

Center Sampling and Dual Frame surveys: comparisons among estimators for the total

Campionamento per Centri e Dual Frame: confronti fra stimatori del totale

Fulvia Mecatti

Dipartimento di Statistica, Università di Milano-Bicocca, fulvia.mecatti@unimib.it

Riassunto: Il campionamento per centri è una recente proposta utile nelle indagini sulla popolazione straniera regolare e irregolare e presenta apprezzabili analogie con la tecnica nota in letteratura come Multiple Frame Surveys. Nel presente lavoro due stimatori proposti per il totale nel campionamento per centri sono adattati al caso del Dual Frame e confrontati, sia sotto il profilo inferenziale che sotto quello operativo, con gli stimatori Dual Frame più noti e che si presentano come possibili *competitors*.

Keywords: Elusive Finite Population, Overlapping Frames; Simulations

1. Introduction

In sampling from finite populations often a single and complete list or frame of all units is not available and two or more incomplete lists have to be used. Particularly, in a dual frame survey, two sampling frames, A and B , together cover the parent population. Two independent simple random samples are selected from A and B and sample data are combined to estimate quantities of interest. For instance frame A could be a list of Italian universities offering a Ph.D. in statistics and frame B could list Italian universities offering a second level or a master's degree in statistics: the frames overlap and together cover the population of Italian universities offering graduate degrees in statistics.

Following the classical notation of Hartley (1962), the target population with size N consisting of the union of the overlapping frames A and B of sizes N_A and N_B , can be separated in three disjoint sets, called "Domains": $a = A \cap \bar{B}$, $b = \bar{A} \cap B$ and $ab = A \cap B$ of size N_a, N_b and N_{ab} respectively so that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ab}$ and $N = N_a + N_b + N_{ab}$. Usually N_A and N_B are known, but N_{ab} is unknown. Further, let Y_i be a value associated with the population unit i and define the totals $Y_a = \sum_a Y_i$, $Y_b = \sum_b Y_i$ and $Y_{ab} = \sum_{ab} Y_i$. Then the population total, taken as the parameter of interest, is $Y = Y_a + Y_b + Y_{ab}$.

Estimation problems for dual frame surveys have been first investigated by Hartley (1962, 1974). Lund (1968) and Fuller and Burmeister (1972) improved Hartley's results. More recently Bankier (1986), Skinner (1991), Skinner and Rao (1996) proposed alternative estimation techniques; Lohr and Rao (2000) discussed variance estimators.

Dual or multiple frame surveys have several analogies with Center Sampling.

Center Sampling has been developed in Italy in connection with surveys of immigrant population legally and illegally resident in the country where the population size N is unknown and exhaustive lists of all units are not available.

Center Sampling relies upon the foreign immigrant habit to congregate in particular places for social contacts, health care, religion and so on, so that every immigrant can be found, with some regularity, at one or more of these places, called “Centers”. For instance a mosque is a center for Muslim people; the train station is a usual meeting point for homeless and illegal immigrants so that it can be considered a center; also incomplete lists, such as official registers of birth and marriages represent a kind of center. Generally, either center sizes and overlapping among centers are unknown.

In a Center Sampling design consisting of L centers ($L \geq 2$) each unit is joined by a vector called “Profile”, which elements are equal to 1 if unit attends center l and are equal to 0 otherwise ($l = 1, \dots, L$).

Each unit has a unique profile in the set of the $2^L - 1$ possible profiles; so that a partition of the population is reached by grouping units with the same profile.

Inferential problems in Center Sampling has been investigated by Mecatti (2000) and Mecatti and Migliorati (2001a, 2001b).

When incomplete lists with known sizes are used as centers, centers are essentially frames and the collections of units with the same profile are domains.

The aim of this paper is to introduce estimators for Center Sampling in the Dual Frame context and to analyze their performance and their inferential properties with respect to major competitors .

In section 2 the most known estimators for the total Y in Dual Frame surveys are outlined and two unbiased estimators derived from Center Sampling framework are proposed. Estimators are compared and discussed in section 3.

2. Estimators for the total in Dual Frame surveys

Let n_A and n_B be the sample sizes for frames and $n_a, n_b, n'_{ab}, n''_{ab}$ for domains where n'_{ab} refers to units belonging to the overlapping domain and sampled from frame A and n''_{ab} refers to units in the overlapping domain selected from frame B . Further define the sample totals $y_A, y_B, y_a, y_b, y'_{ab}$ and y''_{ab} .

In the seminal paper, Hartley (1962) suggests to estimate Y by combining data from domains in an unbiased estimate with the following form:

$$\hat{y}_H = f_A^{-1} y_a + p f_A^{-1} y'_{ab} + (1-p) f_B^{-1} y''_{ab} + f_B^{-1} y_b \quad (1)$$

where $f_A = n_A/N_A$ and $f_B = n_B/N_B$ are the sample fractions and p is a constant to be chosen to minimize $\text{var}(\hat{y}_H)$.

Lund (1968) improved Hartley’s basic estimate by:

$$\hat{y}_L = f_A^{-1} y_a + \left[p f_A^{-1} n'_{ab} + (1-p) f_B^{-1} n''_{ab} \right] \frac{y_{ab}}{n_{ab}} + f_B^{-1} y_b \quad (2)$$

where $y_{ab} = y'_{ab} + y''_{ab}$ and $n_{ab} = n'_{ab} + n''_{ab}$.

$\hat{y}_L$ is also unbiased and $\text{var}(\hat{Y}_L) \leq \text{var}(\hat{Y}_H)$ for all values of p .

Again p is chosen to minimize the variance so that it results in a function of unknown population quantities and it must be estimated. Lund suggests an estimate $\hat{p}$; by substituting $\hat{p}$ in (2) yields a bias of order $O(1)$. Finally, the optimum for $\text{var}(\hat{Y}_L)$ provides also with the optimum allocation of sample size $n = n_A + n_B$ subject to a linear cost function. Lund derives an iterative formula for the optimum ratio n_A/n_B , so that the optimum allocation is always approximated. Furthermore optimum allocation results in a function either of the domains' variances and of the domains' means and sizes.

A further improvement of Hartley's estimate is the estimate by Fuller and Burmeister (1972):

$$\hat{y}_{FB} = (N_A - \hat{N}_{ab}) \frac{y_a}{n_a} + \hat{N}_{ab} \bar{y}_{ab,FB} + (N_B - \hat{N}_{ab}) \frac{y_b}{n_b} \quad (3)$$

where $\bar{y}_{ab,FB} = [n'_{ab}(1 - f_B) + n''_{ab}(1 - f_A)]^{-1} [(1 - f_B)y'_{ab} + (1 - f_A)y''_{ab}]$ and $\hat{N}_{ab}$ is an estimate for the unknown size of the overlapping domain which coincides with the smallest root of a quadratic equation.

An expression for $a \text{var}(\hat{Y}_{FB})$ is given in a suitable asymptotic context; $\hat{y}_{FB}$ has bias of order $O(n^{-1})$ and $a \text{var}(\hat{Y}_{FB}) \leq \text{var}(\hat{Y}_L)$ for all p . Furthermore $\hat{Y}_{FB}$ is heuristically interpretable as a maximum likelihood estimator (Skinner, 1991).

Assuming that units selected from the overlapping domain are identifiable, Bankier (1986) proposes to use the Horvitz-Thompson estimator. If the finite population corrections are negligible also the probabilities of duplicated units are negligible so that the Horvitz-Thompson estimate can be approximated by a "simple" estimate; by applying r successive ratio adjustments to the simple estimate the Raking Ratio estimate $\hat{y}_{RR}^{(r)}$ is obtained. The limiting form of $\hat{y}_{RR}^{(r)}$ as $r \rightarrow \infty$ is (Skinner, 1991):

$$\hat{y}_{RR}^{(\infty)} = (N_A - \tilde{N}_{ab}) \frac{y_a}{n_a} + \tilde{N}_{ab} \frac{y_{ab}}{n_{ab}} + (N_B - \tilde{N}_{ab}) \frac{y_b}{n_b} \quad (4)$$

where $\tilde{N}_{ab}$ is an estimate for the unknown size of the overlapping domain which coincides with the smallest root of a quadratic equation.

In a suitable asymptotic context it is shown that $a \text{var}(\hat{Y}_{RR}^{(\infty)}) \geq a \text{var}(\hat{Y}_{FB})$ but the relative efficiency is close to unity in special cases. Furthermore $\hat{y}_s$ and $\hat{y}_{RR}^{(r)}$ can be easily extended to multiple frame designs while estimates based on Hartley's approach become complex or even fail as the number of frames increases.

Finally, two unbiased estimate which rise from Center Sampling design (Mecatti, 2000; Mecatti-Migliorati, 2001a, 2001b) can be adapted to the Dual Frame context where they assume the simple forms:

$$\hat{y}_m = f_A^{-1} y_a + (2f_A)^{-1} y'_{ab} + (2f_B)^{-1} y''_{ab} + f_B^{-1} y_b \quad (5)$$

and:

$$\hat{y}_c = f_A^{-1} y_a + (f_A + f_B)^{-1} y_{ab} + f_B^{-1} y_b \quad (6)$$

3. Comparisons among estimators

Both Center Sampling estimates $\hat{y}_m$ and $\hat{y}_c$ are unbiased for Y and they satisfy the approach suggested by Hartley. Particularly $\hat{y}_m$ has the Hartley form (1) with fixed $p = 1/2$ (the inverse of the “multiplicity”, i.e. the number of 1’s in the profile, in the Multiple Frame case) while $\hat{y}_c$ has the Lund form (2) with fixed $p = f_A(f_A + f_B)^{-1}$ as well as the initial form of the Raking Ratio estimate $\hat{y}_{RR}^{(0)}$ (i.e. of the Horvitz-Thompson form).

Consequently they describe estimators that are less or equally efficient than Lund’s estimator only with the optimum choice of p while they can result more efficient as $\hat{p}$ affects unbiasedness and variance.

For both estimators $\hat{Y}_m$ and $\hat{Y}_c$ are given the exact variances while for other estimators considered the variance is only approximated (Lund) or given asymptotically (Fuller-Burmeister and Raking Ratio). Hence variance of estimators $\hat{Y}_m$ and $\hat{Y}_c$ can be easily estimated: simple estimates, in a closed form and unbiased for negligible finite population corrections, are proposed in Mecatti-Migliorati (2001a, 2001b). Furthermore $\hat{Y}_m$ allows for the optimum allocation of the sample size $n = n_A + n_B$ subject to a linear cost function as a function of domains variability only.

Finally both estimators $\hat{Y}_m$ and $\hat{Y}_c$ apply, by definition, to any number of frames.

Estimators derived from Center Sampling design then represent feasible alternatives to the more popular Dual Frame estimators.

Comparisons among estimators with regard to the inferential properties are also advisable. Hence a simulation study has been performed in order to analyze the considered estimators either under operational aspects and with regard to unbiasedness, consistency and relative efficiency by exploring a variety of different initial framework and for different values of population and sample sizes.

Essential References

- Hartley H.O. (1962) Multiple Frame Surveys, in *Proceedings of the Social Statistical Section, ASA*, 203-206.
- Mecatti F. (2000) La stima della media nel campionamento per centri, *Statistica*, To appear.
- Mecatti F., Migliorati S. (2001a) Center sampling: theory and estimation, *Tech. Rep. 01-06*, Department of Statistics, Pennsylvania State University.
- _____ (2001b) Confronti fra stimatori per la media nel campionamento per centri, Submitted.
- Skinner C.J. (1991) On the efficiency of Raking Ratio estimation for multiple frame surveys, *Journal of the American Statistical Association*, 86, 779-784.