SEQUENTIAL SAMPLING OF NON-OVER LAPPING CLUSTERS—CLUSTERING AFTER SELECTION

By

G. C. Chawla and M. Rajagopalan*

Indian Agricultural Statistics Research Institute, New Delhi

(Received: January, 1978)

SUMMARY

In this paper, a method of selecting one by one n non-over-lapping clusters of two units each by selecting one unit at random (key unit) from the population and another randomly from the units falling within a distance d from the key unit, selecting another unit randomly from the N-2 units and forming another cluster in the same way and so on has been suggested. An unbiased procedure for estimating the population total has also been suggested, with illustration.

1. INTRODUCTION

Cluster sampling reduces travel cost when clusters are formed by combining near by units, but efficiency is reduced. Thus, a guiding principle behind cluster formation would be to strike a balance between the cost and efficiency.

If maps showing location of all units in the population are available, non-over-lapping clusters may be made before selecting a sample of clusters. But in many situations when this is not so, clusters are made after selecting some key units at random. This procedure is quite convenient from practical stand point and has been actually adopted in several surveys, but it creates difficulties at the estimation stage. It generates over lapping clusters and thus assigns unequal probabilities of selection to different units. Attempts to make non-over-lapping clusters in such procedures lead to complications in estimation. Some studies in cluster sampling for formation of clusters after selection (CAS) have been made by Goel (1973).

*Present addresss : National Academy of Agricultural Research Management Rajendranagar, Hyderabad-30

In this paper, a method of forming clusters of two units and their selection in such a way that the units in a cluster do not fall beyond a pre-assigned distance 'd' and the selection procedure ensures sampling without replacement, is sugested.

2. Cluster Formation and Method of Selection

Let the population consist of N distinct and identifiable units (U_i), $(i=1, 2, \ldots, N)$. A sample of n clusters of size 2, each, is to be selected. The procedure of selecting a cluster consists of the following steps.

Step 1: Select one unit at random from the population and call it a key unit.

Step 2: Select one unit at random from the units falling within a distance 'd' from the key unit. Attach the selected unit to the key unit selected in step 1 to form a cluster.

In between these two steps it is required that all the units falling within a distance 'd' from the key unit should be enumerated. This operation may be conveniently performed in the field. The second cluster is selected after removing the selected cluster from the population following the procedure described in steps 1 and 2. This will, perhaps create no difficulty in the field if the two units falling in the first cluster are known before the formation of second cluster. In fact if the distance between key units is more than $2d$ the problem of removing the two units of the first cluster would not arise at all. Procceding in a similar way, all the n clusters may be selected.

3. Estimation Procedure

Let $M_i^{(r)}$ be the number of units falling within the distance 'd' from U_i at the rth draw presuming $M_i^{(r)} \geqslant 1$ for all $i=1, 2, \ldots, N$ and $r=1, 2, 3, \ldots, n$. The probability of selecting U_i at the first draw as a key unit and U_i' (U_i' being one of the $M_i^{(1)}$ units) as an associate in the cluster is $\frac{1}{NM_i^{(1)}}$

Let this probability be denoted by P_i. It is evident that P_i is same for all i' such that the distance between U_i and U_i' is not more than 'd'. Thus with ith unit $(i=1, 2, \ldots, N)$ in the population a P_i is associated which denotes the probability of getting a specified cluster with U_i as the key unit in the first draw. At the first draw the probability of a cluster of U_i and U_i' is therefore P_i+P_i'. Let there be L_1 possible clusters denoted by $C_j^{(1)}$, $(j=1, 2, \ldots, L_1)$.

Let $P_j^{(1)}$ be the probability of selecting jth cluster at the first draw. Evidently

$$P_j^{(1)} = \sum_{i \epsilon C_j^{(1)}} P_i$$

It may be seen that

$$L_1 = \sum_{i=1}^{N} M_i^{(1)}$$

Let $\pi_i^{(1)}$, $(i=1, 2, \ldots, N)$ be the inclusion probability of ith unit in the cluster selected at the first draw. Clearly

$$\pi_i^{(1)} = \sum_{j} P_j^{(1)} \qquad \ldots(2.1)$$

The summation is taken over the clusters in which ith unit occurs. Let us denote the cluster selected at the rth draw by $C^{(r)}$. We have to find the estimater of the population total

$$Y = \sum_{i}^{N} y_i;$$

where y_i' s are the values of the character under study for N units of the population.

At the first draw, we define

$$t_1 = \sum_{i \epsilon C_{L_1}^{(1)}} \frac{y_i}{\pi_i^{(1)}}$$

$$E(t_1) = E \sum_{i \epsilon C_{(1)}} \left(\frac{y_i}{\pi_i^{(1)}} \right) = \sum_{j=1} \left(\sum_{i \epsilon C_j^{(1)}} \frac{y_i}{\pi_i^{(1)}} \right) P_j^{(1)}$$

$$= \sum_{i \epsilon C_1^{(1)}} \frac{y_i}{\pi_i^{(1)}} \cdot P_1^{(1)} + \left(\sum_{i \epsilon C_2^{(1)}} \frac{y_i}{\pi_i^{(1)}} \right) P_2^{(1)} + \ldots + \left(\sum_{i \epsilon C_{L_1}^{(1)}} \frac{y_i}{\pi_i^{(1)}} \right) P_{L_1}^{(1)}$$

$$=\frac{y_i}{\pi_1^{(1)}}\sum_{1} P_j^{(1)}+\frac{y_2}{\pi_2^{(1)}}\sum_{2} P_j^{(1)}+\ldots+\frac{y_N}{\pi_N^{(1)}}\sum P_j^{(1)}$$

where $\sum_i$ denotes summation over those clusters which contain ith unit ($i=1, 2,\ldots\ldots,N$)

$$=\frac{y_1}{\pi_1^{(1)}}\cdot\pi_1^{(1)}+\frac{y_2}{\pi_2^{(1)}}\cdot\pi_2^{(1)}+\ldots+\frac{y_N}{\pi_N^{(1)}}\cdot\pi_N^{(1)} \text{ from (2.1)}$$

$$=y_1+y_2+\ldots+y_N=Y.$$

Now remove $C^{(1)}$ cluster from the population and we are left with $(N-2)$ units. Again we define clusters of size two for the remaining $N-2$ units and probabilities and inclusion probabilities on the above lines.

Let the clusters formed at the second draw be

$C_1^{(2)}, C_2^{(2)}, \ldots, C_{L_2}^{(2)}$ with probabilities

$P_1^{(2)}, P_2^{(2)}, \ldots, P_{L_2}^{(2)}$, where $L_2=L_1-1$.

Inclusion probabilities, at the second draw of remaining $(N-2)$ units in the population are denoted by

$$\pi_1^{(2)}, \pi_2^{(2)}, \ldots, \pi_{(N-2)}^{(2)} \text{ where } \pi_i^{(2)}=\sum_j P_j^{(2)}$$

The summation is taken over the clusters in which ith unit occurs.

Select a cluster from $C_1^{(2)}, C_2^{(2)}, \ldots, C_{L_2}^{(2)}$ with the same procedure as described in selection of the first cluster. Denote the selected cluster by $C^{(2)}$.

We define t_2 at the second draw.

$$t_2=\sum_{i\epsilon C^{(1)}} y_i+\sum_{i\epsilon C^{(2)}}\frac{y_i}{\pi_i^{(2)}}$$

We build up an estimator similar to that of Des Raj's ordered estimator. It is evident that probability of selection of second cluster depends upon the previous draw.

$$E(t_2)=E_1\ E_2\ (t_2)$$

Where E_2 is the conditional expectation for given first draw and E_1 is the expectation over the random selection of the first draw

$$\text{Now } E_2(t_2) = \sum_{i \epsilon C(1)} y_i + E_2\left(\sum_{i \epsilon C(2)} \frac{y_i}{\pi_i^{(2)}}\right)$$

$$= \sum_{i \epsilon C(2)} y_i + Y - \sum_{i \epsilon C(1)} y_i = Y.$$

Hence a combined estimator $\hat{Y}$ based on both the draws may be considered as

$$\hat{Y} = \frac{1}{2}(t_1 + t_2)$$

Which is an unbiased estimator of the population total Y. Proceeding in the same way, at the rth draw, estimator of Y is given by

$$t_r = \sum_{j=1}^{r-1}\left(\sum_{i \epsilon C(j)} y_i\right) + \sum_{i \epsilon C(r)} \frac{y_i}{\pi_i^{(r)}}$$

Then it can be seen easily that $E(t_r) = Y$

Thus, t_r is an unbiased estimator of the population total. If n clusters are selected as above, it is clear that

$$\bar{t} = \frac{1}{n}\sum_{j=1}^{n} t_j$$

is an unbiased estimater of Y. Also as in the case of Des Raj's estimator t_r and t_k $(r \neq k \neq n)$ may be easily shown to be uncorrelated and an unbiased estimater of the variance of $\bar{t}$ is given by

$$Est.\ V(\bar{t}) = \frac{1}{n(n-1)}\sum_{j}^{n} (t - \bar{t})^2$$

Illustration :

The methodology developed above is illustrated with the help of data on cattle population from Velvadam Firka (Vijayawada Taluk), Krishna Delta Area, Andhra Pradesh, which consist of 17 villages. It is proposed to select a sample of 2 clusters of 2 villages

TABLE 1

(*a*) Villages with code numbers	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Cattle population	85	1184	299	51	690	2561	506	1070	85	417	320	138	000	000	506	159	1166
(*b*) Associated villages within a distance of 4 knıs (Codes of villages)	19 20 25 26 27 33 34	18 20 21 22 25 26	18 19 21 22 25 2ɔ 27	19 20 22 25	19 20 21 25	24	23 25	18 19 20 21 22 24 26 27	18 19 20 25 27 28 33 34	18 20 25 2ɔ 28 29 31 32 33 34	26 27 29 30 31 32 33	27 28 30 31 32 33 34	28 29 31 32 33	27 28 29 30 32 33	27 28 29 30 31 33	18 26 27 28 29 30 31 32 34	18 26 27 29 33
Total Number of Associates	(7)	(6)	(7)	(4)	(4)	(1)	(2)	(8)	(8)	(10)	(7)	(7)	(5)	(6)	(6)	(9)	(5)
(*c*) (P_i)	.0084	.0098	.0084	.0147	.0147	.0008	.0292	.0074	.0074	.0059	.0084	.0084	.0118	.0098	.0098	.0065	.0118
(*d*) Clusters of two villages with code Nos.	18,19 : 18,34	19,18 : 19,26	20,18 : 20,27	21,19 : 21,25	22,19 : 22,25	23,24 :	24,23 :	25,18 : 24,25	26,18 : 25,27	27,18 : 26,34	28,26 : 27,34	29,27 : 28,33	30,28 : 29,34	31,27 : 30,33	32,27 : 31,33	33,18 : 32,33	34,18 : 33,34
(*e*) ($P_i^{(1)}$)	.0182 : .0202	.0182 : .0172	.0168 : .0143	.0245 : .0221	.0245 : .0221	.0882 :	.0882 : .0368	.0158 : .0133	.0158 : .0192	.0143 : .0177	.0158 : .0149	.0143 : .0202	.0202 : .0183	.0157 : .0163	.0157 : .0163	.0149 : .0183	.0202 : .0183
(*f*) ($\pi_i^{(1)}$)	.1160	.1198	.1271	.0991	.0991	.0882	.1250	.1579	.1258	.1453	.1184	.1228	.1016	.1096	.1096	.1402	.095ɔ

Total Cattle Population=9237

each such that the distance between the villages in a cluster is not more than 4 kilometres. The information regarding codes of villages and cattle population with P_i probabilities etc. are given in the Table. Since there are only 17 villages the probability of selecting village of code 18 is 1/17. Its associates within a distance of four kilometres are seven in number and are given in (*b*). Since the probability of selecting any one associate is 1/7, the probability of forming a cluster of two villages with village bearing code 18 as key village is given by 1/17. 1/7=0.0084, as given in (*c*). All the clusters formed are given in (*d*). The first cluster is associated of code numbers 18 and 19. Therefore,

$$P_1^{(1)} = \sum_{i}^{2} P_i = 0.0084 + 0.0098 = 0.0182 \text{ (all these are given in [}e\text{])}.$$

$$\text{Inclusion probability} = \sum_{i}^{6} P_i^{(1)} = 0.0182 + \ldots + 0.0202$$

$$= 0.1160 \text{ (given in [}f\text{])}.$$

Select a cluster say (20, 27) at random. Therefore,

$$t_1 = \frac{y_{20}}{\pi_{20}^{(1)}} + \frac{y_{27}}{\pi_{27}^{(1)}} = \frac{299}{0.1271} + \frac{417}{0.1453} = 5222$$

Remove this cluster and again form a new set of clusters with a new list of probabilities for the remaining ($N-2=15$) villagers. These can be formed similar to that given in the Table. At the second draw, select a cluster say (24, 25), at random.

Now,

$$t_2 = (y_{20} + y_{27}) + \frac{y'_{24}}{\pi'^{(2)}_{24}} + \frac{y'_{25}}{\pi'^{(2)}_{25}}$$

$$= 292 + 417 + \frac{506}{.1111} + \frac{1070}{.1820} = 11149$$

Hence

$$\hat{Y} = \frac{1}{2}(5222 + 11149) = 8186$$

$$Est.\,[V(\hat{Y})]=\frac{1}{n(n-1)}\sum_{r=1}^{n}(t_r-\hat{Y})^2$$

$$=\frac{1}{2\,(2-1)}\,(t_1-\hat{Y})^2+(t_2-\hat{Y})^2$$

$$=8782333$$

(*i*) Estimate of the population total=8186

(*ii*) Estimate of the variance of the estimate=8782333.

ACKNOWLEDGEMENT

Authors are grateful to the referees for their useful comments.

REFERENCES

[1] Singh, D. (1956), : 'On efficiency of cluster sampling' *Jour. of Ind. Soc. of Agril. Stat.* 8.

[2] Des Raj (1956) : 'Some estimators in sampling with varying probabilities without replacement. *Jour, of Amer. Stat. Assoc.* 51.

[3] Goel, B.B.P.S. (1973) : 'efficiency of certain systems of cluster sampling and its application.' Ph.D. thesis submitted to IARI, New Delhi.

[4] Sukhatme, P.V. and Sukhatme, B.V. (1976) : 'Sampling theory of surveys with applications'. *Ind. Soc. Agri. Stat.* New Delhi.