USE OF AUXILIARY INFORMATION IN CLUSTER FORMATION—UNEQUAL CLUSTERS

By

B.B.P.S. Goel and Savita Garg

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

(Received : November, 1978)

1. Introduction

Cluster sampling is commonly used in surveys. It is generally used when a list of elements in the population is neither available nor can be easily prepared and the use of an element as the sampling unit is, therefore, not feasible. Sometimes it is also used when the list of elements in the population is available, on account of cost and efficiency considerations. In such situations the clusters have to be formed articially. Clusters can be formed either before sampling (CBS) or after sampling (CAS) [1].

The procedures of formation of clusters in CBS system suggested in the past are mostly based on the grouping of contiguous units [2], [3]. Generally, the contiguous units are more or less similar. But from the efficiency point of view, units within clusters should not be similar but cluster means or totals should be homogeneous. Thus, a better criterion for the formation of clusters would be that which takes into account both the requirements viz. nearness of units as well as that of homogeneity of cluster means or totals. This can be done by using the sizes of units while forming clusters of neighbouring units. No work has been done so far in CBS system for forming clusters using auxiliary information. Here an attempt has been made to devise a procedure of forming clusters of nearby units using auxiliary information so as to make cluster totals almost equal. On the basis of empirical studies, the suggested procedure has been found to be more efficient than SRS, besides being simple, objective and Convenient.

An attempt to make the cluster totals more or less equal may result in combining unequal number of elements to form clusters. The clusters of unequal sizes are generally not preferred because apart from

the fact that sample size for such clusters is a random variable, the available estimates of population mean based on them are either biased or have a high variance unless cluster sizes and their means vary in such a way that their product is almost constant. The basic requirement of every sampling scheme is to obtain an unbiased estimate of the population mean or total with a low variance. For the suggested method of cluster formation the simple random sampling of clusters satisfies the requirement. When information is available on some auxiliary character, x, highly correlated with the character under study, the clusters should be formed so that :

1. The elements in a cluster should be within a prescribed distance from the key element so that the average cost of travelling from one element to another within clusters is negligible as compared to the average cost of travelling between clusters.

2. The elements combined to form a cluster should be such that the total of x values for them is more or less equal for all the clusters.

2. Method Suggested for the Formation of Clusters

Let the population under consideration consist of N distinct and identifiable elements. Let y be the character under study and 'x' the auxiliary character which is highly correlated with the character under study and the information on which is available for all elements of the population. Further, let

$$X=\sum_{i=1}^{N} x_i$$

Here the first problem is to determine the total value for the auxiliary variable for each of the clusters. This can be determined if we know the total number of clusters into which the population has to be divided, let it be λ ($\lambda \geqslant 2$ so that the variance of the estimate of population mean or total from a sample of clusters can be estimated). Thus $\frac{X}{\lambda}$ is the most desirable total value of x character for each of the clusters to be made.

Having determined the most desirable total value $\frac{X}{\lambda}$ for each cluster, the next problem is how to form the clusters for a given population of elements. Assume that a map showing the location of units in the population is available so that the units in the neighbourhood of every unit of the population are known.

Take an element from any corner of the map and continue the pooling of elements for the formation of clusters which are within a prescribed distance from the key unit till the cumulative totals are just short of (or just exceed) the number X/λ.

The procedure is continued till all the elements of the population are exhausted and no element belongs to more than one cluster. Suppose the λ clusters formed have $M_1, M_2, \ldots\ldots, M_\lambda$ elements each

$$(M_i \geqslant 1,\ i=1, 2, \ldots\ldots, \lambda \text{ and } \sum_{i=1}^{\lambda} M_i = N).$$

3. NOTATIONS

The following notations are used :

y_{ij} —the value of the character under study for the j^{th} element, $(j=1, 2, \ldots\ldots, M_i)$ in the i^{th} cluster, $(i=1, 2, \ldots\ldots, \lambda)$;

x_{ij} —the value of the auxiliary variate for the j-th element in the i-th cluster, $(j=1, 2, \ldots\ldots M_i)$, $(i=1, 2, \ldots\ldots, \lambda)$;

$\overline{M} = N/\lambda$ —the average number of elements per cluster in the population ;

$\bar{y}_{i\cdot} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$ —the mean per element in the i-th cluster ;

$\bar{\bar{y}}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{\lambda} \sum_{j=1}^{M_i} y_{ij} = \frac{1}{N} \sum_{i=1}^{\lambda} M_i \bar{y}_{i\cdot}$ —the mean per element in the population ;

$\bar{\bar{y}}_{n\cdot} = \frac{1}{n} \sum_{i}^{n} \bar{y}_{i\cdot}$ —the mean of cluster means in a random sample of n clusters,

$\bar{\bar{y}}_{\lambda\cdot} = \frac{1}{\lambda} \sum_{i=1}^{\lambda} \bar{y}_{i\cdot}$ —mean of cluster means in the population ;

$$S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})^2$$ —the mean square between elements in the ith cluster, $i = 1, 2, \ldots\ldots, \lambda$;

$$\bar{S}_w^2 = \frac{1}{\lambda} \sum_{i=1}^{\lambda} S_i^2$$ —the mean square within clusters ;

$$S_b^2 = \frac{1}{\lambda - 1} \sum_{i=1}^{\lambda} (\bar{y}_{i.} - \bar{y}_{\lambda.})^2$$ —the mean square between cluster means in the population and

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^{\lambda} \sum_{j=1}^{M_i} (y_{ij} - \bar{\bar{y}}_{..})^2$$ —the mean square between elements in the population.

4. Estimators of the Population Mean, their Variances and Estimates of the Variances

The following estimates of the population mean for a random sample of n unequal clusters are available in literature [3].

(*i*) $$\bar{\bar{y}}_{(1)} = \frac{1}{n\bar{M}} \sum_{i}^{n} M_i \bar{y}_{i.}$$

$$= \frac{1}{n} \sum_{i}^{n} \frac{M_i}{\bar{M}} \bar{y}_{i.}$$

(*ii*) $$\bar{\bar{y}}_{(2)} = \frac{1}{n} \sum_{i}^{n} \bar{y}_{i.}$$

(*iii*) $$\bar{\bar{y}}_{(3)} = \frac{\sum_{i}^{n} M_i \bar{y}_{i.}}{\sum_{i}^{n} M_i}$$

(*iv*) $$\bar{\bar{y}}_{(4)} = \bar{\bar{y}}_{(2)} + \frac{\lambda - 1}{M} S_{M\bar{y}}$$

where $S_{M\bar{y}} = \Sigma (M_i - \bar{M}) \bar{y}_{i.}$

(v) $$\bar{\bar{y}}_{(5)}=\sum_{i}^{n}\bar{z}_{i.} \quad \text{where } \bar{z}_{i.}=\frac{M_i\,\bar{y}_{i.}}{NP_i},$$

$$P_i=\frac{X_i}{X},\ X_i=\sum_{i=1}^{M_i}x_{ij}$$

However, since the method of formation of clusters is based on the equalization of cluster totals the estimators $\bar{\bar{y}}_{(2)}$, $\bar{\bar{y}}_{(3)}$ and $\bar{\bar{y}}_{(4)}$ are bound to be inefficient and need not be considered further.

The variances and estimates of variances for the estimators $\bar{\bar{y}}_{(1)}$ and $\bar{\bar{y}}_{(5)}$ are given by

$$V(\bar{\bar{y}}_{(1)})=\frac{\lambda-n}{\lambda}\,\frac{S^2_{b(1)}}{n}$$

where $$S^2_{b(1)}=\frac{1}{\lambda-1}\sum_{i=1}^{\lambda}\left(\frac{M_i\bar{y}_i}{\bar{M}}-\bar{\bar{y}}..\right)^2$$

and $$\text{Est. } V(\bar{\bar{y}}_{(1)})=\frac{\lambda-n}{\lambda}\,\frac{s^2_{b(1)}}{n}$$

where $$s^2_{b(1)}=\frac{1}{n-1}\sum_{i}^{n}\left(\frac{M_i\bar{y}_{i.}}{\bar{M}}-\bar{\bar{y}}_{(1)}\right)^2$$

$V(\bar{\bar{y}}_{(5)})=\sigma^2_{bz}/n$ where

$$\sigma^2_{bz}=\sum_{i}^{n}P_i\,(\bar{z}_{i.}-\bar{\bar{y}}..)^2 \text{ and}$$

Est- $V(\bar{\bar{y}}_{(5)})=s^2_{bz}/n$ where

$$s^2_{bz}=\frac{1}{n-1}\times\sum^{n}(\bar{z}_{i.}-\bar{\bar{y}}_{(5)})^2$$

The estimator $\bar{\bar{y}}_{(1)}$ is unbiased and will have a very low variance for the clusters formed according to the suggested method. For pps with replacement sampling, here the probabilities are taken proportional to cluster totals which are made more or less equal and thus gives more or less equal probabilities. Therefore, unbiased estimator $\bar{\bar{y}}_{(5)}$ which also depends upon cluster totals is likely to have variance of the same order as $\bar{\bar{y}}_{(1)}$.

5. EFFICIENCY

Efficiency of cluster sampling when clusters are of unequal sizes for fixed sample size is

$E=S^2/\bar{M}\,S^2_{b(i)}$ where $S^2_{b(i)}$ is the mean square between cluster means for the i-th estimator, $(i=1, 5)$

It is known that less cost is involved in cluster sampling than simple random sampling. Suppose the cost structure is such that it enables us to select either n clusters of average size $\bar{M}$ for a given cost C_0 or $n\,\bar{M}$ elements for the same cost. Then the cost efficiency will be given by

$$E_{co}=\frac{n}{n'}\,E$$

Which will be considerably more than E since n' will be much smaller than n.

6. EMPIRICAL INVESTIGATION ON THE SUGGESTED METHOD

The suggested method was used to form clusters with the help of a natural population consisting of 122 villages in Nokha Tehsil of Bikaner district of Rajasthan. The information for each village in respect of geographical area (auxiliary character) and fallow land (character under study) was taken from the District Census Hand Book (1971 population Census). The distribution for both the characters was found to be positively skewed and platykurtic and the correlation coefficient between them was 0.8696.

As in the method of formation of clusters discussed, λ can take any value greater than or equal to 2. So for different values of λ, the average cluster total for the auxiliary character will be different.

Here by taking λ equal to 24, 31 and 36 different average total values of clusters were determined. The average cluster size $\bar{M}$ for different values of λ was 5.1, 3.9 and 3.3 respectively. For these three values of $\bar{M}$ the absolute efficiencies and cost efficiencies of the various estimators were obtained and are given in Table 1. The coefficient of variation of $\bar{M}$ was around 50 per cent for the above values of λ. For different values of λ (24, 31, 36) the cluster sizes vary from 1 to 11, 1 to 10 and 1 to 7 respectively.

From the results, it is seen that for the clusters so formed the unbiased estimate of the population mean based on unequal clusters is even more efficient than the estimate based on simple random sampling of elements or on clusters of equal size (without using

TABLE 1

Absolute efficiencies and cost efficiencies of the various estimates of the population mean for Average Cluster sizes $\overline{M}=5.1$, $\overline{M}=3.9$, $\overline{M}=3.3$

S. No.	*Sampling Method*	*Estimator*	*Cluster size* $\overline{M}$	*Absolute efficiency* %	*Cost efficiency*						
					$\frac{n}{n'}=1.2$	$\frac{n}{n'}=1.4$	$\frac{n}{n'}=1.6$	$\frac{n}{n'}=1.8$	$\frac{n}{n'}=2.0$	$\frac{n}{n'}=2.2.$	$\frac{n}{n'}=2.4$
1.	Simple random sampling	$\bar{y}_{SRS}$	—	100	—	—	100	—	—	—	—
2.	Cluster sampling with suggested method	$\bar{y}_{(1)}$	5.1	162.2	194.6	227.1	259.5	291.9	324.4	356.8	389.3
			3.9	223.1	267.7	312.3	356.9	401.5	446.2	490.8	536.4
			3.3	245.4	294.5	343.6	392.6	441.7	490.8	539.8	588.9
3.	pps sampling of clusters with replacement (Prob. proportional to cluster total)	$\bar{y}_{(5)}$	5.1	166.4	199.6	232.9	266.2	299.5	332.8	366 1	399.3
			3.9	230.4	276.7	322.8	368.9	415.1	461.2	507.3	553.4
			3.3	251.3	301.5	351.8	402.11	452.3	502.6	552.8	603.1
4.	Cluster sampling without using auxiliary information	$\bar{y}_{n.}$	5.1	62.9	75.4	88.1	100.6	113.2	125.8	138.4	150.9
			3.9	68.3	81.9	95.6	109.3	122.9	136.6	150.2	163.9
			3.3	79.9	95.8	111.8	127.8	143.8	159.8	175.7	191.7
5.	pps sampling of elements with replacement.	$\bar{y}$(ppswr)	—	—	—	—	536	—	—	—	—

auxiliary information). For such clusters the efficiency of the estimate of population mean further increases when compared for a fixed cost and in certain situations it may be even more than the efficiency of the usual estimate based on pps with replacement sample of elements.

SUMMARY

In this paper a method of forming clusters before sampling (CBS) has been suggested which takes into account the twin requirements of nearness of the units within a cluster as also the homogeneity of cluster totals by using some auxiliary information. The method will, in general, result in unequal number of units in different clusters. But for the clusters so formed the unbiased estimate of the population mean is very efficient which is not so when we consider natural clusters or form clusters without using auxiliary information. The method has been illustrated with the help of an actual example.

ACKNOWLEDGEMENT

The authors are thankful to the referees for their valuable comments in improving the paper.

REFERENCES

[1] B.B.P.S. Goel and D. Singh, (1977) : 'On the Formation of Clusters' Jour Ind. Soc. Agri. Stat., Vol. 29. No. 1, pp. 53-68.

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

[3] P.V. Sukhatme and B.V. Sukhatme, (1970) : 'Sampling Theory of Survey with Application'. Ind. Soc. Agri. Stat. New Delhi-12.