ON THE FORMATION OF CLUSTERS

By

B. B. P. S. Goel and D. Singh

Institute of Agricultural Research Statistics, New Delhi

(Received : February, 1975)

Introduction

Cluster sampling is widely used in sample surveys. More often the procedure is used because the list of individual elements in the population for sampling purposes is not readily available and preparation of such a list is costly and time-consuming. However, list of such elements may be available in natural groups which we call clusters; therefore it becomes convenient and economical to use such a list as sampling units. Sometimes, even though the list of individual elements is available clusters of elements are formed and adopted as sampling units for cost and efficiency considerations. In the latter case the available procedures for formation of clusters can be divided in two categories *viz.*, (*i*) clustering before sampling (CBS) and (*ii*) clustering after sampling (CAS). The procedures mentioned by Sethi [5] and Sukhatme & Sukhatme [11] and adopted by Jessen [2] and Asthana [1] belong to the first one and those adopted by Mahalanobis [3], Panse et al [4]; Singh, Murty and Goel [6] and Singh, Rajagopalan and Maini [7] fall in the second category. The clusters formed by procedures of Sethi and Sukhatme may statistically be more or equally efficient as compared to individual elements but may not be economical or operationally convenient. The procedure of cluster formation indicated by Jessen and Asthana though likely to be economical and operationally convenient may not be efficient. The theory relating to CAS procedures has not been investigated and the various authors using these procedures analysed their data assuming that the elements of the selected clusters were a random sample from the population. In this paper the problem of formation of clusters has been discussed in detail and an attempt has been made to study the efficiency of various procedures of forming clusters.

2. The Problems Involved in Cluster Formation

Before clusters are formed the following questions will have to be considered :—

(*i*) Whether clustering before sampling (CBS) or clustering after sampling (CAS) would be used ?

(*ii*) Whether clusters would be equal or unequal ?

(*iii*) What would be the size of the clusters ?

(*iv*) How the clusters would be actually formed ?

When the population is large and the cluster size is proposed to be small grouping of elements into clusters before sampling (CBS) will involve considerable cost and efforts since this will involve identification of each element and tagging it to one and only one cluster. Alternatively clustering after sampling (CAS), in which first n elements of the (n being the sample size in terms of clusters). will be selected from the population and with each of these M-1 more elements of the population will be suitably clubbed to form n clusters of size M, will be cheap and convenient. The initial sample of n elements may be selected with equal or unequal probabilities with or without replacement. Such a procedure, however, is likely to give unknown probabilities of inclusion in the sample to different units and thus will introduce complications in estimating the population mean. If these complications can be overcome CAS would be preferable to CBS.

Clusters of unequal sizes are not to be preferred because apart from causing complications in calculations, such a scheme also introduces uncertainties in planning the cost and variance of the sample on account of the sample size being a random variable. Unless the sizes of such clusters are known in advance it may also result in unequal distribution of work load among the field workers.

The question of choice of size of clusters from the point of view of efficiency has been studied by several authors in the case of natural clusters. When the clusters have to be formed artificially there is no unique criterion available for formation of clusters. In this case, therefore, the determination of the optimum size of the clusters will depend upon the composition of clusters and therefore, on the criterion for forming clusters. The relationship between variability within clusters and the cluster size suggested by Fairfield Smith (1938), Mahalanobis [3] and Jessen [2] may not hold good for such clusters.

As is well known, a criterion for forming clusters should be such that the variability within clusters is as large as possible and

that between clusters as small as possible. A criterion for forming clusters *inter-alia* depends upon the nature of the population, the size and shape of the elements, the number of elements in the population and the type of association between elements of the population etc. The population with which we are dealing may be a file of cards or unit areas in a field or villages in a district or tehsils/talukas in a State. In a population of cards in a file the formation of clusters may be a trivial matter whereas in a population of villages or tehsils/talukas it may involve considerable difficulty.

Moreover, a good criterion of forming clusters should be objective, simple and convenient. The average travel cost between elements of a cluster should be very small as compared to the average travel cost between clusters. It is not essential that the clusters formed according to a given criterion should be non-overlapping. If a criterion determines overlapping clusters it should be considered good provided suitable estimates of the population mean or total can be obtained from a sample of such clusters.

3. Overlapping and Non-Overlapping Clusters

Suppose there is a population containing N elements ($E_1, E_2, \ldots, E_N$) and the list of these elements is available. If cluster size is M we can form at the most Nc_M distinct groups and all of them may not be really clusters in the sense that elements of a cluster should necessarily be closely located or associated so that the average travel cost between them is very small as compared to the average travel cost between clusters. Further all of them will not be non-overlapping. Every element will occur in $(N-1)c_{(M-1)}$ clusters. If we want that the clusters formed are all non-overlapping we can form only N' clusters. where N' is such that,

$$\sum_{i=1}^{N'} M_i = N$$

where M_i is the number of elements in the 'i'th cluster. If all the clusters are equal and contain M elements each then N' will be equal to $\frac{N}{M}$. Whenever $\sum_{i=1}^{N'} M_i > N$, all the N' clusters will not be non-overlapping and some of them will have one or more elements in common.

4. Probabilities of Inclusion in the Sample of Different Elements

The probability of selection/inclusion in the sample of every element of a cluster is the same as the probability of selection/inclusion in the sample of the cluster to which it belongs, provided, of course, the clusters are non-overlapping. When the clusters are overlapping this need not be so. Suppose the element E_i belongs to V_i ($V_i > 1$) clusters. Then E_i will be selected whenever any of these V_i clusters is selected. Suppose there are in all N' overlapping clusters in the population consisting of N elements. If these are selected with equal probability the element E_i will have a relative probability $\frac{V_i}{N'}$ of being selected rather than $1/N'$, which would have been the case had the N' clusters been non-overlapping. If the sample consists of 'n' clusters then the probability of inclusion of E_i in the sample will be proportional to $\frac{nV_i}{N'}$ rather than $\frac{n}{N'}$. Similarly if the probabilities of selection of N' clusters are $\alpha_1, \alpha_2, \ldots, \alpha_{N'}$ then the probability of selection of E_i at any draw will be

$$\sum_{s \ni E_i} \alpha_s = p_i \quad i = 1, 2, \ldots, N$$

If the sampling is carried out with replacement then the probability that E_i will get included in a sample of n clusters will be $1-(1-p_i)^n$. If the sampling is carried out without replacement then the probability of E_i being included in a sample of n clusters will be

$$\Pi'_i = \phi_i(\alpha_1, \alpha_2, \ldots\ldots, \alpha_{N'}, n)$$

In overlapping clusters, therefore, when the mean or total of the population is estimated on the basis of the probabilities of selection of the clusters bias enters the estimate which may not be always trivial.

5. Examples of Overlapping Clusters

(*i*) The grid sampling or method of selection of a plot of given shape and size in a field for estimation of crop yield is a well known example of overlapping clusters. A unit area (basic cell) is first located at random and then a given number of adjoining cells are combined with it according to a predetermined rule. This method is known to give higher probabilities of selection to central areas as compared to border areas, Sukhatme [11]. A plot can be regarded as a cluster of unit areas. It can easily be seen that certain basic cells will belong to only a single plot. Others will belong to two or more plots.

(*ii*) Suppose the sampling unit in a survery is a cluster of M households and the clusters are selected as follows. A list of all the households in a village is prepared in the order of their location starting from one end. Let the list be $H_1, H_2, \ldots\ldots, H_N$. A household is selected at random from this list and with this are combined $(M-1)$ more households in the increasing serial order so that we assume that we cannot go backward. If the randomly selected household is H_i then the cluster selected will be $H_i, H_{i+1}, \ldots\ldots, H_{i+m-1}$. Obviously the first $M-1$ and the last $M-1$ households will get a smaller chance of selection than the remaining $N-2M+2$ households. Thus H_1 and H_N will belong to one cluster each, H_2 and H_{N-1} will belong to two clusters each and so on. The households H_M to H_{N-M+1} will each belong to M clusters. Here it is also assumed that households in the beginning and those at the end in the list cannot together form clusters and last $(M-1)$ households or less will not be considered for formation of clusters because for such clusters the size of the cluster will be less than M. Thus the relative probabilities of selection of various households will be

$$P(H_i) = \frac{i}{N-M+1} \qquad 1 \leqslant i \leqslant M-1$$

$$= \frac{M}{N-M+1} \qquad M \leqslant i \leqslant N-M+1$$

$$= \frac{N+1-i}{N-M+1} \qquad N-M+2 \leqslant i \leqslant N$$

(*iii*) Suppose a cluster of M villages is the sampling unit, all the villages in the district being the elements of the population and the clusters are selected as follows. A village is first selected from the list of villages in a directory (District Census Handbook) with equal probability and without replacement. After locating the selected village $M-1$ more adjacent villages are combined with it to form a cluster of M villages. In this case also every village will not belong to the same number of such clusters.

In the above examples, formation of clusters has been suggested with reference to each element of the population. It is not necessary that such clusters should be always formed in this way. However, it is a convenient way and ensures that requirements of probability sampling are satisfied.

Household surveys with multiplicity, Sirken [8], [9] and Sirken and Levy [10] in which sample households report information about their own residents as well as about other persons who live elsewhere also provide examples of overlapping of clusters. However, with a

view to avoid response errors which might creep in such studies it is envisaged here that information about any one element in the sample would be obtained from itself and not from any associated element.

6. CAS—Non-repeating elements

The procedure of formation of clusters in the above examples gives rise to overlapping clusters and if more than one cluster are selected in the sample, some elements may be repeated. Thus this procedure will be similar to sampling with replacement and may therefore be less efficient. CAS may also be done as follows. After selecting one cluster the second may be formed out of the remaining $N-M$ elements and so on. In this way we can select as many clusters as we want and the clusters so selected will be mutually non-overlapping. The sample of clusters of households given in example 5(*ii*) on page 57 can also be selected using this scheme. This scheme calls for defining of clusters at each draw and is thus time consuming and onerous. Therefore this procedure is not considered in the discussions to follow. On the other hand when the number of elements in the population is large and only a few clusters (cluster size also being small) are to be selected the probability of repeating of an element in the sample of overlapping clusters will be small and there may not be much loss of efficiency.

7. CBS and CAS Systems

Now we shall study the CBS and CAS systems of clustering with reference to the population $E_1, E_2 \ldots\ldots, E_N$.

System I—CBS :—Without loss of generality the clusters may be formed as follows :

$$
\begin{aligned}
C_1 &: E_1, E_2, \ldots\ldots, E_M \\
C_2 &: E_{M+1}, E_{M+2}, \ldots\ldots E_{2M} \\
&\vdots \qquad \vdots \qquad \vdots \qquad \vdots \\
C_{N'-1} &: E_{(N'-2)M+1}, E_{(N'-2)M+2}, \ldots, E_{(N'-1)M} \\
C_{N'} &: E_{(N'-1)M+1}, E_{(N'-1)M+2}, \ldots\ldots, E_N
\end{aligned}
$$

one of the following two situations may exist

(*i*) $N = N'M$ and (*ii*) $N = (N'-1)M + M'$

In the first situation all the N' clusters will be equal, while in the other, the 'N''th cluster will have $M'(1 \leqslant M' < M)$ elements and others will have M elements.

This procedure will be simple, objective and convenient provided the starting point can be chosen objectively ; the list of elements in their natural order is available or can be prepared easily and the size of the population in relation to the size of the cluster is not very large. When these conditions are not satisfied this procedure will become cumbersome. The clusters formed according to this criterion will be non-overlapping and the average travel cost between elements within clusters will be small as compared to average travel cost between clusters. A sample of n clusters can now be selected out of the N' clusters using a probability sampling scheme.

System—II. CAS :— Here we will confine ourselves to the criteria of forming clusters with reference to each element of the population. Suppose we have selected a sample of n elements using a certain probability sampling scheme. With each of these elements we want to form clusters according to a certain criterion. In the case of CBS system we formed clusters on the basis of geographical proximity by combining next $M-1$ adjacent elements in the list. In this case the element selected at random, say E_i, is surrounded by other elements of the population on all sides and all of them will be adjacent to it. So the question arises which elements and how many of them should be combined with it to form a cluster. In other words, we are faced with the problem of choice of a criterion for forming clusters.

Distance criterion

One of the considerations in the choice of a criterion is that the average distance between elements of a cluster should be less than the average distance between clusters. If d_w and d_b denote the average distance between elements within clusters and that between clusters respectively then d_w/d_b should be small. The values of this ratio lie between 0 and 1. When the cluster size is of 1 element the value of this ratio is zero. When the cluster size is N the value of this ratio is 1. If we choose a value of this ratio, say a_o, and combine with E_i all elements E_j which satisfy the following inequality

$$d_{ij} \leqslant \delta_o (a_o) \qquad (0 < a_o < 1)$$

where d_{ij} is the distance between E_i and E_j, to form a cluster with E_i this will be an objective criterion to form clusters. According to this criterion all elements which are at a distance δ_o or less from the randomly selected element will form a cluster with it. The value of δ_o will depend upon the value of a_o chosen. Thus if we apply this criterion to all the N elements of the population we will have the following N clusters of CAS type.

$$\begin{array}{lcll} E_1 & \vdots & E_{11} & E_{12}....E_1M_1 \\ E_2 & \vdots & E_{21} & E_{22}....E_2M_2 \\ \vdots & \vdots & \vdots & \\ E_N & \vdots & E_{N1} & E_{N2}...E_{NM_N} \end{array}$$

In the above notation E_1, E_2, ..., E_N the elements of the population have been re-labelled as $E_{11}, E_{21}, E_{31}, \ldots, E_{N1}$. Also the elements E_{ij} $(i=1, ..., N; j=2, ..., M_i)$ are subgroups of elements from the original population suitably re-labelled. In an actual survey all the above N clusters need not be formed and the 'n' *CAS* clusters that will be formed will be a random sample out of the above mentioned 'N' clusters.

It may be noted that the distance criterion is simple, convenient and objective. The average travel cost between elements within clusters will be small in relation to average travel cost between clusters. However, the clusters, formed according to this criterion will be unequal in general and further they will be overlapping. Such clusters will be large or small according as the value of a_0 chosen is large or small. It can be easily verified that the clusters formed according to this criterion have a special property *i.e.* if E_i comprises M_i elements of the population then E_i is associated with M_i clusters also. In other words

$$M_i = V_i$$

for clusters formed according to this criterion.

8. Sampling with equal probability and without replacement—Estimates of the population mean and their variances

(*i*) *System I (CBS)* : We have seen that in this system of sampling the N' clusters which can be formed out of N elements are non-overlapping and therefore the procedures for obtaining the estimates of population mean and their variances from a sample of n such clusters, which may be equal or unequal, available in the standard texts on sampling can be adopted and it does not seem to be necessary to discuss them here. For all practical purposes we may assume such clusters to be equal and estimate the population mean by

$$\bar{y}_I = \frac{1}{n}\sum_{i}^{n} \bar{y}_i. \quad \text{where } \bar{y}_i. = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$$

Its variance will be given by

$$V(\bar{\bar{y}}_I) = \left(\frac{1}{n} - \frac{1}{N'}\right) S_b^2(I) \text{ where } S_b^2(I) = \frac{\sum_{i=1}^{N'} (\bar{y}_i - \bar{y}_N)^2}{N'-1}$$

(*ii*) *System II (CAS)* : It has been seen that the clusters formed according to this system will be overlapping. It was pointed in section 4 that if the population mean/total is estimated on the basis of the probabilities of selection of the clusters bias enters the estimate. Therefore, there can be two alternatives *viz.* (*i*) to study the nature of this bias and if this bias is small then to obtain the estimates and their variances using the usual procedures for non-overlapping clusters and (*ii*) to obtain the probabilities of inclusion for each of the different elements selected in the sample and then obtain the unbiased estimates and their variances corresponding to the scheme of varying/probabilities without replacement.

In the general case when the clusters are unequal the following notation may be adopted :

S. No. of cluster	*Cluster size*	*Values of elements of the clusters*	*Cluster Mean*	*Cluster Total*
1	M_1	y_{11} $y_{12} \cdots y_{1M_1}$	$\bar{y}_{1\cdot}$	Y_1
2	M_2	y_{21} $y_{22} \cdots y_{2M_2}$	$\bar{y}_{2\cdot}$	Y_2
⋮	⋮	⋮	⋮	⋮
N	M_N	y_{N1} $y_{N2} \cdots y_{NM_N}$	$\bar{y}_{N\cdot}$	Y_N

Here $y_{i1} = y_i$ $(i = 1, 2, \ldots, N)$. The following two estimates can be considered

$$(i)\ \bar{\bar{y}} = \frac{1}{n}\sum_{i}^{n} \bar{y}_{i\cdot} \quad \text{and} \quad (ii)\ \bar{\bar{y}}' = \frac{\sum_{i}^{n} M_i\, \bar{y}_{i\cdot}}{\sum_{i}^{n} M_i}$$

It may be mentioned that since M_i's are known only for the n clusters in the sample after these have been selected and

$$\bar{M}\left(= \frac{1}{N}\sum_{i=1}^{N} M_i\right)$$

is not known and therefore we cannot use the estimate

$$\frac{1}{n\bar{M}}\sum_{i}^{n} M_i\, \bar{y}_{i\cdot}$$

Similarly we cannot use Sampford's (1962) method which requires the knowledge of M_i's before selecting the sample. It can be easily seen that

$$E(\bar{\bar{y}}) = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i\cdot} = \bar{\bar{y}}_{N\cdot} \neq \bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$B(\bar{\bar{y}}) = \bar{\bar{y}}_{N\cdot} - \bar{y}_N = \frac{1}{\bar{M}}(\sigma_{vy} - \sigma_{M\bar{y}})$$

where $\sigma_{M\bar{Y}} = \dfrac{\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_{i\cdot} - \bar{\bar{y}}_{N\cdot})}{N}$, $\quad \sigma_{vy} = \dfrac{\sum_{i=1}^{N}(V_i - \bar{V})(y_i - \bar{y}_N)}{N}$

$$\text{M.S.E. } (\bar{\bar{y}}) = V(\bar{\bar{y}}) + \{B(\bar{\bar{y}})\}^2$$

where $V(\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$ and $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i\cdot} - \bar{\bar{y}}_{N\cdot})^2$

Relative bias in $\bar{\bar{y}}$ can be written as

$$\text{R.B. } (\bar{\bar{y}}) = \rho_{vy}\, C_v\, C_y - \rho_{M\bar{y}} C_M \frac{\sigma_{\bar{y}}}{\bar{y}_N} = B(I) + B(II)$$

C_v, C_M and C_y are coefficients of variation of V, M and y respectively. ρ_{vy} and $\rho_{M\bar{y}}$ are the coefficients of correlation between the pairs v, y and M, $\bar{y}$ respectively. Relative bias is the difference of the components $B(I)$ and $B(II)$. Their signs may be opposite and in that case the two will add up. If two are equal the relative bias will be zero.

In $B(I)$ there are three factors *i.e.* ρ_{vy}, Cv and C_y out of which the last one is a population constant and does not depend upon the criterion of clustering. If a criterion can ensure that each element is associated with the same number of clusters C_v and so also ρ_{vy} will be zero and thus $B(I)$ will be zero. If the criterion is such that the values of V do not vary much C_v and ρ_{vy} are expected to be very small and thus for populations with small or moderate variation in the values of y the net value of the component $B(I)$ will be very small.

Similarly if the criterion of clustering is such that all clusters are equal, C_M and $\rho_{M\bar{y}}$ will be both zeros and hence $B(II)$ will be zero. If the criterion is such that the variation in cluster sizes is very small so that C_M and $\rho_{M\bar{y}}$ are very small, the value of $B(II)$ will not be appreciable. Further if the criterion of clustering is such

that cluster means/totals are equal so that $\sigma_{\bar{y}}$ is zero even then $B(II)$ will be zero. Thus we see that the relative bias in $\bar{\bar{y}}$ will depend upon the criterion of clustering and by suitable choice of the criterion this relative bias can be made zero or very small.

Generally in practice we may be using only such criteria for which the values of the various quantities involved in the expression of relative bias will be very small, say C_v and C_M between 0.1 and 0.3., ρ_{vy} and $\rho_{M\bar{y}}$ between -0.2 to 0.2, $\sigma_{\bar{y}}/\bar{y}_N$ between 20 to 30 per cent. Within these ranges of values in populations with small or moderate variability it can be easily seen that the relative bias will not exceed 7.2 per cent and in most cases it will be less than 1 or 2 per cent in either direction.

Now we will consider the estimate $\bar{\bar{y}}'$. This, being a ratio type estimate, is evidently biased. Its bias will be given by

$$B(\bar{\bar{y}}') = E(\bar{\bar{y}}')-\bar{y}_N=\frac{\sigma_{vy}}{\bar{M}}-\frac{Cov\,(\bar{\bar{y}}',\ \bar{M}_N)}{\bar{M}} \quad \text{where } \bar{M}_N=\frac{\sum_{i}^{n} M_i}{n}$$

To the first degree of approximation, thus, the bias in $\bar{\bar{y}}'$ can be written as

$$B(\bar{\bar{y}}')=\frac{\sigma_{vy}}{\bar{M}}+\frac{N-n}{(N-1)n}\,(C_M^2-\rho_{M\bar{y}}\,C_M\,C_{\bar{M}\bar{y}})\bar{\bar{y}}_N$$

And therefore the relative bias will be given by

$$\text{R.B.}\,(\bar{\bar{y}}')=\rho_{vy}\,C_v\,C_v+\frac{N-n}{(N-1)n}\,(C_M^2-\rho_{M,M\bar{y}}\,C_M\;C_{M\bar{y}})$$
$$= B(I)+B'(II)$$

where $B(I)$ is the same as was before and $B'(II)$ is approximately equal to $(C^2_M-\rho_{M,M\bar{y}}\,C_M\,C_{M\bar{y}})/n$

For large values of n, $B(II)$ will be negligible and thus R.B. $(\bar{\bar{y}})$ will be approximately of the same order as $B(I)$. But for small values of n the component $B'(II)$ may be appreciable and thus the overall relative bias may also become appreciable. In particular component $B(II)$ or $B'(II)$ will vanish if all the clusters are of equal size and in that case the range of values of relative bias will become further narrow and CAS procedure may be used in population with small or moderate variability without any hesitation.

9. Efficiency of CAS in Relation to CBS

The absolute efficiency (or the efficiency for a fixed sample size) of cluster sampling in most natural populations is known to be low in rela-

tion to the sampling of elements. This applies to both CAS and CBS. We shall compare the efficiencies of CAS and CBS procedures in the general case when the clusters are unequal. The expected value of the sample size in CAS clusters *i.e.* $\sum_{i}^{n} M_i$ will be equal to $n\,\overline{M}$. We will consider the estimate based on the mean of the cluster means *i.e.* $\overline{\overline{y}}$ for the n clusters of CAS type. The efficiency of $\overline{\overline{y}}_1$ will be given by

$$E=\frac{\left(\frac{1}{n}-\frac{1}{N'}\right) S_b^2\ (I)}{\left[\left(\frac{1}{n}-\frac{1}{N}\right) S_b^2 +\{B(\overline{\overline{y}})\}^2\right]}$$

$$=\frac{S_b^2\ (I)}{\left[S_b^2 +n\{B(\overline{\overline{y}})\}^2\right]}$$

$S_b^2\ (I)$ and S_b^2 are expected to be of the same order and therefore when the sample size as also the bias are small the two procedures will be more or less equally efficient. In case, however, n and or $B(\overline{\overline{y}})$ is large the efficiency of CAS may be low in relation to CBS. Further in case some trend is present in the value of the variable under study in a particular direction and order of listing coincides with that direction S_b^2, may be much less than, $S_b^2\ (I)$ and CAS may be more efficient than CBS.

Cluster sampling procedure is preferred to sampling of elements in practice on account of convenience and its low cost. In other words this procedure is more efficient when the cost of the survey is fixed and the cost on travel is high. In the case of natural clusters the cost function that is usually considered is

$$C=c_1nM+c_2d$$

where d is the total distance between n clusters and the distance between elements within clusters are neglected, c_1 is the cost of enumeration per element and c_2 is the cost of travel per unit distance.

In the cluster sampling procedures where clusters are to be formed artificially one more component of cost *i.e.* cost of formation

of clusters becomes relevant. Further when the elements are spread up and distances between elements within clusters are also large the cost of travel between elements within clusters will also have to be taken into account. Therefore the cost function appropriate to CBS will be

$$C=c_1nM+c_2(d+d')+c_3N^{\prime}$$

and appropriate to CAS will be $C=c_1\sum_{i}^{n} M_i+c_2\ (d+d')+c_3n$ where l is the total distance between n clusters d' is the total distance between elements within clusters in the sample and c_3 is the cost of formation of a cluster. Suppose the total amount available for the survey is C_0 and with this amount either we can have n clusters of size $\bar{M}$ of the CBS type or we can have n^* clusters of expected size $\bar{M}$ of the CAS type. In CBS type clusters the cost component C_3N' will be much larger than the corresponding component C_3n in CAS type clusters and therefore n^* will be much larger than n. Then the cost efficiency of CAS as compared to CBS can be written as :

$$E(C_0)=\frac{n^*(N-n)\,S_b^2\,(I)}{n\left[\ (N-n^*)S_{\ b}^{\ 2}+Nn^*\{B(\bar{\bar{y}})\}^2\ \right]}$$

$$=\frac{n^*\ S_b^2\,(I)}{n\left[\ S_b^2+n^*\{\text{Bias}\ (\bar{\bar{y}})\}^2\ \right]}$$

For small values of bias $(\bar{\bar{y}})$, n and n^*, $E(C_0)$ is expected to be more than 1 and thus CAS procedure will be more efficient than CBS procedure. In case the bias is large and/or sample size is also large $E(C_0)$ may be low.

10. Estimation of Variance and Bias

For the system CBS estimate of variance can be obtained as usual. It can be easily proved that for system CAS an unbiased estimate of $V(\bar{\bar{y}})$ will be

$$\hat{V}(\bar{\bar{y}})=\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2$$

where $$s_b^2 = \frac{\sum_i^n (\bar{\bar{y}}_{i\cdot} - \bar{\bar{y}})^2}{n-1}$$

Similarly an unbiased estimate of the bias $(\bar{\bar{y}})$ will be

$$\hat{B}(\bar{\bar{y}}) = \bar{\bar{y}} - \bar{y}_n$$

where $$\bar{y}_n = \frac{1}{n}\sum_i^n y_{t_1} = \frac{1}{n}\sum_i^n y_i$$

11. A Device for Avoiding Bias

It can be seen from section 8 that even if the clusters are equal the bias in $\bar{\bar{y}}$ will not be zero. The bias in $\bar{\bar{y}}$ will be zero if in addition to M_i's, V_i's are also equal *i.e.* if each element belongs to the same number of clusters. Now we will define a criterion for CAS by which bias in $\bar{\bar{y}}$ can be avoided.

Circular listing of elements: Suppose the list of elements in the population is prepared not by starting from one end but by starting from somewhere in the middle in a circular order so that elements 1 to $M-1$ and $(N-M+1)$ to N are located in adjacent positions and are eligible to form clusters of size M. If we select a sample of n elements with equal probability and without replacement and combine with each one of these $(M-1)$ more elements in increasing or decreasing serial order we will have n clusters of M elements each of the CAS type. For these clusters it can be easily verified that size of association of every element will be equal to M. Thus we will have

$$M_1 = M_2 = \ldots = M_N = M$$

and $$V_1 = V_2 = \ldots = V_N = M$$

For this criterion of CAS $\bar{\bar{y}}$ will be an unbiased estimate of $\bar{y}_N$. This procedure may be called circular CAS. The probability of inclusion of every element in the sample of n clusters of this type, will be $\frac{nM}{N}$ which is proportional to $\frac{1}{N}$, the probability of selection of various elements before clustering. Thus, although the clusters formed by this method of CAS are overlapping these can be regarded as non-overlapping for practical purposes.

Efficiency of circular CAS : If the list of elements in a circular order is available/or can be prepared without much difficulty the cost efficiency of circular CAS will be more as compared to CBS. Otherwise the two procedures will be more or less equally efficient. The absolute efficiency of the circular CAS will, however, be of the same order as that of CBS.

12. Summary and conclusions

In some situations although the list of elements (and not of clusters) in a population is available cluster sampling is used in surveys for the sake of convenience or economy and the clusters have to be formed artificially. The available procedures for cluster formation are of two categories, (*i*) clustering before sampling (CBS) in which N' non-overlapping clusters of M elements each are first formed from N elements of the population and then a random sample of n clusters is selected and (*ii*) clustering after sampling (CAS) in which first a random sample of n elements is selected from the population and then clusters are formed with each of these n elements according to some suitable criterion and these n clusters constitute the sample. When the population is large and cluster size is small CAS will be much more cheap and convenient as compared to CBS. But clusters formed by CAS procedure are overlapping and therefore the estimate of population mean based on the mean of means of such clusters is biased. An algebraic expression has been obtained for this bias and its nature has also been investigated. In populations with small or moderate variability CAS may be used with advantage over CBS for small samples. A criterion has also been suggested for CAS by which bias can be completely avoided.

Acknowledgements

The authors are thankful to the referees for their useful suggestions.

References

[1] Asthana, R.S. (1950) : The size of sub-sampling unit in area estimation, Unpublished Thesis for Diploma, ICAR, New Delhi.

[2] Jessen, R.J. (1942) : Statistical Investigation of a sample survey for obtaining farm facts, Iowa Agricultural Research Station, *Research Bulletin*, 304.

[3] Mahalanobis, P.C. (1940) : A sample survey for acreage under Jute in Bengal *Sankhya*. 4, pp. 511-530.

[4] Panse, V.G., Singh, D. and Murty V.V.R. (1964-66) : Sample survey for estimation of milk production, Reports I.C.A.R. New Delhi.

[5] Sethi V.K., (1965] : On optimum paring of units *Sankhya* (B), 27, pp. 315-320.

[6] Singh, D., Murty V.V.R., and Goel B.B.P.S. (1970) : Monograph on estimation of milk production, I.C.A.R., New Delhi.

[7] Singh, D., Rajagopalan, M. and Maini, J.S. (1970) : Monograph on estimation of wool production, I.C.A.R,, New Delhi.

[8] Sirken, M.G. (1970) : Household Surveys with multiplicity, *Journal of American Statistical Association*, 65, pp. 257-266.

[9] „ „ (1972) : Stratified Sample Surveys with multiplicity, *Journal of American Statistical Association*, 67, pp. 224-27.

[10] Sirken, M.G. and Levey, P.S. (1974) : Multiplicity estimation of proportions based on ratios of random variables, *Journal of American Statistical Association*, 69, pp. 68-73.

[11] Sukhatme, P. V. and Sukhatme, B.V. (1976) : Sampling Theory of Surveys with Applications, Indian Society of Agricultural Statistics, New Delhi.