International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 3, 2004, pp 29–36

On the Estimation of a Population Mean in Two-Phase Sampling

S. C. Senapati, L. N. Sahoo
Department of Statistics
Utkal University
Bhubaneswar-751004, India

G. N. Singh
Department of Applied Mathematics
Indian School of Mines
Dhanbad-826004, India

[Received March 17, 2004; Revised July 27, 2004; Accepted November 3, 2004]

Abstract

In this paper, we consider the problem of estimation of a population mean under two-phase sampling when the population mean of the main auxiliary variable x is unknown but that of a second auxiliary variable z is known. Here, we propose a new class of estimators replacing mean by regression estimators. Numerical examples are also included for comparisons among the available estimators.

Keywords and Phrases: Asymptotic variance, auxiliary variable, minimum variance bound, two-phase sampling.

AMS Classification: 62D05.

1 Introduction

Let y_i, x_i and z_i, denote the ith observations on the survey variable y and two auxiliary variables x and z respectively. We consider a practical situation, where the population mean of x, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is unknown but the population mean of z, $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ is known accurately and we seek to estimate the population mean of y, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ using a two-phase sampling mechanism. For instance, if the elements of U 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 $\bar{Z}$ can easily be available from the official records of the Health Department.

Considering simple random sampling without replacement for sample selection and using standard notations as given in Cochran (1977), our two-phase sampling is described as follows:

a. A first phase sample $s'(s' \subset U)$ of fixed size n' is drawn to observe x and z.

b. Given s', a second phase sample $s(s \subset s')$ of fixed size n is drawn to observe y only.

Define $\bar{y} = \frac{1}{n}\sum\limits_{i\in s} y_i$, $\bar{x} = \frac{1}{n}\sum\limits_{i\in s} x_i$, $\bar{z} = \frac{1}{n}\sum\limits_{i\in s} z_i$, $\bar{x}' = \frac{1}{n'}\sum\limits_{i\in s'} x_i$ and $\bar{z}' = \frac{1}{n'}\sum\limits_{i\in s'} z_i$.

In the above scenarios, the basic work on estimation was initiated by Chand(1975) and subsequently studied by several authors producing a huge stock of estimators in the survey sampling literature during the last two decades. But, a common technique adopted by Chand(1975) and his followers to construct an estimator is the replacement of $\bar{x}'$ by an improved estimator of $\bar{X}$ in the standard two-phase ratio estimator $\bar{y}_R = \bar{y}\bar{x}'/\bar{x}$ or product estimator $\bar{y}_P = \bar{y}\bar{x}/\bar{x}'$ or regression estimator $\bar{y}_{RG} = \bar{y} - b_{yx}(\bar{x} - \bar{x}')$, where $b_{yx} = \sum\limits_{i\in s}(y_i - \bar{y})(x_i - \bar{x})/\sum\limits_{i\in s}(x_i - \bar{x})^2$. An improved estimator of $\bar{X}$ is usually defined in term of the auxiliary variable z, utilizing data on s'. For example, using a ratio estimator $\bar{x}'\bar{Z}/\bar{z}'$ for $\bar{x}'$ in $\bar{y}_R$, Chand(1975) defined a ratio-in-ratio estimator $\bar{y}_{CR} = \bar{y}\frac{\bar{x}'}{\bar{x}}\frac{\bar{Z}}{\bar{z}'}$. Similarly, considering a regression estimator $\bar{x}' - b'_{xz}(\bar{z}' - \bar{Z})$, where $b'_{xz} = \sum\limits_{i\in s'}(x_i - \bar{x}')(z_i - \bar{z}')/\sum\limits_{i\in s'}(z_i - \bar{z}')^2$, Kiregyera(1984) obtained a regression-in-regression estimator $\bar{y}_{CRG} = \bar{y} - b_{yx}\left[\bar{x} - \left\{\bar{x}' - b'_{xz}(\bar{z}' - \bar{Z})\right\}\right]$ from $\bar{y}_{RG}$. In this paper, on the availability of the same auxiliary information, we consider an alternative approach to estimate $\bar{Y}$ and also construct a class of estimators for the purpose.

2 The Proposed Class of Estimators

As stated earlier, an estimator is developed from $\bar{y}_R$ or $\bar{y}_P$ or $\bar{y}_{RG}$, with replacement of $\bar{x}'$ by a better estimate of $\bar{X}$ than $\bar{x}'$, one should think that $\bar{x}$ provides a less efficient estimate of $\bar{X}$ than $\bar{x}'$. Therefore, he can also hope for a better estimate of $\bar{X}$ than $\bar{x}$ by taking advantage of the correlation between x and z. This philosophy encourages us to develop a number of estimators from $\bar{y}_R$, $\bar{y}_P$ and $\bar{y}_{RG}$ replacing $\bar{x}$ and $\bar{x}'$ simultaneously by some improved estimators of $\bar{X}$ treating z as an auxiliary variable in many alternative ways. But, for simplicity, we consider difference estimators $t_x = \bar{x} - d(\bar{z} - \bar{z}')$ and $t'_x = \bar{x}' - d'(\bar{z}' - \bar{Z})$ in places of $\bar{x}$ and $\bar{x}'$ respectively, where the coefficients d and d' are constants. These difference estimators are not only simple to handle mathematically, but reduce to many well-known estimators when their coefficients are rightly chosen.

Thus, applying our technique of formulating an estimator, we may consider the following generalized estimators from $\bar{y}_R$, $\bar{y}_P$ and $\bar{y}_{RG}$:

$$t_R^* = \bar{y}\frac{\bar{x}' - d'(\bar{z}' - \bar{Z})}{\bar{x} - d(\bar{z} - \bar{z}')}$$

$$t_P^* = \bar{y}\frac{\bar{x} - d(\bar{z} - \bar{z}')}{\bar{x} - d'(\bar{z}' - \bar{Z})}$$

$$t_{RG}^* = \bar{y} - b_{yx}[\{\bar{x} - d(\bar{z} - \bar{z}')\} - \{\bar{x}' - d'(\bar{z}' - \bar{Z})\}].$$

An infinite number of estimators can be obtained from t_R^*, t_P^* and t_{RG}^* for various selections of d and d'. But, we focus attention on the creation of a general class for estimators having a greater scope than the system of estimators generated from t_R^*, t_P^* and t_{RG}^*.

Whatever be the samples s and s' chosen, let $w = \frac{t_x}{t'_x}$, and $h = (\bar{y}, w)$ assumes values in a closed convex subspace, Q, of two-dimensional real space containing the point $H = (\bar{Y}, 1)$.

Let $\lambda(h) = \lambda(\bar{y}, w)$ be a known function of $\bar{y}$ and w, such that $\lambda(\bar{y}, 1) = \bar{y}$ and satisfying the following conditions:

(i) $\lambda(h)$ is continuous in Q, and

(ii) the first and second order partial derivatives of $\lambda(h)$ exist and are continuous in Q.

These conditions taken together are called regularity conditions and any situation where these conditions hold is called a regular estimation case. Hence, in any regular estimation case a general class of estimators of $\bar{Y}$ may be defined by

$$t = \lambda(h). \tag{1}$$

We note that the generalised estimators t_R^*,t_P^* and t_{RG}^*, and the estimators like $t_1 = \bar{y} + \alpha(w-1)$, $t_2 = \bar{y}w^\alpha$, $t_3 = \frac{\bar{y}}{1+\alpha(w^\beta - 1)}$, $t_4 = \bar{y}(2 - w^\alpha)$, $t_5 = r\exp(\alpha(w-1))$ etc. where α and β are suitably chosen constants, come out as special cases of t.

3 Properties of the Proposed Estimator

Here, $\lambda(h)$ is a composite function of various statistics. It is therefore impossible to obtain exact results on its bias and variance. We use the Taylor linearization technique to yield an approximate expression for the variance of t under the assumptions (i) and (ii). So, on expanding $\lambda(h)$ around the point H upto second order, we obtain

$$\begin{aligned} t = \lambda(h) &= \lambda(H) + (\bar{y} - \bar{Y})\lambda_1(H) + (w-1)\lambda_2(H) \\ &+\frac{1}{2}[(\bar{y} - \bar{Y})^2\lambda_{11}(\tilde{h}) + 2(\bar{y} - \bar{Y})(w-1)\lambda_{12}(\tilde{h}) + (w-1)^2\lambda_{22}(\tilde{h})], \end{aligned} \tag{2}$$

where $\tilde{h} = (\tilde{y}, \tilde{w})$, $\tilde{y} = \bar{Y} + \eta(\bar{y} - \bar{Y})$, $\tilde{w} = 1 + \eta(w - 1)$, for $0 < \eta < 1$; λ_1, λ_2 denote the first order partial derivatives of λ, and λ_{11}, λ_{12}, λ_{22} its second order partial derivatives w.r.t. first and second arguments respectively.

Noting that $\lambda(H) = \bar{Y}$, $\lambda_1(H) = 1$ and simplifying (2) under the assumption $|\delta\, t'_x| < 1$, we obtain

$$E(t) = \bar{Y} + 0(n^{-1})$$

and $t - \bar{Y} \cong \bar{Y}(\delta\,\bar{y}) + \lambda_2(H)(\delta\, t_x - \delta\, t'_x)$, where $\delta\,\bar{y} = \frac{\bar{y}-\bar{Y}}{\bar{Y}}$, $\delta\, t_x = \frac{t_x - \bar{X}}{\bar{X}}$, $\delta\, t'_x = \frac{t'_x - \bar{X}}{\bar{X}}$. Hence, we have the following theorem:

Theorem 3.1. Under the regularity conditions, the first order variance of t is given by

$$\begin{aligned} V(t) &= \phi S_y^2 + (\phi - \phi')[\lambda_2^2(H)\{C_x^2 - 2dG\rho_{xz}C_xC_z + d^2G^2C_z^2\} \\ &\quad +2\lambda_2(H)\bar{Y}\rho_{yx}C_yC_x - 2\lambda_2(H)\bar{Y}dG\rho_{yz}C_yC_z] \\ &\quad +\phi'[\lambda_2^2(H)d'^2G^2C_z^2 + 2\lambda_2(H)\bar{Y}d'G\rho_{yz}C_yC_z] \end{aligned}$$

where $G = \frac{\bar{Z}}{\bar{X}}$, $\phi = \left(\frac{1}{n} - \frac{1}{N}\right)$, $\phi' = \left(\frac{1}{n'} - \frac{1}{N}\right)$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$, $C_x = S_x/\bar{X}$, ρ_{xz} = correlation coefficient between x and z etc.

The asymptotic variance of t is a function of $\lambda_2(H)$, d and d' and is minimized for

$$\lambda_2(H) = \frac{-\bar{Y}\frac{C_y}{C_x}(\rho_{yx} - \rho_{xz}\rho_{yz})}{1 - \rho_{xz}^2},\ d = \frac{C_x}{GC_z}\left[\frac{\rho_{xz}\rho_{yx} - \rho_{yz}}{\rho_{yx} - \rho_{xz}\rho_{yz}}\right],\ d' = \frac{C_x}{GC_z}\left[\frac{\rho_{yz}(1 - \rho_{xz}^2)}{\rho_{yx} - \rho_{xz}\rho_{yz}}\right].$$

Thus, these results lead to the following theorem:

Theorem 3.2. The minimum asymptotic variance of t is given by,

$$\min V(t) = [(\phi - \phi')(1 - \rho_{y.xz}^2) + \phi'(1 - \rho_{yz}^2)]S_y^2, \tag{3}$$

where $\rho_{y.xz} = \sqrt{\frac{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{xz}\rho_{yz}}{1 - \rho_{xz}^2}}$ is the multiple correlation coefficient of y on x and z.

The expression (3) may be called as the asymptotic minimum variance bound (MVB) of the class of estimators defined by t. An estimator (not necessarily unique) attaining this bound may be termed as a MVB estimator. For example, a regression-type estimator of the form

$$t_{RG} = \bar{y} - \beta_{yx.z}(\bar{x} - \bar{x}') - \beta_{yz.x}(\bar{z} - \bar{z}') - \beta_{yz}(\bar{z}' - \bar{Z}),$$

where $\beta_{yx.z}, \beta_{yz.x}$ are the partial regression coefficients of y on x and z respectively and β_{yz} is the regression coefficient of y on z, suggested by Tripathi and Ahmed(1995) is a MVB estimator of this class. So, the proposed general class of estimators $t$can not provide any more imporved estimators over Tripathi and Ahmed (1995), but it is an another class.

The optimum values of $\lambda_2(H)$, d and d' are usually unknown. Thus in practice, one has to use their guessed values either available through one's past experience or through a pilot sample survey. Over the history of survey sampling, most of the considerable experience gathered can also be useful for this purpose under a variety of survey situations [*cf.*, Reddy (1978)]. However, when these parametric functions are completely unknown or difficult to obtain accurately, we can replace them by their consistent estimates. Hence, the class of estimators of $\bar{Y}$ may be expressed as

$$\hat{t} = \lambda\left(\bar{y}, \frac{\hat{t}_x}{\hat{t}'_x}, \hat{\lambda}_2(H)\right),$$

where $\hat{t}_x = \bar{x} - \hat{d}(\bar{z} - \bar{z}')$, $\hat{t}'_x = \bar{x}' - \hat{d}'(\bar{z}' - \bar{Z})$, $\hat{d}$ and $\hat{\lambda}_2(H)$ are respectively estimates of d and $\lambda_2(H)$ based on s, and $\hat{d}'$ is an estimate of d' based on s'. But, the first order of approximation $V(t)$ and $V\hat{t}$ as well as their minimal values are equal.

4 Comparison of Estimators

To study the effectiveness of the proposed estimation technique, let us now examine the precision of t in comparison with a few specific classes discussed below.

If an estimation of $\bar{Y}$ is carried out with the involvement of x only, then a class of estimators, covering $\bar{y}_R$, $\bar{y}_P$ and $\bar{y}_{RG}$ as its special cases can be defined as $\bar{y}_f = f(\bar{y}, \bar{x}, \bar{x}')$, where $f(.,.,.)$ is a known function of $\bar{y}$, $\bar{x}$ and $\bar{x}'$ satisfying certain regularity conditions. $\bar{y}_f$ may be considered as an extension of Srivastava's (1980) class of estimators into two-phase sampling procedure. The assymptotic *MVB* of the class is

$$\min V(\bar{y}_f) = [(\phi - \phi')(1 - \rho_{yx}^2) + \phi']S_y^2 \tag{4}$$

and the corresponding MVB estimator is $\bar{y}_{RG}$.

Replacing $\bar{x}'$ in $\bar{y}_f$ by $h(\bar{x}', \bar{z}')$, a class of estimators for $\bar{X}$ based on s', Sahoo and Sahoo (1993) developed a class of estimators for $\bar{Y}$ defined by $l_h = f(\bar{y}, \bar{x}, h(\bar{x}', \bar{z}'))$. Using concept developed by Singh *et al.*(1994), we may also consider a class $l_p = p\left(\bar{y}, \frac{\bar{x}}{\bar{x}'}, \frac{\bar{z}'}{\bar{Z}}\right)$. Recently, Sahoo and Sahoo (1999) composed an alternative class defined by $l_q = q_1(q_2(\bar{y}, \bar{x}), \bar{x}', \bar{z}')$ where $q_2(\bar{y}, \bar{x})$ serves as a class of estimators of $\bar{y}'$ based on s. An analysis of the properties of l_h, l_p and l_q shows that these classes are not necessarily disjoint but attain the same *MVB* given by

$$\min V(l_h) = \min V(l_p) = \min V(l_q) = [(\phi - \phi')(1 - \rho_{yx}^2) + \phi'(1 - \rho_{yz}^2)]S_y^2 \tag{5}$$

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

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

where $b_{yz} = \sum\limits_{i \in s}(y_i - \bar{y})(z_i - \bar{z}) / \sum\limits_{i \in s}(z_i - \bar{z})^2$, considered earlier by Sahoo *et al.* (1993).

Our aim is to compare the precision of t with that of $\bar{y}_f$,l_h, l_p and l_q. But, it is not possible to 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. However, for simplicity, if we accept MVB as an intrinsic measure of the precision of a class, our attention will be concentrated on the MVB estimators only. Thus from (3), (4) and (5), we have

$$\min V(t) \leq \min(l_h) \leq \min(\bar{y}_f)$$

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

i.e., t is superior to $\bar{y}_f$, l_h, l_p and l_q in respect of MVB criterion.

5 Numerical Illustrations and Conclusions

In order to study the gain in precision of the proposed estimation technique over others numerically, we compute relative precision of different estimators with respect to mean per unit estimator $\bar{y}$ using data of 4 natural populations described in table 1. Here, the relative precision of an estimator e is defined by $RP = \frac{V(\bar{y})}{V(e)} \times 100$, where $V(e)$ is the first order variance of e.

For comparison purpose we have taken the three MVB estimators viz., $\bar{y}_{RG}$, l_{RG} and t_{RG}; Chand's(1975) regression-in-regression estimator $\bar{y}_{CRG}$ and a regression-type estimator $\bar{y}^*_{RG}$ obtained from $\bar{t}^*_{RG}$ on considering

$$d = b_{xz} = \sum_{i \in s}(x_i - \bar{x})(z_i - \bar{z}) / \sum_{i \in s}(z_i - \bar{z})^2 \quad \text{and} \quad d' = b'_{xz}$$

.

Table 1: Description of Populations

Pop. No.	Source	N	y	x	z
1.	Srivastava *et al.*(1990)	1000 jute	yield	height	diameter
2.	Tripathi and Ahmed (1995)	278 villages	no. of agricultural labourers	population size	no. of cultivators
3.	Tripathi (1980)	225 persons	persons employed	persons in service	educated persons
4.	Sukhatme and Chand (1977)	120 trees	bushels of apples harvested in 1964	apple trees of bearing age in 1964	bushels of apples harvested in 1959

Table 2: Relative Precision of Different Estimators

Pop. No.	n'	n	Estimators					
			$\bar{y}$	$\bar{y}_{RG}$	$\bar{y}_{CRG}$	l_{RG}	$\bar{y}^*_{RG}$	t_{RG}
1	100	75	100	117	135	162	166	177
2	60	40	100	130	164	165	167	170
3	90	68	100	110	107	111	115	118
4	30	15	100	197	394	491	459	549

Based on the expressions (3), (4) and (5), we compute $V(t_{RG})$, $V(\bar{y}_{RG})$ and $V(l_{RG})$ respectively . To compute $V(\bar{y}_{CRG})$ we refer to its formula given in Kiregyera(1984). However, to compute $V(\bar{y}^*_{RG})$, we consider its formula given by

$$V(\bar{y}^*_{RG}) = [(\phi - \phi')\{1 - \rho^2_{yx}(1 + \rho^2_{xz}) + 2\rho_{yx}\rho_{yz}\rho_{xz}\} + \phi'\{1 + \rho^2_{yx}\rho^2_{xz} - 2\rho_{yx}\rho_{yz}\rho_{xz}\}]S^2_y.$$

Relative precision of the comparable estimators are displayed in table 2. It is seen that the performance t_{RG} over others is quite appreciable. On the other hand $\bar{y}^*_{RG}$ is also better than $\bar{y}_{RG}$ and $\bar{y}_{CRG}$ for all cases, and l_{RG} for three cases.

Analytical as well as empirical findings of this paper indicate that the proposed class is capable of producing estimators of $\bar{Y}$ which are no way inferior to those of $\bar{y}_f$, l_h, l_p and l_q. The estimators so obtained are also 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. Hence, proposed estimation procedure is one of optimum classes.

Acknowledgement

The authors are thankful to the referee for some valuable comments on the earlier draft of the paper.

References

[1] Chand, L.(1975). *Some ratio-type estimators based on two or more auxiliary variables.* Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

[2] Cochran, W.G.(1977). *Sampling Techniques,* 3rd ed. New York: Wiley.

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

[4] Reddy, V.N.(1978). A study on the use of prior knowledge on certain population parameters in estimation, *Sankhya*, **C40**, 29-34.

[5] 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.

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

[7] Sahoo, J., Sahoo, L.N. and Mohanty, S. (1993). A regression approach to estimation in two-phase sampling using two auxiliary variables. *Current Science,* **65**(1), 73-75.

[8] 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 finite populations. *Comm. Stat -Theo Meth.,* **23**, 1341-1355.

[9] Srivastava, Rani S., Khare, B.B. and Srivastava, S.R.(1990). A generalized chain ratio estimator for mean of a finite poulation. *Jour. Indian Soc. Agric. Stat.,* **42**, 108-117.

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

[11] Sukhatme, B.V. and Chand, L.(1977). Multivariate ratio-type estimators. *Proceedings of Social Statistics Section, Amer. Stat. Assoc.,* 927-931.

[12] Tripathi, T.P.(1980). A general class of estimators for population ratio. *Sankhya,* **C42**, 63-75.

[13] 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.