

Estimation for finite population mean in double sampling

Stima della media di una popolazione finita nel campionamento doppio

Giancarlo Diana, Chiara Tommasi

Department of Statistics

University of Padova

diana@stat.unipd.it – microbo@stat.unipd.it

Riassunto: In questo lavoro vengono proposte delle classi generali di stimatori per la media di una popolazione finita, quando si usa uno schema di campionamento in due fasi. Tali classi dipendono da quali medie campionarie, di due variabili ausiliarie, vengono considerate. Per ciascuna classe si fornisce il minimo MSE e si dimostra che esiste almeno uno stimatore tipo regressione a catena che lo raggiunge.

Keywords: two-phase sampling, auxiliary variable, regression type estimator.

1. Introduction

In this paper we propose general classes of estimators for the finite population mean in double sampling, when two auxiliary variables are considered. In that case the most commonly used auxiliary quantities are the first and the second phase sample means and a specific class of estimators can be identified by the involved auxiliary quantities. Thus, we may have as many classes as the possible choices of auxiliary quantities.

In each class we can obtain an estimator which minimizes the MSE (up to terms of order n^{-1}). This estimator has a good interpretation, it is a chain regression type estimator and it will be denoted as the “best” estimator in the class. Actually, any other estimator which is equivalent, at the first order of approximation, to the best estimator is optimum as well. Many authors dealt with the problem of estimating a population mean using auxiliary information. Thus, a lot of different estimators were proposed following different ideas but they can be classified as members of some specific general class. Sometimes these estimators are optimal in this class but other times they are not. Some examples will be provided in Section 3 while a quite good review of the proposed estimators in double sampling is given in Cuscunà (1997).

In the second section we define three general classes of estimators, which depend on different auxiliary quantities. For each class we provide the best estimator expression and the minimum MSE.

2. General classes of estimators

Let $\mathcal{U} = \{1, \dots, i, \dots, N\}$ be a finite population, Y the study variable and X, Z two auxiliary variables taking values Y_i, X_i and Z_i for the i -th population unit. We are interested in estimating the population mean of Y , $\bar{Y} = \sum_{i=1}^N Y_i / N$.

When X and Z are related with Y but no information is available about the population

mean of X , the estimation of $\bar{Y}$ can be based on a double sample. In this scheme, we assume that a preliminary large sample of n' ($n' < N$) units is drawn by a simple random sample without replacement (SRSWOR). At this phase only X and Z can be measured. In a second phase a smaller sample of size n ($n < n'$) is drawn from the first phase sample, by a SRSWOR as well. At this phase all the variables Y , X and Z can be observed. Let $\bar{x}'$, $\bar{z}'$ be the sample means of X and Z at the first phase, while $\bar{y}$, $\bar{x}$ and $\bar{z}$ denote the sample means of Y , X and Z at the second phase. Finally, let $\bar{U} = \sum_{i=1}^N U_i / N$ and $S_U^2 = \sum_{i=1}^N (U_i - \bar{U})^2 / (N - 1)$, where U denotes any variable.

If the whole sample information is available about both X and Z , then a general class of estimators is

$$\bar{y}_4 = g(\bar{y}, t) \quad (1)$$

where $t = (\bar{x}, \bar{z}, \bar{z}', \bar{x}')$ and g is a function such that

- a. $g : \mathcal{S} \rightarrow \mathbb{R}$ where $\mathcal{S} \in \mathbb{R}^5$ is a convex and bounded set which contains the point $(\bar{Y}, T)$, where $T = E(t) = (\bar{X}, \bar{Z}, \bar{Z}, \bar{X})$.
- b. It is a continuous and bounded function in $\mathcal{S}$.
- c. Its first and second partial derivatives are continuous and bounded in $\mathcal{S}$.
- d. $g(\bar{y}, T) = \bar{y}$.

Let

$$g_0 = \left. \frac{\partial g(\bar{y}, t)}{\partial \bar{y}} \right|_{(\bar{y}, t) = (\bar{Y}, T)} \quad \text{and} \quad g_i = \left. \frac{\partial g(\bar{y}, t)}{\partial t_i} \right|_{(\bar{y}, t) = (\bar{Y}, T)} \quad i = 1, \dots, 4$$

be the partial derivatives of g with respect to the first component, $\bar{y}$, and other components, t_i , $i = 1, \dots, 4$, respectively. From point (d) above, we have that $g(\bar{Y}, T) = \bar{Y}$ and $g_0 = 1$.

Expanding $\bar{y}_4$ at the point $(\bar{Y}, T)$ in a second order Taylor's series we have

$$\begin{aligned} \bar{y}_4 &\cong g(\bar{Y}, T) + (\bar{y} - \bar{Y}) g_0 + \sum_{i=1}^4 (t_i - T_i) g_i \\ &= \bar{y} + (\bar{x} - \bar{X}) g_1 + (\bar{z} - \bar{Z}) g_2 + (\bar{z}' - \bar{Z}) g_3 + (\bar{x}' - \bar{X}) g_4. \end{aligned} \quad (2)$$

Since the population mean of X is unknown, we have to impose the constraint $g_4 = -g_1$ and equation (2) becomes

$$\bar{y}_4 \cong \bar{y} + (\bar{x} - \bar{x}') g_1 + (\bar{z} - \bar{Z}) g_2 + (\bar{z}' - \bar{Z}) g_3 \quad (3)$$

Minimizing the first order approximation of $MSE(\bar{y}_4)$ with respect to g_i , $i = 1, \dots, 3$, we get the following "best" values

$$\begin{bmatrix} g_1^* \\ g_2^* \\ g_3^* \end{bmatrix} = - \begin{bmatrix} \beta_{Y,X|Z} \\ \beta_{Y,Z|X} \\ \beta_{Y,X|Z} \beta_{X,Z} \end{bmatrix}, \quad \text{where} \quad \beta_{U,V|W} = \frac{S_U}{S_V} \frac{\rho_{UV} - \rho_{UW}\rho_{VW}}{1 - \rho_{VW}^2}$$

and $\beta_{X,Z}$ is the regression coefficient of X on Z . Here $\rho_{.,.}$ denotes the correlation coefficient between the specified variables.

Replacing g_i^* , $i = 1, \dots, 3$, in (3) we get the following chain regression type estimator,

$$\bar{y}_{4reg} = \bar{y} - \beta_{Y,X|Z} \left\{ \bar{x} - \left[\bar{x}' - \beta_{X,Z} (\bar{z}' - \bar{Z}) \right] \right\} - \beta_{Y,Z|X} (\bar{z} - \bar{Z}),$$

whose bias is of order n^{-1} . We call $\bar{y}_{4reg}$ the “best” estimator in the class g but any other estimator which is equivalent to $\bar{y}_{4reg}$, at the first order of approximation, is optimum as well.

Let us denote the minimum first order approximation of $\text{MSE}(\cdot)$ as $\text{MSE}^*(\cdot)$.

In class g this minimum MSE (except terms of $O(n^{-2})$) has a simple form,

$$\text{MSE}^*(\bar{y}_4) = S_Y^2 \left[\left(\frac{1}{n'} - \frac{1}{N} \right) (1 - \rho_{Y,Z}^2) + \left(\frac{1}{n} - \frac{1}{n'} \right) (1 - \rho_{Y,\tilde{V}}^2) \right], \quad (4)$$

where $\tilde{V} = (X, Z)^T$ and $\rho_{Y,\tilde{V}}$ is a multiple correlation coefficient.

When only X is measured in both phases, then $t_3 = (\bar{x}, \bar{z}', \bar{x}')$ can be used instead of t and we get another class of estimators. In this case the best regression type estimator is

$$\bar{y}_{3reg} = \bar{y} - \beta_{YX}(\bar{x} - \bar{x}') - \beta_{YZ}(\bar{z}' - \bar{Z}), \quad (5)$$

and

$$\text{MSE}^*(\bar{y}_3) = S_Y^2 \left[\left(\frac{1}{n'} - \frac{1}{N} \right) (1 - \rho_{Y,Z}^2) + \left(\frac{1}{n} - \frac{1}{n'} \right) (1 - \rho_{Y,X}^2) \right]. \quad (6)$$

Finally, if no information about Z is available, we have the class based on $t_2 = (\bar{x}, \bar{x}')$. In this class the standard regression estimator $\bar{y}_{2reg} = \bar{y} - \beta_{YX}(\bar{x} - \bar{x}')$ is the best estimator and

$$\text{MSE}^*(\bar{y}_2) = S_Y^2 \left[\frac{1}{n'} - \frac{1}{N} + \left(\frac{1}{n} - \frac{1}{n'} \right) (1 - \rho_{Y,X}^2) \right]. \quad (7)$$

It is easy to prove that $\text{MSE}^*(\bar{y}_4) \leq \text{MSE}^*(\bar{y}_3) \leq \text{MSE}^*(\bar{y}_2)$, therefore $\bar{y}_{4reg}$ is the preferable estimator.

Note that population regression coefficients involved in the best estimators $\bar{y}_{4reg}$, $\bar{y}_{3reg}$ and $\bar{y}_{2reg}$, are usually unknown. Anyway, for practical purposes, they can be replaced with suitable estimates without efficiency loss.

3. Conclusions

The main aim of this paper is to show that no estimator based on t , t_3 or t_2 , can be more efficient than the corresponding best estimators. As a matter of fact, a lot of the proposed estimators are less efficient than the best ones and hence they are completely uninteresting. On the other hand, those estimators which reach the minimum MSE^* are equivalent to the corresponding best estimators, at the first order of approximation.

Let us give some examples for the three previously described classes.

The best estimator $\bar{y}_{4reg}$ for the class $\bar{y}_4$ was proposed by Tripathi and Ahmed (1995) as a regression type estimator, by Gopabandhu and Kulamani (1997) as the optimum estimator in a linear class and finally by Ahmed *et al.* (1998) as the best estimator in a partially defined class

$$\hat{\bar{y}}_4 = h \left(\bar{y}, \frac{\bar{x}}{\bar{x}'}, \frac{\bar{z}}{\bar{z}'}, \frac{\bar{z}}{\bar{Z}} \right).$$

Anyway nobody shows that $\bar{y}_{4reg}$ is the best estimator based on the auxiliary vector t , since $\text{MSE}^*(\bar{y}_4)$ is the MSE lower bound, at the first order of approximation.

On the other hand, about the class based on t_3 , Srivastava *et al.* (1990) propose an estimator which is optimum in a biparametric class and reaches $\text{MSE}^*(\bar{y}_3)$. It turns out to

be equivalent to $\bar{y}_{3reg}$, up to terms of $O(n^{-1})$. Sahoo *et al.* (1994) get $\bar{y}_{3reg}$ as the optimum estimator in a linear one-parametric class while Sahoo and Sahoo (1993) and Sahoo and Sahoo (1999) get it as the optimum estimator in the class

$$\hat{y}_s = d(\bar{y}, \bar{x}, f(\bar{x}', \bar{z}')) .$$

Only this last approach is equivalent to that one followed in the previous section. Finally, even for the class defined by t_2 many other estimators, which are equivalent to the standard estimator, can be found. An example is given by Singh and Gangele (1999).

Therefore, the main suggestion we wish to give is to stop proposing new estimators based on t , t_3 or t_2 since we cannot get estimators more efficient than the “best” estimators, at least at the first order of approximation.

References

- Ahmed M.S., Rahman M.S. and Ahmed R. (1998) General class of chain estimators for a finite population mean using double sampling, *J. Appl. Statist. Sci.*, 7, 185–190.
- Cuscunà G. (1997) *Stimatori Ottimi nel Campionamento a più Fasi*, Tesi di Laurea, Univ. Padova.
- Gopabandhu M. and Kulamani R. (1997) A regression estimator in two-phase sampling in presence of two auxiliary variables, *Metron*, 55, 177–186.
- Sahoo J. and Sahoo L.N. (1993) A class of estimators in two-phase sampling using two auxiliary variables, *J. Indian. Statist. Ass.*, 31, 107–114.
- Sahoo J. and Sahoo L.N. (1999) An alternative class of estimators in double sampling procedures, *Calcutta Statist. Ass. Bull.*, 49, 79–83.
- Sahoo J., Sahoo L.N. and Mohanty S. (1994) An alternative approach to estimation in two-phase sampling using two auxiliary variables, *Biom. J.*, 36, 293–298.
- Singh H.P. and Gangele R.K. (1999) Classes of almost unbiased ratio and product-type estimators in two phase sampling, *Statistica*, 1, 109–124.
- Srivastava S.R., Srivastava S.R. and Khare B.B. (1990) A generalized chain ratio estimator for mean of a finite population, *J. Ind Soc. Ag. Statistics*, 42, 108–117.
- Tripathi T.P. and Ahmed M.S. (1995) A class of estimators for a finite population mean based on multivariate information and general two phase sampling, *Calcutta Statist. Ass. Bull.*, 45, 203–218.