Improved Class of Estimators using Multi-Auxiliary Information in Successive Sampling

Sunil Kumar
Department of Statistics
University of Jammu
Jammu – 180004
J & K, India
sunilbhougal@gmail.com

Sandeep Bhougal
School of Mathematics
Shri Mata Vaishno Devi University
Katra, J & K, India
sandeep.bhougal@smvdu.ac.in

J. P. Singh Joorel
Department of Statistics
University of Jammu
Jammu – 180004
J & K, India
jroorel@rediffmail.com

Abstract

The present study propounded an improved class of estimators using multi-auxiliary information in successive sampling at both the design and the estimation stages. The proposed method is an extension of the work by Tankou and Dharmadhikari (1989) in successive sampling. The present study is compared with the work carried out by Olkin (1958). The numerical illustration is carried out to show the performance of the proposed strategy.

Keywords: Auxiliary Variable, Study Variable, Successive sampling, Ratio-type estimator.

2000 AMS Classification: 62D05.

1. Introduction

If the value of study character of a finite population is subject to change (dynamic) over time, a survey carried out on a single occasion will provide information about the characteristics of the surveyed population for the given occasion only and can not give any information on the nature or the rate of change of the characteristic over different occasions and the average value of the characteristic over all occasions or most recent occasion. To meet these requirements, three possible repetitive survey sampling procedures may be used:

i)   Extracting a new sample on each occasion (repeated sampling),

ii)   Using the same sample every occasion (panel sampling),

iii) Performing a partial replacement of units from one occasion to another (sampling on successive occasions, which is also called rotation sampling when the units constructed in the number of stages in which they are to become part of the sample, as it happens with the EPA-Spanish Survey of Working Population-which are performed quarterly and most of the family surveys carried out by the INE-Spanish Statistics Institute).

The third possibility has been examined extensively by several authors in case of estimating the population mean (total) Jessen (1942), Patterson (1950), (Rao and Mudholkar (1967), Das (1982), Okator and Arnab (1987), Okafor (1992), Artes and Garcia (2000 a, 2001 a, b, 2005), Singh and Vishwakarma (2007), Singh and kumar (2008), etc).

It has been shown, in this context, that the combined estimator that uses a double sampling ratio estimator for the matched part of the sample is more accurate than the simple estimator when the auxiliary variable is positively related to the study variable.

In this paper, we have proposed an improved class of estimators using multi-auxiliary variable for estimating the population mean $\bar{Y}$ of the study variable y in two occasions successive sampling.

Suppose that the samples are of size n on both occasions, we use a simple random sampling and the size of the population N is sufficiently large for the factor of correction be ignored.

Let a simple random sample of size n be selected on the first occasion from a universe of size N. For selecting the second sample, we assume that $m = pn, (0 < p < 1)$ of the units of the selected sample on the first occasion are retained for the second occasion (matched sample) and the remaining $u = n - m = qn, (q = 1 - p)$ units are replaced by a new selection from the universe $N - m$ left after omitting the m units. Here, we have a population of N identifiable units on which $(p+1)$ characteristics $y, x_1, ..., x_p$ are defined. Here y is the study variable and $x_j; j = 1, 2, ..., p$ are the auxiliary characteristics. The symbols Y_i and $X_{jj'}$, respectively, denote the values of y and x_j on the unit j'. Suppose that a probability sample is chosen by using some sampling design d and let $\bar{y}_{dm}$ and $\bar{x}_{jdm}$, respectively, be some estimators of the population means $\bar{Y}$ and $\bar{X}_j$ of the study variable y and the auxiliary variables x_j, respectively, on matched portion of second occasion. Thus, we construct an estimator following the lines of Tankou and Dharmadhikari (1989) in successive sampling as

$$T_R = \sum_{j=1}^{p} k_j \bar{y}_{dm} \left(\frac{\bar{x}_{j(n)}}{\bar{x}_{jdm}} \right) + k_{p+1} \bar{y}_{dm} , \tag{2.1}$$

where $\bar{y}_{dm} = \dfrac{1}{m} \sum_{i=1}^{m} y_{di}$; $\bar{x}_{jdm} = \dfrac{1}{m} \sum_{i=1}^{m} x_{jdi}$; $j = 1, 2, ... p$; $\bar{x}_{j(n)} = \dfrac{1}{n} \sum_{i=1}^{n} x_{j(i)}$; $\sum_{j=1}^{p+1} k_j = 1$.

For $k_j = 0$, $j = 1,2,....,p$; the proposed estimator reduces to $\bar{y}_{dm}$, which is the simplest unbiased estimator with variance

$$Var(\bar{y}_{dm}) = \left(\frac{1}{m}\right)S_y^2 \tag{2.2}$$

When $k_{p+1} = 0$, the proposed estimator reduces to t_{Ro}, which we find on the classical method of Olkin (1958) as

$$t_{Ro} = \sum_{j=1}^{p} k_j \bar{y}_{dm}\left(\frac{\bar{x}_{j(n)}}{\bar{x}_{jdm}}\right) \tag{2.3}$$

with minimum *MSE* as given by

$$min.MSE(t_{Ro}) = S_0^2\left[\frac{1}{m} + \left(\frac{1}{m} - \frac{1}{n}\right)\frac{(p-1)\rho_{ij}\rho_{oi}^2 - (2p-1)\rho_{oi}^2}{p}\right], \tag{2.4}$$

where ρ_{oj} and $\rho_{jj'}$ are the Pearson correlation coefficient between (x_j and y) and (x_j and x_j'); $j, j' = 1,2,....,k$ ($j \neq j'$), respectively, $p = \dfrac{m}{n}$; the matching fraction; S_0^2 is the population variance of the study variable y, S_j^2 is the population variance of the jth auxiliary variable x_j; $(j = 1,2,...,p)$; $S_j'^2$ is the population variance of the j'^{th} auxiliary variable x_j' for matched portion, ρ_{0j} is the correlation coefficient between study variable y and j^{th} auxiliary variable x_j for the matched portion; $\rho_{jj'}$ is the correlation coefficient between j^{th} auxiliary variable x_j and j'^{th} auxiliary variable x_j' for matched portion.

Thus, for the optimal choice of the weights k_j, the estimator T_R is better than both the estimators t_{Ro} and $\bar{y}_{dm}$.

The estimator T_R is biased. Also the constants k_j are to be chosen so as to minimize the mean squared error (*MSE*) of T_R. Denote $t_{om} = \bar{y}_{dm} - \bar{Y}$ and

$$t_{Rj} = \bar{y}_{dm} - \bar{y}_{dm}\frac{\bar{x}_{j(n)}}{\bar{x}_{jdm}}; j = 1,2,....,p .$$

Since $k_{p+1} = 1 - \sum_{j=1}^{p} k_j$, one can easily obtain

$$T_R - \bar{Y} = t_{om} - \sum_{j=1}^{p} k_j t_{Rj} \tag{2.5}$$

Thus, the optimum MSE of T_R is equivalent to obtaining the best homogeneous linear predictor of t_{om} in terms of $t_{1m},......,t_{pm}$. For this reason, regression type calculations are needed to find the optimal values of the k_j.

Pak.j.stat.oper.res. Vol.VI No.2 2010 pp117-124

119

Let k be the column vector $(k_1,......,k_p)'$. Let Z be the $p \times p$ matrix whose $(j,u)^{th}$ entry is $E(t_{Rj} t_{Ru})$. Write $z_j = E(t_{0m} t_{Rj})$ and let z be the column vector formed by $z_1, z_2,....,z_p$. Then, the mean squared error (MSE) of T_R is

$$MSE(T_R) = E(T_R - \overline{Y})^2 = E(t_{om}^2) - 2 z' k + k' Z k = f(k) \text{, say}$$

The function f is quadratic and convex. Therefore, the critical point corresponds to the minimum value of ' f '. The optimal choice of k which minimizes $MSE(T_R)$ is easily obtained as $k = Z^{-1} z$, and

$$min.MSE(T_R) = E(t_{0m}^2) - z' Z^{-1} z . \tag{2.6}$$

Since the minimum value cannot be negative, we can write

$$min.MSE(T_R) = E(t_{0m}^2)(1 - \gamma^2)$$

where $\quad \gamma^2 = \dfrac{z' Z^{-1} z}{E(t_{0m}^2)}$.

The non-negativity of (2.6) shows that $\gamma^2 \leq 1$. We note that γ is just like the multiple correlation coefficient of t_{0m} on $t_{im},.....,t_{pm}$, the only change being that $Cov(t_{Rj} t_{Ru})$ is replaced by $E(t_{Rj} t_{Ru})$. For this reason, we write γ^2 more suggestively as $\gamma^2_{t_{0m},t_{1m},....t_{pm}}$. Thus, the minimum *MSE* of T_R is given as

$$min.MSE(T_R) = E(t_{0m}^2)(1 - \gamma_{t_{0m},t_{1m},....t_{pm}}) \tag{2.7}$$

2. First Order Approximation To $MSE(T_R)$

Suppose that the introductory estimators $\overline{y}_{dm}, \overline{x}_{jdm}$ and $\overline{x}_{j(n)}$ are unbiased estimators of $\overline{Y}$ and $\overline{X}_j$, respectively. Thus, we can obtain a first order approximation to $MSE(T_R)$ as follows

For $j = 1,2,..., p$, let $\varepsilon_0 = \dfrac{\overline{y}_{dm} - \overline{Y}}{\overline{Y}}$; $\varepsilon_j = \dfrac{\overline{x}_{jdm} - \overline{X}_j}{\overline{X}_j}$; $\varepsilon_j' = \dfrac{\overline{x}_{jdn} - \overline{X}_j}{\overline{X}_j}$;

such that

$$E(\varepsilon_0) = E(\varepsilon_j) = E(\varepsilon_j') = 0;$$

$$E(\varepsilon_0^2) = \frac{1}{m}S_0^2; \ E(\varepsilon_j^2) = \frac{1}{m}S_j^2; \ E(\varepsilon_j'^2) = \frac{1}{n}S_{j'}^2; \ E(\varepsilon_0\varepsilon_j) = \frac{1}{m}\rho_{0j}S_0 S_j;$$

$$E(\varepsilon_0\varepsilon_j') = \frac{1}{n}\rho_{0j'}S_0 S_{j'}; \ E(\varepsilon_j\varepsilon_j') = \frac{1}{n}\rho_{jj'}S_j S_{j'}.$$

120

Pak.j.stat.oper.res. Vol.VI No.2 2010 pp117-124

Expanding $t_{Rj} = \bar{y}_{dm} - \bar{y}_{dm}\left(\dfrac{\bar{x}_{j(n)}}{\bar{x}_{jdm}}\right)$ in terms of $\varepsilon's$ to the first degree of approximation, we have

$$t_{Rj} = \bar{Y}(1+\varepsilon_0) - \bar{Y}(1+\varepsilon_0)(1+\varepsilon'_j)(1-\varepsilon_j)^{-1}$$
$$\cong \bar{Y}(\varepsilon_j - \varepsilon'_j) \tag{2.8}$$

Thus

$$E(t_{Rj}t_{Ru}) \cong \bar{Y}^2[Cov(\varepsilon_j,\varepsilon_u) - Cov(\varepsilon'_j,\varepsilon_u)]$$

where j, u both go from 0 to p.

Let H denotes the covariance matrix of the variables $(\varepsilon^*_1, \varepsilon^*_2,....,\varepsilon^*_p)$ with $\varepsilon^*_j = \varepsilon_j - \varepsilon'_j$ and let $\underset{\sim}{h}$ denotes the column p-vector formed by the quantities $Cov(\varepsilon_0, \varepsilon^*_j)$. Then $Z \cong \bar{Y}^2 H$ and $\underset{\sim}{z} \cong \bar{Y}^2 \underset{\sim}{h}$. Therefore, $\underset{\sim}{k} = Z^{-1} \underset{\sim}{z} \cong H^{-1} \underset{\sim}{h}$ Further, (2.6) shows that

$$min.MSE(T_R) \approx \bar{Y}^2[Var(\varepsilon_0) - \underset{\sim}{h}' H^{-1} \underset{\sim}{h}]$$

$$= \bar{Y}^2 Var(\varepsilon_0)\left(1 - \frac{\underset{\sim}{h}' H^{-1} \underset{\sim}{h}}{Var(\varepsilon_0)}\right)$$

$$= Var(\bar{y}_{dm})\left[1 - \rho^2_{\varepsilon_0.\varepsilon^*_1,\varepsilon^*_2,....,\varepsilon^*_p}\right]$$

$$\text{or } min.MSE(T_R) = Var(\bar{y}_{dm})\left[1 - \rho^2_{\bar{y}_{dm}.\bar{x}^*_{1dm},\bar{x}^*_{2dm},.....,\bar{x}^*_{pdm}}\right] \tag{2.9}$$

where $\rho^2_{\bar{y}_{dm}.\bar{x}^*_{1dm},\bar{x}^*_{2dm},.....,\bar{x}^*_{pdm}} = \left(\dfrac{1}{m} - \dfrac{1}{n}\right)\dfrac{R^2_{0.1,2,...p}}{Var(\bar{y}_{dm})}S^2_0$

Then, (2.9) reduces to

$$min.MSE(T_R) = Var(\bar{y}_{dm})\left[1 - \left(\frac{1}{m} - \frac{1}{n}\right)\frac{R^2_{0.1,2,...p}}{Var(\bar{y}_{dm})}S^2_0\right]$$

$$= S^2_0\left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(1 - R^2_{0.1,2,...p}\right) + \frac{1}{n}\right] \tag{2.10}$$

For the sake of simplicity, we assume that $\rho_{0j} = \rho_0; \rho_{jj'} = \rho; R_j = \dfrac{\bar{Y}}{\bar{X}_j}$, thus we obtain

$$\left(1 - R^2_{0.1,2,...p}\right) = \frac{\left[1 + (p-1)\rho - p\rho^2_0\right]}{1 + (p-1)\rho}.$$

Thus, the minimum $MSE(T_R)$ at (2.10) reduces to

$$min.MSE(T_R) = S^2_y\left[\left(\frac{1}{m} - \frac{1}{n}\right)\frac{\left\{1 - \rho^2_0 + (p-1)\left(\rho - \rho^2_0\right)\right\}}{1 + (p-1)\rho} + \frac{1}{n}\right] \tag{2.11}$$

Pak.j.stat.oper.res. Vol.VI No.2 2010 pp117-124

121

3. Comparison Of Estimators

From (2.1), (2.4) and (2.11), we get

$$Var(\bar{y}_{dm}) - min.MSE(T_R) = \left[\left(\frac{1}{m}+\frac{1}{n}\right) - \left(\frac{1}{m}-\frac{1}{n}\right)\frac{\{1-\rho_0^2+(p-1)(\rho-\rho_0^2)\}}{1+(p-1)\rho}\right] \geq 0 \quad (3.1)$$

$$min.MSE(t_{R0}) - min.MSE(T_R) = \left(\frac{1}{m}-\frac{1}{n}\right)\frac{\rho_0^2(p-1)^2(1-\rho)^2}{p\{1+(p-1)\rho\}}S_y^2 \geq 0 \quad (3.2)$$

It follows that the proposed estimator T_R is more efficient than the usual unbiased estimator $\bar{y}_d$ and the estimator by Olkin (1958) if the equations (3.1) and (3.2) are satisfied. Thus, it is to be mentioned that the proposed estimator T_R is recommended for the further research in this field.

4. Numerical Illustration

To see the performance of the proposed estimator T_R and the estimator t_{R0} by Olkin (1958), over the direct estimator $\bar{y}_{dm}$, we calculate the percent relative efficiencies of the proposed estimator T_R and the estimator t_{R0} for different choices of ρ, ρ_0, p, S_y^2 and n, respectively, as shown in Table 1.

Table 1: Percent relative efficiencies (*PRE's*) of different estimators with respect to unbiased estimator $\bar{y}_{dm}$ for different values of ρ, ρ_0, p, and S_0^2, when n is fixed i.e. $(n=40)$.

ρ	ρ_0	p	S_0^2	E_1	E_2
$\rho > \rho_0$					
0.6	0.5	0.7	0.6	106.61	106.33
0.7	0.3	0.7	0.6	102.02	101.98
0.8	0.4	0.7	0.6	104.23	104.19
$\rho < \rho_0$					
0.5	0.6	0.7	0.6	109.45	108.95
0.3	0.7	0.7	0.6	112.49	111.19
0.2	0.8	0.7	0.6	116.50	114.21
$\rho = \rho_0$					
0.4	0.4	0.7	0.6	103.56	103.29
0.6	0.6	0.7	0.6	109.85	109.52
0.8	0.8	0.7	0.6	121.40	121.21
p					
0.5	0.7	0.6	0.8	117.31	112.77
0.5	0.7	0.5	0.8	119.54	109.59
0.5	0.7	0.4	0.8	120.19	107.93
S_0^2					
0.4	0.5	0.3	0.7	108.27	107.94
0.4	0.5	0.3	0.5	108.27	107.94
0.4	0.5	0.3	0.3	108.27	107.94

It is to be explicit from the Table 1 that the percent relative efficiency of the proposed estimator T_R over unbiased estimator $\bar{y}_{dm}$ is maximum over the estimator t_{R0} by Olkin (1958) over direct estimator $\bar{y}_{dm}$ in all cases. For the case $\rho > \rho_0$, the efficiencies of both the estimators, firstly decreases then increases. For the case $\rho < \rho_0$ and $\rho = \rho_0$, the percent relative efficiencies of both estimators i.e. T_R and t_{R0} over the direct estimator $\bar{y}_{dm}$ increases. For the different values of p, it increases for T_R while decreases for t_{R0}. Further, for the different values of S_0^2, it remains constant for both the cases.

On the conclusion, the proposed estimator T_R in successive sampling is more efficient than the estimator t_{R0} by Olkin (1958) and the unbiased estimator $\bar{y}_{dm}$. So, it is recommended to use the proposed estimator in further research in this field.

Acknowledgement

References

1.   Artes, E. and Garcia, A. (2005). Multivariate Indirect Methods of Estimation in Successive Sampling. J. Ind. Soc. Agril. Statist., 59 (2), 97-103.

2.   Artes, E. and Garcia, A. (2001 b). Successive sampling for the ratio of population parameters. Journal of the Portuguese Nacional Statistical Institute, 2, 43-44.

3.   Artes, E. and Garcia, A. (2001 a). Metodo diferencia multivariate en muestreo en dos ocasiones. VIII conferencia Espanola de Biometria, 199-200.

4.   Artes, E. and Garcia, A. (2000 a). A note on successive sampling using auxiliary information. Proceedings of the 15th International Work-shop on Statistical Modeling, 376-379.

5.   Das, A.K. (1982). Estimation of population ratio on two occasions. J. Indian Soc. Agricultural Statist., 34, 1-9.

6.   Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa agricultural experiment statistical research bulletin, 304.

7.   Okafor, F.C. (1992). The Theory and application of sampling over two occasions for the estimation of current population ratio. Statistica, 1, 137-147.

8.   Okafor, F.C. and Arnab, R. (1987). Some strategies of two-stage sampling for estimating population ratios over two occasions. Austral. J. Statist., 29(2), 128-142.

9.   Olkin, I. (1958). Multivariate ratio estimation for finite population. Biometrika, 45, 154-165.

10.   Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. J. Roy. Statist. Soc. Ser B, 12, 241-255.

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

12.   Singh, H.P. and Kumar, S. (2008). Estimation of population product in presence of non-response in successive sampling. Statistical Papers. DOI 10. 1007/s00362-008-0193-5.

13.   Singh, H.P. and Vishwakarma, G.K. (2007). A general class of estimators in successive sampling. Metron, LXV, 2, 201-227.

14.   Tankou, V. and Dharmadhikari, S. (1989). Improvement of ratio-type estimators. Biometrical Journal, 37, 7, 795-802.