EFFICIENCY OF CLUSTER SAMPLING IN CONJUNCTION WITH RATIO AND REGRESSION METHODS OF ESTIMATION*

G.K. MISHRO[1] AND B.V. SUKHATME

(*Received in December, 1971*)

INTRODUCTION

Cluster sampling is often used in the designing of large scale surveys, mainly on account of low cost and operational convenience in the collection of data as also organization and supervision of field work. Its use is more appealing when the population is widely scattered and repeated observations have to be made on the selected units. However, cluster sampling provides a less efficient estimate of the population mean or total than simple random sampling whenever the intra-class correlation coefficient is positive.

When data on an ancillary characteristic X correlated with Y, the characteristic under study are available for all the units in the population, it is customary to use the data to provide a more efficient of the population mean $\bar{Y}$ by using ratio or regression method of estimation. The resulting gain in efficiency will naturally depend on the correlation between X and Y. Zarkovich and Krane (1965) have shown that the correlation between the two characteristics X and Y increases if a cluster of units is taken as a sampling unit and that the correlation increases with the size of the cluster especially when the ancillary characteristic is the value of the characteristic under study on some previous occasion. Based on these findings, we shall consider the problem of estimating the poplation mean $\bar{Y}$, derive conditions under which cluster sampling in conjunction with ratio and regression methods of estimation is more efficient than simple random sampling in conjunction with ratio and regression methods of estimation even if the intra class correlation coefficient is positive and illustrate the results with the help of two sets of data collected in sample surveys conducted by the Institute of Agricultural Research Statistics, New Delhi.

[1] Bureau of Statistics and Economics, Cuttack, India

[2] Iowa State University, Ames, Iowa

* Contributed for the Silver Jubilee Session of the Indian Society of Agricultural Staristics.

2. Notation and preliminaries

We shall assume that a finite population is composed of N clusters of M units each and that a sample of n clusters is drawn with equal probabilities and without replacement. Denote by Y_{ij} the value of the characteristic under study corresponding to the j-th unit of the i-th cluster $j=1, 2, \ldots, M$ and $i=1, 2, \ldots, N$.

Let

$$\bar{Y}=\frac{1}{M}\sum_{j=1}^{M} Y_{ij}\ , \text{ the mean per unit of the } i\text{-th cluster}$$

$$\bar{Y}_{..}=\frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{i.}=\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}, \text{ the population mean}$$

$$S^2_{by}=\frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_{i.}-\bar{Y}_{..})^2\ , \text{ the mean square between cluster means } \bar{Y}_{i.}$$

$$S^2_{Y}=\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_{..})^2, \text{ the mean square between } Y_{ij}$$

$$\bar{Y}_c=\frac{1}{n}\sum^{n}\bar{y}_{i.}\ , \text{ the mean of cluster means in the sample}$$

with similar definitions for $\bar{X}_{i.}$, $\bar{X}_{..}$, S^2_{bX} S^2_{X} and $\bar{x}_c$.

Several ratio and regression type estimators can be formed. We shall only consider two of these estimators; namely the ratio estimator

$$\bar{y}_{Rc}=\frac{\bar{y}_c}{\bar{x}_c}\cdot\bar{X}_{..} \qquad \ldots(1)$$

and the regression estimator

$$\bar{y}_{lc}=\bar{y}_c+\hat{\beta}_c\,(\bar{X}_{..}-\bar{x}_c) \qquad \ldots(2)$$

where $\hat{\beta}_c$ is the estimated regression cofficient of $\bar{Y}_{i.}$ on $\bar{X}_{i}$.

It is well known that both the estimators are biased, the bias in $\bar{y}_{lc}$ vanishing when the relationship between $\bar{Y}_i$ and $\bar{X}_{i\cdot}$ is linear. If in addition, the relationship also passes through the origin, the bias in $\bar{y}_{Rc}$ is also zero. Further, it can be seen that the mean square errors of $\bar{y}_{Rc}$ and $\bar{y}_{lc}$ to the first degree of approximation are given by

$$\text{M.S.E. } (\bar{y}_{Rc}) = \frac{N-n}{N} \cdot \frac{1}{n}[S^2_{bY} + R^2 S^2_{bX} - 2R\rho_{bX} S_{bY}] \quad \ldots(3)$$

and
$$\text{M.S.E. } (\bar{y}_{lc}) = \frac{N-n}{N} \; \frac{1}{n} \; (1-\rho_b^2) \; S^2_{b_y} \quad \ldots(4)$$

where ρ_b is the correlation coefficient between $\bar{Y}_i$ and $\bar{X}_i$. and $R = \bar{Y}../X..$.

3. Efficiency of cluster sampling

If an equivalent sample of size nM were drawn by simple random sampling. the corresponding ratio and regression estimators of $\bar{Y}..$ are

$$\bar{y}_R = \frac{\bar{y}(nM)}{\bar{x}\,(uM)} \,.\, \bar{X}.. \quad \ldots(5)$$

and

$$\bar{y}_l = \bar{y}_{(nM)} + \hat{\beta}\,(\bar{X}.. - \bar{x}_{(nM)}) \quad \ldots(6)$$

where $\bar{z}_v$ denotes the mean based on v units in the sample and $\hat{\beta}$ is the estimated regression coefficient of y on X.

Further, the mean square errors of these estimators to the first degree of approximation (Sukhatme and Sukhatme, 1970) are given by

$$\text{M.S.E. } (\bar{y}_R = \frac{N-n}{N} \cdot \frac{1}{nM}[S^2_Y + R^2 S^2_X - 2R\rho S_X S_Y] \quad \ldots(7)$$

and

$$\text{M.S.E. } (\bar{y}_l) = \frac{N-n}{M} \cdot \frac{1}{nM} S^2_y (1-\rho^2) \quad \ldots(8)$$

where ρ is the correlation coefficient between y and X.

To discuss the efficiency of cluster sampling, it is convenient to express the mean square errors of $\bar{y}_{Rc}$ and $\bar{y}_{lc}$ in terms of intra class correlation coefficients ρ_x and ρ_y by relations of the type

$$S^2_{bY} = \frac{MN-1}{M(N-1)}\left[1+(M-1)\,\rho_Y\right]\frac{S^2_Y}{M} \qquad \ldots(9)$$

In addition, we shall assume that $\rho_X = \rho_Y = \rho'$ $\frac{NM-1}{M(N-1)} \doteq 1$

and $$\frac{s^2_Y}{\bar{Y}^2..} = \frac{s^2_X}{\bar{X}^2..} = C^2$$

It can then be seen after some manipulation that

$$\text{M.S.E. } (\bar{y}_{Rc}) \doteq \frac{N-n}{N}\cdot\frac{1}{nM}\, 2C^2\bar{Y}^2..[1+(M-1)\rho'][1-\rho_b] \qquad \ldots(10)$$

and

$$\text{M.S.E. } (y_{lc}) \doteq \frac{N-n}{N}\cdot\frac{1}{nM}\, C^2\,\bar{Y}^2..[1+M-1)\,\rho'][1-\rho^2_c] \qquad \ldots(11)$$

Also,

$$\text{M.S.E. } (\bar{y}_R) \doteq \frac{N-n}{N}\cdot\frac{1}{nM}\cdot 2\,C^2\bar{Y}^2..(1-\rho) \qquad \ldots(12)$$

and

$$\text{M.S.E. } (\bar{y}_l) = \frac{N-n}{N}\cdot\frac{1}{nM}\,C^2\bar{Y}^2..(1-\rho^2) \qquad \ldots(13)$$

On comparing (10) and (12), it can be seen that

$$\text{M.S.E. } (\bar{y}_{Rc}) < \text{M.S.E. } (\bar{y}_R)$$

If

$$\rho_b > \rho + \frac{(M-1)\rho(1-\rho')}{[1+(M-1)\,\rho']} \qquad \ldots(14)$$

Similarly, it can be seen from (11) and (13) that

$$\text{M.S.E. } (\bar{y}_{lc}) < \text{M.S.E. } (\bar{y}_l)$$

If

$$\rho_b^2 > \rho^2 + \frac{(M-1)\rho'(1-\delta^2)}{[1+(M-1)\,\rho']} \qquad \ldots(15)$$

Curves showing values of ρb corresponding to different values of $(M-1)\rho'$ for ρ ranging from 0.5 to 0.8 for which M.S.E. $(\bar{y}_{Rc})$=M.S.E. $(\bar{y}_R)$ are shown in Figure 1. Corresponding curves for regression method of estimation are shown in Figure 2. It follows that for values of ρ_b greater than the plotted value, cluster sampling

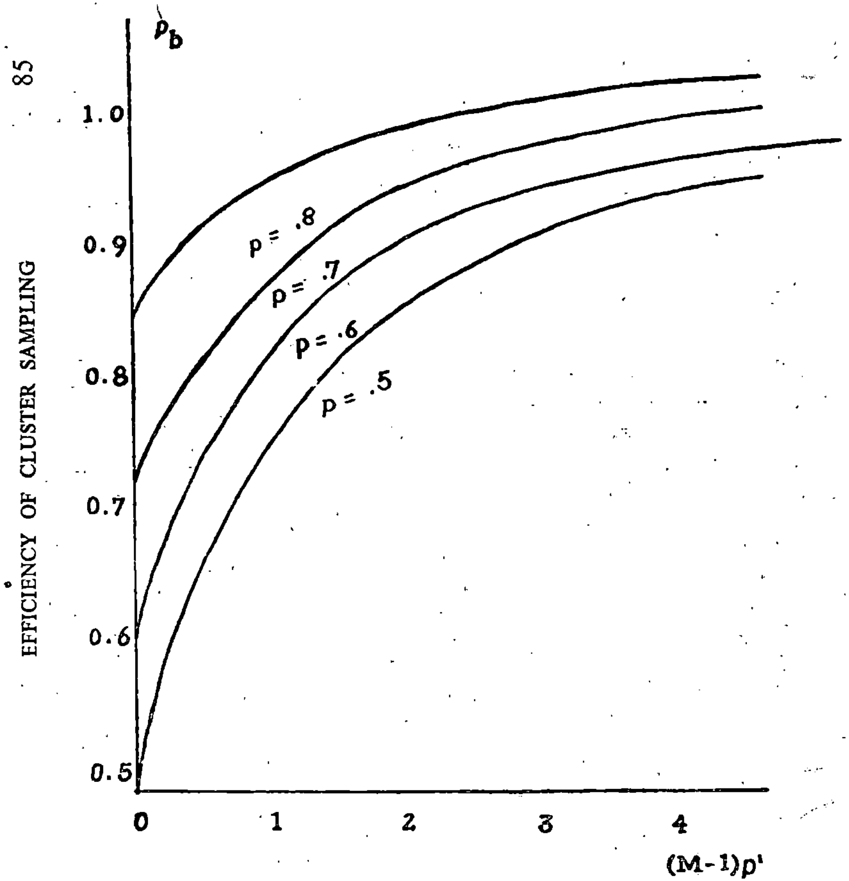

Figure 1

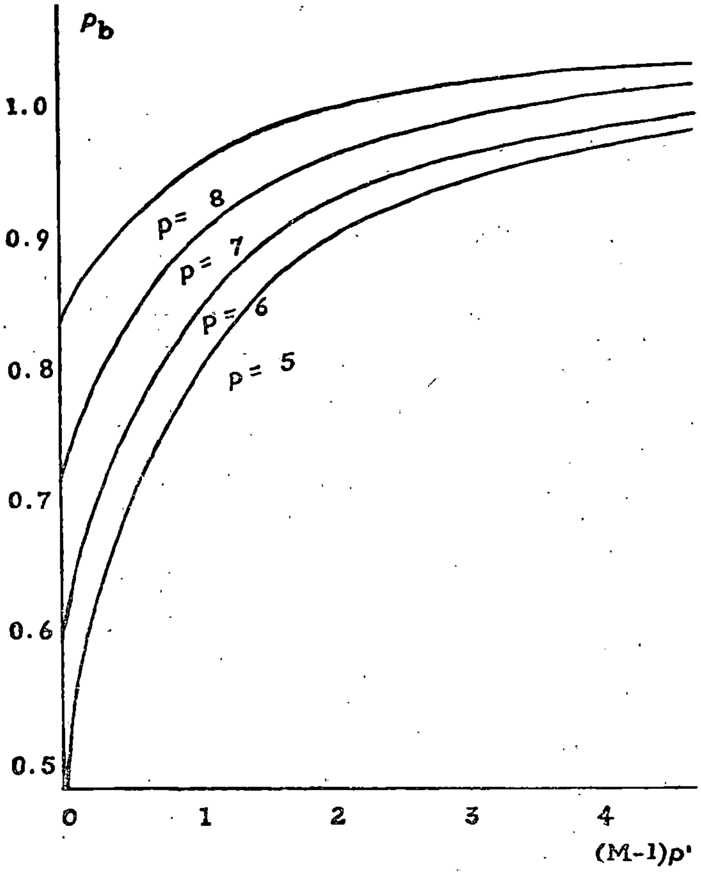

Figure 2

in conjunction with ratio or regression method of estimation will be more efficient than simple random sampling in conjunction with ratio or regression method of estimation.

We shall briefly consider the case when the population mean $\bar{X}_{..}$ is not known. In this situation, the usual technique consists in drawing a simple random sample of n' clusters to estimate $\bar{X}_{..}$ while a sub-sample of n clusters is drawn from n' to estimate the ratio and regression estimators of $\bar{Y}_{..}$ now take the form

$$\bar{y}'_{Rc} = \frac{\bar{y}_c}{\bar{x}_c} \cdot \hat{\bar{X}}_{..} \qquad .. (16)$$

and

$$\bar{y}'_{lc} = \bar{y}_c + \hat{\beta}_c (\hat{\bar{X}}_{..} - \bar{x}_c) \qquad ...(17)$$

where

$$\hat{\bar{X}}_{..} = \frac{1}{n^1} \sum^{n'} \bar{x}_i \qquad ...(18)$$

The corresponding estimators in simple random sampling when a sample of size $n'M$ is drawn to estimate $\bar{X}_{..}$ and a sub-sample of size nM is drawn from $n'M$ to estimate the ratio R are

$$\bar{y}'_R = \frac{\bar{y}_{(nM)}}{\bar{x}_{(nM)}} \cdot \bar{x}_{(n'M)} \qquad ...(19)$$

and

$$\bar{y}'_b = \bar{y}_{(nM)} + \hat{\beta}(\bar{x}_{(n'M)} - \bar{x}_{(nM)}) \qquad ...(20)$$

Proceeding as before, it can then be seen that to the first order of approximation

$$\text{M.S.E.}\ (\bar{y}'_{Rc}) < \text{M.S.E.}\ (\bar{y}'_R)$$

If

$$\rho_b > \rho + \frac{(M-1)\rho'[\frac{1}{1-f} - \rho]}{1+(M-1)\rho'} \qquad ...(21)$$

and

$$\text{M.S.E.}\ (\bar{y}'_{lc}) < \text{M.S.E.}\ (\bar{y}'_l)$$

if

$$\rho_b^2 > \rho^2 + \frac{(M-1)\rho'[\frac{1}{1-f} - \rho^2]}{1+(M-1)\rho'} \qquad ...(22)$$

where $f=n/n'$. If $n^1=N$ and N is so large that $\frac{N-n}{N} \doteq 1$, then the conditions (21) and (22) reduce to (14) and (15) as is to be expected.

4. Numerical illustration

In some surveys, it is necessary to take repeated observations on the units selected in the sample. For example, in surveys to estimate the total productions of fruits and vegetables, a field assistant has to attend to all the pickings in the selected orchards or fields of the sampled villages. In a survey to estimate the total production of milk and the cost of production of milk, a field assistant has to visit the selected stalls at regular intervals. In such cases if the sample units are very much scattered, it will be difficult to carry out the work smoothly and efficiently without spending considerable amounts of money by way of time and effort. In such cases, cluster sampling under certain conditions is likely to prove useful and efficient.

We have seen in Section 3 that cluster sampling in conjunction with ratio or regression methods of estimation will provide a more efficient estimate of the population mean or total as compared to simple random sampling provided the correlation coefficient ρ_b between the characteristic under study and the ancillary characteristic for clusters is greater than a certain quantity defined in equations (14) and (15). We shall therefore examine the possibility of using cluster sampling in the type of surveys mentioned above. To examine the possibility of using a cluster of villages as a unit of sampling in surveys to estimate the production of vegetables and livestock products, correlation coefficients were calculated for varying cluster sizes between (a) the areas under vegetables for two successive years (1963-64 and 64-65) in Najafgarh block of Delhi and Delhi state [3] and (b) the number of households and the bovine population in Hissar tehsil of Panjab [4]. These results are presented in Tables 1 and 2.

TABLE 1

Correlation coefficient between areas under vegetables for 1963-64 and 1964-65 for varying cluster size

Cluster size M	*1*	*2*	*4*
Najafagarh Block	0.788	0.942	0.970
Delhi State	0.668	0.709	0.725

TABLE 2

Correlation coefficient between number of households and the characteristic under study for verying cluster size

Cluster size M / *Characteristic under study*	*1*	*2*	*4*
Number of cattle	0.767	0 791	0.855
Number of buffaloes	0.764	0.789	0.792
Bovine population	0.820	0.846	0.881

It will be seen that the correlation coefficient increases with the size of the cluster. If the increase in correlation is more than $\frac{(M-1)\rho'(1-\rho),}{[1+(M-1)\rho']}$ it follows that cluster sampling in conjunction with ratio method of estimation will provide a more efficient estimate of the population mean or total as compared to simple random sampling in conjunction with ratio method of estimation. It appears that the use of cluster sampling will not only improve the efficiency of the estimate but also reduce the cost and make the operational and administrative work easier.

For a fixed cost C_0, the efficiency of cluster sampling with respect to simple random sampling, using ratio method of estimation is given by

$$E_1=\frac{(1-\rho)}{(1+\rho_b)\,[1+(M-1)\rho']}\cdot\frac{n^*}{n} \tag{23}$$

while that using regression method of estimation is given by

$$E_2=\frac{(1-\rho^2)}{(1-\rho_b^2)[1+(M-1)\rho']}\cdot\frac{n^*}{n} \tag{24}$$

where the cost C_0 permits a sample of size nM units under simple random sampling and a sample n^*M units $(n^*>n)$ under cluster sampling. The values of E_1 and E_2 have been calculated for varying cluster size utilizing values of the corresponding correlation coefficient from tables 1 and 2 and different values of ρ' and are given in tables

3 and 4. It will be seen that whenever n^*/n is greater than 1.10, cluster sampling is always more efficient than simple random sampling, the percentage gain in efficiency being very considerable at times.

TABLE 3

Efficiency of cluster sampling in conjunction with ratio method of estimation for varying cluster size and fixed cost

		Characteristic under study			
		Area under vegetables		*Bovine population*	
n^*/n	M / ρ'	2	4	2	4
1.10	0.00	1.25	1.33	1.28	1.66
	0.05	1.20	1.16	1.22	1.44
	0.10	1.14	1.02	1.17	1.24
	0.15	1.09	0.91	1.11	1.14
1.20	0.00	1.37	1.45	1.39	1.81
	0.05	1.31	1.26	1.33	1.57
	0.10	1.25	1.17	1.27	1.36
	0.15	1.19	1.00	1.21	1.25
1.30	0.00	1.48	1.57	1.51	1.96
	0.05	1.42	1.37	1.44	1.70
	0.10	1.35	1.21	1 38	1.47
	0.15	1.29	1.08	1.32	1.35
1.40	0.00	1.60	1.69	1.62	2.11
	0.05	1.53	1.47	1.55	1.83
	0.10	1.46	1.30	1.48	1.58
	0.15	1.39	1.16	1.41	1.46
1.50	0.00	1.71	1.82	1.72	2.27
	0.05	1.64	1.58	1.67	1.97
	0.10	1.56	1.40	1.59	1.70
	0.15	1.49	1.25	1.52	1.56

TABLE 4

Efficiency of cluster sampling in conjunction with regression method of estimation for varying cluster size and fixed cost

		Characteristic under study			
		Area under vegetables		*Bovine population*	
n^*/n	M / ρ'	2	4	2	4
1.10	0.00	1.22	1.29	1.27	1.61
	0.05	1.17	1.11	1.20	1.40
	0.10	1.11	0.99	1.14	1.23
	0.15	1.07	0.88	1.10	1.11
1.20	0.00	1.33	1.40	1.38	1.75
	0.05	1.27	1.21	1.31	1.52
	0.10	1.21	1.08	1.24	1.34
	0.15	1.16	0.96	1.20	1.21
1.30	0.00	1.44	1.52	1.50	1.90
	0.05	1.38	1.31	1.42	1.65
	0.10	1.31	1.17	1.35	1.46
	0.15	1.26	1.04	1.30	1.31
1.40	0.00	1.55	1.64	1.61	2.04
	0.05	1.48	1.41	1.53	1.78
	0.10	1.41	1.26	1.46	1.57
	0.15	1.36	1.12	1.40	1.41
1.50	0.00	1.67	1.76	1.73	2.19
	0.05	1.60	1.52	1.64	1.91
	0.10	1.52	1.35	1.56	1.68
	0.15	1.46	1.20	1.50	1.52

REFERENCES

I.C.A.R., New Delhi (1963)	: Sample surveys for estimation of milk production (Punjab). Bulletin.
I.C.A.R., New Delhi (1964-66)	: Sample surveys to study yield and cultivation practices of vegetables in Delhi, Unpublished report.
Sukhatme, P.V. and Sukhatme, B.V. (1970)	: Sampling Theory of Surveys with Applications, Iowa State Press, Ames, Iowa.
Zarkovich, S.S. and Krane, J. (1965)	: 'Some efficient ways of cluster sampling', Proceedings of 35th Session of International Statistical Institute, Belgrade.