Aligarh Journal of Statistics
Vol. 26 (2006), 49-56

OPTIMAL ESTIMATION OF MEANS OF SEVERAL VARIABLES USING MULTIVARIATE AUXILIARY INFORMATION UNDER STRATIFIED SAMPLING

Mohammad Vaseem Ismail, Abdul Razzaq and T.P. Tripathi

ABSTRACT

In this paper, we define the estimator of the finite population mean vector of several principal variables under stratified sampling design, in the situations where mean vector of the auxiliary variables is known. An optimum estimator by using the criterion of preference given by Tripathi and Chaubey (2000) has been obtained.

1. INTRODUCTION

Most of the sample surveys are devoted to collect information on several variables simultaneously. The usual problem in multipurpose surveys is to estimate the population means or totals of several variables simultaneously by using a number of auxiliary variables the information on which may be available through the past census data or it may be collected through diverting a part of the survey budget. In a land survey, for instance the estimates of the total number of agricultural labourers, literates and schedule casts for a certain administrative block may be easily available through past census data and the information on the variables such as the number of households, number of male workers and number of cultivators of the villages may not be readily available but may be known through diverting a part of the survey budget to it.

The problem of estimation of the population mean (or total) of a single survey variable in the situation where population means (or totals) of several auxiliary variables are known has been considered by several authors including Olkin (1958), Raj (1965), Srivastava (1965, 1966), Rao and Mudholkar (1967), Singh (1967), Srivastava (1971), Tripathi (1970, 1976, 1987) and Mukherjee *et al.* (1987).

The use of information on several auxiliary variables for estimating the population means of more than one principal variable has also been considered by several authors. Tripathi and Khattree (1989) discussed the estimation of means of principal variables $y_1, \cdots, y_p$ under simple random sampling, in the situations where means of auxiliary variables $x_1, \cdots, x_q$ are known. Further, Tripathi (1989) extended the result to the case of two occasions. Tripathi and Chaubey (1993) have considered the problem of obtaining the optimum

probabilities of selection based on $x_1,\cdots,x_q$ in *pps* sampling for estimating the means of $y_1,\cdots,y_p$. Recently, Tripathi and Chaubey (2000) discussed the problem of estimating the mean of a vector variable $\underline{y}=(y_1,\cdots,y_p)'$ based on a general sampling design and on the knowledge of means of several variables $\underline{x}=(x_1,\cdots,x_q)'$ for a finite population. They also gave the criterion of preference of one estimation procedure over the others in a quite general form stronger than customary criteria.

In this paper, we discuss the estimation of finite population mean vector $(\overline{Y}_1,\cdots,\overline{Y}_p)=\underline{\overline{Y}}'$ of the principal variables $(Y_1,\cdots,Y_p)=\underline{Y}'$, under stratified sampling design, in the situations where mean vector $(\overline{X}_1,\cdots,\overline{X}_q)=\underline{\overline{X}}'$ of the auxiliary variables $(X_1,\cdots,X_q)=\underline{X}'$ is known.

2. NOTATION

Consider a finite population $U=\{1,2,\cdots,N\}$. The population is divided into L strata.

Let

$y_{ijh}=$ the value of $i-$th unit for $j-$th estimation character in the $h-$th stratum.

and

$x_{ikh}=$ the value of $i-$th unit for $k-$th auxiliary character in the $h-$th stratum $(j=1,2,\cdots,p;\ k=1,2,\cdots,q;\ h=1,2,\cdots,L)$

Let $\underline{y}_{ik}$ be the observed value of the vector of estimation variables $y_1,\cdots,y_p$ on the $i-$th unit in the $h-$th stratum and similarly

let $\underline{x}_{ih}$ be the observed value of the vector of auxiliary variables $x_1,\cdots,x_q$ on the $i-$th unit in the $h-$th stratum.

The population mean vectors of the estimation variables and of the auxiliary variables in the $h-$th stratum are given respectively as

$$\underline{\overline{Y}}_h=\frac{1}{N_h}\sum_{i=1}^{N_k}\underline{y}_{ik}$$

and $\underline{\overline{X}}_h=\dfrac{1}{N_h}\sum_{i=1}^{N_h}\underline{x}_{ih}$.

Denote by $\underline{\overline{Y}} = \sum_{h=1}^{L} W_h \underline{\overline{Y}}_h$

and $\underline{\overline{X}} = \sum_{h=1}^{L} W_h \underline{\overline{X}}_h$.

Consider a random sample of size n from a finite population U. On each of the sample unit, the measurement for p estimation variables $y_1, \cdots, y_p$ and the q auxiliary variables $x_1, \cdots, x_q$ are obtained as

$$\begin{pmatrix} y_{11} & \cdots & y_{p1} \\ y_{12} & \cdots & y_{p2} \\ \vdots & \vdots & \vdots \\ y_{1n} & \cdots & y_{pn} \end{pmatrix} \text{ and } \begin{pmatrix} x_{11} & \cdots & x_{q1} \\ x_{12} & \cdots & x_{q2} \\ \vdots & \vdots & \vdots \\ x_{1n} & \cdots & x_{qn} \end{pmatrix}$$

Let the population be stratified into L strata and denote by $\underline{y}_{ih}$ the vector of sample values of estimation variables on the $i-$th unit in the $h-$th stratum, $i = 1, 2, \cdots, n$; $h = 1, 2, \cdots, L$ and denote by $\underline{x}_{ih}$ the vector of sample values of auxiliary variables on the $i-$th unit in the $h-$th stratum, $i = 1, 2, \cdots, n_h$; $h = 1, 2, \cdots, L$.

The customary unbiased estimators of $\underline{\overline{Y}}_h$ and $\underline{\overline{X}}_h$ are given by

$$\underline{\hat{\overline{Y}}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \underline{y}_{ih} \quad \text{and} \quad \underline{\hat{\overline{X}}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \underline{x}_{ih}$$

Denote by $\underline{\hat{\overline{Y}}} = \sum_{h=1}^{L} W_h \underline{\hat{\overline{Y}}}_h$ and $\underline{\hat{\overline{X}}} = \sum_{h=1}^{L} W_h \underline{\hat{\overline{X}}}_h$.

3. THE PROPOSED CLASS OF ESTIMATORS

For $h-$th stratum, let us define

$$\underline{\tilde{\overline{Y}}}_h = \underline{\hat{\overline{Y}}}_h + T_h (\underline{\overline{X}}_h - \underline{\hat{\overline{X}}}_h), \; h = 1, 2, \cdots, L$$

where $\underline{\hat{\overline{Y}}}_h = (\hat{\overline{Y}}_{h1}, \cdots, \hat{\overline{Y}}_{hp})'$ and $\underline{\hat{\overline{X}}}_h = (\hat{\overline{X}}_{h1}, \cdots, \hat{\overline{X}}_{hk})'$ are the customary unbiased estimators of $\underline{\overline{Y}}_h$ and $\underline{\overline{X}}_h$ respectively, and $T_h = (t_{jk}^{h})$ is a $p \times q$ matrix of statistics.

The class of estimators for the vector of population mean $\underline{\overline{Y}}$ may be defined as

$$\underline{\tilde{\bar{Y}}}_{(st)} = \sum_{h=1}^{L} W_h \,[\underline{\hat{\bar{Y}}}_h + T_h \,(\underline{\bar{X}}_h - \underline{\hat{\bar{X}}}_h)] \tag{3.1}$$

where $T_h = \begin{pmatrix} t_{11}^h & \cdots & t_{1q}^h \\ \vdots & \vdots & \vdots \\ t_{p1}^h & \cdots & t_{pq}^h \end{pmatrix}_{p\times q}$ and t_{jk}^h are suitably chosen statistics such that their means exits. It may be noted that parallel to random sampling case several interesting estimators may be generated from $\underline{\tilde{\bar{Y}}}_{(st)}$ for specific choices of T_h.

We will consider only the class of estimators (3.1) when T_h is a pre-specified non-random matrix.

4. CRITERION OF OPTIMIZATION

For fixed T_h, $\underline{\tilde{\bar{Y}}}_{(st)}$ is unbiased for $\underline{\bar{Y}}$ and its *MSE* matrix $M(\underline{\tilde{\bar{Y}}}_{(st)})$ is obtained below. We have

$$(\underline{\tilde{\bar{Y}}}_{(st)} - \underline{\bar{Y}}) = \sum_{h=1}^{L} W_h [(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h) - T_h \,(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)] .$$

On squaring both sides,

$$(\underline{\tilde{\bar{Y}}}_{(st)} - \underline{\bar{Y}})^2 = \sum_{h=1}^{L} W_h^2 [\,(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)^2 + T_h^2 \,(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)^2$$

$$- 2T_h \,(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)] .$$

Taking expectation on both sides, we have

$$E\,(\underline{\tilde{\bar{Y}}}_{(st)} - \underline{\bar{Y}})^2 = \sum_{h=1}^{L} W_h^2 \; E\,(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)'$$

$$+ T_k^2 \; E\,(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)' - 2T_k \; E\,(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)'$$

or $$M\,(\underline{\tilde{\bar{Y}}}_{(st)}) = \sum_{h=1}^{L} W_h^2 \,(V_{yy}^h + T_h \, V_{xx}^h \, T_h' - T_h \, C_{yx}'^h - C_{yx}^h \, T_h') \tag{4.1}$$

where

$$V_{yy}^h = E\,(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)', \qquad V_{xx}^h = E\,(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)'$$

and $C_{yx}^{h} = E(\underline{\hat{\bar{Y}}}_h - \underline{\bar{Y}}_h)(\underline{\hat{\bar{X}}}_h - \underline{\bar{X}}_h)'$.

Now, we consider the following criterion of preference given by Tripathi and Chaubey (2000):

Let $\underline{M}(\underline{Z}_y) = E[(\underline{Z}_y - \underline{\bar{Y}})(\underline{Z}_y - \underline{\bar{Y}})']$ denote the mean square error (MSE) matrix of an estimator $\underline{Z}_y$ of $\underline{\bar{Y}}$.

C.P. (1): An estimator $\underline{Z}_y$ is said to be better than another estimators $\underline{Z'}_y$ of $\underline{\bar{Y}}$ if and only if $M(\underline{Z}'_y) - M(\underline{Z}_y)$ is non negative definite whatever be the value of $\underline{y}_1, \cdots, \underline{y}_N$.

C.P. (2): Let $C = \{\underline{Z}_y\}$ be a class of estimators of $\underline{\bar{Y}}$. An estimator $\underline{Z}_{oy} \in C$ is said to be optimum for $\underline{\bar{Y}}$ in the class C if and only if $M(\underline{Z}_y) - (\underline{Z}_{oy})$ is non-negative definite $(n.n.d)$ for all $\underline{Z}_y (\neq \underline{Z}_{oy})$ in the class C and for all possible values of $\underline{y}_1, \cdots, \underline{y}_N$.

We will find the optimum value of T_h in (3.1) under the criterion C.P. (2).

5. OPTIMUM CHOICE OF T_h

For obtaining the optimum choice of T_h, we differentiate (4.1) w.r.t. T_h and equate to zero.

$$\frac{\partial M(\underline{\tilde{\bar{Y}}}_{(st)})}{\partial T_h} = \sum_{h=1}^{L} W_h^2 [-2C'^{h}_{yx} + 2T_h^{'} V_{xx}^{h}] = 0$$

$$\Rightarrow \; -2C'^{h}_{yx} + 2T'_h V_{xx}^{h} = 0$$

$$2T_h^{'} V_{xx}^{h} = 2C'^{h}_{yx}$$

$$T_h^{opt} = C_{yx}^{h}(V_{xx}^{h})^{-1} \tag{5.1}$$

Substituting the optimum value of T_h in (4.1), we have

$$M(\underline{\tilde{\bar{Y}}}_{(st)}) = \sum_{h=1}^{L} W_h^2 [V_{yy}^{h} + C_{yx}^{h}(V_{xx}^{h})^{-1} V_{xx}^{h} C^{'h}_{yx}(V_{xx}^{h})^{-1} - C_{yx}^{h}(V_{xx}^{h})^{-1} C^{'h}_{yx}]$$

$$- C_{yx}^{h}(V_{xx}^{h})^{-1} C^{'h}_{yx}$$

$$= \sum_{h=1}^{L} W_h^2 [V_{yy}^h + C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx} - 2C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx}]$$

$$= \sum_{h=1}^{L} W_h^2 [V_{yy}^h - C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx}].$$

Hence, optimum MSE Matrix of $\underline{\bar{Y}}$ is given by

$$M(\underline{\tilde{\bar{Y}}}^{opt}_{(st)}) = \sum_{h=1}^{L} W_h^2 [V_{yy}^h - C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx}] \tag{5.2}$$

Now, consider the difference

$$M(\underline{\tilde{\bar{Y}}}_{(st)}) - M(\underline{\tilde{\bar{Y}}}^{opt}_{(st)}) = \sum_{h=1}^{L} W_h^2 [V_y^h + T_h V_{xx}^h T_h' - T_h C^{'h}_{yx} - C_{yx}^h T_h']$$

$$- \sum_{h=1}^{L} W_h^2 [V_{yy}^h - C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx}]$$

$$= \sum_{h=1}^{L} W_h^2 [T_h V_{xx}^h T_h' - T_h C^{'h}_{yx} - C_{yx}^h T_h' + C_{yx}^h (V_{xx}^h)^{-1} C^{'h}_{yx}]$$

$$= \sum_{h=1}^{L} W_h^2 [T_h V_{xx}^h T_h^{'} - T_h^{opt} V_{xx}^h T_h^{opt'} + T_h^{opt} V_{xx}^h T_h^{opt'}]$$

$$- T_h C^{'h}_{yx} - C_{yx}^h T_h^{'} + C_{yx}^h V_{xx}^{-1} C^{'h}_{yx}$$

$$= \sum_{h=1}^{L} W_h^2 [T_h V_{xx}^h T_h' - T_h^{opt} V_{xx}^h T_h^{opt'} + T_h^{opt} C^{'h}_{yx} - T_h C'^{h}_{yx}]$$

$$- C_{yx}^h T_h^{'} + C_{yx}^h T_h^{opt'}$$

$$= \sum_{h=1}^{L} W_h^2 [T_h V_{xx}^h T_h^{'} - T_h^{opt} V_{xx}^h T_h^{opt'} - (T_h - T_h^{opt}) C^{'h}_{yx}]$$

$$- C_{yx}^h (T_h - T_h^{opt})'$$

$$= \sum_{h=1}^{L} W_h^2 [(T_h - T_h^{opt}) V_{xx}^h (T_h - T_h^{opt})' + (T_h - T_h^{opt})(V_{xx}^h T_h^{opt} - C^{'h}_{yx})]$$

$$+ (T_h^{opt} V_{xx}^h - C^{'h}_{yx})(T_h - T_h^{opt})' \tag{5.3}$$

Since the first term on the *RHS* of (5.3) is non-negative definite $(n.n.d)$, the difference on the *LHS* for $T_h \neq T_h^{opt}$ can be made *n.n.d.* if and only if

$$T_h = C_{yx}^h (V_{xx}^h)^{-1} \tag{5.4}$$

Hence the optimum choice of T_h *w.r.t.* the criterion C.P.(2) is as given in (5.4).

REFERENCES

Mukherjee, R., Rao, T.J. and Vijayan, K. (1987): Regression type estimators using multi-auxiliary information. *Aust. J. Statist*., **29**, 244-254.

Olkin, I. (1958): Multivariate ratio estimation for finite populations. *Biometrika,* **45**, 154-165.

Raj, D. (1965): On a method of using multi-auxiliary information in sample surveys. *J. Amer. Statist. Assoc.,* **60**, 270-277.

Rao, P.S.R.S. and Mudholkar, G.S. (1967): Generalized multivariate estimators for the mean of a finite population. *J. Amer. Statist. Assoc.*, **62**, 1009-1012.

Singh, M.P. (1967): Multivariate product method of estimation for the finite population. *J. Indian Soc. Agricultural Statist.,***19**, 1-10.

Srivastava, S.K. (1965): An estimate of the mean of a finite population using several auxiliary character. *J. Indian Statist. Assoc.,* **3**, 189-194.

Srivastava, S.K. (1966): On ratio and linear regression method of estimation with several auxiliary variables. *J. Indian Statist. Assoc.,* **4**, 66-72.

Srivastava, S.K. (1971): A generalized estimator for the mean of a finite population using multiauxiliary information. *J. Amer. Statist. Assoc.,* **66**, 404-407.

Tripathi, T.P. (1973): Double sampling for inclusion probabilities and regression method of estimation. *J. Indian Statist. Assoc.,* **10**, 33-46.

Tripathi, T.P. (1976): On double sampling for multivariate ratio and difference method of estimation. *J. Indian Statist. Assoc,* **33**, 33-54.

Tripathi, T.P. (1987): A class of estimators for population mean using multivariate auxiliary information under general sampling designs. *Aligarh J. Statist.,* **7**, 49-62.

Tripathi, T.P. (1989): *Optimum estimation of mean vector for dynamic population*. Invited paper in the Proceedings of the International Symposium on Optimization and Statistics, held at AMU, Aligarh, 8-21 Dec. 1989.

Tripathi, T.P. and Chaubey, Y.P. (1993): *Optimum probabilities of the selection in PPS sampling based on super population model and multivariate information*. Tech. Report No. **2/ 93**; Stat-Math, ISI, Calcutta.

Tripathi, T.P. and Chaubey, Y.P. (2000): *Estimators for finite Population mean vector based on multivariate auxiliary information*. Current Development in Survey Sampling, Editor A.K.P.C. Swain, The Modern Book Depot. Bhubaneshwar, India.

Tripathi, T.P. and Khattree, R. (1989): *Simultaneous estimation of several means using multivariate auxiliary information*. Tech. Report No. **13/ 89,** Stat-Math. Unit, ISI, Calcutta, India.

Received : 02-06-2004

Revised : 15-02-2006

Mohammad Vaseem Ismail
Department of Mathematics
Integral University, Lucknow, India
e-mail: vaseem_ismail2@rediffmail.com

Abdul Razzaq
Department of Statistics and Oper. Res.
AMU, Aligarh-202 002, India

T.P. Tripathi
Math-Stats Division,, ISI, Calcutta, India