BULLETIN *of the*
Malaysian Mathematical
Sciences Society
http://math.usm.my/bulletin

Bull. Malays. Math. Sci. Soc. (2) **29***(1) (2006), 89–94*

An Alternative Class of Estimators in Double Sampling

[1]S.C. Senapati and [2]L.N. Sahoo
[1]Department of Statistics, Ravenshaw College, Cuttack – 753003, India
[2]Department of Statistics, Utkal University, Bhubaneswar – 751004, India
[1]scsenapati2002@rediffmail.com

Abstract. Using double sampling procedure, this paper presents a general class of estimators for the finite population mean when the population mean of the main auxiliary variable x is unknown but that of an additional auxiliary variable z is known. The proposed class of estimators is superior to some of the previously studied classes under minimum variance criterion.

2000 Mathematics Subject Classification: 62 D05

Key words and phrases: Auxiliary variable, bias, double sampling, minimum variance bound, variance.

1. Introduction

Let y and x denote study variable and auxiliary variable taking values y_i and x_i $(1 \leq i \leq N)$ respectively for the ith unit of a finite population Ω. When the two variables are strongly related but no information is available on the population mean $\bar{X}$ of x, we seek to estimate the population mean $\bar{Y}$ of y using a double sampling (two-phase sampling) mechanism. Allowing simple random sampling without replacement (SRSWOR) for sample selection, this scheme is described as follows:

a. A large preliminary sample s' $(s' \subset \Omega)$ of fixed size n' is drawn from Ω to observe only x in order to compose an estimate of $\bar{X}$.
b. Given s', a sub-sample s $(s \subset s')$ of fixed size n is drawn to observe y only.

A class of estimators in this context, covering standard ratio, product and regression estimators as its special cases, can be defined as $\bar{y}_d = d(\bar{y}, \bar{x}, \bar{x}')$, where $d(\cdot, \cdot, \cdot)$ is a known function of $\bar{y}$, $\bar{x}$ and $\bar{x}'$ satisfying certain regularity conditions, where $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$, $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$ and $\bar{x}' = \frac{1}{n'}\sum_{i\in s'} x_i$. $\bar{y}_d$ may be considered as an extension of Srivastava's [10] class of estimators into double sampling procedure. The minimum asymptotic variance (may be called as the minimum variance bound (MVB)) of the class, is given by

$$\min V(\bar{y}_d) = [(f - f')(1 - \rho_{yx}^2) + f']S_y^2, \tag{1.1}$$

Received: November 4, 2004; *Revised:* September 9, 2005.

where $f = \frac{1}{n} - \frac{1}{N}$, $f' = \frac{1}{n'} - \frac{1}{N}$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and ρ_{yx} is the correlation coefficient between y and x. An estimator attaining this bound, may be called as an MVB estimator, is the traditional double sampling regression estimator

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

where $\hat{\beta}_{yx}$ is the sample regression coefficient of y on x based on s.

In many practical situations even if $\bar{X}$ is unknown, information on a second auxiliary variable z, closely related to x, is readily available on all units of Ω such that z_i denotes its value on unit$i$ and $\bar{Z}$ as its known mean. For instance, if the elements of Ω are hospitals, and y_i, x_i and z_i are respectively the number of deaths, number of patients admitted and number of available beds relating to the ith hospital, then information on z_i's can be collected easily from the official records of the Health Department.

In the above practical scenarios, the data available on s' can be used to furnish a good estimate of $\bar{X}$ treating z as an auxiliary variable. As argued by Chand [1], substitution of such an estimate of $\bar{X}$ in place of $\bar{x}'$ in a standard double sampling estimator will lead to achieve an acceptable gain in precision under certain specified conditions. Motivated by this, Chand [1] and Kiregyera [2,3] developed many estimators taking ratio or regression estimator as the base.

Instead of considering a particular estimate of $\bar{X}$ for $\bar{x}'$, Sahoo and Sahoo [5] considered $h_2(\bar{x}', \bar{z}')$,a class of estimators of $\bar{X}$ and developed a class of estimators for $\bar{Y}$ defined by $l_h = h_1\left(\bar{y}, \bar{x}, h_2(\bar{x}', \bar{z}')\right)$, where $\bar{z}' = \frac{1}{n'}\sum_{i\in s'} z_i$. Using the concept developed by Singh *et al.* [9], we may also consider a class of estimators $l_p = p\left(\bar{y}, \frac{\bar{x}}{\bar{x}'}, \frac{\bar{z}'}{\bar{Z}}\right)$, where $p(\cdot,\cdot,\cdot)$ is a function of $\bar{y}$, $\frac{\bar{x}}{\bar{x}'}$ and $\frac{\bar{z}'}{\bar{Z}}$ satisfying some regularity conditions. Recently, Sahoo and Sahoo [6] composed an alternative class defined by $l_q = q_1\left(q_2(\bar{y}, \bar{x}), \bar{x}', \bar{z}'\right)$, where $q_2(\bar{y}, \bar{x})$ serves as a class of estimators of $\bar{y}' = \frac{1}{n'}\sum_{i\in s'} y_i$ based on s. An analysis of the properties of l_h, l_p and l_q shows that the classes are not necessarily disjoint but attain the same minimum variance bound given by

$$\min V(l_h) = \min V(l_p) = \min V(l_q) = \left[(f - f')(1 - \rho_{yx}^2) + f'(1 - \rho_{yz}^2)\right] S_y^2, \tag{1.2}$$

which is equal to the asymptotic variance of a regression-type estimator

$$l_{RG} = \bar{y} - \hat{\beta}_{yx}(\bar{x} - \bar{x}') - \hat{\beta}_{yz}(\bar{z}' - \bar{Z}),$$

considered earlier by Sahoo *et al.* [7], where ρ_{yz} is the correlation coefficient between y and z, and $\hat{\beta}_{yz}$ is the sample regression coefficient of y on z based on s.

In this paper, with the same available auxiliary information, we consider an alternative approach to the estimation of $\bar{Y}$ and also construct a class better than l_h, l_p and l_q for the purpose.

2. Alternative approach and proposed class of estimators

Inspired by Chand [1], when an attempt has been made to develop a class with replacement of $\bar{x}'$ by a class of estimators of $\bar{X}$ based on s', one should think that $\bar{x}$ provides a less efficient estimate of $\bar{X}$ than $\bar{x}'$. Hence, one should hope for a

better estimate of $\bar{X}$ than $\bar{x}$ using data on s, by taking advantage of the correlation between x and z. With this spirit, we now develop a general class of estimators for $\bar{Y}$ which has a greater scope than the system of estimators generated from $\bar{y}_d$ or l_h or l_p or l_q.

Motivated by Srivastava [10] using z as an auxiliary variable, suppose $t_x = u(\bar{x}, \bar{z}, \bar{z}')$ and $t'_x = v(\bar{x}', \bar{z}')$ are two different classes of estimators of $\bar{X}$ through s and s' respectively such that $u(\bar{X}, \bar{Z}, \bar{Z}) = v(\bar{X}, \bar{Z}) = \bar{X}$. Let $(\bar{y}, t_x, t'_x)$ assumes values in a closed convex subspace R_3 of 3-dimensional real space containing the point $(\bar{Y}, \bar{X}, \bar{X})$. Suppose that $g(\bar{y}, t_x, t'_x)$ is a known function of $\bar{y}$, t_x and t'_x such that $g(\bar{Y}, \bar{X}, \bar{X}) = \bar{Y}$ and three functions u, v and g satisfy the regularity conditions stated by Srivastava [10]. Then, a general class of estimators for $\bar{Y}$ may be defined as

$$t_g = g(\bar{y}, t_x, t'_x).$$

It may be noted that $t_g = \bar{y}_d$ when $t_x = \bar{x}$, $t'_x = \bar{x}'$ *i.e.*, if the information on z is not taken into account; and $t_g = l_h$ when $t_x = \bar{x}$. Thus, the system of estimators generated from $\bar{y}_d$ and l_h come out as special cases of t_g. But l_p, l_q and t_g are overlapping classes because many estimators belonging to l_p or l_q may come out as particular cases of t_g. t_g can also be reduced to a number of independent estimators involving information on both x and z when the functions u, v and g are properly selected. For instance,

$$t_g = t_{11} = \bar{y}\frac{\bar{x}'}{\bar{x}}\frac{\bar{z}}{\bar{z}'}\frac{\bar{Z}}{\bar{z}'} = \bar{y}\frac{t'_x}{t_x}$$

for $t_x = \bar{x}\frac{\bar{z}'}{\bar{z}}$, $t'_x = \bar{x}'\frac{\bar{Z}}{\bar{z}'}$,

$$t_g = t_{22} = \bar{y}\frac{\bar{x}}{\bar{x}'}\frac{\bar{z}}{\bar{z}'}\frac{\bar{Z}}{\bar{z}'} = \bar{y}\frac{t_x}{t'_x}$$

for $t_x = \bar{x}\frac{\bar{z}}{\bar{z}'}$, $t'_x = \bar{x}'\frac{\bar{z}'}{\bar{Z}}$,

$$t_g = t_{33} = \bar{y} - \hat{\beta}_{yx}\left[\{\bar{x} - \hat{\beta}_{xz}(\bar{z} - \bar{z}')\} - \{\bar{x}' - \hat{\beta}_{xz}(\bar{z}' - \bar{Z})\}\right] = \bar{y} - \hat{\beta}_{yx}(t_x - t'_x)$$

for $t_x = \bar{x} - \hat{\beta}_{xz}(\bar{z} - \bar{z}')$, $t'_x = \bar{x}' - \hat{\beta}_{xz}(\bar{z}' - \bar{Z})$, where $\hat{\beta}_{xz}$ is the sample regression coefficient of x on z computed using data on s. However, we observe that the estimators obtained from t_g are easy to apply in practice as they are simple to compute without any appreciable increase in cost as compared to the estimators developed in the line of Chand's approach.

We now analyze the properties of t_g in some depth by obtaining approximate expressions for its bias and variance. For this on expanding $t_x = u(\bar{x}, \bar{z}, \bar{z}')$ and $t'_x = v(\bar{x}', \bar{z}')$ around the points $(\bar{X}, \bar{Z}, \bar{Z})$ and $(\bar{X}, \bar{Z})$ respectively by the first order Taylor's series and neglecting the remainder terms we get

$$t_x = \bar{X} + u_{11}(\bar{x} - \bar{X}) + u_{12}(\bar{z} - \bar{Z}) + u_{13}(\bar{z}' - \bar{Z})$$

and $t'_x = \bar{X} + v_{11}(\bar{x}' - \bar{X}) + v_{12}(\bar{z}' - \bar{Z})$ where $u_{1i}, i = 1, 2, 3$ $(v_{1j}, j = 1, 2)$ is the first order partial derivative of $u(\bar{x}, \bar{z}, \bar{z}')$ $(v(\bar{x}', \bar{z}'))$ with respect to ith (jth) argument when evaluated at $(\bar{X}, \bar{Z}, \bar{Z})$ $((\bar{X}, \bar{Z}))$. Here we note that $v_{11} = 1$ because $v(\bar{X}, \bar{Z}) = \bar{X}$, and $u_{11} = 1$, $u_{12} = -u_{13}$ because $u(\bar{x}, \bar{z}, \bar{z}')$ and $u(\bar{x}, \bar{z}', \bar{z})$ assume

the same value *i.e.*, $\bar{X}$ at $(\bar{X}, \bar{Z}, \bar{Z})$. Hence, we have

$$t_x = \bar{X} + (\bar{x} - \bar{X}) + u_{12}\{(\bar{z} - \bar{Z}) - (\bar{z}' - \bar{Z})\} \tag{2.1}$$

and

$$t'_x = \bar{X} + (\bar{x}' - \bar{X}) + v_{12}(\bar{z}' - \bar{Z}). \tag{2.2}$$

Similarly, observing that $g_{11} = 1$, $g_{12} = -g_{13}$, an expansion of $g(\bar{y}, t_x, t'_x)$ about $(\bar{Y}, \bar{X}, \bar{X})$ in a first order Taylor's series gives

$$t_g \cong \bar{Y} + (\bar{y} - \bar{Y}) + g_{12}\{(t_x - \bar{X}) - (t'_x - \bar{X})\} \tag{2.3}$$

where g_{11}, g_{12} and g_{13} are the first order partial derivatives of $g(\bar{y}, t_x, t'_x)$ *w.r.t.* the corresponding arguments about $(\bar{Y}, \bar{X}, \bar{X})$. Hence, from (2.1), (2.2) and (2.3) we get

$$\begin{aligned} t_g - \bar{Y} \quad \cong \quad & (\bar{y} - \bar{Y}) + g_{12}\{(\bar{x} - \bar{X}) - (\bar{x}' - \bar{X})\} \\ & + g_{12}u_{12}\{(\bar{z} - \bar{Z}) - (\bar{z}' - \bar{Z})\} - g_{12}v_{12}(\bar{z}' - \bar{Z}) \end{aligned} \tag{2.4}$$

which shows that bias of t_g is of order n^{-1} contributing terms of order n^{-2} to its variance.

From (2.4), after a considerable simplification, we obtain the approximate expression for the variance of t_g to terms of order n^{-1} as

$$\begin{aligned} V(t_g) = (f - f')\Big(S_y^2 + g_{12}^2 S_x^2 + 2g_{12}S_{yx} + g_{12}^2 u_{12}^2 S_z^2 + 2g_{12}u_{12}S_{yz} \\ + 2g_{12}^2 u_{12}S_{xz}\Big) + f'\Big(S_y^2 + g_{12}^2 v_{12}^2 S_z^2 - 2g_{12}v_{12}S_{yz}\Big) \end{aligned} \tag{2.5}$$

where $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$, $S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})$ etc. The variance of t_g is thus is a function of u_{12}, v_{12} and g_{12}, and is minimized for $g_{12} = -\beta_{yx.z} = \hat{g}_{12}$ (say), $u_{12} = \beta_{yz.x}/\beta_{yx.z} = \hat{u}_{12}$ (say), and $v_{12} = -\beta_{yz}/\beta_{yx.z} = \hat{v}_{12}$ (say), where $\beta_{yx.z}$ and $\beta_{yz.x}$ are the population regression coefficients of y on x and y on z respectively, and β_{yz} is the population regression coefficient of y on z.

The optimum values $\hat{g}_{12}, \hat{u}_{12}$ and $\hat{v}_{12}$ can thus be determined uniquely in the sense that they do not depend on each other for their computation. Use of these optimum values in (2.5) yields the minimum asymptotic variance (the MVB of the class) as

$$\min V(t_g) = \left[(f - f')(1 - \rho_{y.xz}^2) + f'(1 - \rho_{yz}^2)\right] S_y^2 \tag{2.6}$$

where $\rho_{y.xz}$ is the multiple correlation coefficient of y on x and z. An estimator attaining this bound (*i.e.*, MVB estimator of t_g) is a regression-type estimator of the form

$$t_{RG} = \bar{y} - \hat{\beta}_{yx.z}(\bar{x} - \bar{x}') - \hat{\beta}_{yz.x}(\bar{z} - \bar{z}') - \hat{\beta}_{yz}(\bar{z}' - \bar{Z})$$

suggested by Tripathi and Ahmed [11], where $\hat{\beta}_{yx.z}$ and $\hat{\beta}_{yz.x}$ are respectively estimators of $\beta_{yx.z}$ and $\beta_{yz.x}$ computed using data on s. Thus, one can not improve upon t_{RG} by using both x and z simultaneously for the situation under consideration.

3. Precision of t_g

To study the effectiveness of the suggested estimation technique, it is desirable to compare the precision of t_g with that of $\bar{y}_d, l_h, l_p$ and l_q. But, in practice one can not draw any meaningful conclusion by comparing all estimators belonging to two different classes. Because, an estimator has its own limitation and is suitable only for a particular situation in terms of the relationship between the variables under consideration. However, for simplicity, if we accept MVB as an intrinsic measure of precision of a class then our attention will be concentrated on the MVB estimators only. Thus, from (1.1), (1.2) and (2.6) we have

$$\min V(t_g) \leq \min V(l_h) \leq \min V(\bar{y}_d), \Rightarrow V(t_{RG}) \leq V(l_{RG}) \leq V(\bar{y}_{RG})$$

showing that t_g is superior to others in respect of MVB criterion.

In order to investigate relative performance of t_{RG} over $\bar{y}$ as well as over $\bar{y}_{RG}$ and l_{RG}, we carry out a simulation study that involves repeated draws of random (double) samples from two natural populations described below:

Population I [Murthy [4], p.127] Consists of data on number of cultivators in 1961 (y), number of persons in 1961 (x) and cultivated area in 1951 (z) for 128 villages in a tehsil.

Population II [Sarndal *et al.* [8], p.662] Provides data on 1983 military expenditure (y), 1983 population (x) and 1982 gross national product (z) for 124 countries.

The following performance measures of an estimator $\hat{\bar{Y}}$ are taken into consideration:

(i) Relative bias (RB) $= 100\,|bias|\,/\bar{Y}$

(ii) Relative efficiency (RE) $= 100V(\bar{y})/MSE(\hat{\bar{Y}})$

Bias of $\bar{y}$ when calculated by considering all possible samples is identically equal to zero. But, its simulated values *i.e.*, the values computed from a long series of independent samples, are usually different from zero. Therefore, in our definition of RE, we used $MSE(\bar{y})$ instead of $V(\bar{y})$. However, these two measures are not the same under finite sample performance. 5000 independent first phase samples each of size 25 are selected from a population by SRSWOR. From every selected first phase sample, a second phase sample of size 10 is again selected by SRSWOR. For each combination (n', n) with $n' = 25$ and $n = 10$, values of $\bar{y}, \bar{y}_{RG}, l_{RG}$ and t_{RG} are computed. Then, considering 5000 such combinations, simulated RB and RE of different estimators are calculated and their values are displayed in Table 1. As argued above, here we also note that the simulated RB values for $\bar{y}$ are not exactly equal to zero. It is seen from the table that the performance of t_{RG} over others is quite appreciable. Thus, our simulation study, though of limited scope, shows that there are practical situations which can favor for the use of t_{RG} as well as the suggested estimation methodology.

Acknowledgment. The authors wish to thank the referee whose valuable suggestions improved the presentation of the paper.

Table 1. RB and RE of different estimators

Estimator	Population I		Population II	
	RB	RE	RB	RE
$\bar{y}$	0.035	100	0.087	100
$\bar{y}_{RG}$	5.281	102	7.129	135
l_{RG}	9.993	135	5.368	167
t_{RG}	4.689	181	5.102	203

References

[1] L. Chand, *Some Ratio-type Estimators Based on Two or More Auxiliary Variables*, Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa, 1975.

[2] B. Kiregyera, A chain ratio-type estimator in finite population double sampling using two auxiliary variables, *Metrika* **27** (1980), 217–223.

[3] B. Kiregyera, Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations, *Metrika* **31** (1984), 215–226.

[4] M. N. Murthy, *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, 1977.

[5] J. Sahoo and L. N. Sahoo, A class of estimators in two-phase sampling using two auxiliary variables, *Jour. Ind. Stat. Assoc.* **31** (1993), 107–114.

[6] J. Sahoo and L. N. Sahoo, An alternative class of estimators in double sampling procedures, *Calcutta Stat. Assoc. Bull.* **49** (1999), 79–83.

[7] J. Sahoo, L. N. Sahoo and S. Mohanty, A regression approach to estimation in two-phase sampling using two auxiliary variables, *Current Science* **65** (1993), 73–75.

[8] C. E. Sarndal, B. Swensson and J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, Berlin, 1992.

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

[10] S. K. Srivastava, A class of estimators using auxiliary information in sample surveys, *Cand. Jour. Stat.* **8** (1980), 253–254.

[11] T. P. Tripathi and M. S. Ahmed, A class of estimators for a finite population mean based on multivariate information and general two-phase sampling, *Calcutta Stat. Assoc. Bull.* **45** (1995), 203–218.