A Class of Estimators for Predictive Estimation of Population Mean in Two-Phase Sampling

L.N. Sahoo
Department of Statistics, Utkal University
Bhubaneswar 751004, India
lnsahoostatuu@rediffmail.com

B.C. Das
Department of Statistics, Utkal University
Bhubaneswar 751004, India

S. C. Senapati
Department of Statistics, Ravenshaw College
Cuttack 753003, India

Summary

The present paper deals with the estimation of a finite population mean in the presence of two auxiliary variables under a two-phase sampling set up. It is assumed that the population mean of one auxiliary variable is known while that of other is unknown. Using the concept of predictive estimation developed by Basu (1971), a general class of estimators has been proposed. Analytical as well as simulation studies have been undertaken for evaluating performance of the suggested class.

Keywords & Phrases: Asymptotic variance, Auxiliary variable, Predictive approach, Two-phase sampling.

AMS 1991 Subject Classification: 62D05.

1. Introduction

Let y_i, x_i and z_i $(1 \le i \le N)$ be the values of the study variable y and two auxiliary variables x and z respectively for the ith unit of a finite population U with means $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\overline{Z} = \frac{1}{N}\sum_{i=1}^{N} z_i$. Suppose that $\overline{X}$ is unknown but $\overline{Z}$ is known and our aim is to estimate $\overline{Y}$ from a random sample obtained through a two-phase selection. Allowing simple random sampling without replacement (SRSWOR) in each phase, we consider a two-phase sampling: At phase one, a sample s' $(s' \subset U)$ of n' units is drawn from U to observe x and z. Then a sub-sample s $(s \subset s')$ of n units is drawn from s' at the second phase to observe y only. Let us define $\bar{y} = \sum_{i \in s} y_i$, $\bar{x} = \frac{1}{n}\sum_{i \in s} x_i$, $\bar{z} = \frac{1}{n}\sum_{i \in s} z_i$, $\bar{x}' = \frac{1}{n}\sum_{i \in s'} x_i$ and $\bar{z}' = \frac{1}{n}\sum_{i \in s'} z_i$.

The basic work on estimation in this area was initiated by Chand (1975) and subsequently studied by Kiregyera (1980, 1984) among others. But, Sahoo and Sahoo (1993), Singh *et al.* (1994) and Sahoo and Sahoo (1999) developed three different classes of estimators for $\overline{Y}$ defined by $\lambda_g = \phi(\bar{y}, \bar{x}, g(\bar{x}', \bar{z}'))$, $\lambda_p = p\left(\bar{y}, \frac{\bar{x}}{\bar{x}'}, \frac{\bar{z}'}{\overline{Z}}\right)$ and $\lambda_q = \phi(q(\bar{y}, \bar{x}), \bar{x}', \bar{z}')$ respectively, such that various functions considered for composing of the classes satisfy certain regularity conditions. However, an analysis of the properties of these classes clearly shows that they are not necessarily disjoint, but attain the same asymptotic minimum variance bound (MVB), given by

$$\begin{aligned}\min V(\lambda_g) &= \min V(\lambda_p) = \min V(\lambda_q) \\ &= \overline{Y}^2\left[(\eta - \eta')(1-\rho_{yx}^2) + \eta'(1-\rho_{yz}^2)\right]C_y^2,\end{aligned} \tag{1}$$

where $\eta' = \frac{1}{n'} - \frac{1}{N}$, $\eta = \frac{1}{n} - \frac{1}{N}$, C_y is the coefficient of variation of y; ρ_{yx} and ρ_{yz} are respectively the correlation coefficients between y, x and y, z. An estimator attaining this MVB (may be called as an MVB estimator) is a regression-type estimator

$$\lambda_{RG} = \bar{y} - b_{yx}(\bar{x} - \bar{x}') - b_{yz}(\bar{z}' - \overline{Z}),$$

considered earlier by Sahoo *et al.* (1994), where b_{yx} and b_{yz} are respectively the sample regression coefficients of y on x and y on z computed using data available on s. λ_g, λ_p and λ_q are also more efficient than $\bar{y}_\phi = \phi(\bar{y}, \bar{x}, \bar{x}')$, a two-phase sampling extension of Srivastava's (1980) class of estimators, in respect of MVB criterion. It may be mentioned here that the MVB of $\bar{y}_\phi$ and the resulting MVB estimator are given by

$$\min V(\bar{y}_\phi) = \overline{Y}^2\left[(\eta - \eta')(1-\rho_{yx}^2) + \eta'\right]C_y^2 \tag{2}$$

and

$$\bar{y}_{RG} = \bar{y} - b_{yx}(\bar{x} - \bar{x}'),$$

the classical two-phase sampling regression estimator of $\overline{Y}$.

In this paper, we utilize the same available information on both auxiliary variables as have been used to compose λ_g or λ_p or λ_q and construct a general class of estimators adopting the prediction criterion given by Basu (1971, p. 212, example 3).

2. Prediction Criterion in Two-Phase Sampling

Decomposing U into three mutually exclusive domains s, $r_2 = \bar{s} \cap s'$ and $r_1 = U - s'$ of n, $n' - n$ and $N - n'$ units respectively, where $\bar{s} = U - s$ denotes the collection of units in U which are not included in s, it is possible to express

$$\overline{Y} = \frac{1}{N}\left[\sum_{i \in s} y_i + \sum_{i \in r_2} y_i + \sum_{i \in r_1} y_i\right]. \tag{3}$$

Writing $(n' - n)\overline{Y}_2 = \sum_{i \in r_2} y_i$ and $(N - n')\overline{Y}_1 = \sum_{i \in r_1} y_i$, we have

$$\overline{Y} = f\bar{y} + (f' - f)\overline{Y}_2 + (1 - f')\overline{Y}_1, \tag{4}$$

where $f = \dfrac{n}{N}$ and $f' = \dfrac{n'}{N}$. The first component of the right hand side of (4) is exactly known. Hence, predicting unknown quantities $\bar{Y}_1$ and $\bar{Y}_2$ by T_1 and T_2 respectively on the basis of sample data, a predictor $\hat{\bar{Y}}$ of $\bar{Y}$ can be represented by the equation

$$\hat{\bar{Y}} = f\bar{y} + (f' - f)T_2 + (1 - f')T_1 . \tag{5}$$

It may be noted here that when $s = s' = U$, $\hat{\bar{Y}} = \bar{Y}$ which is the target of our prediction.

For different choices of the predictors, the predictive equation (5) generates a family of estimators. Accordingly, considering $T_1 = \bar{y}\bar{Z}_1 / \bar{z}'$ and $T_2 = \bar{y}\bar{X}_2 / \bar{x}$, where $\bar{X}_2 = \dfrac{n'\bar{x}' - n\bar{x}}{n' - n}$ and $\bar{Z}_1 = \dfrac{N\bar{Z} - n'\bar{z}'}{N - n'}$, Sahoo and Sahoo (2001) developed a ratio-type estimator defined by

$$l_{1R} = f'(\bar{y}_R - \bar{y}) + \frac{\bar{y}\bar{Z}}{\bar{z}'} ,$$

where $\bar{y}_R = \bar{y}\dfrac{\bar{x}'}{\bar{x}}$, the classical two-phase sampling ratio estimator. On the other hand, considering $T_1 = \bar{y} - b_{yz}(\bar{z}' - \bar{Z}_1)$ and $T_2 = \bar{y} - b_{yx}(\bar{x} - \bar{X}_2)$, Sahoo *et al.* (2003) obtained a regression-type estimator

$$l_{1RG} = \bar{y} - f'b_{yx}(\bar{x} - \bar{x}') - b_{yz}(z' - \bar{Z}).$$

The authors also derived sufficient conditions for the superiority of l_{1R} and l_{1RG} over the estimators

$$\lambda_{11} = \bar{y}\frac{\bar{x}'}{\bar{x}}\frac{\bar{Z}}{\bar{z}'} ,$$

proposed by Chand (1975), and

$$\lambda_{22} = \bar{y} - b_{yx}\left[\bar{x} - \left\{\bar{x}' - b'_{xz}(\bar{z}' - \bar{Z})\right\}\right],$$

proposed by Kiregyera (1984), where b'_{xz} is the sample regression coefficient of x on z based on s'.

3. The Proposed Class of Estimators

For given s and s', let $t_1 = (\bar{y}, \bar{z}', \bar{Z}_1)$ and $t_2 = (\bar{y}, \bar{x}, \bar{X}_2, \bar{z}, \bar{Z}_2)$, where $\bar{Z}_2 = \dfrac{n'\bar{z}' - n\bar{z}}{n' - n}$, assume values in R_3 and R_5, the 3- and 5-dimensional real spaces containing the points $\theta_1 = (\bar{Y}, \bar{Z}, \bar{Z})$ and $\theta_2 = (\bar{Y}, \bar{X}, \bar{X}, \bar{Z}, \bar{Z})$ respectively. Further, let $h_1(t_1)$ and $h_2(t_2)$ be some known functions of t_1 and t_2 respectively such that $h_1(\theta_1) = h_2(\theta_2) = \bar{Y}$. Further assume that,

(a) the functions h_1 and h_2 are continuous in R_3 and R_5 respectively, and

(b) first and second order partial derivatives of these functions *w.r.t.* their arguments exist and are also continuous in their respective range spaces.

Thus, based on information available on z in the domain r_1, $h_1(t_1)$ clearly defines a class of estimators for $\overline{Y}$. Similarly, availing information on both x and z in the domain r_2, $h_2(t_2)$ also defines a class of estimators for $\overline{Y}$. Hence, substituting $T_1 = h_1(t_1)$ and $T_2 = h_2(t_2)$ in (5), we now define the following class of predictive estimators for $\overline{Y}$:

$$l_h = f\bar{y} + (f' - f)h_2(t_2) + (1 - f')h_1(t_1).$$

Note that when $s' = U$ and z is not involved, our result corresponds to single phase sampling *i.e.*, selection of s from U by SRSWOR with one auxiliary variable x. Then we have

$$l_h = f\bar{y} + (1 - f)h_2(\bar{y}, \bar{x}, \overline{X}_2),$$

which defines a class of estimators of $\overline{Y}$. On the other hand, if no auxiliary information is taken into account, $h_1 = h_2 = \bar{y}$, which implies that $l_h = \bar{y}$, the simple expansion estimator of $\overline{Y}$.

Many estimators of $\overline{Y}$ using x and z may turn out as particular cases of l_h. For example, l_{1R}, λ_{11}, l_{1RG} and λ_{22} are particular cases of l_h for some specific selections of h_1 and h_2. However, we also consider the following two noteworthy cases:

(i) Let $h_1 = \bar{y}\dfrac{\overline{Z}_1}{\bar{z}'}$ and $h_2 = \bar{y}\dfrac{\overline{X}_2}{\bar{x}}\dfrac{\overline{Z}_2}{\bar{z}}$, then

$$l_h \to l_{2R} = f(1+\delta)\left(\bar{y} - \bar{y}\frac{\bar{z}'}{\bar{z}}\right) + f'\left[(1+\delta)\bar{y}_R\frac{\bar{z}'}{\bar{z}} - \delta\bar{y}_R - \bar{y}\right] + \bar{y}\frac{\overline{Z}}{\bar{z}'},$$

where $\delta = \dfrac{n}{n' - n}$.

(ii) Let $h_1 = \bar{y} - b_{yz}(\bar{z}' - \overline{Z}_1)$ and $h_2 = \bar{y} - b_{yx}(\bar{x} - \overline{X}_2) - b_{yz}(\bar{z} - \overline{Z}_2)$, then

$$l_h \to l_{2RG} = \bar{y} - f'b_{yx}(\bar{x} - \bar{x}') - f'b_{yz}(\bar{z} - \bar{z}') - b_{yz}(\bar{z}' - \overline{Z}).$$

4. Asymptotic Variance of the Class

Expanding $h_1(t_1)$ and $h_2(t_2)$ around the points θ_1 and θ_2 respectively in a first order Taylor's series and then neglecting the remainder terms, we get

$$h_1(t_1) = h_1(\theta_1) + h_{11}(\bar{y} - \overline{Y}) + h_{12}(\bar{z}' - \overline{Z}) + h_{13}(\overline{Z}_1 - \overline{Z}) \qquad (6)$$

and

$$h_2(t_2) = h_2(\theta_2) + h_{21}(\bar{y} - \overline{Y}) + h_{22}(\bar{x} - \overline{X}) + h_{23}(\overline{X}_2 - \overline{X}) + h_{24}(\bar{z} - \overline{Z}) + h_{25}(\overline{Z}_2 - \overline{Z}), \qquad (7)$$

where h_{ij} is the partial derivative of $h_i(t_i)$ *w.r.t.* its *j*th argument around the point θ_i, i = 1,2 ; $j = 1,2,3,4,5$. Further, noting that $h_{11} = h_{21} = 1, h_{12} = -h_{13}, h_{22} = -h_{23}$ and $h_{24} = -h_{25}$, we have after a considerable simplification

$$l_h - \overline{Y} = (\bar{y} - \overline{Y}) + f'h_{22}(\bar{x} - \bar{x}') + f'h_{24}(\bar{z} - \bar{z}') + h_{12}(\bar{z}' - \overline{Z}). \qquad (8)$$

This shows that the bias of l_h is of order n^{-1} and hence its contribution to the mean square error is of order n^{-2}.

From (8), after a few algebraic steps suppressed to save space, the asymptotic variance of l_h is obtained as

$$V(l_h)=(\eta-\eta')\left[S_y^2+f'^2h_{22}^2S_x^2+f'^2h_{24}^2S_z^2+2f'h_{22}S_{yx}+2f'h_{24}S_{yz}+2f'^2h_{22}h_{24}S_{xz}\right]$$
$$+\eta'\left(S_y^2+h_{12}^2S_z^2+2h_{12}S_{yz}\right), \qquad (9)$$

where $S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})^2$, $S_{yx}=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})(x_i-\overline{X})$ etc. Hence, $V(l_h)$ attains its minimum value when

$$h_{22}=-\beta_{yx.z}/f'=h_{22}^0 \text{ (say)}$$
$$h_{24}=-\beta_{yz.x}/f'=h_{24}^0 \text{ (say)}$$

and $h_{12}=-\beta_{yz}=h_{12}^0$ (say),

where β_{yz} is the regression coefficient of y on z; $\beta_{yx.z}$ and $\beta_{yz.x}$ are respectively the partial regression coefficients of y on x and y on z. The minimum value of $V(l_h)$ *i.e.*, the asymptotic MVB of l_h is given by

$$\min V(l_h)=\overline{Y}^2\left[(\eta-\eta')(1-\rho_{y.xz}^2)+\eta'(1-\rho_{yz}^2)\right]C_y^2, \qquad (10)$$

where $\rho_{y.xz}$ is the multiple correlation coefficient of y on x and z. The corresponding MVB estimator is a regression-type estimator

$$l_{RG}=\bar{y}-b_{yx.z}(\bar{x}-\bar{x}')-b_{yz.x}(\bar{z}-\bar{z}')-b_{yz}(\bar{z}'-\overline{Z}),$$

suggested by Tripathi and Ahmed (1995), where $b_{yx.z}$ and $b_{yz.x}$ are respectively sample partial regression coefficients of y on x and y on z based on s.

4. Precision of the Class

Our objective is to study precision of the predictive method of estimation developed in this work compared to the classical method. For this we need to compare efficiency of l_h with that of $\bar{y}_\phi, \lambda_g, \lambda_p$ and λ_q. But, the task of drawing a meaningful conclusion by comparing all estimators belonging to two different classes is not easy. Because, an estimator has its own limitations and can also be suitable for a particular situation in terms of the relationship between the variables under consideration. However, for simplicity, here we concentrate on the MVB estimators only *i.e.*, we accept MVB as a precision measure of a class. On the basis of this consideration, from (1), (2) and (10) we see that,

$$\min V(l_h)\le \min V(\lambda_g)\le \min V(\bar{y}_\phi)$$

i.e., $V(l_{RG})\le V(\lambda_{RG})\le V(\bar{y}_{RG})$.

Hence, we conclude that on the ground of MVB, the class of estimators represented by l_h is definitely superior to the classes of estimators represented by $\bar{y}_\phi$ and λ_g or λ_p or λ_q. To study efficiency of some specific predictive estimators belonging to l_h compared to some similar type of classical estimators belonging to $\bar{y}_\phi$ and λ_g or λ_p or λ_q numerically, we have carried out a simulation study that involved repeated draws of random samples from the following populations:

Population I [Murthy (1977), p. 228]: Consists of data on output (y), number of workers (x) and fixed capital (z) for 80 factories. We consider $n' = 20$ and $n = 10$.

Population II [Sarndal, Swensson and Wretman (1992), p. 662]: Consists of data on 1983 import (y), 1983 population (x) and 1980 population (z) for 124 countries. We consider $n' = 30$ and $n = 12$.

5000 independent first phase samples each of size n' were selected from a population by SRSWOR. Then, from each selected first phase sample, a second phase sample of size n was again selected by SRSWOR. For each combination (n', n), values of several estimators were computed. Then, considering 5000 such combinations simulated mean square errors of the estimators were calculated. Relative efficiencies of different estimators compared to the expansion estimator $\bar{y}$ are displayed in table 1.

From table 1 it is observed that l_{RG} attains the maximum precision and the suggested regression-type estimator l_{2RG} has better precision than other predictive and non-predictive estimators. The suggested ratio-type estimator l_{2R} also leads to substantial increase in precision over other ratio and ratio-type estimators.

Table 1: Relative Efficiency of Different Estimators *w.r.t.* $\bar{y}$ (in %)

Estimators	Population I	Population II
$\bar{y}_R$	115	143
λ_{11}	123	164
l_{1R}	124	168
l_{2R}	135	171
$\bar{y}_{RG}$	125	165
λ_{22}	139	167
λ_{RG}	140	175
l_{1RG}	142	178
l_{2RG}	144	191
l_{RG}	149	201

5. Conclusions

We have made a successful attempt in constructing a class of estimators for the finite population mean that is decidedly better than some other classes gathering information on two auxiliary variables in a two-phase sampling. Because, our analytical comparison shows that the suggested class is more efficient than others on the ground of MVB criterion. On the other hand, from the simulation study it is also seen that some estimators of the proposed class are more efficient than some similar type estimators belonging to other classes. These findings clearly indicate that there are situations which can favor for the application of the suggested estimation methodology. Of course, our analytical and simulation studies have limited scope and cannot able to reveal essential features of different classes in a straightforward manner. Further investigations may be made for arriving at better conclusions.

References

1. Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. *Foundations of Statistical Inferences,* V.P. Godambe and D.A. Sprott (eds), Holt, Rinehart and Winston, Toronto, Canada, 203-204.
2. Chand, L. (1975). *Some Ratio-type Estimators Based on Two or More Auxiliary Variables*. Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa.
3. Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, 27, 217-223.
4. Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika*, 31, 215-226.
5. Muthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.
6. Sahoo, J. and Sahoo, L.N. (1993). A class of estimators in two-phase sampling using two auxiliary variables. *Jour. Indian Stat Assoc*., 31, 107-114.
7. Sahoo, J. and Sahoo, L.N. (1999). An alternative class of estimators in double sampling procedures. *Calcutta Stat. Assoc. Bull*., 49, 79-83.
8. Sahoo, J. Sahoo, L.N. and Mohanty, S. (1994). An alternative approach to estimation in two-phase sampling using two auxiliary variables. *Biometrical Jour*., 36, 293-298.
9. Sahoo, L.N. and Sahoo, R.K. (2001). Predictive estimation of finite population mean in two-phase sampling using two auxiliary variables. *Jour. Indian Soc. Agric. Stat*., 54, 250-254.
10. Sahoo, L.N., Sahoo, R.K. and Senapati, S.C. (2003). Predictive estimation of finite population mean using regression-type estimators in double sampling procedures. *Jour. Stat. Res*., 37, 291-296.
11. Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

12. Singh, V.K., Singh, Hari P., Singh, Housila P. and Shukla, D. (1994). A general class of chain estimators for ratio and product of two means of a finite population. *Comm. Stat. – Thoe. Meth*., 23, 1341-1355.

13. Srivastava, S.K. (1980). A class of estimators using auxiliary information in sample surveys. *Canad. Jour. Stat*., 8, 253-254.

14. 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 Stat. Assoc. Bull*., 45, 203-218.