Jour. Ind. Soc. Ag. Statistics
50(2), 1997 : 199-208

Stratified PPS Sampling and Allocation of Sample Size

B.K. Gupt and T.J. Rao[1]
North-Eastern Hill University, Shillong
(Received : August, 1996)

SUMMARY

In this paper, the problem of optimum allocation of sample size to strata is considered when sampling with probability proportional to size with replacement (ppswr) within each stratum. Observing that this allocation depends on population parameters, 'near optimum' allocations based on auxiliary information is compared. Finally, these results are illustrated by numerical examples from live data.

Key words: Stratified sampling, Probability proportional to size selection, Auxiliary information, Neyman-optimum allocation, Super population models.

1. Introduction

Consider a finite population of size N divided into k strata of sizes N_i, i = 1, 2, ..., k. Let y be the study variate parametric functions of which we are interested in estimating. We also have information on a positive valued auxiliary characteristic x closely related to the characteristic y under study. Let Y_{ij} and X_{ij} denote the y and x-values respectively of the j-th unit in the i-th stratum, j = 1, 2, ..., N_i, i = 1, 2, .., k. When using simple random sampling with replacement (SRSWR) design within each stratum, it is well known that the optimum allocation of a total sample size n to strata is given by (Neyman [8]).

$$n_i^{opt} = n N_i \sigma_i / \sum_{i=1}^{k} N_i \sigma_i$$

where $\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2$ is the within stratum variance. Computation of n_i^{opt} requires at least the proportionate values of σ_i^2's which are unknown. In

1 Indian Statistical Institute, Calcutta

practice, when pilot surveys and past data are not satisfactory, some estimates α_i^2's of σ_i^2's based on the prior information on related x-values are substituted. These estimates usually are the within stratum variances of the auxiliary information x which is assumed to be positively correlated with y.

The justification for assuming that the unknown proportionate values of σ_i^2's are usually not very different from the proportionate values of known α_i^2's was examined in the light of prior distributions using a super population model approach first by Hanurav [3] in his thesis which has been the basis for all subsequent generalizations of Rao [9], [10]. For the special case of the model with model variance proportional to x_i^2, it was shown that Neyman optimum allocation reduces to allocation proportional to stratum totals of the x-values, provided the coefficients of variation (c.v.) of x-values in each stratum are equal. Earlier, Mahalanobis [5] proposed equalization of the strata totals together with equal allocation as an approximation to optimum allocation while Kitagawa [4] provided the justification for equipartition. Hansen *et al.* [2] gave several illustrations to show that allocation proportional to stratum totals is a 'near-optimum' allocation.

However, when the x-values within a stratum are far from being equal, it may be more prudent to sample with a 'probability proportional to x' (ppx) selection method within each stratum.

Furthermore, the gains due to ppx selection method over simple random selection in many practical situations are well established and during the last five decades this method has been successfully employed in many surveys (cf. Rao [12]). Motivated by this, we consider the allocation of sample size to strata when sampling by ppswr design within each stratum.

2. *Stratified PPSWR Sampling*

Let $p_{ij} = x_{ij}/X_i$ be the selection probability for the jth unit of the ith stratum. Then as an estimator of the population total $Y = \sum_{i=1}^{k} \sum_{j=1}^{N_i} Y_{ij}$, consider

$$\hat{Y} = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{y_{ij}}{p_{ij}} \tag{2.1}$$

with

$$V(\hat{Y}) = \sum_{i=1}^{k} \frac{A_i^2(y)}{n_i} \tag{2.2}$$

where $A_i^2(y) = \sum_{j=1}^{N_i} \frac{Y_{ij}^2}{p_{ij}} - Y_i^2$

Neyman's optimum allocation of a total sample size of n to strata which minimizes (2.2) is given by

$$n_i^{opt} = n \frac{A_i(y)}{\sum_{1}^{k} A_i(y)} \tag{2.3}$$

Here the values of $A_i(y)$ are unknown. In practice suitable estimates of $A_i(y)$ based on auxiliary information are used. Cochran [1] showed that whenever auxiliary information on a characteristic x closely related to y is available, this information could be used to set up a criterion of optimality by regarding $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{kN_k})$ as a realization of N-length random vector with distribution depending on $\mathbf{X} = (X_{11}, X_{12}, \ldots, X_{kN_k})$ and some unknown parameters. Thus given **X**, we explicitly formulate the super population model $\theta(g)$ given by

$$\begin{aligned} \xi_{\theta(g)}(Y_{ij}/X_{ij}) &= \beta X_{ij} \\ \upsilon_{\theta(g)}(Y_{ij}/X_{ij}) &= \sigma^2 X_{ij}^g \\ \zeta_{\theta(g)}(Y_{ij}, Y_{i'j'}/X_{ij}, X_{i'j'}) &= 0 \end{aligned} \tag{2.4}$$

where the script letters ξ, υ and ζ denote the conditional expectation, variance and covariance given X_{ij}'s respectively. Here the super-population parameter g lies mostly between 1 and 2 and it is more often close to 2.

Following Rao [9], we shall now derive the expected value of $A_i^2(y)$ which can be used in obtaining the optimum allocation. We thus have

$$\begin{aligned} \xi_{\theta(g)} A_i^2(y) &= \sum_{j=1}^{N_i} \frac{1}{p_{ij}} \xi(Y_{ij}^2) - \xi Y_i^2 \\ &= \sum_{j=1}^{N_i} \frac{1}{p_{ij}} (\sigma^2 X_{ij}^g + \beta^2 X_{ij}^2) - \left\{ \sigma^2 \sum_{j} X_{ij}^g + \beta^2 X_i^2 \right\} \end{aligned}$$

$$= \sigma^2 \sum_j X_{ij}^g \left(\frac{1}{p_{ij}} - 1 \right) + \beta^2 A_i^2(x) \qquad (2.5)$$

where $A_i^2(x)$ is the same as $A_i^2(y)$ with known x values substituted in place of unknown y values.

With the motivation given in the introduction let us take the values of P_{ij} equal to X_{ij}/X_i. Then (2.5) reduces to

$$\xi_{\theta(g)} A_i^2(y) = \sigma^2 \left\{ X_i \sum_j X_{ij}^{g-1} - \sum_j X_{ij}^g \right\} \qquad (2.6)$$

Thus Neyman's optimum allocation based on auxiliary variate is justified if allocation is made proportional to $(X_i \sum_j X_{ij}^{g-1} - \sum_j X_{ij}^g)^{1/2}$ instead of the unknown $A_i(y)$. For the particular cases of interest, optimum allocation is achieved when

$$n_i^{opt} \propto (X_i^2 - \Sigma X_{ij}^2)^{1/2} \quad \text{if} \quad g = 2$$

and

$$\propto ((N_i - 1)X_i)^{1/2} \quad \text{if} \quad g = 1 \qquad (2.7)$$

Remark 2.1. For $g = 2$, the optimum allocation can be rephrased as $n_i \propto X_i(1 - \delta_i)^{1/2}$, where $\delta_i = \sum_j \frac{X_{ij}^2}{X_i^2} = \frac{(1 + C_i^2(x))}{N_i}$, with $C_i^2(x)$ = square of the coefficient of variation of the x-values in the ith stratum given by $\alpha_i^2/\bar{X}_i^2$ where $N_i\alpha_i^2 = \sum_j X_{ij}^2 - (X_i^2/N_i)$. When strata sizes are large and $C_i^2(x)$ is relatively quite small, the correction factor can be ignored and allocation can be taken as proportional to strata totals of x-values. This resembles an earlier result of Rao's [9] for SRS situation.

Observe that when δ_i's are equal in all strata, optimum allocation also reduces to allocation proportional to stratum totals X_i.

Remark 2.2. Here we shall consider the case when $P_{ij} \neq X_{ij}/X_i$ with $\sum_j P_{ij} = 1$. From (2.5), for the special case of $g = 2$ we have

$$\xi_{\theta(2)} A_i^2(y) = \sigma^2 \sum_j X_{ij}^2 \left(\frac{1}{p_{ij}} - 1 \right) + \beta^2 A_i^2(x)$$

$$= (\sigma^2 + \beta^2) A_i^2(x) + \sigma^2 (X_i^2 - \sum_i X_{ij}^2) \qquad (2.8)$$

As mentioned earlier, the normal practice for obtaining optimum allocation is to take n_i proportional to known $A_i(x)$ instead of unknown $A_i(y)$. Thus under the super population model approach allocation proportional to $A_i(x)$ is justified from (2.8), if $A_i^2(x)$, in turn, is proportional to $(X_i^2 - \Sigma X_{ij}^2)$. This means that $n_i^2 = A_i^2(x)/(X_i^2 - \Sigma X_{ij}^2)$ or $A_i^2(x)/X_i^2(1 - \delta_i)$ where $\delta_i = (1 + C_i^2(x))/N_i$, as before, should be equal in all strata. Noting that $A_i^2(x)/X_i^2$ is the square of the C.V. $(\hat{X}_i)$ with $X_i = X_{ij}/P_{ij}$, one can interpret that n_i^2 are squares of corrected C.V. As remarked before the correction factor δ_i is relatively very small and allocation proportional to $A_i(x)$, in turn, can be taken to be equivalent to allocation proportional to the stratum totals of x-values X_i.

Remark 2.3. The results in Remark (2.3) can easily be derived for the model with a general value of g which we shall omit here.

Remark 2.4. By using Moors and Muilwijk [6] inequality an upper bound for δ_i is given by

$$(1 + t_i)^2/4N_i t_i \text{ where } t_i = \max_j X_{ij} / \min_j X_{ij}$$

3. *Illustrations*

We shall illustrate the above results first with live data on 44 countries with the lowest GNP (Sarndal *et al.* [13]).

When the selection probabilities p_{ij} are proportional to X_{ij}, it is clear from Tables 3.2 and 3.3 below that the allocation proportional to stratum totals X_i, $i = 1, 2$ is almost as efficient as optimum allocation. The near-optimality of this allocation is more striking for the type A of stratification based on equalization of X_i values (cf. Mahalanobis [5] and Kitagawa [4]), compared to stratification type B based on equalizing $\sum_j X_{ij}^2$ and type C based on equalizing N_i.

Table 3.1. Data on 44 countries with lowest GNP

Country	G.N.P.	Import	Export	Country	G.N.P.	Import	Export
i	x	y	z	i	x	y	z
1	32	125	33	23	234	807	977
2	36	109	58	24	239	1219	494
3	47	283	256	25	247	342	80
4	76	127	109	26	262	767	692
5	81	227	305	27	262	799	411
6	93	1527	779	28	266	250	200
7	95	391	177	29	266	993	867
8	99	422	464	30	299	540	316
9	122	131	63	31	300	1531	745
10	122	194	76	32	315	695	708
11	123	344	167	33	325	293	345
12	123	438	373	34	330	724	2161
13	123	171	119	35	356	891	735
14	128	484	240	36	371	1521	39
15	135	288	57	37	375	3342	3200
16	136	312	230	38	386	690	831
17	144	279	79	39	415	1412	304
18	158	461	154	40	416	705	873
19	165	655	436	41	419	3030	579
20	172	351	428	42	465	635	132
21	178	442	333	43	465	875	403
22	212	330	199	44	491	1786	1123

Table 3.2. Showing optimum and near-optimum allocations

Stratification type	Stratum size N_i	Allocation proportional to			
		$A_i(y)$	X_i	$(X_i^2 - \Sigma X_{ij}^2)^{1/2}$	∂_i
A	31	0.4850 n	0.4924 n	0.4975 n	.0397
	13	0.5150 n	0.5076 n	0.5025 n	.0784
B	36	0.5993 n	0.6603 n	0.6713 n	.0351
	8	0.4007 n	0.3397 n	0.3287 n	.1260
C	22	0.3176 n	0.2573 n	0.2569 n	.0520
	22	0.6824 n	0.7427 n	0.7431 n	.0478

It may also be noted that the relative (to the optimum) efficiencies (r.e) of allocations (i) proportional to $(X_i^2 - \Sigma X_{ij}^2)^{1/2}$ corresponding to $g = 2$ and (ii) proportional to X_i are quite close. We have also given in Table 3.3 r.e.

Table 3.3. Relative efficiencies (r.e.) of near-optimum allocations compared with optimum allocation

Stratification type	R.e. of allocations		
	(i)	(ii)	(iii)
A	0.9994	0.9998	0.9991
B	0.9771	0.9837	0.9609
C	0.9811	0.9813	0.8825

for (iii) equal allocation. It was also demonstrated by Murthy [7] that allocation proportional to stratum totals compare very favourably with optimum allocation (see Table 3.4)

Table 3.4. Showing variances for stratified PPS sampling with fixed costs*

Cost is proportional to	Allocation proportional to Geographical area	Optimum allocation
No. of villages	0.002922	0.002854
Expected no. of Geographical area of sample villages	0.003065	0.003028

* *Source:* M.N. Murthy [7] referring to total area under autumn paddy in Nadia

The above table once again confirms the near optimality of allocation proportional to stratum totals.

Remark 3.1. When the selection probabilities p_{ij} are proportional to some other characteristic Z_{ij} different from X_{ij}, as is done in certain situations (cf. Remark 2.2), we find that for the data of Table 3.1, when stratification is based on equal X_i's and strata sizes are 31 and 13 respectively, the relative efficiency of allocation proportional to $A_i(x)$ compared to allocation proportional to $A_i(y)$ was 91%.

We notice that the values of x, y and z in the above data are somewhat erratic and do not satisfy the conditions required. Even then the efficiency of using $A_i(x)$ for allocation instead of $A_i(y)$ was about 91% though not higher.

We next turn our attention to the data given in Table 3.5 relating to the population figures of 16 districts of the West Bengal State of India.

Table 3.5. Showing population figures of districts of West Bengal state

District	1951 (z in '000s)	Population in 1961 (x in '000s)	1971 (y)
1	460	625	765677
2	671	1020	1412148
3	938	1222	1614570
4	979	1324	1846215
5	915	1359	1752171
6	1169	1360	1610577
7	1067	1446	1779805
8	1319	1665	2035273
9	1145	1713	2229022
10	1611	2038	2420095
11	1604	2231	2873779
12	1716	2290	2942125
13	2698	2927	3141180
14	2192	3083	3920395
15	3359	4342	5515320
16	4459	6281	8581743

We stratify the population into 2 strata by each of three types of stratifications considered earlier. Then we obtain allocations proportional to (i) $A_i(y)$, (ii) $A_i(x)$ and (iii) X_i values (see Table 3.6). The relative efficiency of allocation (ii) compared to (i) is quite close to unity for all the three stratification types (see Table 3.7) thereby confirming our prescription of allocation proportional to $A_i(x)$ instead of $A_i(y)$. However, for this data, allocation proportional to X_i is not as efficient for stratification types A and B, since the data on ∂_i and n_i do not satisfy the required conditions (cf. Table 3.6). It is believed that the assumptions hold good when we have data on x, y, z variables which relate to the three different epochs of time which are not too far apart.

Table 3.6. Comparing optimum and near-optimum allocations

Type of stratification	Stratum i	Stratum size N_i	Allocation of Sample Size			Condition	
			(i) $\propto A_i$ (y) (optimum)	(ii) $\propto A_i$ (x)	(iii) $\propto X_i$	∂_i	n_i
A	1	12	.4175 n	.4519 n	.5238 n	.0913	.0057
	2	4	.5825 n	.5481 n	.4762 n	.2760	.0128
B	1	14	.8121 n	.8477 n	.6958 n	.0823	.0103
	2	2	.1879 n	.1523 n	.3042 n	.5166	.0037
C	1	8	.25485 n	.2713 n	.2869 n	.1318	.0075
	2	8	.74515 n	.7287 n	.7131 n	.1510	.0090

Table 3.7. Relative efficiencies of near-optimum allocations compared with optimum allocation

Type of Stratification	Optimum allocation	r.e. of near optimum allocations	
		$\propto A_i(x)$	$\propto X_i$
A	1.0000 $(3791 \times 10^6)^*$	.9953	.9967
B	1.0000 $(3406 \times 10^6)^*$	.9943	.9395
C	1.0000 $(3955 \times 10^6)^*$	.9986	.9950

* Figures in parenthesis indicate the variances under respective optimum allocations.

REFERENCES

[1] Cochran, W.G, 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.*, **17**, 164-177.

[2] Hansen, M.H., Hurwitz, W.N. and Madow, W.G., 1953. *Sample Surveys Methods and Theory.* John Wiley and Sons, New York.

[3] Hanurav, T.V., 1965. *Optimum sampling strategies* and some *related problems.* Ph.D. Thesis submitted to the Indian Statistical Institute.

[4] Kitagawa, T., 1956. Some contributions to the design of sample surveys. *Sankhya*, **17**, 1-36.

[5] Mahalanobis, P.C., 1952. Some aspects of the design of sample surveys. *Sankhya*, **12**, 1-7.

[6] Moors, J.J.A. and Muilwijk, H., 1971. An inequality for the variance of a discrete random variable. *Sankhya*, **B33**, 385-388.

[7] Murthy, M.N., 1967. *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.

[8] Neyman, J., 1934. On two different aspects of the representative method, the method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.*, **97**, 558-606.

[9] Rao, T.J., 1968. On the allocation of sample size in stratified sampling. *Ann. Inst. Statist. Math.*, **20**, 159-166.

[10] Rao, T.J., 1977. Optimum allocation of sample size and prior distributions: A review. *Int. Statist. Rev.*, **45**, 173-179.

[11] Rao, T.J., 1984. Allocation of sample size to strata and related problems. *Biometrical J.*, **26**, 517-526.

[12] Rao, T.J., 1993. Fifty years of PPS sampling. In: *Proceedings of the Mahalanobis Birth Centenary Conference* (ed.) A.M. Mathai Centre for Mathematical Sciences, Trivandrum.

[13] Sarndal, C.E., Swenson, B. and Wretman, J., 1991. *Model Assisted Survey Sampling.* Springer-Verlag.