Ratio type estimators for estimating proportions in a dual frame approach

Hemilio Coelho
Federal University of Paraiba, Department of Statistics, CCEN, Cidade Universitária s/n, João Pessoa - PB, 58.051-900, Brazil, hemilio@de.ufpb.br

Cristiano Ferraz
Federal University of Pernambuco, Department of Statistics, Av. Prof. Luiz Freire, s/n, Cidade Universitária,Recife - PE 50740-540, Brazil, cferraz@de.ufpe.br

Diogo Candido
Federal University of Paraiba, Department of Statistics, CCEN, Cidade Universitária s/n, João Pessoa - PB, 58.051-900, Brazil, diogovasconcelos.17@gmail.com

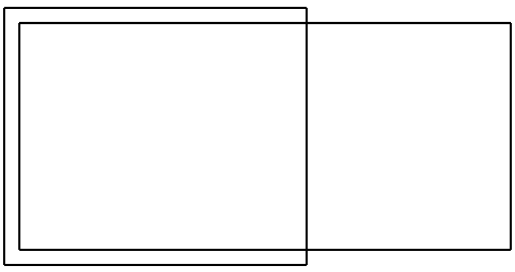
Abstract

Dual frame surveys use two overlapping frames to simultaneously cover target populations, and independent samples are drawn from each frame. Several types of surveys have been used dual frame designs successfully. When auxiliary information is available from at least one of the frames, it is reasonable to take advantage of them for deriving ratio type estimators. In this paper we investigate how the availability of auxiliary information can be used for building ratio type estimators for the populational proportion. The statistical properties of the proposed estimators are numerically investigated and a comparison of their statistical performances provided using the Monte Carlo simulation method.
Keywords: **Dual frame surveys; auxiliary information.**

1. Introduction

A dual frame approach is defined as a survey sample design with two sets of elements identified in two overlapping frames denoted by A and B providing coverage for the same target population, U. Hartley (1962) proposed parameter estimation under the situation of two frames with elements in common, considering the case where simple random sampling was used in both frames and noted that the use of a dual frame approach may result have better results, comparing with the use of a single frame survey. This approach may be motivated by several circunstances, such as when there is no single frame that can provide full coverage for the target population, but the simultaneous use of both frames can solve the problem (Figure 1a). It also happens situations where one of the frames provide adequate coverage but is expensive to sample from, while another one is less than perfect but represents an economical source for sampling units (Figure 1b).

Figure 1: Examples of Dual frame Scenarios



The simultaneous use of both frames, in a dual frame design generate three domains mutually exclusive: $a = A \cap B^c$, $b = B \cap A^c$ and $ab = A \cap B$. Thus, if N, N_A, N_B, N_a, N_b and N_{ab} denote the sizes of the sets U, A, B, a, b and ab, then $N = N_A + N_B - N_{ab}$. Further, let y_k be a response value of a dichotomous variable associated with the k-th element of U and define the population proportions $P = \sum_{k \in U} y_k/N$, $P_A = \sum_{k \in A} y_k/N_A$, $P_B = \sum_{k \in B} y_k/N_B$, $P_a = \sum_{k \in a} y_k/N_a$, $P_b = \sum_{k \in b} y_k/N_b$ and $P_{ab} = \sum_{k \in ab} y_k/N_{ab}$. Let S_A and S_B be the samples obtained from each frame according to specific sampling designs and let n, n_A, n_B, n_a, n_b and n_{ab} denote the sample sizes for the sets described. The notation used in a dual frame approach is presented in table (1). Estimation in a dual frame approach can be considered in any of three cases (1) N_A, N_B and N_{ab} are known; (2) N_A, N_B are known, but N_{ab} is unknown; (3) N_A, N_B and N_{ab} are unknown.

Table 1: Notation for population and sample quantities under a dual frame approach

Information	Frame		Domain		
	A	B	a	b	ab
Population set	U_A	U_B	U_a	U_b	U_{ab}
Population size	N_A	N_B	N_a	N_b	N_{ab}
Population total	t_{yA}	t_{yB}	t_{ya}	t_{yb}	t_{yab}
Population proportion	P_A	P_B	P_a	P_b	P_{ab}
Sample set	S_A	S_B	S_a	S_b	S_{ab}
Sample size	n_A	n_B	n_a	n_b	n_{ab}^A n_{ab}^B
Sample total	$\widehat{t}_{yA}$	$\widehat{t}_{yB}$	$\widehat{t}_{ya}$	$\widehat{t}_{yb}$	$\hat{t}_{yab}^A$ $\hat{t}_{yab}^B$
Sample proportion	$\hat{P}_A$	$\hat{P}_B$	$\hat{P}_a$	$\hat{P}_b$	$\hat{P}_{ab}^A$ $\hat{P}_{ab}^B$

2. Objective

We considered the problem of incorporating auxiliary information, available from at least one of the frames, into a ratio type estimator for population proportion. In this paper, we consider case 1 scenario above to investigate the performance of ratio type estimators under a dual frame approach, and we consider that response variable y is dichotomous. Under this context, strategies of building ratio type estimators for a dual frame survey are proposed to estimate a overall population proportion, which can be used to calculate indicators in health, as the prevalence for example. These estimators will be investigated under a random sampling design in both frames and are built over Ezzati et al. (1995), Hartley (1962) Bankier (1986), Kalton and Anderson (1986) and Skinner (1991) strategies of estimation and compared through a Monte Carlo simulation study.

3. Dual Frame estimation issues

3.1. Hartley's approach

Let A and B two overlapping frames used to cover the target population. U_A and U_B are the set of identificable elements of target population in each frame. The dual frame estimators for the population total, as example, can be written in the following general form, as originally proposed by Hartley(1962):

$$\hat{t}_H = \hat{t}_a + \hat{t}_b + \theta \hat{t}_{ab(A)} + (1-\theta)\hat{t}_{ab(B)} \quad (1)$$

where $\hat{t}_H$ denotes a dual frame estimator for the population total, $\hat{t}_a$ and $\hat{t}_b$ denote estimators for the totals of domains a and b, $\hat{t}_{ab(A)}$ and $\hat{t}_{ab(B)}$ denote estimators for the total of domain ab based on samples obtaneid

of each frame, and θ is a weighting constant, where $0 \leq \theta \leq 1$, chosen to minimize the variance of (1). The proposed estimator of this paper for overall proportion have the general form under Hartley's strategy:

$$\hat{P}_H = w_a\hat{P}_a + \theta w_{ab}\hat{P}^A_{ab} + w_b\hat{P}_b + (1-\theta)w_{ab}\hat{P}^B_{ab}, \quad (2)$$

where $w_a = N_a/N$, $w_b = N_b/N$, $w_{ab} = N_{ab}/N$ and $\hat{P}_a$, $\hat{P}_b$, $\hat{P}^A_{ab}$ and $\hat{P}^B_{ab}$ denote the estimators for the population proportion of domains a, b and ab, respectively.

3.2. Single frame estimators

Bankier (1986), Kalton and Anderson (1986) and Skinner (1991) proposed a general form to estimate the population total by treating all observations of two frames, A and B, as though they had been sample from a single frame, with modified weights for observations observed from domain ab. We have that the single frame estimator for the population total, as example, is given by the following expression:

$$\hat{t}_{SF} = \hat{t}^A_y + \hat{t}^B_y = \sum_{k\in S_A} w_k y_k + \sum_{k\in S_B} w_k y_k, \quad (3)$$

where $w_k = \frac{1}{\pi^A_k}$ if $k \in a$; $w_k = \frac{1}{\pi^B_k}$ if $k \in b$; $w_k = \frac{1}{\pi^A_k + \pi^B_k}$. This estimator does not use auxiliary information about N_A and N_B (Lohr and Rao, 2000). When we consider the estimation of the population proportion, this single frame estimator can be written as follows:

$$\hat{P}_{SF} = \frac{\sum_{k\in S_A} w_k y_k + \sum_{k\in S_B} w_k y_k}{N} = \theta_A\hat{P}_A + \theta_B\hat{P}_B, \quad (4)$$

where $\theta_A = N_A/N$, $\theta_B = N_B/N$ and $\hat{P}_A$ and $\hat{P}_B$ denote the estimators for the population proportion of frames A and B, respectively, and considering the same weights w_k defined above.

3.3. Ezzati et al. approach

The second estimator under the strategy proposed by Ezzati et al. (1995) considered the estimation of proportions of some characteristics in a subdomain of interest. Let two frames: an area frame denoted by A and a list frame denoted by B and the associated subdomain of them, denoted by a. The estimator of the overal proportion is given by

$$\hat{P} = \hat{P}_a\left[\lambda\hat{P}_{Ra} + (1-\lambda)\hat{P}_B\right] + \left(1-\hat{P}_a\right)\hat{P}_{RNa}, \qquad \text{where} \quad \lambda = \frac{n_A\hat{P}_D\hat{P}_a}{n_B + n_A\hat{P}_D\hat{P}_a} \quad (5)$$

where $\hat{P}_a$ is an estimate of the proportion of one particular subdomain a of interest obtained from administrative sources; $\hat{P}_{Ra}$ is the area estimate of the proportion of a specific carachteristic of interest whithin the portion of the subdomain a; $\hat{P}_B$ is the corresponding estimate from the special list sample; $\hat{P}_{RNa}$ is the area estimate of the proportion of a non-specific carachteristic of interest whithin the portion of the subdomain a.

4. Ratio type estimators for populational proportion in a dual frame design

When one or more auxiliary information is available from at least one of the frames, it is reasonable to design an e□cient sample scheme, or derive e□cient ratio type estimators. When the relationship between y and only one auxiliary information, denoted by x, is assisted by a linear model without intercept, where $E(y_k) = \beta x_k$ and $Var(y_k) = \sigma^2_k = \sigma^2 x_k$ for each domain, it is reasonable to consider the ratio estimator to estimate the population proportion. In this work we investigate how the availability of auxiliary information can be used for building some dual frame ratio type estimators for the population proportion, based on the approaches showed in the previous section.

4.1. Ratio type estimator 1

When we consider the Hartley's approach, it is possible to propose the following form of ratio type estimator for the population proportion:

$$\hat{P}_1 = \left(\frac{\overline{x}_{U_a}}{\overline{x}_{S_a}}\right)\hat{P}_a + \left(\frac{\overline{x}_{U_b}}{\overline{x}_{S_b}}\right)\hat{P}_b + p\left(\frac{\overline{x}_{U_{ab}}}{\overline{x}_{S^A_{ab}}}\right)\hat{P}^A_{ab} + (1-p)\left(\frac{\overline{x}_{U_{ab}}}{\overline{x}_{S^B_{ab}}}\right)\hat{P}^B_{ab}, \quad (6)$$

where $\hat{P}_a$, $\hat{P}_b$, $\hat{P}^A_{ab}$ and $\hat{P}^B_{ab}$ are the HT estimators for the domains proportions. The value of p is chosen to minimize Var($\hat{P}_1$). We have that $\overline{x}_{U_a}$. $\overline{x}_{U_b}$, $\overline{x}_{U_{ab}}$ are the population domain means of the auxiliary information in each frame.

4.2. Ratio type estimator 2

When we consider the single frame estimator approach, it is possible to propose the following form of ratio type estimator for the population proportion:

$$\hat{P}_2 = \frac{1}{N}\left\{\frac{\sum\limits_{k\in S_A} w_k y_k}{\sum\limits_{k\in S_A} w_k x_k} t^A_x + \frac{\sum\limits_{k\in S_B} w_k y_k}{\sum\limits_{k\in S_B} w_k x_k} t^B_x\right\}. \quad (7)$$

where t^A_x and t^B_x are the population total of auxiliary information in each frame.

4.3. Ratio type estimator 3

When we consider the Ezzati et al. approach, we have that the form of ratio type estimator for the population proportion is given by

$$\hat{P}_3 = \left(\frac{\overline{x}_{U_a}}{\overline{x}_{S_a}}\right)\hat{P}_a\left[\lambda\left(\frac{\overline{x}_{U_{Ra}}}{\overline{x}_{S_{Ra}}}\right)\hat{P}_{Ra} + (1-\lambda)\left(\frac{\overline{x}_{U_B}}{\overline{x}_{S_B}}\right)\hat{P}_B\right] + \left[1 - \left(\frac{\overline{x}_{U_a}}{\overline{x}_{S_a}}\right)\hat{P}_a\right]\left(\frac{\overline{x}_{U_{RNa}}}{\overline{x}_{S_{RNa}}}\right)\hat{P}_{RNa}, \quad (8)$$

where $\hat{P}_a$, $\hat{P}_{Ra}$, $\hat{P}_B$ and $\hat{P}_{RNa}$ are the same quantities shown in the expression (8). We have that a reasonable approximation for the value of λ in expression (8) is given by

$$\lambda = \frac{n_A\left[\left(\frac{\overline{x}_{U_D}}{\overline{x}_{S_D}}\right)\hat{P}_D\right]\left[\left(\frac{\overline{x}_{U_a}}{\overline{x}_{S_a}}\right)\hat{P}_a\right]}{n_B + n_A\left[\left(\frac{\overline{x}_{U_D}}{\overline{x}_{S_D}}\right)\hat{P}_D\right]\left[\left(\frac{\overline{x}_{U_a}}{\overline{x}_{S_a}}\right)\hat{P}_a\right]} \quad (9)$$

5. Preliminary results

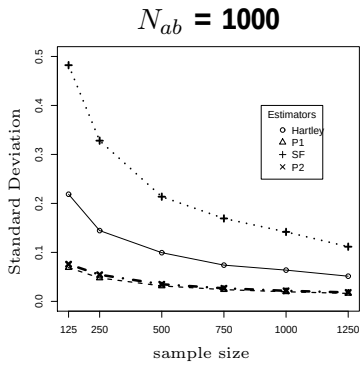
The performance of the proposed estimators 1 and 2, based on the population proportion squared error and relative bias was evaluated through a Monte Carlo simulation study, and we aim to present results based on r = 10000 replicates. The performance of these estimators is compared with original strategies showed previously, and investigated for some dual frame scenarios, as shown in the table 1, for N = 4500. The study focused on the situation where a simple random sampling design is used in each frame.

The estimators were evaluated in terms of relative bias, standard deviation and mean square error (MSE). The initial analysis showed that the use of auxiliary information for estimating proportions, as was expected, contributed to a decrease in variance, when compared with the estimators that do not use the information of auxiliary variables. We expected that the proportion estimator 3 shows the same results when we use the auxiliary information in the next step of this work. We verify that the estimator $\hat{P}_2$ remains better than estimator $\hat{P}_1$ when samples sizes in each frame are increased. The standard deviations of estimators for sample are shown in figure 2 for N_{ab} = 1000.

Table 2: Simulation scenarios

Scenarios	N_a	N_b	N_{ab}
1	2000	2000	500
2	1750	1750	1000
3	1500	1500	1500
4	1250	1250	2000
5	1000	1000	2500

Figure 2: Standard deviations of estimators for proportion with N_{ab} = 1000



5.1 Properties of proposed estimators

The often-ocurring problem of estimating a population parameter, as shown by Särndal (1992), that can be expressed as a function of q population total. In this study we considered q = 2, because we have two populational parameters for response variable and the auxiliary variable and our main concern is to consider the Taylor linearization technique to find approximated variances of all proposed estimators, because they are non-linear functions of the estimators of the parameters of the response variable and auxiliary variable. The approximated variances will be used to find asymptotic properties for all proposed estimators in this work.

6 Conclusions

The initial simulation results shows that some of the proposed ratio type estimators have better performance than original estimators. We intend to do the same study with the estimator 3 and compare the results in order to determine which is the best estimator when we consider the scenarios considered in the simulation study. Furthemore, we intend to show all asymptotic properties of proposed estimators, using the results presented by Isaki and Fuller (1982) and Robinson and Särndal (1983) in a general scenario when we consider a general regression model for estimating a population

References

Bankier (1986). Estimators based on several stratified samples with applications to multiple frame surveys. Journal of the American Statistical Association.

Kalton, G. & Anderson, D. W (1986). Sampling Rare Populations. Journal of Royal Statistical Society, Ser. A, 149,65-82.

Hartley, H. O. (1962). Multiple Frame Surveys. Proceeding of the Social Statistics Association - ASA.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of American Statistical Association.

Robinson, P. M. & Särndal, C. E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. Sankhya: The Indian Journal of Statistics, 45, Series B, 240-248.

Skinner, C. J. (1991), On the e□ciency of raking ratio estimator for dual frame surveys, Journal of the American Statistical Association, 1991,86,779-784.

Ezzati, T. M., Ho□man, K., Judkins, D. R., Massey, J. T., & Moore, T. F. (1995), A dual frame design for sampling elderly minorities and persons with disabilities, Statistics in Medicine, 14, 571-583.

Lohr, S. L. and Rao, J. N. K. (2000), Inference from dual frame surveys, Journal of the American Statistical Association, 2000, 95, 449, 271-280.