

New Developments in Center Sampling

Nuovi Sviluppi nel Campionamento per Centri

Sonia Migliorati

Dipartimento di Statistica, via Bicocca degli Arcimboldi 8, 20126 Milano
sonia.migliorati@unimib.it

Riassunto: il campionamento per centri (Blangiardo, 1996) risulta essere una delle poche soluzioni al problema delle indagini relative a popolazioni straniere o, in generale, popolazioni per le quali non si dispone di una lista oppure di più liste incomplete. In tale ambito assume particolare rilevanza il “profilo” che informa circa la struttura di afferenza ai centri da parte delle unità statistiche. Tramite le numerosità di profilo è possibile risalire sia alle numerosità di centro sia alla dimensione dell’intera popolazione. Stante l’impossibilità, dovuta al disegno di campionamento, di pervenire a stime delle numerosità di profilo senza ricorrere ad ipotesi ausiliarie, il presente lavoro propone uno stimatore per le versioni relative delle medesime, versioni che entrano in via naturale nella stima della media di un carattere quantitativo oggetto d’interesse.

Keywords: Elusive Populations, Detectability, Profile.

1. Introduction

Center Sampling (CS) (Blangiardo, 1996) proposes to deal with surveys of immigrant populations. In such a framework the population size N is unknown and also there does not exist an exhaustive list of the target population. The units show both problems of anonymity and of detectability (Thompson, 1992). The only way to sample immigrants (as well as many other undetectable populations) is to pick them out at particular places, call centers, where they gather for different reasons (e.g. social, religious and so on).

The number L of centers is known and finite; moreover it is assumed that the centers cover the whole population, but they can overlap, i.e. a unit can obviously attend many of them. Notice that nothing is known about the center sizes N_l ($l = 1, \dots, L$) and about overlapping as well. Since units are not identifiable through a label, a fundamental tool for each of them in CS is the “profile”. This is a vector of length L whose l^{th} entry equals 1 if the unit attends the l^{th} center and 0 otherwise (and, according to the above hypotheses, with at least one non-null entry). Note that the $2^L - 1 \times L$ matrix $\mathbf{U}$ of all possible ordered L -tuples of zeroes and ones, i.e. the collection of all the profiles, is perfectly known, and that each unit has a unique profile. Therefore profiles enable the partitioning the population while centers do not. Letting $\mathbf{u}_r = [u_{r1}, \dots, u_{rL}]$ denote the r^{th} row of $\mathbf{U}$ ($r = 1, \dots, 2^L - 1$) and $N_{\mathbf{u}_r}$ the unknown number of units with profile $\mathbf{u}_r$, it is clear that $\sum_r N_{\mathbf{u}_r} = N$ while in general $\sum_l N_l \geq N$. Hence estimating the profile sizes $N_{\mathbf{u}_r}$ becomes a fundamental issue since they represent a basic source of

information both on the overall population size N and on center sizes N_l . In fact it is immediate to show that

$$N_l = \sum_r N_{\mathbf{u}_r} u_{rl} = \mathbf{N}'_{\mathbf{u}} \mathbf{u}_{.l} \quad (1)$$

where $\mathbf{N}'_{\mathbf{u}} = [N_{\mathbf{u}_1}, \dots, N_{\mathbf{u}_{2^L-1}}]$ is the vector of profile sizes and $\mathbf{u}_{.l}$ is the l^{th} column of $\mathbf{U}$.

In CS a simple random sample of size n_l ($l=1, \dots, L$) is chosen from each center independently under the hypothesis that the N_l units who attend the l^{th} center are present almost surely. The overall sample size $n = \sum_l n_l$ is a fixed number. Such a sample design bars any chance of estimating profile sizes (and consequently center as well as population sizes) (Migliorati and Terzera, 2001), without assuming at least one center being an incomplete list, i.e. a center whose size is known.

Yet, it is possible to estimate the relative profile sizes $N_{\mathbf{u}_r} / N$ and hence the “weight” $\alpha_l = N_l / N$ of each center with respect to the population size. This is a necessary step to attain an operative estimate of the mean of a quantitative characteristic of interest present on the population (Mecatti and Migliorati, 2001).

The aim of this paper is to provide a new estimate of the relative profile sizes.

2. The new proposal

Let $f_{\mathbf{u}_r, l}$ ($r=1, \dots, 2^L-1; l=1, \dots, L$) be the frequency of profile $\mathbf{u}_r$ in the sample from the l^{th} center and $f_{\mathbf{u}_r} = \sum_l f_{\mathbf{u}_r, l}$ the overall frequency of such a profile. It is straightforward to show that the corresponding estimator $F_{\mathbf{u}_r}$ is biased for $N_{\mathbf{u}_r}$ since we have:

$$E(F_{\mathbf{u}_r}) = N_{\mathbf{u}_r} \sum_l \frac{n_l u_{rl}}{N_l} \quad (2)$$

i.e. profiles with many entries equal to one are over-represented in the sample. In fact the bias factor depends on the unknown center sizes N_l which can not be easily estimated. A first proposal for estimating $N_{\mathbf{u}_r} / N$ (Blangiardo, 1996) relies on the sample quantities $f_{k/l} = \sum_r f_{\mathbf{u}_r, l} u_{rk}$ ($l, k=1, \dots, L; l \neq k$). Such quantities involve couples of centers and indicate the number of units sampled in the l^{th} center and attending the k^{th} too. Once a “base” center l^* has been chosen, i.e. one for which $f_{l/l^*} \neq 0$ for each $l \neq l^*$, the estimate for the relative size of profile $\mathbf{u}_r$ has the form:

$$\left[f_{\mathbf{u}_r} \left(\sum_l \frac{n_{l*} u_{rl} f_{l*/l}}{f_{l/l*}} \right)^{-1} \right] / \left[\sum_r f_{\mathbf{u}_r} \left(\sum_l \frac{n_{l*} u_{rl} f_{l*/l}}{f_{l/l*}} \right)^{-1} \right] \quad (3)$$

Estimate (3) is rather complicated to deal with and it involves many random quantities. In fact recalling (1) it is clear that the center sizes N_l which appear in (2) are themselves function of the profile sizes $N_{\mathbf{u}_r}$. Thus there is no need to involve the sample frequencies $f_{l*/l}$ and $f_{l/l*}$ and (2) can be rewritten as:

$$E(F_{\mathbf{u}_r}) = N_{\mathbf{u}_r} \sum_l n_l u_{rl} \left(\sum_r N_{\mathbf{u}_r} u_{rl} \right)^{-1} = N_{\mathbf{u}_r} (\mathbf{n}' \mathbf{D} \mathbf{u}_r') \quad (4)$$

($r = 1, \dots, 2^L - 1$) where $\mathbf{n}' = [n_1, \dots, n_L]$ is the vector of the sample sizes and $\mathbf{D}$ is the diagonal matrix of dimension $L \times L$ whose l^{th} entry on the diagonal is $(\mathbf{N}'_{\mathbf{u}_l} \mathbf{u}_l')^{-1}$.

As r varies from 1 to $2^L - 1$, equations (4) can be simultaneously solved with respect to the unknown $N_{\mathbf{u}_r}$. As a matter of fact if $\mathbf{u}_r$ is the profile with all one entries, then the corresponding equation is not independent from the others: substituting all the remaining equations in it we obtain the identity $\sum_r E(F_{\mathbf{u}_r}) = \sum_l n_l$. Hence only a solution for the relative profile sizes $N_{\mathbf{u}_r} / N$ can be found. Notice also that the system of equations (4) is not linear so that an unbiased estimate of $N_{\mathbf{u}_r} / N$ is not available. In any case, once the system has been solved, an estimate can be achieved substituting the expected values with the values of the corresponding random variables.

Let us consider, for example, the case of $L = 2$ centers. Then solving equations (4) we have:

$$\begin{aligned} \frac{N_{\mathbf{u}_1}}{N} &= \frac{n_1 n_2 - n_2 E(F_{\mathbf{u}_3}) - n_1 E(F_{\mathbf{u}_2}) + E(F_{\mathbf{u}_3}) E(F_{\mathbf{u}_2})}{n_1 n_2 - E(F_{\mathbf{u}_3}) E(F_{\mathbf{u}_2})} \\ \frac{N_{\mathbf{u}_2}}{N} &= \frac{E(F_{\mathbf{u}_2}) [n_1 - E(F_{\mathbf{u}_3})]}{n_1 n_2 - E(F_{\mathbf{u}_3}) E(F_{\mathbf{u}_2})} \quad \frac{N_{\mathbf{u}_3}}{N} = \frac{E(F_{\mathbf{u}_3}) [n_2 - E(F_{\mathbf{u}_2})]}{n_1 n_2 - E(F_{\mathbf{u}_3}) E(F_{\mathbf{u}_2})} \end{aligned} \quad (5)$$

Thus an estimate of $N_{\mathbf{u}_r} / N$ ($r = 1, 2, 3$) is:

$$\frac{n_1 n_2 - n_2 f_{\mathbf{u}_3} - n_1 f_{\mathbf{u}_2} + f_{\mathbf{u}_3} f_{\mathbf{u}_2}}{n_1 n_2 - f_{\mathbf{u}_3} f_{\mathbf{u}_2}}, \quad \frac{f_{\mathbf{u}_2} [n_1 - f_{\mathbf{u}_3}]}{n_1 n_2 - f_{\mathbf{u}_3} f_{\mathbf{u}_2}}, \quad \frac{f_{\mathbf{u}_3} [n_2 - f_{\mathbf{u}_2}]}{n_1 n_2 - f_{\mathbf{u}_3} f_{\mathbf{u}_2}} \quad (6)$$

Notice that estimates (6) show a simpler structure than (3), they demand no choice of a “base” center and they are positive.

If $L > 2$ a numerical solution of equations (4) is needed, yet the case of two centers is quite relevant as it emerges from the literature devoted to dual frame surveys (Lohr and Rao, 2000; Fuller and Burmeister, 1972) which show some evident analogies with CS.

3. Results and conclusions

Delta method enables one to approximate the expected values and the variances of the estimators given by (6). Note that the random variables $F_{u_r,l}$ described by $f_{u_r,l}$ are Binomials with parameters n_l and

$$p_{rl} = N_{u_r} u_{rl} / N_l \quad (7)$$

($r = 1, \dots, 2^L - 1; l = 1, \dots, L$), $F_{u_r,l}$ and $F_{u_r,k}$ are independent ($l, k = 1, \dots, L; l \neq k$) while $(F_{u_1,l}, \dots, F_{u_{2^L-1},l})$ are Multinomials with parameters n_l and $(p_{1l}, \dots, p_{2^L-1,l})$ where p_{rl} are given by (7). Hence it is trivial to show that the following relations hold:

$$\begin{aligned} E(F_{u_2}) &= O(n_2); \quad E(F_{u_2} F_{u_3}) = O(n_1 n_2); \quad \text{Var}(F_{u_2}) = O(n_2); \\ E[(F_{u_2})^2] &= O(n_2^2); \quad \text{Var}(F_{u_2} F_{u_3}) = O(n_1 n_2); \\ E[(F_{u_2} F_{u_3})^2] &= O(n_1^2 n_2^2); \quad E[(F_{u_2} F_{u_3})^3] = O(n_1^3 n_2^3) \end{aligned}$$

This ensures that the estimators given by (6) are asymptotically unbiased and consistent as both sample sizes n_1 and n_2 diverge. The case of $L > 2$ centers has been investigated through a simulation study based on real data (AA.VV., 2000). The new estimates have been obtained numerically with the procedure described in Section 2. Their performances have been compared with the ones of the estimators given by (3). The comparison has been carried out for different overlapping structures among centers. The study of the consequences of the new proposal on the estimate of the mean of a quantitative characteristic of interest will be a future matter of concern.

References

- AA.VV. (2000) L'immigrazione straniera nell'area milanese, *Rapporto Statistico dell'Osservatorio Fondazione Cariplo-I.S.MU.*, Provincia di Milano.
- Blangiardo G. C. (1996) Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera in Italia, *Studi in Onore di Giampiero Landenna*, Giuffrè, Milano, 15-30.
- Fuller W. A., Burmeister L. F. (1972) Estimators for samples selected from two overlapping frames, in *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.
- Lohr S. L., Rao J. N. K. (2000) Inference from dual frame surveys, *Journal of the American Statistical Association*, 95, 271-280.
- Mecatti F., Migliorati S. (2001) Center sampling: theory and estimation, *Tech. Rep. 01-06*, Department of Statistics, Pennsylvania State University.
- Migliorati S., Terzera L. (2001) Una proposta di stima della numerosità nel campionamento per centri, *Convegno "Metodi di Inferenza Statistica per Problemi Complessi"* Bressanone 2-4 novembre 2001.
- Thompson S. K. (1992) *Sampling*, Wiley, New York.