USE OF CERTAIN TYPE OF MULTI-AUXILIARY INFORMATION FOR INCREASING THE EFFICIENCY OF CLUSTER SAMPLING WITH RATIO AND REGRESSION ESTIMATORS

By

S. K. Agarwal[1] and B.B.P.S Goel[2]

(Received : May, 1978)

1. Introduction

Cluster sampling is often used in sample surveys for its well known advantages. However, cluster sampling is less efficient than simple random sampling (s. r. s.) for a given sample size, whenever intra-class correlation coefficient for the character under study is positive and which is almost always so in actual practice.

Zarkovich and Krane [5] have shown that the correlation between the cluster means of two characters increases as the cluster size increases. Mishro and Sukhatme [3] utilized these findings and gave the conditions under which cluster sampling in conjunction with ratio/regression estimators is more efficient than s. r. s. in conjunction with the corresponding estimators even if the intra-class correlation coefficient is positive.

It is shown in this paper that using multi-auxiliary information this efficiency can be increased further.

Mishro and Sukhatme [3] showed that cluster sampling with regression and ratio estimators is more efficient than s.r.s. with corresponding estimators even if the intra-class correlation coefficient for the variable under study is positive provided

$$\rho_b^2 > \rho^2 + \frac{(M-1)\rho'(1-\rho^2)}{1+(M-1)\rho'} \qquad ...(1.1)$$

1. University of Jodhpur, Jodhpur-342 001

2. I.A.S.R.I., New Delhi-110 012

for regression method of estimation and

$$\rho_b > \rho + \frac{(M-1)\,\rho'\,(1-\rho)}{1+(M-1)\,\rho'} \quad \text{...(1.2)}$$

for ratio method of estimation,

where,

ρ_b is the correlation coefficient between cluster means of the character under study (y) and auxiliary character (x) for clusters of size M.

ρ is the correlation cocfficient between y and x

ρ' is the intra-class correlation for character y

The efficiencies E_1 and E_2 of cluster sampling with regression and ratio estimators relative to that of s. r.s. with the corresponding estimators, for a fixed cost, are given by

$$E_1 \doteq \frac{1-\rho^2}{(1-\rho_b^2)\,(1+\overline{M-1}\,\rho')} \times \frac{n^*}{n} \quad \text{...(1.3)}$$

and

$$E_2 = \frac{1-\rho}{(1-\rho_b)\,\{1+(M-1)\rho'\}} \times \frac{n^*}{n} \quad \text{...(1.4)}$$

respectively, where the fixed cost (say C_0) permits a s. r. s. of nM elements or a s. r. s. of n^* ($n^* > n$) clusters of size M.

2. Construction of Indices

Agarwal [1] gave several methods for constructing indices of sizes of units of a population as linear functions of p-auxiliary variates whenever such information is available for all units of a population. Let $x_1, x_2 \ldots\ldots\ldots x_p$ be p-auxiliary variates. Consider the Index I given by

$$I = \underline{a}'\ \underline{X} \quad \text{...(2.1)}$$

where

$$\underline{X}' = (x_1, x_2, \ldots\ldots\ldots, x_p)$$

and

$$\underline{a}' = (a_1, a_2, \ldots\ldots\ldots, a_p)$$

$\underline{a}$, is chosen such that

(*i*) I has a correlation with y higher than the correlation which any of the x's has with it.

(*ii*) The regression hyper plane given by (2.1) is nearer to the origin than the regression line based on any of the x's.

There exist several criteria to obtain vector $\underline{a}$ which will maximize the correlation co-efficient between y and I. However, all such indices will be linear functions of one another and give the same value of the correlation coefficient and also the same intercept on the y axis, Rao [4].

3. Use of P-Auxiliary Variates to Increase the Efficiency of Cluster Sampling

From (1.3) and (1.4) it can be seen that the efficiencies E_1 and E_2 of cluster sampling, depend upon ρ, ρ', ρ_b and M. For a given set of clusters of size M, ρ' is fixed and also ρ is fixed for given x and y and therefore the values of E_1 and E_2 can be increased by increasing ρ_b. By constructing an index $I_{(M)}$ similar to the one given in section 2 as linear combination of cluster means of p-auxiliary variates, the correlation coefficient, ρ'_b, between cluster means of y and those of the index can be increased considerably. The proposed index is

$$I_{(M)}=\underline{C}'\underline{\bar{X}}_{(M)} \qquad \ldots(3.1)$$

where $\underline{\bar{X}}'_{(M)}=(\bar{x}_{1(M)}, \bar{x}_{2(M)}, \ldots\ldots\ldots, \bar{x}_{p(M)})$

and $\bar{x}_{1(M)}, \bar{x}_{2(M)}, \ldots\ldots\ldots \bar{x}_{p(M)}$ are means of clusters of size M and

$$\underline{C}'=(c_1, c_2, \ldots\ldots, c_p)$$

is the vector of constants to be chosen so that

$$\rho'_b > \rho_b.$$

The efficiency of cluster sampling with regression and ratio estimators based on $I_{(M)}$ is given by

$$E_1'=\frac{1-\rho_I^2}{\left(1-\rho_b'^2\right)\{1+(M-1)\rho'\}}\times\frac{n^*}{n} \qquad \ldots(3.2)$$

and

$$E_2'=\frac{1-\rho_I}{(1-\rho_b')\{1+(M-1)\rho'\}}\times\frac{n^*}{n} \qquad \ldots(3.3)$$

respectively. Here ρ_I is the correlation coefficient between y and I. It can be shown that the efficiencies E_1' and E_2' will be considerably higher than E_1 and E_2 respectively. E_1' will be greater than E_1

if
$$\frac{\left(1-\rho_I^2\right)\left(1-\rho_{ib}^2\right)}{\left(1-\rho_i^2\right)\left(1-\rho_b'^2\right)}-1>0 \quad \forall i=1, 2, \ldots p. \qquad \ldots(3.4)$$

where ρ_i and ρ_{ib} are correlation coefficients between y and x_i and their respective cluster means.

This result can be shown to hold good in the case of two auxiliary variates ($p=2$) x_1 and x_2. In this case (3.4) reduces to

$$\frac{\left(\rho_b'^2-\rho_{1b}^2\right)-\left(\rho_I^2-\rho_1^2\right)+\rho_I^2\ \rho_{1b}^2-\rho_1^2\ \rho_b'^2}{\left(1-\rho_1^2\right)\left(1-\rho_b'^2\right)}>0$$

Substituting the values of ρ_b' and ρ_{1b} in some of the terms in the above we get

$$\frac{\dfrac{(\rho_{1b}-\rho_{2b}\,\rho'_{12})^2}{1-\rho_{12}'^2}-\dfrac{(\rho_1-\rho_1\,\rho_{12})^2}{1-\rho_{12}^2}+\rho_I^2\ \rho_{1b}^2-\rho_1^2\ \rho_b^2}{\left(1-\rho_1'^2\right)\left(1-\rho_b'^2\right)}>0 \quad \text{...(3.5)}$$

where ρ_{12} and ρ'_{12} are the correlation coefficients between x_1 and x_2 and their respective cluster means.

For the sake of simplicity let,

$$\rho_1=\rho_2=\rho$$

and

$$\rho_{1b}=\rho_{2b}=\rho_b$$

under these assumption (3.5) reduces to

$$\frac{\dfrac{\rho_b^2\,(1-\rho'_{12})}{(1+\rho'_{12})}-\dfrac{\rho^2(1-\rho_{12})}{1+\rho_{12}}+2\ \rho_b^2\ \rho^2\left(\dfrac{1}{1+\rho_{12}}\right)-\dfrac{1}{1-\rho'_{12}}\Bigg)}{(1-\rho^2)\left(1-\rho_b'^2\right)}>0$$

After some algebraic manipulation this further simplifies to

$$\rho_b(1+\rho_{12})\,(1-\rho'_{12})+\rho(1+\rho'_{12})\,(1-\rho_{12})>0$$

which is obviously true. Thus E_1' will be greater than E_1.

Similarly it can be proved that $E_2'>E_2$.

4. Empirical Illustration

Table 1 presents the correlation coefficients ρ_1 (between number of cultivators, y, and number of male cultivators, x_1), ρ_2 (between number of cultivators and number of female cultivators, x_2), ρ_{12} (between x_1 and x_2) and ρ_b' (between means of y and $I_{(M)}$ for clusters of 2 and 4 villages *i.e.* $M=2$, 4) in Lodha block of Aligarh district (U.P.)

TABLE 1

Values of various correlation coefficients for clusters of 1,2 and 4 villages.

Correlation coefficients \ Cluster Size (M)	1	2	4
ρ_1	0.701	0.832	0.882
ρ_2	0.670	0.782	0.812
ρ_{12}	0.321	0·542	0.611
ρ'_b	0.844	0·921	0.947

Using the values of $\rho_i (i=1, 2)$ and ρ'_b from Table 1, the percentage gain in efficiency of cluster sampling with regression and ratio estimator using $I_{(M)}$ over that using x_1 or x_2 is given in Table 2.

TABLE 2

Percentage gain in efficiency of cluster sampling with regression/ratio estimators using $I_{(M)}$ in place of x_1 or x_2.

Cluster size (M)	$\frac{E_1'-E_1}{E_1} \times 100$		$\frac{E_2'-E_2}{E_2} \times 100$	
	Using $I_{(M)}$ in place of		*Using $I_{(M)}$ in place of*	
	x_1	x_2	x_1	x_2
2	12.2	19.4	10.9	30.5
4	28.8	13 2	65.4	67.7

SUMMARY

Following Mishro and Sukhatme [3] use of multi-auxiliary information has been suggested for building up ratio and regression estimates for increasing the efficiency of cluster sampling further.

REFERENCES

[1] Agarwal, S. K. (1977) : Use of Multi-auxiliary Information in Probability Sampling, unpublished Ph.D. thesis—University of Jodhpur, Jodhpur (Rajasthan]

[2] District Census (1961) : Handbook Aligarh (U.P.)

[3] Mishro, G. K. and Sukhatme, B. V. (1972) : Efficiency of cluster sampling in conjunction with ratio and regression methods of estimation, *J.I S.A S.* Vol. 24(2) pp 81-90.

(4] Rao, C. R. (1965) : Linear Statistical Inference and its Applications, John Wiley & Sons Inc. (Chapter 4 (g) pp 223).

[5] Zarkovich, S. S. and Krane, J. (1965) : Some efficient ways of Cluster Sampling Proc. of 35th session of I. S. I. Belgrade.