STATISTICA, anno LXIV, n. 1, 2004

ON EFFICIENCY OF CLUSTER SAMPLING ON SAMPLING ON TWO OCCASIONS

Bijoy Kumar Pradhan

1. INTRODUCTION

In sample surveys cluster or area sampling is widely practised because of its low cost and time saving device to conduct large scale and complicated surveys. Its use becomes more desirable when a list of elements is not available or units of the population are widely scattered and it is required to take repeated observations on the selected units. However, the cluster sampling is more efficient than the comparable simple random sampling only when the intra-class correlation within the same cluster is negative and less than $-\frac{1}{(M-1)}$.

As, the relative efficiency of the cluster sampling is controlled by both the size of the cluster and the intra-class correlation coefficient, it decreases if the size of the cluster increases substantially (Sukhatme and Sukhatme, 1970).

In practice the intra-class correlation is usually positive and further increase in size of the cluster leads to substantial decrease in the relative efficiency. Zarkovich and Krane (1965) have shown that the correlation between two characters in cluster sampling with clusters as sampling units increases with increase in cluster size and correlation coefficient between the clusters as units is expected to be larger than correlation coefficient in element sampling.

In the following sampling on two occasions is considered to estimate population mean on second occasion when the sampling units are clusters and the observations on the first occasion are regarded as ancillary information for the observations on the second or current occasion.

2. NOTATION AND PRELIMINARIES

Consider a simple random sample of n clusters drawn from a population which consists of N clusters of M units each.

Let x and y be the characters under study in the first and second occasion respectively. In the second occasion m of the n clusters selected on the first occasion are retained at random and the remaining u=n-m of the clusters are replaced by a fresh selection. x and y are supposed to be correlated when they are observed on the same unit repeatedly.

Let

X_{ij}= Value of the chatacter under study in the first occasion, for the j^{th} unit of the i^{th} cluster ($i=1, 2, \ldots, N$ and $j=1, 2, \ldots, M$)

Y_{ij}= Value of the chatacter under study in the second occasion, for the j^{th} unit of the i^{th} cluster ($i=1, 2, \ldots, N$ and $j=1, 2, \ldots, M$)

Define,

i. $\overline{X}_{i.} = \frac{1}{M}\sum_{j=1}^{M} X_{ij}$ and $\overline{Y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ as means of the i^{th} cluster on the first and second occasion respectively.

ii. $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_{i.}$ and $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_{i.}$ as cluster population mean of x and y respectively.

iii. $\overline{X}_{nM} = \frac{1}{nM}\sum^{nM} x_{ij}$ and $\overline{Y}_{nM} = \frac{1}{nM}\sum^{nM} y_{ij}$ as sample means based on a simple random sample of nM units.

iv. $\overline{X}_{uM} = \frac{1}{uM}\sum^{uM} x_{ij}$ and $\overline{Y}_{uM} = \frac{1}{uM}\sum^{uM} y_{ij}$ as sample means based on a simple random sample of uM units.

v. $\overline{X}_{mM} = \frac{1}{mM}\sum^{mM} x_{ij}$ and $\overline{Y}_{mM} = \frac{1}{mM}\sum^{mM} y_{ij}$ as sample means based on a simple random sample of mM units.

3. A GENERALIZED ESTIMATOR IN CLUSTER SAMPLING ON TWO SUCCESSIVE OCCASIONS

Consider a generalized estimator $\overline{t}$ of the population mean $\overline{Y}$ on second or current occasion as

$$\overline{t} = a\,\overline{x}_{uM} + b\,\overline{x}_{mM} + c\,\overline{y}_{uM} + d\,\overline{y}_{mM} \tag{3.1}$$

where a, b, c and d are suitable constants.

We have,

$$\mathrm{E}(\overline{t}) = (a+b)\,\overline{X}_{NM} + (c+d)\,\overline{Y}_{NM} \tag{3.2}$$

In order that $\overline{t}$ is an unbiased estimator of $\overline{Y}_{NM}$, we have

$$(a+b)=0 \text{ and } (c+d)=1 \tag{3.3}$$

Hence, the estimator (3.1) takes the form

$$\overline{t} = a\,(\,\overline{x}_{uM} - \overline{x}_{mM}\,) + c\,\overline{y}_{uM} + (1 - c)\,\overline{y}_{mM} \tag{3.4}$$

The variance of estimator $\overline{t}$ is

$$\begin{aligned} \mathrm{V}(\overline{t}) = {} & a^2\,\mathrm{V}\,(\,\overline{x}_{uM}\,) + a^2\mathrm{V}\,(\,\overline{x}_{mM}\,) + c^2 V\,(\,\overline{y}_{uM}\,) \\ & + (1-c)^2\mathrm{V}\,(\,\overline{y}_{mM}\,) - 2\,a(1 - c)\,\mathrm{Cov}\,(\,\overline{y}_{mM}\,,\overline{x}_{mM}\,) \end{aligned} \tag{3.5}$$

other covariance terms being zero.

Minimising the variance of $\overline{t}$ with respect to *a* and *c* when *N* is sufficiently large, the optimum values of *a* and *c* are

$$a = \frac{S_y}{S_x}\left[\frac{1+\overline{M-1}\rho_y}{1+\overline{M-1}\rho_x}\right]^{1/2} .\frac{\mu(1-\mu)\rho_b}{1-\mu^2\rho_b^2} \text{ and } c = 1 - \left[\frac{1-\mu}{1-\mu^2\rho_b^2}\right] \tag{3.6}$$

where $\frac{u}{n} = \mu\,,\rho_y$ and ρ_x are intra-class correlation coefficients defined by

$$\rho_y = \frac{\sum\sum_{i,j<k}(Y_{ij} - \overline{Y}_{NM})(Y_{ik} - \overline{Y}_{NM})}{(M - 1)(NM - 1)S_y^2},$$

$$\rho_x = \frac{\sum\sum_{i,j<k}(X_{ij} - \overline{X}_{NM})(X_{ik} - \overline{X}_{NM})}{(M - 1)(NM - 1)S_x^2},$$

$$S_y^2 = \frac{\sum_i\sum_j(Y_{ij} - \overline{Y}_{NM})^2}{NM - 1},$$

$$S_x^2 = \frac{\sum_i\sum_j(X_{ij} - \overline{X}_{NM})^2}{NM - 1},$$

and the simple correlation coefficient between cluster means on both occasions is defined by

$$\rho_b = \frac{\sum_{i=1}^{N}(\overline{Y}_i - \overline{Y}_{NM})(\overline{X}_i - \overline{X}_{NM})}{\left[\sum_{i=1}^{N}(\overline{Y}_i - \overline{Y}_{NM})^2\sum_{i=1}^{N}(\overline{X}_i - \overline{X}_{NM})^2\right]^{1/2}}$$

Using the optimum values of *a* and *c* given by (3.6), the estimator $\bar{t}$ reduces to

$$\bar{t} = \frac{S_y}{S_x}\left[\frac{1+\overline{M-1}\rho_y}{1+\overline{M-1}\rho_x}\right]^{1/2} \frac{\mu(1-\mu)\rho_b}{1-\mu^2\rho_b{}^2}\left(\bar{x}_{uM}-\bar{x}_{mM}\right)+\frac{1-\mu}{1-\mu^2\rho_b{}^2}\bar{y}_{mM}+\left[1-\frac{1-\mu}{1-\mu^2\rho_b{}^2}\right]\bar{y}_{mM} \qquad (3.7)$$

with variance

$$\mathrm{V}_{opt}(\bar{t}) = (1+\overline{M-1}\rho_y)\left[\frac{1-\mu\rho_b^2}{1-\mu^2\rho_b^2}\right]\frac{S_y^2}{nM} \qquad (3.8)$$

Now, the optimum value of μ is given by further minimising $V_{opt}(\bar{t})$ with respect to μ, which gives

$$\mu = \left[1+\sqrt{1-\rho_b^2}\right]^{-1} \qquad (3.9)$$

Thus,

$$\mathrm{V}_{opt(opt)}(\bar{t}) = \frac{1+\overline{M-1}\rho_y}{M}\left[1+\sqrt{1-\rho_b^2}\right]\frac{S_y^2}{2n} \qquad (3.10)$$

4. EFFICIENCY OF CLUSTER SAMPLING ON TWO OCCASIONS

If the samples on both occasions are drawn using SRSWOR, the variance of the optimum estimator $\bar{t}_1$ neglecting the finite population correction factor is given by

$$\mathrm{V}_{opt}(\bar{t}) = \left[1-\sqrt{1-\rho^2}\right]\frac{S_y^2}{2Mn}, \qquad (4.1)$$

where ρ is the simple correlation coefficient between values of units on first and second occasion.

The relative efficiency of $\bar{t}$ compared to $\bar{t}_1$ is

$$\frac{\mathrm{V}_{opt}(\bar{t}_1)}{\mathrm{V}_{opt(opt)}(\bar{t})} = \frac{1+\sqrt{1-\rho^2}}{(1+\overline{M-1}\rho_y)\left(1+\sqrt{1-\rho_b^2}\right)} \qquad (4.2)$$

$\bar{t}$ would be more efficient than $\bar{t}_1$ if

$$\rho_y \le -\frac{1}{M-1} + \frac{1}{M-1}\left[\frac{1+\sqrt{1-\rho^2}}{1+\sqrt{1-\rho_b^2}}\right]$$

i.e. $\rho_y \le \frac{\delta\text{ - }1}{M\text{ - }1}$ where $\delta = \left[\frac{1+\sqrt{1-\rho^2}}{1+\sqrt{1-\rho_b^2}}\right]$ (4.3)

Further, in order that $\overline{t}$ would be more efficient than $\overline{t}_1$ if

$$M \le \frac{1}{\rho_y}\left[\frac{\sqrt{1-\rho^2}-\sqrt{1-\rho_b^2}}{1-\sqrt{1-\rho_b^2}}\right]+1 \tag{4.4}$$

which gives the upper limit of *M*.

As ρ_b^2 is expected to be greater than ρ^2 (Zarkovich and Krane, 1965) δ is likely to be greater than unity and as such, even if ρ_y is positive, the cluster sampling on both occasions provides more efficient estimate than the simple random sampling on both occasions.

Tables: 1-3 have been computed below to show the relative efficiency of cluster sampling in sampling on two occasions compared to simple random sampling of elements for some specified values of ρ_y, ρ , ρ_b and *M*.

TABLE 1

Relative efficiency of cluster sampling over simple random sampling

(M=2, ρ_y =0.01)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	1.0585	1.0973	1.1523	1.2345	1.3756	1.5052	1.6474
0.2	1.0505	1.0890	1.1435	1.2251	1.3651	1.4938	1.6349
0.3	1.0367	1.0748	1.1286	1.2091	1.3473	1.4742	1.6135
0.4	1.0169	1.0542	1.1070	1.1860	1.3215	1.4460	1.5826
0.5	0.9901	1.0264	1.0778	1.1547	1.2867	1.4079	1.5409

(M=4, ρ_y =0.01)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	1.0380	1.0760	1.1299	1.2105	1.3489	1.4760	1.6154
0.2	1.0301	1.0678	1.1234	1.2013	1.3386	1.4648	1.6031
0.3	1.0166	1.0539	1.1067	1.1856	1.3211	1.4456	1.5822
0.4	0.9971	1.0337	1.0855	1.1629	1.2958	1.4179	1.5519
0.5	0.9708	1.0065	1.0569	1.1323	1.2617	1.3806	1.5110

(M=5, ρ_y =0.01)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	1.0280	1.0656	1.1190	1.1989	1.3359	1.4618	1.5999
0.2	1.0202	1.0575	1.1126	1.1897	1.3257	1.4507	1.5877
0.3	1.0068	1.0438	1.0960	1.1742	1.3084	1.4317	1.5670
0.4	0.9875	1.0238	1.0751	1.1517	1.2833	1.4043	1.5370
0.5	0.9615	0.9968	1.0467	1.1214	1.2496	1.3673	1.4965

TABLE 2

Relative efficiency of cluster sampling over simple random sampling

(M=2, ρ_y =0.05)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	1.0182	1.0555	1.1084	1.1875	1.3232	1.4478	1.5846
0.2	1.0105	1.0475	1.0999	1.1784	1.3131	1.4369	1.5726
0.3	0.9972	1.0338	1.0856	1.1630	1.2960	1.4180	1.5520
0.4	0.9782	1.0140	1.0648	1.1408	1.2711	1.3909	1.5223
0.5	0.9524	0.9873	1.0367	1.1107	1.2377	1.3543	1.4822

(M=4, ρ_y =0.05)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	0.9297	0.9637	1.0120	1.0842	1.2081	1.3220	1.4468
0.2	0.9226	0.9564	1.0062	1.0759	1.1989	1.3119	1.4358
0.3	0.9105	0.9439	0.9912	1.0619	1.1832	1.2947	1.4171
0.4	0.8930	0.9258	0.9722	1.0415	1.1606	1.2699	1.3900
0.5	0.8695	0.9015	0.9466	1.0141	1.1300	1.2365	1.3533

(M=5, ρ_y =0.05)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	0.8909	0.9235	0.9698	1.0390	1.1578	1.2669	1.3866
0.2	0.8842	0.9165	0.9642	1.0311	1.1490	1.2573	1.3760
0.3	0.8726	0.9046	0.9499	1.0177	1.1340	1.2408	1.3581
0.4	0.8558	0.8873	0.9317	0.9982	1.1123	1.2171	1.3321
0.5	0.8333	0.8639	0.9071	0.9719	1.0830	1.1850	1.2970

TABLE 3

Relative efficiency of cluster sampling over simple random sampling

(M=2, ρ_y =0.1)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	0.9719	1.0075	1.0580	1.1335	1.2631	1.3820	1.5126
0.2	0.9645	1.0000	1.0499	1.1249	1.2534	1.3716	1.5011
0.3	0.9519	0.9860	1.0363	1.1102	1.2371	1.3536	1.4815
0.4	0.9337	0.9679	1.0164	1.0890	1.2134	1.3277	1.4531
0.5	0.9091	0.9424	0.9896	1.0602	1.1814	1.2927	1.4148

(M=4, ρ_y =0.1)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	0.8224	0.8525	0.8952	0.9591	1.0687	1.1694	1.2799
0.2	0.8161	0.8460	0.8901	0.9518	1.0606	1.1606	1.2701
0.3	0.8055	0.8350	0.8768	0.9394	1.0467	1.1454	1.2536
0.4	0.7900	0.8190	0.8600	0.9214	1.0267	1.1234	1.2296
0.5	0.7692	0.7974	0.8374	0.8971	0.9996	1.0939	1.1972

(M=5, ρ_y =0.1)

ρ/ρ_b	0.5	0.6	0.7	0.8	0.9	0.95	0.98
0.1	0.7850	0.7916	0.8312	0.8906	0.9924	1.0859	1.1885
0.2	0.7579	0.7856	0.8265	0.8838	0.9848	1.0777	1.1794
0.3	0.7479	0.7754	0.8142	0.8723	0.9715	1.0635	1.1640
0.4	0.7336	0.7605	0.7986	0.8555	0.9533	1.0432	1.1418
0.5	0.7142	0.7405	0.7775	0.8330	0.9283	1.0157	1.1117

COMMENTS:

(a) For fixed ρ_y (intra-class correlation coefficient) and ρ , the efficiency increases with large increase in $\rho_b(\rho_b > \rho\ )$ (correlation coefficient between cluster means).

(b) For fixed ρ_b and ρ , the efficiency decreases with increase in ρ_y .

Note: The same results are obtained if optimum estimator is formed by a weighted combination of double sampling regression estimator using variate values of the entire clusters in the first occasion and of the clusters relating to matched part in the second occasion and simple estimator using unmatched clusters in the second occasion. (Appendix - A).

Department of Statistics
Utkal University, Bhubaneswar

BIJOY KUMAR PRADHAN

APPENDIX-A

For estimating $\overline{Y}_{NM}$, let us consider the following regression estimate of $\overline{Y}_{NM}$ as

$$\bar{t}_1^* = \bar{y}_{mM} + b\,(\bar{x}_{uM} - \bar{x}_{mM}\,)$$

where *b* is the sample regression coefficient.

Hence,

$$\mathrm{V}(\bar{t}_1^*) = \mathrm{V}\,(\,\bar{y}_{mM}\,) + b^2\,\mathrm{V}\,(\,\bar{x}_{uM}\,) + b^2\,\mathrm{V}\,(\bar{x}_{mM}\,) - 2\,b\,\mathrm{COV}\,(\bar{x}_{mM}\,,\bar{y}_{mM}\,),$$

other covariance terms being zero.
Therefore,

$$\mathrm{V}(\bar{t}_1^*) = (\,1 + \overline{M-1}\,\rho_y\,)(\,1-\rho_b^2\,)\frac{S_y^2}{mM} + (1 + \overline{M-1}\,\rho_y)\,\rho_b^2\,\frac{S_y^2}{nM}$$

The estimate of $\overline{Y}_{NM}$ from the unmatched portion in the second occasion is given by

$$\bar{t}_2^* = \bar{y}_{uM}$$

with variance

$$\mathrm{V}(\overline{t}_2^{\,*}) = (1+\overline{M-1}\rho_y)\frac{S_y^2}{uM} \text{ for large } N.$$

The weighted estimate of $\overline{Y}_{NM}$, say, $\overline{t}^{\,*}$, is given by weighting $\overline{t}_1^{\,*}$ and $\overline{t}_2^{\,*}$ inversely proportional to their population variance.

Hence $$\overline{t}^{\,*} = \frac{w_1\overline{t}_1^{\,*} + w_2\overline{t}_2^{\,*}}{w_1 + w_2}$$

where $$w_1 = \frac{1}{V(\overline{t}_1^{\,*})} \text{ and } w_2 = \frac{1}{V(\overline{t}_2^{\,*})}$$

Therefore, the variance of $\overline{t}^{\,*}$ is given by

$$\mathrm{V}(\overline{t}^{\,*}) = (1+\overline{M-1}\rho_y)\left[\frac{1+\mu\rho_b^2}{1+\mu^2\rho_b^2}\right]\frac{S_y^2}{nM}$$

where $\mu = \dfrac{u}{n}$.

The optimum value of μ which minimizes $V(\overline{t}^{\,*})$ is

$$\mu = \left[1+\sqrt{1-\rho_b^2}\right]^{-1}$$

Hence,

$$\mathrm{V}_{opt}(\overline{t}^{\,*}) = (1+\overline{M-1}\rho_y)\left[1+\sqrt{1-\rho_b^2}\right]\frac{S_y^2}{2nM}.$$

ACKNOWLEDGEMENTS

Author is thankful to the referee for many useful suggestion for improving the manuscript.

REFERENCE

M.H. HANSEN, W.N. HURWITZ, W.G. MADOW (1953), *Sample Survey Methods and Theory*, Wiley, New York.

P.V. SUKHATME, B.V. SUKHATME (1970), *Sampling theory of Surveys with Applications, Food and Agriculture Organisation*, Rome, Second Edition.

S.S. ZARKOVICH, J. KRANE (1965), *Some efficient ways of cluster sampling*. Proceedings of 35th session of International Statistical Institute, Belgrade.

RIASSUNTO

Sull'efficienza del campionamento a grappolo nel campionamento a due occasioni

In questo lavoro si mostra come anche quando il coefficiente di correlazione intraclasse tra le unità dello stesso cluster è positivo, il campionamento a grappolo su due occorrenze (occasions) è più efficiente che il campionamento casuale semplice su due occorrenze (occasions) se si vuole stimare la media di popolazione di un carattere studiato sulla (seconda) occorrenza in esame.

SUMMARY

On efficiency of cluster sampling on sampling on two occasions

In this paper it has been shown that even if the intra-class correlation coeffcient among the units in the same cluster is positive, under certain condition, the cluster sampling on two occasions is likely to be more efficient than the simple random sampling on two occasions to estimate the population mean of the character under study on current (second) occasion.