Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 3 (1986), pp. 352-362

OPTIMUM ALLOCATION FOR CLUSTER SAMPLING ON TWO OCCASIONS

R. R. CHANDAK and O. P. KATHURIA
IASRI, New Delhi
(Received : May, 1980)

SUMMARY

An estimator of population mean and its variance have been obtained for sampling on two occasions when a fixed proportion of clusters of units drawn on the first occasion is retained on the second occasion. A cost function for the sampling design has been considered and the problem of optimum allocation of sample clusters between matched and unmatched samples has been studied for varying sample sizes on each occasion. The efficiency of matching of clusters of units has been examined in relation to the matching of an equivalent simple random sample of units and the results illustrated with the data of area estimation enquiry on rice in Assam state.

Keywords : Optimum allocation, cluster sampling, Successive occasions.

Introduction

Patterson [2] considered the problem of sampling on successive occasions and obtained efficient estimator of population mean on the h-th occasion. Kulldorff [1] examined the problem of optimum allocation of units on the second occasion for a matching scheme in simple random sampling. In this paper we examine the problem of optimum allocation of clusters of units by taking an appropriate cost function and examine its efficiency with respect to simple random sampling (SRS).

Estimate of Mean and Its Variance

Suppose that the population consists of a finite number of N clusters each of M units. On the first occasion a simple random sample of n

clusters is drawn without replacement and the value of the character under study, say x, is observed. On the second occasion we draw two random samples as follows :

(i) a simple random sample without replacement of n_1 clusters out of n clusters drawn on the first occasion.
(ii) a simple random sample without replacement of n_2 clusters afresh from the remaining $(N - n)$ clusters in the population.

Observe the value of the *character* under the study for each unit in the sample of $(n_1 + n_2)$ clusters. Further $(n_1 + n_2)$ need not be equal to n. The best weighted estimator of the population mean $\bar{Y}$ on the second occasion may be written as

$$\bar{y} = \frac{n'[\bar{y}_1 + (\rho S_{by}/S_{bx})(\bar{x} - \bar{x}_1)] + n_2 . \bar{y}_{f2}}{n' + n_2}$$

where

$$\frac{1}{n'} = \frac{\rho^2}{n} + \frac{1 - \rho^2}{n_1}$$

$\bar{x}_1$ and $\bar{y}_1$ are the means of n_1 clusters on first and second occasion respectively which are common on both the occasions and $\bar{x}$ and $\bar{y}_{f2}$ are the means based on n and n_2 clusters on first and second occasions respectively.

ρ is correlation coefficient between cluster means on first and second occasion, while S_{bx}^2 and S_{by}^2 are the mean squares between cluster means in the population on first and second occasion respectively.

The variance of the estimate $\bar{y}$ is given by

$$V(\bar{y}) = [n' + n_2)^{-1} - N^{-1}]\, S_{by}^2 \tag{2}$$

If ρ_c be the intra-class correlation coefficient defined as

$$\rho_c = E(Y_{ij} - \bar{Y})\,(Y_{ik} - \bar{Y})/E(Y_{ij} - \bar{Y})^2$$

and

$$S_y^2 = (NM - 1)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y})^2$$

then following Sukhatme and Sukhatme (1970), we can write

$$S_{by}^2 = M^{-2}(N - 1)^{-1}\,(NM - 1)\, S_y^2\,[1 + (M - 1)\,\rho_c]$$

Minimum Variance Allocations of n_1 and n_2 for Fixed Costs

We consider the following cost function for sampling on the second occasion.

$$C = c_0 + c_1 n_2 + c_2(n_1 + n_2)\, M \tag{3}$$

where

C = total cost of the survey for the second occasion; c_0 = overhead cost; c_1 = Cost per cluster on enumeration (including travel costs) and preparation of frame; and c_2 = Cost per element of enumeration and data collection from ultimate sampling units within clusters.

Denoting $(C - c_0)/c_2 = R_1$ and $c_1/c_2 = R_2$, the above cost function may be written as

$$R_1 = R_2 n_2 + (n_1 + n_2)\, M \tag{4}$$

Minimizing (2) for a given (4), the optimum values of n_1 and n_2 may be obtained as

$$n_1 = \frac{n}{\rho^2}\left[\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right]$$

$$n_2 = \eta^{-1}\,[R_1 M^{-1} - n_1] \tag{5}$$

where,

$$\eta = (R_2 + M)/M$$

To find the optimum sample sizes (under different conditions) we shall disregard the fact that the sample sizes must be integers. Also, the values so obtained must be under the following restrictions :

(i) $n_1 \geqslant 0$ (ii) $n_2 \geqslant 0$ i.e. $n_1 \leqslant R_1/M$

If $\rho \neq 0$, we get three distinct cases depending upon the values of various cost components and consequently on R_1 and R_2 as given in Table 1.

Minimum Cost Allocation of n_1 and n_2 for Fixed Variance

Let $Q = [n_1^{-1} + \rho^2(n^{-1} - n_1^{-1})]^{-1} + n_2 > 0$

be a constant quantity and hence the variance $V(\bar{y}) = (Q^{-1} - N^{-1})\, S_{by}^2$ is fixed. If we minimise the total cost given by (4) subject to (2), then it can be easily verified that the optimum solution for n_1 remains the same as given by equation (5) and n_2 is given by the relation

$$n_2 = Q - n\,\rho^{-2}\,[1 - \sqrt{\eta^{-1}(1-\rho^2)}] \tag{6}$$

TABLE 1

Case	Condition	Optimum value of n_1	Optimum value of n_2	$V(y) = (Q^{-1} - N^{-1})\, S_{by}^2$ where Q is given by
I	(i) $\eta \geqslant (1-\rho^2)^{-1}$ (ii) $R_1 \geqslant n\eta$	n	$\eta^{-1}[R_1 M^{-1} - n]$	$n^{-1}[R_1 M^{-1} - n]$
II	(i) $\delta < \eta < (1-\rho^2)^{-1}$ (ii) $R_1 \geqslant \eta \frac{n}{\rho^2}\left[\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right]$ where, $\delta = (M + R_2\ (\text{min.}))/M$	$\frac{n}{\rho^2}\left[\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right]$	$\eta^{-1}\left[\frac{R_1}{M} - \frac{n}{\rho^2}\left\{\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right\}\right]$	$\frac{n}{\rho^2}\left(1 - \sqrt{\frac{1-\rho^2}{\eta}}\right)^2 + \frac{R_1}{M\eta}$
III	(i) $R_1 \leqslant n\eta$ (ii) $R_1 \leqslant \eta \frac{n}{\rho^2}\left[\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right]$	$R_1\eta^{-1}$	$\frac{R_1}{\eta}\left[\frac{1}{M} - \frac{1}{\eta}\right]$ if $M < \eta$ 0 if $M \geqslant \eta$	$\frac{\left[1 + \frac{R}{\eta}\left(\frac{1}{M} - \frac{1}{\eta}\right)\left(\frac{\rho^2}{n} + \frac{1-\rho^2}{R_1}\right)\right]\eta}{\left[\frac{\rho^2}{n} + \frac{1-\rho^2}{R_1}\eta\right]}$ if $M < \eta$ $\left(\frac{\rho^2}{n} + \frac{1-\rho^2}{R_1}\eta\right)^{-1}$ if $M \geqslant \eta$

δ should be closed to 1, but it cannot be equal to 1 since c_1 will not be zero.

As the sample size n_1c annot be greater than n and $n_2 \geqslant 0$, we again get three distinct cases providing optimum solution for n_1 and n_2 as given in Table 2.

TABLE 2—OPTIMUM ALLOCATION OF n_1 AND n_2 WHICH MINIMISES COST FUNCTION (4) FOR A FIXED VARIANCE (2)

Case	*Condition*	n_1	n_2
I			
	(i) $\eta \geqslant (1-\rho^2)^{-1}$	n	$Q - n$
	(ii) $Q > n$		
II			
	(i) $\delta \leqslant \eta < (1-\rho^2)^{-1}$	$\frac{n}{\rho^2}\left[\sqrt{\eta(1-\rho^2)} - (1-\rho^2)\right]$	$Q - \frac{n}{\rho^2}\left[1 - \sqrt{\frac{1-\rho^2}{\eta}}\right]$
	(ii) $Q \geqslant \frac{n}{\rho^2}\left[1 - \sqrt{\frac{1-\rho^2}{\eta}}\right]$		
III			
	(i) $Q < n$	$(1-\rho^2)\left(\frac{1}{Q} - \rho^2/n\right)^{-1}$	0
	(ii) $Q < \frac{n}{\rho^2}\left[1 - \sqrt{\frac{1-\rho^2}{\eta}}\right]$		

Numerical Illustration

We obtain optimum values of n_1 and n_2 and relative efficiency of cluster sampling for varying sizes as compared to an equivalent simple random sample of $(n_1 + n_2)\,M$ units following Kulldorff's scheme for an area estimation survey on high yielding varieties of rice crop conducted during 1976-77 and 1977-78 by I.A.S.R.I. in Sibsagar district of Assam state. The sample sizes during the 2 years consisted of 300 and 228 cultivators respectively of which 108 cultivators constituted 'matched' units. For purpose of this study neighbouring cultivators were combined to form clusters of sizes 2, 3 and 4 respectively, both for matched as well as unmatched units. The relative efficiency of matching of clusters w.r.t. matching of an equivalent simple random sample for different cluster sizes is shown in Table 3, the character studied being the cultivated area under winter rice.

TABLE 3

M	$\hat{S}_b^2$	$\hat{S}_w^2$	$\hat{\rho}_c$	ρ	*Relative efficiency*
2	47460	56705	0.248	0.764	0.75
3	37421	57701	0.235	0.867	0.71
4	34194	55736	0.261	0.885	0.60

$\hat{S}^2 = 75728$ and $\hat{\rho}_1 = 0.826$, $\hat{S}_b^2$, $\hat{S}_w^2$ and $\hat{S}^2$ have their usual meanings. $\hat{\rho}_c$, $\hat{\rho}$ are estimates of ρ_c, ρ defined in section 2. $\hat{\rho}_1$ is estimate of correlation on per unit basis between the sampling units on first and second occasions.

Consider now a matching scheme in *SRS* without replacement for an equivalent sample of nM units on the first occasion. On the second occasion n_1M units are selected with *SRS* without replacement from nM units on the first occasion and n_2M units are selected afresh from the remaining $(N - n)M$ population units again with *SRS* without replacement. If the estimator based on these $(n_1 + n_2)M$ units be denoted by $\bar{Y}_{SRS}$, it may be verified that $V(\bar{y}_{SRS})$ will be

$$V(\bar{y}_{SRS}) = M^{-1}(\phi^{-1} - N^{-1})S^2$$

where

$$\phi = \left(\frac{\rho_1^2}{n} + \frac{1 - \rho_1^2}{n_1}\right)^{-1} + n_2 \tag{7}$$

For the above matching scheme in *SRS* the cost function for sampling on the second occasion may be written as

$$C = c_0 + c_1 n_2 M + c_2(n_1 + n_2)M \tag{8}$$

The optimum values of n_1/n and n_2/n may be obtained by minimising (7) for a given cost function (8). For the area estimation survey using $\hat{\rho}_c$, $\hat{\rho}_1$ and $\hat{\rho}$ as given in Table 3, we obtain optimum n_1/n and n_2/n for matching scheme in *SRS* and for clusters of sizes 2, 3 and 4 and their relative efficiencies with respect to *SRS* by using arbitrary value of R_1 and R_2. These are presented in Table 4(a) and 4(b) respectively. We do not assume that $(n_1 + n_2) = n$ on the second occasion.

It may be seen that if the funds available for the survey are not restricted, there is scope for taking a larger sample of cultivators on the second occasion as compared to the first occasion. When funds are meagre the sample on the second occasion may be even smaller than that on the first

TABLE 4(*a*)—OPTIMUM n_1/n, n_2/n FOR DIFFERENT VALUES OF R_1 AND R_2 FOR MATCHING IN *SRS*

R_1/R_2	n_1/n						n_2/n					
	0.2	0.5	1.0	2.0	5.0	10.0	0.2	0.5	1.0	2.0	5.0	10.0
100	0.33	0.33	0.33	0.33	0.33	0.33	0.58	0.58	0.58	0.58	0.58	0.58
500	0.34	0.55	0.70	0.96	1.00	1.00	0.57	0.56	0.54	0.52	0.51	0.51
1000	0.34	0.55	0.70	0.96	1.00	1.00	0.57	0.56	0.54	0.52	0.51	0.51
2000	0.34	0.55	0.70	0.96	1.00	1.00	0.57	0.56	0.54	0.52	0.51	0.51

occasion. As may be seen in case of $M = 2$, and 4 (Table 4(*b*)), no fres sample need be taken under some cases on the second occasion when $R_1 = 100$. Also when R_1 increases the efficiency of matching of clusters increases as compared to matching of an equivalent *SRS*.

Remark 1: When $\rho = 0$ then the case II in Table 1 can not occur, only I or III would occur and the optimum values of n_1 and n_2 will be given by

$$\begin{aligned} n_1 &= \text{Min}\ [n, R_1/\eta] \\ n_2 &= \text{Max.}\ [\eta^{-1}(R_1M^{-1} - n),\ R_1\eta^{-1}(M^{-1} - \eta^{-1}),\ 0] \end{aligned} \tag{9}$$

Remark 2: When $n_1 + n_2 = n$, i.e. the sample size remains same on both the occasions, then the optimum replacement fraction in terms of intra-class correlation coefficient and other constants is given by the following fourth degree equation in q:

$$\begin{aligned} & q^4(R_2^2\,\rho_c\,\rho^4) + q^3[2R_2\,\rho^2\,\rho_c(2 - \rho^2)\,R + R_2\,\rho^4(\rho_c - 2) \\ & - 2R_2^2\,\rho_c\,\rho^2)] + q^2[3R_2\,\rho^2(1 - \rho_c) + R\,\rho^4(1 + \rho_c R) \\ & + R_2\,\rho_c \cdot \rho^2(\rho^2 + R_2)] + q[R\,\rho_c\,\rho^2)\,(2 + \rho^2) - 2R\,\rho_c\,\rho^2 \\ & (R_2 + R) - 2R\,\rho^2] + [(1 - \rho_c)\,(R\,\rho^2 - R_2) + R^2\,\rho_c\,\rho^2] = 0 \end{aligned} \tag{10}$$

where

$$n_2 = nq \text{ and } n_1 = n(1 - q),\ R = R_1/n.$$

Table 5 gives the values of optimum q for some values of ρ, ρ_c, c_1, c_2 and $C^1 = C - c_0$ obtained by solving equation (10)

It may be seen that depending on the relative magnitudes of costs c_1 and c_2 and the total funds available optimum q can be even far below 1/2 which is the minimum replacement fraction in *SRS*. In Table 6 the relative efficiency of matching of clusters of sizes 2, 3 and 4 in relation to matching of an equivalent *SRS* for different values of ρ_c, ρ and ρ_1 has been worked out.

TABLE 4(*b*)—OPTIMUM n_1/n, n_2/n AND RELATIVE EFFICIENCY OF MATCHING OF CLUSTERS OF SIZES 2, 3 AND 4 AGAINST MATCHING OF EQUIVALENT SAMPLE OF *SRS*

R_1/R_2	n_1/n						n_2/n						Relative efficiency					
	0.2	0.5	1.0	2.0	5.0	10.0	0.2	0.5	1.0	2.0	5.0	10.0	0.2	0.5	1.0	2.0	5.0	10.0
For $M = 2$																		
100	0.45	0.52	0.44	0.33	0.19	0.11	0.00	0.00	0.00	0.00	0.06	0.06	0.44	0.49	0.44	0.37	0.27	0.18
500	0.45	0.52	0.64	0.85	0.95	0.56	1.19	0.91	0.68	0.41	0.20	0.18	1.10	0.97	0.84	0.71	0.62	0.49
1000	0.45	0.52	0.64	0.85	1.00	1.00	2.62	2.25	1.79	1.24	0.67	0.39	2.05	1.76	1.46	1.15	0.88	0.73
2000	0.45	0.52	0.64	0.85	1.00	1.00	7.65	4.91	4.02	2.91	1.62	0.94	3.94	3.34	2.70	2.03	1.38	1.02
For $M = 3$																		
100	0.35	0.39	0.43	0.53	0.37	0.23	0.00	0.00	0.00	0.00	0.00	0.03	0.39	0.40	0.43	0.46	0.40	0.32
500	0.35	0.39	0.43	0.53	0.75	1.00	1.23	1.10	0.92	0.68	0.34	0.15	1.01	0.91	0.80	0.67	0.56	0.51
1000	0.35	0.39	0.43	0.53	0.75	1.00	2.79	2.53	2.17	1.68	0.97	0.54	1.84	1.62	1.39	1.12	0.84	0.68
2000	0.35	0.39	0.43	0.53	0.75	1.00	5.92	5.38	4.67	3.68	2.22	1.31	3.49	3.05	2.57	2.02	1.40	1.03
For $M = 4$																		
100	0.33	0.35	0.39	0.45	0.59	0.38	0.00	0.00	0.00	0.00	0.00	0.05	0.32	0.33	0.35	0.37	0.41	0.34
500	0.33	0.35	0.39	0.45	0.62	0.83	1.27	1.17	1.02	0.81	0.47	0.24	0.85	0.77	0.69	0.59	0.49	0.44
1000	0.33	0.35	0.39	0.45	0.62	0.83	2.85	2.65	2.36	1.92	1.21	0.71	1.54	1.38	1.21	1.00	0.76	0.61
2000	0.33	0.35	0.39	0.45	0.62	0.83	6.03	3.61	5.02	4.14	2.68	1.67	2.92	2.60	2.24	1.01	1.30	0.96

TABLE 5—q_{opt} FOR GIVEN VALUES OF ρ, ρ_c, C_1, C_2, AND C^1

	$C^1 = 100$						$C^1 = 200$					
	$\rho = 0.5$		0.7		0.9		0.5		0.7		0.9	
ρ_c	$c_1 = 10$ $c_2 = 20$	$c_1 = 20$ $c_2 = 10$	$c_1 = 10$ $c_2 = 20$	$c_1 = 20$ $c_2 = 10$	$C_1 = 10$ $c_2 = 20$	$C_1 = 20$ $c_2 = 10$	$c_1 = 10$ $c_2 = 20$	$c_1 = 20$ $c_2 = 10$	$c_1 = 10$ $c_2 = 20$	$c_1 = 20$ $c_2 = 10$	$c_1 = 10$ $c_2 = 20$	$c_1 = 20$ $c_2 = 10$
0.0	0.37	0.11	0.50	0.42	0.87	0.78	0.45	0.38	0.53	0.49	0.89	0.78
0.1	0.43	0.31	0.51	0.47	0.87	0.61	0.48	0.48	0.55	0.51	0.88	0.63
0.3	0.46	0.41	0.60	0.49	0.90	0.62	0.50	0.48	0.57	0.53	0.90	0.54

TABLE 6—EFFICIENCY OF MATCHING OF CLUSTERS OF UNITS IN RELATION TO MATCHING OF *SRS* OF UNITS, WHEN $n_1 + n_2 = n$, FOR DIFFERENT VALUES OF ρ, ρ_1, ρ_c AND M

		$M = 2$					3					4				
ρ_1	ρ $\rho_c = (-)\,0.3$	−0.1	0.0	0.1	0.3	−0.3	−0.1	0.0	0.1	0.3	−0.3	−0.1	0.0	0.1	0.3
	0.5	1.43	1.11	1.00	0.91	0.77	2.50	1.25	1.00	0.83	0.62	10.00	1.43	1.00	0.77	0.53
	0.7	1.56	1.21	1.09	0.99	0.81	2.72	1.36	1.09	0.91	0.68	10.88	1.56	1.09	0.81	0.57
0.5	0.9	1.86	1.44	1.30	1.18	1.00	3.24	1.62	1.30	1.08	0.82	12.99	1.86	1.30	1.00	0.68
	0.95	2.43	1.89	1.70	1.55	1.31	4.25	2.12	1.70	1.42	1.06	17.00	2.43	1.70	1.31	0.89
	0.5	1.31	1.02	0.92	0.84	0.71	2.30	1.15	0.92	0.77	0.57	9.18	1.31	0.92	0.71	0.48
	0.7	1.43	1.11	1.00	0.91	0.77	2.50	1.25	1.00	0.83	0.63	10.00	1.43	1.00	0.77	0.53
0.7	0.9	1.70	1.33	1.19	1.08	0.92	2.98	1.49	1.19	1.00	0.75	11.93	1.71	1.19	0.92	0.63
	0.95	2.23	1.74	1.56	1.42	1.20	3.90	1.95	1.56	1.30	0.98	15.61	2.23	1.56	1.20	0.82
	0.5	1.10	0.85	0.77	0.70	0.59	1.92	0.96	0.77	0.64	0.48	7.69	1.10	0.77	0.59	0.41
	0.7	1.20	0.93	0.84	0.76	0.64	2.09	1.05	0.84	0.70	0.52	8.38	1.20	0.84	0.64	0.44
0.9	0.9	1.43	1.11	1.00	0.91	0.77	2.50	1.25	1.00	0.83	0.63	10.00	1.43	1.00	0.77	0.53
	0.95	1.87	1.45	1.31	1.19	1.01	3.27	1.64	1.31	1.09	0.82	13.08	1.87	1.31	1.01	0.69
	0.5	0.84	0.65	0.59	0.54	0.45	1.47	0.74	0.54	0.49	0.37	5.88	0.84	0.59	0.45	0.31
	0.7	0.92	0.71	0.64	0.59	0.49	1.60	0.81	0.04	0.53	0.40	6.40	0.92	0.64	0.49	0.34
0.095	0.9	1.09	0.85	0.76	0.70	0.59	1.91	0.96	0.76	0.64	0.48	7.64	1.09	0.76	0.59	0.40
	0.95	1.43	1.11	1.00	0.91	0.77	2.50	1.25	1.00	0.83	0.63	10.00	1.43	1.00	0.77	0.53

ACKNOWLEDGEMENT

The authors are grateful to the referees for some useful suggestions.

REFERENCES

[1] Kulldorff, G. (1963) : Some problems of optimum allocation for sampling on two occasions, *Rev. Inter. Statist. Inst*, **31** : 29-50.

[2] Patterson, H. D. (1950) : Sampling on successive occasions with partial replacement of units, *JRSS. Series B*, **12** : 241-255.

[3] Sukhatme, P. V. and Sukhatme, B. V. (1970) : Sampling theory of surveys with applications, Asia Publishing House, Bombay.