ON EFFICIENCY OF CLUSTER SAMPLING

By D. Singh

Indian Council of Agricultural Research, New Delhi

Introduction

The principal advantage of a sub-sampling design over the one stage random sampling is that it is more flexible, operationally more convenient, and involves comparatively less travelling expenditure. In the limiting case of sub-sampling procedure when all the second-stage units in the selected first-stage sampling units are included in the sample, the system is defined as the cluster sampling. The latter method further introduces operational convenience and cuts down the travelling costs considerably. But it suffers from a serious defect that its statistical efficiency goes down as compared with the sub-sampling system when the intraclass correlation coefficient between the second-stage units is positive. Thus from the statistical point of view the efficiency of a sub-sampling system may lie between the efficiencies obtained by the above two systems, namely, (*i*) one-stage sampling, and (*ii*) cluster sampling whenever the intraclass correlation coefficient between the second-stage units is positive.

Sometimes, the average distance between the second-stage units may not be small and even if it is small there may not be transport facilities to travel it. In such cases much of the advantage of the sub-sampling method is lost. For example, if in a survey, tehsils or talukas whose area varies from 500–1000 square miles be first-stage sampling units and villages the second-stage units, the distance between the two villages in the sample may not be small; particularly in a country like India whose rural areas are so undeveloped in matters of communications that it is not unusual to come across the paradoxical situation that under certain circumstances sample villages separated by big distances will be much "nearer in time" and more convenient of approach as regards physical exhaustion due to travelling (which indirectly affects the quality of fieldwork for which no quantitative measurements can be given) than villages which are apparently quite close to one another. In the former case when villages are favourably situated in respect of train or bus services, the distance can be covered comparatively easily within a short period of time. In the latter case one has sometimes to plod the whole distance on foot and this possibly in tracts with practically no roads and the movements have to be

confined to daytime only. Moreover, the enumerator may have to make very large detours to avoid forests, hills, swamps or rivers. The human factor is not only important for the enumerator but also for the respondent or interviewee. An enumerator's stay for a reasonably long period is conducive to the establishment of good relations between the enumerator and the householders. Within limits, longer the enumerator's stay at a sample spot, the easier will it be for him to secure the necessary co-operation of the householders and better will be his chances of collecting more complete and more accurate information by call-backs if necessary to the appropriate persons who are in a position to give the correct information. Another consideration of great importance needs mention. An increase in the number of sample spots (a village in case of sub-sampling design and a group of villages in the other case) is fraught with certain difficulties. Perhaps the greatest drawback is that the enumerator is a human being and he cannot be expected to move very frequently from one point to another, which would be necessary if the total number of sample spots to be covered by him in each round is to be large, specially under very trying conditions prevailing in the rural tracts of this country. Under such circumstances it will be worth considering the one-stage cluster sampling of villages against the sub-sampling procedure where tehsils are first-stage units and villages as second-stage sample units.

The main effect of the elimination of large first-stage units like tehsil would appear to be an increase in travel time. But it is important to point out that increase is not so much as one would expect at first thought. Tehsils are generally connected by rails, roads and travelling between tehsils may not be very expensive 'in unit of time'.

Suppose, for simplicity, that each of the first-stage units contains the same number of second-stage units and let the population be composed of N first units of M second-stage units each. Let n denote the number of first-stage units in the sample and m the number of second-stage units to be drawn from each selected first-stage unit. Further, assume that the units of each stage are drawn with equal probability. For example, N may be the number of tehsils or talukas, each consisting of M villages. It may also be assumed that N and M are large so that n/N and m/M may be neglected.

A further stage of sampling, *viz.*, third-stage, may also be considered but it is not very relevant here as distances between two-third-stage units within second-stage units is negligible and involves no physical exhaustion in travelling it.

Now let

y_{ij} = the value of the j-th second-stage unit within i-th first-stage unit ($j = 1, 2, \ldots\ldots\ldots\ldots M$, and $i = 1, 2, \ldots\ldots\ldots\ldots N$),

$\bar{y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ = the mean per second-stage unit in the i-th first-stage unit,

$\bar{y}_{NM} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$ = the mean per second-stage unit in the population,

$\bar{y}_{nm} = \frac{1}{nm}\sum_{i}^{n}\sum_{j}^{m} y_{ij}$ = the mean per second-stage unit in the sample,

$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{y}_{NM})^2$ = mean square between first-stage units,

$S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{i.})^2$ = mean square per second-stage unit within the first-stage units, and

$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{NM})^2.$

It can be easily shown that when N and M are large,

$$S^2 \cong S_b^2 + S_w^2. \tag{1}$$

Further

$$E(\bar{y}_{nm}) = \bar{y}_{NM}$$

and

$$V_1(\bar{y}_{nm}) = \frac{S_b^2}{n} + \frac{S_w^2}{nm}. \tag{2}$$

Suppose n' clusters of size m' each are selected by the method of simple random sampling at one stage so that

$$n'm' = nm,$$

it can be shown that

$$\bar{y}_{n'm'} = \frac{1}{n'm'} \sum_{i}^{n'} \sum_{j}^{m'} y_{ij}$$

is an unbiased estimate of $\bar{y}_{NM}$, and

$$V_2(\bar{y}_{n'm'}) = \frac{S^2}{n'm'} \{1 + (m' - 1)\rho_{m'}\}, \tag{3}$$

where $\rho_{m'}$ is intraclass correlation coefficient between second-stage units within the cluster of size m'.

It is well known that whenever intraclass correlation coefficient is negative, cluster sampling is more efficient than the simple random sampling. Therefore, subsequent discussion will be applicable to the case where intraclass correlation coefficient is positive. The cluster sampling as described above is more or equal or less efficient than sub-sampling according as

$$V_1(\bar{y}_{nm}) - V_2(\bar{y}_{n'm'}) \gtrless 0$$

or

$$\frac{S_b^2}{n} + \frac{S_w^2}{nm} - \frac{S^2}{n'm'} \{1 + (m' - 1)\rho_{m'}\} \gtrless 0$$

or

$$\frac{(m-1)}{nm} S_b^2 - \frac{S^2(m'-1)}{n'm'} \rho_{m'} \gtrless 0.$$

Now for the one-stage cluster sampling to be as efficient as the two-stage sub-sampling

$$m' = (m - 1)\frac{S_b^2}{S^2\rho_{m'}} + 1. \tag{4}$$

If S_b^2/S^2 and $\rho_{m'}$ for given m' are known from the previous census or from some pilot sample surveys conducted for estimating these constants, it is easy to determine the size of the cluster.

It may be noted that for large M

$$S_b^2 \cong \rho_M S^2,$$

and

$$S_w^2 \cong (1 - \rho_M) S^2. \tag{5}$$

Now since as M increases ρ_M decreases it is easy to see that

$$\rho_{m'} > \rho_M \tag{6}$$

for

$$m' < M.$$

Now, from (5) and (6) it is observed that

$$\frac{S_b^2}{\rho_{m'} S^2} \leqslant 1. \tag{7}$$

Similarly

$$\frac{S_b^2}{S^2} \leqslant 1. \tag{8}$$

Now Table I below gives the value of θ/ρ_m, where

$$\theta = \frac{S_b^2}{S^2}.$$

TABLE I

Value of $\theta/\rho_{m'}$

$\rho_{m'}$ \ θ	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·00
·1	1·00									
·2	0·50	1·00								
·3	0·33	0·67	1·00							
·4	0·25	0·50	0·75	1·00						
·5	0·20	0·40	0·60	0·80	1·00					
·6	0·17	0·33	0·50	0·67	0·83	1·00				
·7	0·14	0·28	0·43	0·57	0·71	0·85	1·00			
·8	0·12	0·25	0·38	0·50	0·62	0·75	0·88	1·00		
·9	0·11	0·22	0·33	0·44	0·55	0·67	0·78	0·89	1·00	
1·0	0·10	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·90	1·00

For given value of m it is easy to determine the size of cluster of size m', so that efficiency in both cases, *viz.*, two-stage design with n first stage units and m second-stage units from each of the first-stage units

in the sample and one-stage design with n' clusters of size m' each may be same.

For example, when

$$m = 5, \quad \text{and} \quad \frac{\theta}{\rho_{m'}} = 0{\cdot}5,$$

from (4) it can be seen that

$$m' = 3.$$

Therefore the cluster size will be of 3 second-stage units and required number of clusters will be

$$n' = \frac{mn}{m'} = \frac{mn}{\left\{\frac{S_b^2}{S^2\rho_{m'}}(m-1)+1\right\}} = \frac{2mn}{m+1},$$

when

$$\frac{\theta}{\rho_{m'}} = 0{\cdot}5.$$

Similarly for any value of m the size of the cluster may be determined.

We have so far considered the sizes of the cluster for given mn, total samples. Now we shall consider what the effect on the expenditure of the survey will be if we adopt cluster sampling as described above instead of the two-stage sub-sampling design.

Assume that cost function for sub-sampling takes the form

$$C = c_0\sqrt{n} + c_1 n + c_2 mn + c_3 n\sqrt{m} + c_4\, mnp, \qquad (9)$$

where

C = total expected cost of the survey (exlusive of fixed overhead and expenditure at the headquarters).

c_0 will depend on the distance to be travelled between first-stage units. If n units are selected, total distance to be travelled in one round will be approximately equal to $\sqrt{An}$, where A is the area to be covered. Suppose an investigator is paid $1\frac{1}{2}$ annas per mile as travelling allowance and 8 annas per hour for his services and in one hour he travels 16 miles, then total expenditure per mile will come to 2 annas, therefore

$$c_0\sqrt{n} = 2\sqrt{An}.$$

or

$$c_0 = 2\sqrt{A}$$

= square-root of the area covered multiplied by cost incurred per mile towards the total expenditure of an investigator.

Average cost per first-stage unit will be

$$2\sqrt{A/n} \text{ annas}$$

or

$$\text{Rs. } \tfrac{1}{8}\sqrt{A/n}\,.$$

If the survey consists in collecting information in several rounds say 4 rounds

$$c_0 = \text{Rs. } \tfrac{1}{2}\sqrt{A}\,.$$

c_1 is the fixed cost associated with a first-stage unit in the sample. It will include among other things the cost of identifying fsu (first-stage units) in the sample and of assembling the materials with which to draw the sample, etc.

c_2 is the average cost per ssu (second-stage unit) included in the sample and includes the cost of selection of the ssu's, locating the sample, field identification, listing, etc.

c_3 will depend upon the distance to be travelled between the ssu's. Supposing an investigator is paid 4 annas per mile (which is not unusual because to travel the distance between ssu, he will not usually get transport like bus or rail and he will go either on foot or hire some conveyance like bullock-cart, horse-cart, etc.) and annas 8 for his service and in one hour he travels 4 miles, the total expenditure will come to 6 annas per mile; therefore,

$c_3 = 6\sqrt{A'}$ annas, where A' is the average area of a fsu.

c_4 is the cost of enumerating an element within the cluster. It includes the travel cost within ssu.

p = total number of elements per ssu on which observations are to be taken. (It may happen that p may be even a sub-sample of all the elements in the ssu. However, it does not change the structure of cost function whether all the elements of the selected ssu's are taken or only a sub-sample.)

The cost function for the cluster sampling may be of the form

$$C' = c_0' \sqrt{n'} + c_1 n' + c_2 m' n' + c_4 m' n' p$$

$$= c_0 \sqrt{\frac{mnS^2\rho_{m'}}{(m-1)S_b^2 + S^2\rho_{m'}}}$$

$$+ c_1 \frac{mnS^2\rho_{m'}}{(m-1)S_b^2 + S^2\rho_{m'}} + c_2 mn + c_4 mnp. \qquad (10)$$

C' and c_0' have the same meaning as C and c_0 in (9). The value of c_0' may be slightly higher than c_0 since c_0' is proportional to the average distance between the two clusters whereas c_0 is proportional to the average distance between two fsu's in the sample. The average distance to be travelled between two clusters will generally be higher in terms of time-unit than that to be travelled between two fsu's as the former will be equal to the sum of the average distance to be travelled between two fsu's and the distance to be travelled within the fsu to reach the cluster.

Now comparing C and C' given by the equations (9) and (10) we find

$$C' - C = c_0' \sqrt{n'} + c_1 n' - c_0 \sqrt{n} - c_1 n - c_3 n \sqrt{m}$$

$$= c_0 \sqrt{\frac{mnS^2\rho_{m'}}{(m-1)S_b^2 + S^2\rho_{m'}}}$$

$$+ c_1 \frac{mnS^2\rho_{m'}}{(m-1)S_b^2 + S^2\rho_{m'}} - c_0 \sqrt{n} - c_1 n$$

$$- c_3 n \sqrt{m}. \qquad (11)$$

Cluster sampling will be more efficient with respect to expenditure of the survey than the sub-sampling system provided (11) is negative. Assume

$$c_0' : c_0 : c_1 : c_3 = 5 : 4 : 3 : 2 \qquad (12)$$

(which is not very unrealistic as this type of relation generally holds good for most of the surveys); the value of $C' - C$ for $\theta/\rho_{m'} = 0{\cdot}5$, $0{\cdot}4$, and $0{\cdot}3$, and $m = 4$, 5, and 8 are given in Table II as function of n.

From Table II it will be observed that under the condition (12) and for $m > 4$, and $\theta/\rho_{m'} = 0{\cdot}5$, expenditure in cluster sampling will always be less than that in two-stage sampling whatever the value of n may be. But if $\theta/\rho_{m'} < 0{\cdot}5$, the efficiency of cluster sampling over that of sub-sampling with respect to cost will depend on m and n. For

TABLE II

Value of $C' - C$

$\theta/\rho_{m'}$	m = 4	m = 5	m = 8
0·5	$2{\cdot}4\sqrt{n} - 2{\cdot}2\,n$	$2{\cdot}4\sqrt{n} - 2{\cdot}5\,n$	$2{\cdot}7\sqrt{n} - 3{\cdot}3\,n$
0·4	$2{\cdot}7\sqrt{n} - 1{\cdot}5\,n$	$2{\cdot}9\sqrt{n} - 1{\cdot}7\,n$	$3{\cdot}2\sqrt{n} - 2{\cdot}3\,n$
0·3	$3{\cdot}2\sqrt{n} - 0{\cdot}7\,n$	$3{\cdot}5\sqrt{n} - 0{\cdot}7\,n$	$4{\cdot}0\sqrt{n} - 0{\cdot}9\,n$

example, if $m = 8$, and $\theta/\rho_{m'} = 0{\cdot}3$, the cluster sampling will be more efficient than sub-sampling provided

$$4{\cdot}0\sqrt{n} - 0{\cdot}9n < 0$$

or

$$n \geqslant 20 \qquad \text{(approximately)}.$$

This limit on n may change if the relationship among the cost items in (12) changes.

It is worthwhile to compare the two-stage sub-sampling with two-stage cluster sampling, cluster selection at the second stage.

Using the notation of the previous section we find that for n fixed

$$V_3(\bar{y}_{n(km')}) = \frac{S_b^2}{n} + \left(\frac{1}{k} - \frac{1}{K}\right)\frac{1}{n} S_o^2, \tag{13}$$

where

S_o^2 = mean square between clusters within the fsu,
K = total number of clusters in each fsu,
k = number of clusters selected from each selected fsu in the sample, and
m' = size of the cluster.

Neglecting the finite correction factor it is easy to see that

$$S_o^2 = \frac{S_w^2}{m'}\{1 + (m' - 1)\,\rho_{m'}\},$$

and

$$V_3(\bar{y}_{n(m'k)}) = \frac{S_b^2}{n} + \frac{S_w^2}{nm'k}\{1 + (m' - 1)\,\rho_{m'}\}. \tag{14}$$

Now comparing (2) and (14) we find that

$$V_1(\bar{y}_{nm}) - V_3(\bar{y}_{n(m'k)}) = \frac{S_w^2}{nm} - \frac{S_w^2}{nm'k}\{1 + (m' - 1)\rho_{m'}\}. \quad (15)$$

Now if the two systems have the same statistical efficiency

$$\frac{S_w^2}{nm} - \frac{S_w^2}{nm'k}\{1 + (m' - 1)\rho_{m'}\} = 0,$$

or

$$\frac{km'}{1 + (m' - 1)\rho_{m'}} = m. \quad (16)$$

At first thought it may appear that the size of the cluster, m' is independent of variance components but it is not so. It depends on intra-class correlation coefficient which is not independent of variance components. The Table III gives the value of k for given value of m, m' and $\rho_{m'}$.

TABLE III

Value of k

$\rho_{m'}$	$m' = 2$			$m' = 3$			$m' = 4$		
	← m →			← m →			← m →		
	4	5	8	4	5	8	4	5	8
·1	2·20	2·75	4·40	1·60	2·00	3·20	1·30	1·62	2·60
·3	2·60	3·25	5·20	2·13	2·67	4·27	1·90	2·38	3·80
·5	3·00	3·75	6·00	2·67	3·33	5·33	2·50	3·12	5·00

Thus it is observed from Table III that for fixed number of fsu, the number of ssu's selected in form of cluster will be much larger than those necessary for attaining the same efficiency in the system when ssu's within each fsu are selected by the simple random method, if intraclass correlation coefficient is positive and high. It may also be noted that if the intraclass correlation coefficient is high, size of the cluster should be as small as possible.

Now consider the efficiency of the two procedures with respect to expenditure of the survey. For sub-sampling system we can easily adopt the cost function as given by (9). For cluster sampling the cost function may be of the form

$$C'' = c_0\sqrt{n} + c_1 n + c_2 nm'k + c_3 n\sqrt{k} + c_4 nm'kp'. \quad (17)$$

If the variance per ultimate-stage sampling unit is relatively not so important as those for the fsu and ssu, we may make

$$mnp = m'nkp'$$

by suitably choosing p'.

Subtracting (9) from (17) we obtain

$$\begin{aligned} C'' - C &= c_2 n \{m'k - m\} + c_3 n \{\sqrt{k} - \sqrt{m}\} \\ &= n \{c_2 A + c_3 B\}, \end{aligned} \tag{18}$$

where

$$A = m'k - m; \qquad B = \sqrt{k} - \sqrt{m}.$$

For all positive values of $p_{m'}$, the value of A will be positive and that of B will be negative. But the absolute relative value of A will be much larger than that of B and therefore the value of (18) will be usually positive unless c_3 is relatively much larger than c_2. In many surveys preparation of the frame, enlisting of second-stage sampling units, etc., may not take much time; in such cases c_2 may be relatively smaller than c_3.

SUMMARY

Efficiency of cluster sampling has been examined in relation to that of sub-sampling procedure. It has been shown that in many surveys where the travelling expenditure between two second-stage units is considerable it is worthwhile to go for the one-stage cluster sampling (cluster consisting of second-stage units). There appears to be not much advantage in adopting two-stage cluster sampling (selection of cluster of ssu's at second stage) over that of sub-sampling procedure. But if the average travelling expenditure between two ssu's in the sample is relatively much larger than the average cost per ssu such as the cost of selection of the ssu's, locating the sample, field identification, listing, etc., and intraclass correlation is not high, two-stage cluster sampling may be more efficient for the fixed cost of the survey.

ACKNOWLEDGEMENT

I am thankful to Dr. V. G. Panse for going through the manuscript and making useful suggestions.

REFERENCES

1. Hansen, M. H., Hurwitz, W. N. and Madow, W. G. — *Sample Survey Methods and Theory*, Vol. 1 & 2, 1953
2. Sukhatme, P. V. — *Sampling Theory of Surveys with Applications*, 1954.