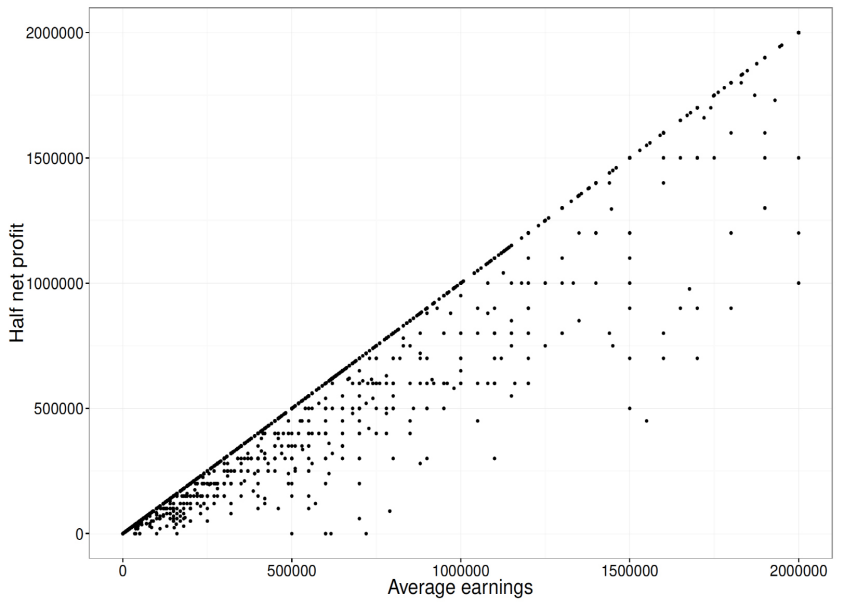


Figure 4

Labor-based income and Net Income dispersion

Furthermore, since 1996, the DANE has a Master Sample of Households that is used as a sampling frame of all those surveys related to the social subject whose information source is private households. This sample was designed according to the requirements and objectives set by the DANE and is constantly updated through the recounting of buildings and homes that is made through the same surveys. With this in mind, in order to carry out the December 2014 GEIH the DANE adopted a multi staged, stratified, cluster probability sampling conditioned by the objectives and characteristics of the sampling frame.

Since the Master Sample of Household (sampling frame) is not available, it is impossible to replicate the inclusion probabilities of each of the elements in the different stages of the sampling design, however, the global expansion factors for the households contained in the sample are available in the data base, which are registered in the base with the label Fex_c_2011, these were obtained by calculating the calibration adjustment using the Clan 97 macro (Andersson and Nordberg (1998)) through a set of routines in the SAS program. Now, by taking the expansion factors we can define π_i as the multiplicative inverse of the expansion factor

($\pi_i = F_i^{-1}$). When calculating the point estimate and errors in sample surveys, the DANE uses as auxiliary information the 2005 Census Population Projections in Colombia. In this case, we have used auxiliary information in groups of age and sex.

The results published by the DANE indicate that, in December 2014, Colombia had a total of 17'188.465 employed people with an average net income of $\hat{\bar{y}} = 300139.6$ per household, with an estimated standard deviation of 8964.297, this way, a confidence interval at 95% for the estimated mean is (282569.9, 317709.3). The computation of the average net income is performed using BCNP and BCP methodologies presented in the sections 2.1 y 2.2. Firstly, the prior distribution for $\bar{y}$ should be established. For this purpose, the estimated averages by DANE for the last eleven months are used (see Table) to obtain the parameters of the normal prior distribution. After that, using the results presented in the section 2.1 and assuming that the calibration estimator is asymptotically normal distributed, that is,

$$\hat{\bar{y}}_{cal} \sim N(\bar{y}, V)$$

this way, the prior distributions are as follow:

$$\bar{y} \sim N(309097.5, 13458.55^2) \text{ and } V \sim inv-gamma\left(\frac{11}{2}, \frac{11v_0}{2}\right)$$

where v_0 denotes the variance of the calibration estimator obtained from the historical data. Assuming independence between t_y and V in the prior distribution, the joint posterior distribution of t_y and V given $\hat{t}_{y,cal}$ is obtained using the equation 6. Finally, by means of the Gibbs sampler, it is found that $\bar{y}_{BCP} = 305990.90$ and the 95% credible interval is (291151.80, 320778.40). Now, for the nonparametric estimation, the normal prior distribution mentioned above is used, as well as the Gamma distribution obtained from the historical data.

Using the averages estimated by the DANE for the previous eleven months averages (see Table 3) Normal and Gamma prior distributions parameters were obtained. In Figure 5 we can see the prior density functions, as well as the likelihood function and the posterior distribution resulting from the use of the methodologies proposed for each one of the estimators. Results obtained with these methodologies are shown in Table 4, where we see that when building a confidence (or credibility) interval at 95% the traditional calibration estimator is the one with the greatest length,

Table 3

Prior information for the mean income based in recent surveys

	n	$\hat{N}$	$\hat{\bar{y}}$
January	25689	16214054	337018
February	26564	16214690	295821
March	26387	16341298	328770
April	26468	16789044	309509
May	27455	16762644	328373
June	25969	16807730	313522
July	26727	16759336	302302
August	27675	17090492	310757
September	27253	17199890	311530
October	28302	17605392	310627
November	27522	17512403	321694

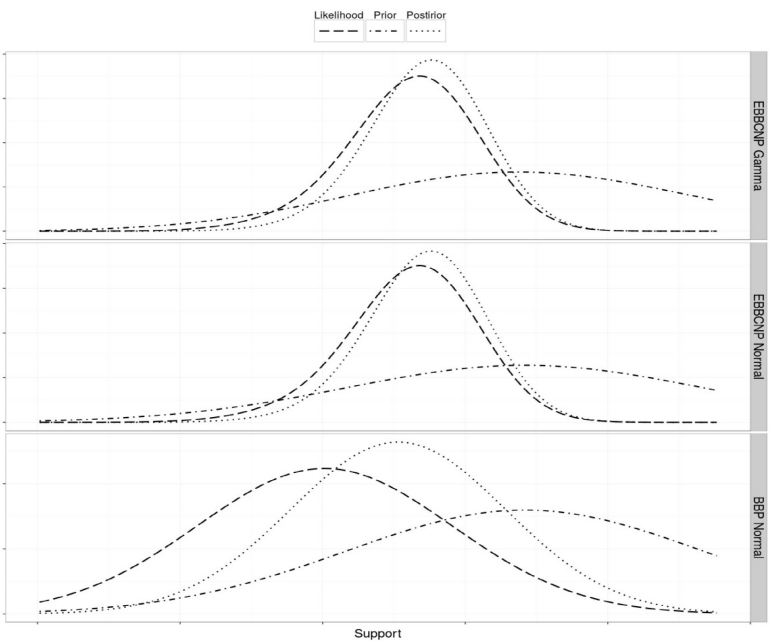


Figure 5

Density function for the prior, likelihood and posterior distributions for the estimation of the average net income.

Table 4

Estimation of the mean income with different estimators.

Estimator	Prior disribution	$\hat{\bar{y}}$	*IC*	*Length*
Calibrated DANE	-----	300139	(282569, 317709)	35139
BBCNP	*N*(309097.5, 13458.5)	307346	(298918, 315874)	16956
	Gamma(564.518, 0.001826)	307490	(299393, 315672)	16279
BCP	N(309097.5, 13458.5)	305990	(291151, 320778)	29626

in contrast, and being consistent with the simulation studies, BCNP is the one with the shortest length when the prior *Gamma*(α, β) is used.

5. Conclusions and Recommendations

The first finding to be highlighted in this paper is that we obtained closed-form expressions for the Calibrated Bayesian parametric estimator, which we were able to demonstrate that, under certain regularity assumptions, complies with the restrictions of a traditional calibration estimator. Furthermore, in the simulation study the Bayesian Calibrated parametric estimator showed better results than traditional estimators. As for the methodology developed for the Bayesian Calibrated nonparametric estimator we can say that is a much more efficient alternative to its parametric version to estimate the total. An outstanding feature the Bayesian Calibrated nonparametric estimator has is that as it does not have a closed-form expression it facilitates the use of prior distributions different from the Normal distribution, distributions that may or may not be specified by an equation.

Another contribution in this paper that should be pointed out is that thanks to the Bayesian Calibrated parametric and nonparametric estimators it has been established the methodological foundations for the building of calibration estimators for totals that rely on Bayesian statistics and can improve the estimates results in terms of bias, coverage and credibility intervals.

In the simulation results it was observed that when incorporating an non-informative prior distribution to the parametric and nonparametric calibrated Bayesian estimators the estimates obtained for the population total have similar performances to those presented by the linear calibration and the Horvitz-Thompson estimators, furthermore if the prior

distributions are informative the methodologies proposed in this research generate estimates with negligible biases in all tested scenarios. In addition, the estimators obtained here smaller RMSE and small amplitudes with excellent coverage in comparison to traditional estimators.

When making comparisons among the proposed estimators with informative prior (Normal) we can see that the nonparametric estimator has better results. Furthermore, in the example we were able to see that when we used an prior Gamma for the nonparametric estimator it showed the best result. Another result to highlight is that the proposed methodology allows us to include auxiliary information (which does not necessarily have to be contained within the sampling frame) in the estimation of the parameters of interest to get better estimates, as it was proved in the example. The next step is to adapt this methodology to the case of non-linear calibration estimators and parameters different to the population total.

Appendix

A Mathematical Appendix

Here we present details of mathematical derivations of the results presented in the text.

A.1 *Marginal Distribution for the Parameter of Interest*

Indeed, we know that,

$$P(t_y, V \mid \hat{t}_{y,cal}) \propto V^{-\left(\frac{n_0+2}{2}\right)-1} \exp\left\{-\frac{1}{2V}\left[n_0 v_0 + (\hat{t}_{y,cal} - t_y)^2 + c_0(t_y - \mu)^2\right]\right\}$$

By completing squares we have that,

$$P(t_y, V \mid \hat{t}_{y,cal}) \propto V^{-\left(\frac{n_0+2}{2}\right)-1} \exp\left\{-\frac{1}{2V}\left[(c_0+1)\left(t_y - \frac{c_0 + \hat{t}_{y,cal}}{c_0+1}\right)^2 - \frac{(c_0 + \hat{t}_{y,cal})^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}^2_{y,cal}\right]\right\}$$

Having said that, the marginal distribution of t_y is

$$P(t_y \mid \hat{t}_{y,cal}) \propto \int_0^{\infty} V^{-\left(\frac{n_0+2}{2}\right)-1} \exp\left\{-\frac{1}{2V}\left[(c_0+1)\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1}\right)^2 - \frac{(c_0+\hat{t}_{y,cal})^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2\right]\right\} dV \tag{7}$$

Being

$$Z = \frac{1}{2V}\left[(c_0+1)\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1}\right)^2 - \frac{(c_0+\hat{t}_{y,cal})^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2\right]$$

$$\rightarrow V = \frac{1}{2Z}\left[(c_0+1)\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1}\right)^2 - \frac{(c_0+\hat{t}_{y,cal})^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2\right]$$

$$= \frac{A}{2Z}, A = (c_0+1)\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1}\right)^2 - \frac{(c_0+\hat{t}_{y,cal})^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2$$

$$\rightarrow dV = -\frac{A}{2Z^2} dZ$$

Replacing in 7 we have

$$P\left(t_y \mid \hat{t}_{y,cal}\right) \propto \int_{\infty}^{0}\left(\frac{2Z}{A}\right)^{\frac{n_0+2}{2}+1}\left(-\frac{A}{2Z^2}\right)\exp\{-Z\}dZ$$

$$= \int_0^{\infty}\left(\frac{2Z}{A}\right)^{\frac{n_0}{2}+2}\left(\frac{A}{2Z^2}\right)\exp\{-Z\}dZ$$

$$= 2^{\frac{n_0}{2}+1} A^{-\left(\frac{n_0}{2}+1\right)} \int_0^{\infty} Z^{\left(\frac{n_0}{2}+1\right)-1}\exp\{-Z\}dZ$$

$$= 2^{\frac{n_0}{2}+1} A^{-\left(\frac{n_0}{2}+1\right)} \Gamma\left(\frac{n_0}{2}+1\right)$$

$$\propto A^{-\left(\frac{n_0}{2}+1\right)}$$

Now then, replacing the value of A we have that,

$$P\left(t_y \mid \hat{t}_{y,cal}\right) \propto \left[\left(c_0+1\right)\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1} \right)^2 - \frac{\left(c_0+\hat{t}_{y,cal}\right)^2}{c_0+1} + n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal} \right]^{-\left(\frac{n_0}{2}+1\right)}$$

Then, dividing by $n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal} - \frac{(c_0+\hat{t}_{y,cal})^2}{c_0+1}$ the following is obtained:

$$P\left(t_y \mid \hat{t}_{y,cal}\right) \propto \left[1 + \frac{c_0+1}{n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2 - \frac{\left(c_0+\hat{t}_{y,cal}\right)^2}{c_0+1}} \left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1} \right)^2 \right]^{-\left(\frac{\left[n_0+1\right]+1}{2}\right)}$$

$$= \left\{ 1 + \frac{1}{n_{0+1}} \left[\frac{\left(t_y - \frac{c_0+\hat{t}_{y,cal}}{c_0+1} \right)}{t_n} \right]^2 \right\}^{-\left(\frac{\left[n_0+1\right]+1}{2}\right)}$$

with

$$t_n^2 = \frac{\frac{n_0+1}{c_0+1}}{n_0 V_0 + c_0\mu^2 + \hat{t}_{y,cal}^2 - \frac{\left(c_0+\hat{t}_{y,cal}\right)^2}{c_0+1}}$$

From which it is concluded that,

$$t_y \mid \hat{t}_{y,cal} \sim t_{n_0+1}\left(\frac{c_0+\hat{t}_{y,cal}}{c_0+1}, t_n^2 \right)$$

A.2 *About Calibration Constraints*

We verify that the parametric calibration parametric complies with the constraint equation. Since calibration estimators satisfy that $t_x = \hat{t}_{x,cal}$, then, the following must be shown for the estimator proposed in this research

$$E\left[P\left(t_x \mid \hat{t}_{x,cal}\right)\right] = t_x$$

Since we have full knowledge of variable X then the likelihood for $\hat{t}_{x,cal}$, is:

$$L\left(\hat{t}_{x,cal} \mid t_x\right) = \begin{cases} t_x & \text{if}, \hat{t}_{x,cal} = t_x, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, let's use as prior distribution for t_x any distribution positive support any distribution $p(\cdot)$. This way, the posterior distribution of t_x is:

$$P\left(t_x \mid \hat{t}_{x,cal}\right) = \frac{1}{\hat{t}_{x,cal} P\left(\hat{t}_{x,cal}\right)} \times t_x P(\cdot) \times I_{\left\{t_x = \hat{t}_{x,cal}\right\}}$$

After calculating the expected value for $P(t_x \mid \hat{t}_{x,cal})$, we get:

$$\begin{aligned} E\left[P\left(t_x \mid \hat{t}_{x,cal}\right)\right] &= \sum_{t_x = \hat{t}_{x,cal}} t_x \times P\left(t_x \mid \hat{t}_{x,cal}\right) \\ &= \hat{t}_{x,cal} P\left(\hat{t}_{x,cal} \mid \hat{t}_{x,cal}\right) = \hat{t}_{x,cal} \times 1 = \hat{t}_{x,cal} = t_x \end{aligned}$$

References

[1] Aitkin, M. (2008). Applications of the bayesian bootstrap in finite population inference. *Journal of Official Statistics*, 24(1):21.

[2] Andersson, C. and Nordberg, L. (1998). A User's Guide to CLAN 97. Statistics Sweden, 299:300.

[3] Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67(2):209–226.

[4] Binder, D. A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society*. Series B (Methodological), pages 388–393.

[5] Box, G. y. T. G. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, Massachusetts.

[6] Carota, C. (2009). Beyond objective priors for the Bayesian bootstrap analysis of survey data. *Journal of Official Statistics*, 25(3):405.

[7] Chen, Q., Elliott, M. R., and Little, R. J. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. Survey Methodology, 36(1):23–34.

[8] DANE (2014). Gran Encuesta Integrada de Hogares. Departamento Administrativo Nacional de Estadï¿oestica.

[9] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418):376–382.

[10] Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423):1013–1020.

[11] Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*. Series B (Methodological), pages 195–233.

[12] Ericson, W. A. (1988). 9 Bayesian inference in finite populations. Handbook of Statistics, 6:213–246.

[13] Gambino, J. G. (2012). pps: Functions for PPS sampling. R package version 0.94.

[14] Gilks, W. R. (2005). Markov chain monte carlo. Wiley Online Library.

[15] Gutiérrez, H. A. (2009). Estrategias de muestreo: diseno de encuestas y estimación de parámetros. Universidad Santo Tomás.

[16] Hamner, M. S., Seaman Jr, J. W., and Young, D. M. (2001). Bayesian Methods in Finite Population Sampling.

[17] Hardle, W. (1990). Applied nonparametric regression, volume 27. Cambridge Univ Press.

[18] Hollander, M. y. D. A. W. (1999). Nonparametric Statistical Methods. Cambridge University Press, Unite State of America.

[19] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

[20] Lazar, R., Meeden, G., and Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. Survey Methodology, 34(1):51.

[21] Little, R. et al. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. Statistical Science, 26(2):162–174.

[22] Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. The American Statistician, 60(3):213–223.

[23] Lumley, T. (2014). survey: analysis of complex survey samples. R package version 3.30.

[24] Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. Sankhyā: *The Indian Journal of Statistics*, Series B, pages 133–144.

[25] Migon, H. and Gamerman, D. (1999). Statistical Inference: An Integrated Approach, Arnold. London, UK.

[26] Piñerez, C. F. T., Guerrero, S. Y., and Pacheco, M. (2014). Inferencia Bootstrap bayesiana para una proporción en muestreo con probabilidades desiguales. Comunicaciones en Estadística, 7(1):31–48.

[27] R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[28] Rojas, H. A. G. (2014). TeachingSampling: Selection of samples and parameter estimation in finite population. R package version 3.2.1.

[29] Särndal, C.-E., Swensson, B., and Wretman, J. (2003). Model assisted survey sampling. Springer Science & Business Media.

[30] Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24(4):495–506.

[31] Shao, J. and Tu, D. (1995). The jackknife and bootstrap. Springer.

[32] Wand, M. P. and Jones, M. C. (1994). Kernel smoothing. Crc Press.

Received November, 2016