

Estimation for finite population mean under multiple-phase sampling

Stima della media di una popolazione finita nel campionamento multifase

Giancarlo Diana, Paolo Preo¹
Department of Statistical Sciences
University of Padova
giancarlo.diana@unipd.it, paolo.preo@unipd.it

Chiara Tommasi²
Department of Economics
University of Milano
chiara.tommasi@unimi.it

Riassunto: In questo lavoro viene proposta una classe generale di stimatori per la media di una popolazione finita, quando si usa uno schema di campionamento multifase. Tale classe dipende dal numero di fasi considerate e da quali variabili ausiliarie si decide di rilevare in ciascuna fase. Fissati il numero delle fasi e l'ordine di rilevazione delle variabili ausiliarie, per tale classe si fornisce il minimo MSE e uno stimatore di tipo regressione che lo raggiunge. A parità di fasi, cambiando l'ordine di rilevazione delle variabili ausiliarie si ottengono stimatori ottimali diversi. Si determina quindi anche l'ordine ottimale, a cui corrisponde lo stimatore il cui MSE minimo è il più piccolo possibile rispetto a quest'ulteriore criterio di ordinamento.

Keywords: auxiliary variable, regression type estimator, part correlation.

1. Introduction

In this paper we propose a general class of estimators for the finite population mean in the $(k + 1)$ -phase sampling, when k auxiliary variables are observed. At each phase a new variable is observed up to the last phase when the objective variable and all the auxiliary variables are recorded. Therefore, given a set of auxiliary variables, there may be several criteria for choosing which one is the variable to be observed from the first phase, which one from the second phase and so on. One of these criteria is called “observing order”.

The multiple-phase sampling scheme was not widely studied. Mukerjee *et al.* (1987) consider mainly the three-phase case. They do not suggest the best estimator in that case, even if they give an optimal estimator in a linear class which is equivalent to our general class (at the first order of approximation). Ahmed (2003) also proposes a parametric class which is equivalent to our general class. Thus any optimal estimator in the linear or parametric class is also optimal in a more general setting. For any observing order of the auxiliary variables, the minimum MSE (at the first order of approximation) for estimators in the general class is computed. From now on AMSE will denote the first order of approximation for the MSE. Let us recall that in a finite population framework asymptotic has the sense established in Cicchitelli *et al.* (1997). A regression type estimator which reaches the minimum AMSE is also found. In general, the optimal estimator depends on which auxiliary variables are observed at each phase. Thus, the optimal observing order can be found: the auxiliary variables are observed with an order such that the minimum AMSE is the least as possible. Section 2 concerns with computations for getting the minimum AMSE and the optimal regression estimator in the general class. The best choice for the observing order of the auxiliary variables is treated in the last section.

2. A general class of estimators

Let $\mathcal{U} = \{1, \dots, j, \dots, N\}$ be a finite population, Y and X_i , $i = 1, \dots, k$, be the study variable and k auxiliary variables taking values Y_j and X_{ij} for the j -th population unit. We are interested in estimating the population mean of Y , $\bar{Y} = \sum_{j=1}^N Y_j / N$, under the multiple-phase sampling scheme. Specifically, we assume that a first sample of n_1 units ($n_1 < N$) is drawn by a simple random sampling without replacement (SRSWOR). At this phase only X_1 is measured. Then, in the second phase a smaller sample of size n_2 ($n_2 < n_1$) is drawn from the first phase sample, by a SRSWOR as well. At this phase both variables X_1 and X_2 are observed. This procedure goes on up to the last phase when the smallest sample of size n_{k+1} ($n_{k+1} < n_k < \dots < n_1$) is drawn. At this phase all the auxiliary variables and also Y are observed. Let $\bar{x}_i^{(u)}$, $i = 1, \dots, k$ and $u = i, \dots, k+1$ be the sample mean of X_i at the u -th phase and $\bar{y}$ denote the sample mean of Y at the last phase.

The general class of estimator can be defined as a function of $\bar{y}$ and $\bar{x}_i^{(u)}$, $i = 1, \dots, k$ and $u = i, \dots, k+1$. For computational reasons it is convenient to give the following definition

$$\bar{y}_k = g(\bar{y}, w_{11}, w_{12}, \dots, w_{1k}, w_{22}, \dots, w_{2k}, \dots, w_{kk}) \quad (1)$$

where w_{iu} is the difference between the sample means at two subsequent phases, i.e. $w_{iu} = \bar{x}_i^{(u+1)} - \bar{x}_i^{(u)}$ for $i = 1, \dots, k$ and $u = i, \dots, k$. The function g is such that

- a. $g : \mathcal{S} \rightarrow \mathbb{R}$ where $\mathcal{S} \in \mathbb{R}^{1+k(k+1)/2}$ is a convex and bounded set which contains the vector $(\bar{Y}, 0^T)^T$ where 0 denotes the $[k(k+1)/2] \times 1$ null vector.
- 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. $g(\bar{y}, 0^T) = \bar{y}$.

Let

$$g_0 = \left. \frac{\partial g(\bar{y}, w_{11}, w_{12}, \dots, w_{kk})}{\partial \bar{y}} \right|_{(\bar{y}, w_{11}, \dots, w_{kk}) = (\bar{Y}, 0^T)}$$

be the partial derivatives of g with respect to the first component $\bar{y}$ and

$$g_{iu} = \left. \frac{\partial g(\bar{y}, w_{11}, w_{12}, \dots, w_{kk})}{\partial w_{iu}} \right|_{(\bar{y}, w_{11}, \dots, w_{kk}) = (\bar{Y}, 0^T)} \quad u = 1, \dots, k; \quad i = 1, \dots, u$$

the partial derivatives of g with respect to the other components, w_{iu} , $u = 1, \dots, k$ and $i = 1, \dots, u$. From point (d), we have that $g(\bar{Y}, 0^T) = \bar{Y}$ and $g_0 = 1$.

Expanding $\bar{y}_k$ at the point $(\bar{Y}, 0^T)$ in a second order Taylor's series we have

$$\bar{y}_k \approx g(\bar{Y}, 0^T) + (\bar{y} - \bar{Y}) g_0 + \sum_{u=1}^k \sum_{i=1}^u w_{iu} g_{iu} = \bar{y} + \sum_{u=1}^k w_u^T g_u, \quad (2)$$

where $w_u = (w_{1u}, w_{2u}, \dots, w_{uu})^T$ and $g_u = (g_{1u}, g_{2u}, \dots, g_{uu})^T$. Any estimator which minimizes the AMSE($\bar{y}_k$) is an "optimum" estimator in the general class. In order to introduce the AMSE($\bar{y}_k$) let us define

$$\bar{U}_i = \frac{\sum_{j=1}^N U_{ij}}{N} \quad \text{and} \quad S_{U_i, U_l} = \frac{\sum_{j=1}^N (U_{ij} - \bar{U}_i)(U_{lj} - \bar{U}_l)}{N-1}, \quad i, l = 1, 2$$

and let $S_{U_i}^2 = S_{U_i, U_i}$, where $U_i, i = 1, 2$, is any variable. In addition, let S_{Y_u} be the $u \times 1$ vector whose r -th element is the population covariance S_{Y, X_r} with $r = 1, \dots, u$ and S_{uu} be the $u \times u$ symmetric matrix whose (r, s) -th element is the population covariance S_{X_r, X_s} with $r, s = 1, \dots, u$. With this notation, the AMSE($\bar{y}_k$) is

$$\text{AMSE}(\bar{y}_k) = \gamma_0 S_Y^2 + \sum_{u=1}^k \gamma_u g_u^T S_{uu} g_u + 2 \sum_{u=1}^k \gamma_u g_u^T S_{Y_u}, \quad (3)$$

where

$$\gamma_0 = \frac{1}{n_{k+1}} - \frac{1}{N} \quad \text{and} \quad \gamma_u = \frac{1}{n_{u+1}} - \frac{1}{n_u}, \quad u = 1, \dots, k.$$

Minimizing (3) with respect to g_u , we get the following “best” coefficients

$$g_u^* = -S_{uu}^{-1} S_{Y_u}, \quad u = 1, \dots, k.$$

Except for the sign g_u^* is the vector of the partial regression coefficients of Y on the auxiliary variables $X_1, X_2, \dots, X_u$. Replacing g_u^* in (3), we have the minimum AMSE($\bar{y}_k$),

$$\text{AMSE}^*(\bar{y}_k) = \gamma_0 S_Y^2 - \sum_{u=1}^k \gamma_u S_{Y_u}^T S_{uu}^{-1} S_{Y_u} = S_Y^2 \left(\gamma_0 - \sum_{u=1}^k \gamma_u \rho_{Y,1,\dots,u}^2 \right), \quad (4)$$

where $\rho_{Y,1,\dots,u}^2 = S_{Y_u}^T S_{uu}^{-1} S_{Y_u} / S_Y^2$ denotes the multiple correlation coefficient between Y and $X_1, X_2, \dots, X_u$. Thus, AMSE($\bar{y}_k$) is the lower bound for the AMSE of any other estimator based on all the auxiliary variables. A consistent estimator for such lower bound may be got by replacing S_Y^2 and $\rho_{Y,1,\dots,u}^2$ with their sampling estimates. Consistency is meant in the sense established in Cicchitelli *et al.* (1997).

Finally, replacing g_u^* in (2) a regression type estimator which is an optimum estimator in the class (1) can be found. Of course, any other estimator whose first order approximation is equal to $\bar{y} + \sum_{u=1}^k w_u^T g_u^*$ will be optimum as well. Hereafter two estimators are defined to be equivalent when they are equal at the first order of approximation. The optimum coefficients g_u^* may be unknown, since they depend on usually unknown population quantities. However, using suitable estimates of such coefficients in (2), $\hat{g}_u^*$, we get a new estimator which is equivalent to an optimum estimator in the class (1).

In a three-phase sampling scheme, Mukerjee *et al.* (1987) proposed four estimators which cannot be optimal since no one of them coincides with $\bar{y} + \sum_{u=1}^k w_u^T \hat{g}_u^*$. However the authors propose a linear class of estimators which is equivalent to our class. In a similar way, Ahmed (2003) provides a parametric class which is also equivalent to our class. In both cases an optimal estimator in the specific class is optimal in our class too.

3. The observing order of the auxiliary variables

Which auxiliary variables must be observed at each phase is an unsolved question. A possible choice is to take the least expensive auxiliary variable as X_1 , the second least expensive auxiliary variable as X_2 and so on, since $n_1 > n_2 > \dots > n_{k+1}$. However, the auxiliary variables are usually observed following a “correlation order”. Specifically, X_k is the auxiliary variable most correlated with Y , that is

$$\rho_{Y,k} = \max_{i=1,\dots,k} \{\rho_{Y,i}\},$$

where $\rho_{Y,i}$ denotes the correlation coefficient between Y and X_i . Then, X_{k-1} is the auxiliary variable most correlated with X_k , that is $\rho_{k,k-1} = \max_{i < k} \{\rho_{k,i}\}$ and so on.

In this paper we propose an “optimal” observing order of the auxiliary variables. Expression (4) shows how the minimum $\text{AMSE}(\bar{y}_k)$ depends on the multiple correlation coefficients. This suggests the possibility of taking the auxiliary variables in such a way that the $\text{AMSE}^*(\bar{y}_k)$ is the least as possible, i.e. in such a way that the involved multiple correlation coefficients are the largest as possible. The idea is to set X_k as the auxiliary variable that makes $\rho_{Y,1,\dots,k-1}^2$ the largest as possible, X_{k-1} as the one that makes $\rho_{Y,1,\dots,k-2}^2$ the largest as possible, and so on. In order to do that, we recall that

$$\rho_{Y,1,\dots,i}^2 = \rho_{Y,1,\dots,i-1}^2 + \rho_{Y(i-1,\dots,i-1)}^2,$$

where $\rho_{Y(i-1,\dots,i-1)}$ is the part correlation coefficient between Y and X_i given $X_1, \dots, X_{i-1}$. Therefore we compute the part correlation coefficient between Y and X_i given all the other auxiliary variables and X_k is the auxiliary variable with the least part correlation coefficient. Then, we drop X_k and compute the part correlation coefficient between Y and X_i given all the remaining auxiliary variables. Again X_{k-1} is the auxiliary variable with the least part correlation coefficient. We proceed in the same way for all the other variables.

In the following table we compare this optimal order with other orders, in the three-phase case and using $f_{i+1} = f_i/2$, $i = 1, 2$, where f_i denotes the sampling fraction at the i -th phase. The data come from Mukerjee *et al.* (1987). Let us denote the AMSE corresponding to the optimal order as $\text{AMSE}_o^*(\bar{y}_k)$ and recall that $\rho_{Y(i,j)}^2 = \rho_{Y,12}^2 - \rho_{Y,j}^2$, $i, j = 1, 2$. The subscripts of the correlation coefficients in the table correspond to the order used by Mukerjee *et al.* (1987). From the last line of the table, we have that the chosen order is optimal only for the fifth data set. Actually, for the first four data sets X_1 and X_2 are the same variable observed in two different periods. In this case the observing order is necessarily chronological and the optimal order cannot be used. Here we compare chronological and optimal orders only for explaining purposes.

Data sets	I	II	III	IV	V
$\rho_{Y,1}$	0.97	0.89	0.92	0.988	0.941
$\rho_{Y,2}$	0.99	0.92	0.99	0.995	0.915
$\rho_{Y,12}$	0.99	0.922	0.99	0.995	0.946
$\frac{\text{AMSE}_o^*(\bar{y}_k)}{\text{AMSE}_n^*(\bar{y}_k)}$	1.041	1.04	1.14	1.015	1

References

- Ahmed M.S. (2003) General chain estimators under multi phase sampling, *J. Applied Statist. Science*, 17, to appear.
- Cicchitelli G., Herzel A. and Montinari G. (1997) *Il campionamento statistico*, Il Mulino, Bologna.
- Mukerjee R., Rao T.J. and Vijayan K. (1987) Regression type estimators using multiple auxiliary information, *Austral. J. Statist.*, 29 (3), 244–254.