Jour. Ind. Soc. Ag. Statistics
Vol. XXXVIII, No. 1 (1986), pp. 92-99

ON THE ALLOCATION OF SAMPLE SIZE IN STRATIFIED SAMPLING UNDER A FINITE POPULATION MODEL

A. M. PEDGAONKAR AND S. G. PRABHU-AJGAONKAR
Marathwada University, Aurangabad

(Received: January, 1984)

SUMMARY

The paper considers the problem of optimum allocation of sample size to the strata when the finite population model furnished by Avadhani and Sukhatme [1] holds.

Introduction

One of the designs frequently used in sample surveys is stratified random sampling. Consider the population under study to be divided into k strata each of size N_i, $i = 1, 2, \ldots, k$ and a stratified random sample of size n_i, $i = 1, 2, \ldots, k$ is drawn without replacement from the ith stratum so that $\sum_{i=1}^{k} n_i = n$. Let Y denote the characteristic of interest and X a characteristic which is highly correlated with Y. Let Y_{ij} and X_{ij} be the Y and X characteristic values respectively of the jth unit in the ith stratum ($j = 1, 2, \ldots, N_i$, $i = 1, 2, \ldots, k$).

The population mean $\bar{Y}_N$ can be expressed as

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{Y}_{N_i} = \sum_{i=1}^{k} p_i \bar{Y}_{N_i} \tag{1.1}$$

where

$$p_i = \frac{N_i}{N} \text{ and } \bar{Y}_{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}.$$

It is known that in simple random sampling the sample mean is an unbiased estimator of the population mean. Here

$$\bar{y}_{st} = \sum_{i=1}^{k} p_i \, \bar{y}_{n_i}, \text{ where } \bar{y}_{n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \tag{1.2}$$

is an estimator of the population mean and its variance is given by

$$\text{Var}(\bar{y}_{st}) = \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 \tag{1.3}$$

where

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_{N_i})^2.$$

2. Optimum Allocation of Sample Size Under Finite Population Model

In the theory of sampling techniques the results developed are purely concerned to the particular finite population under consideration. Consider a finite population which is changing with time. Here the knowledge of stochastic process which has engendered it is very informative. The actual finite population with which the statistician is dealing with, may be regarded as one particular set of realization of infinite values. Cochran [2] was the first to visualize the process by assuming the particular finite population to be a random sample from a super-population. Rao [4] has employed this super-population model of Cochran for allocating sample sizes in stratified sampling.

Recently, another model suggested for the finite population by Cochran [2] has been further developed by Avadhani and Sukhatme [1]. Commenting on their model Avadhani and Sukhatme [1] have rightly remarked that their model is confined solely to the finite population under consideration and no assumptions of any distributional form are involved. Tacitly, in sampling from finite populations, this approach appears to be more comprehensible.

Consider the finite population model furnished by Avadhani and Sukhatme [1].

$$\left.\begin{array}{l} Y_{ij} = \beta X_{ij} + e_{ij}, \; i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, N_i \\ \sum\limits_{j=1}^{n_i} e_{ij} = 0 = \sum\limits_{j=1}^{n_i} e_{ij} X_{ij} \\ E(e_{ij}^2) = r \, X_{ij}^g, \; r > 0, \qquad 0 \leqslant g \leqslant 2 \end{array}\right\} \tag{2.1}$$

Under the finite population model, $\text{Var}(\bar{y}_{st})$ reduces to

$$\overset{*}{\text{Var}}(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) P_i^2 \left(\beta^2 S_{i\infty}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N_i} X_{ij}^g \right) \tag{2.2}$$

where

$$S_{i\infty}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_{N_i})^2,$$

and

$$\overline{X}_{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}.$$

The cost function may be noted as

$$C = \sum_{i=1}^{k} c_i n_i \tag{2.3}$$

where c_i is the cost per unit in the ith stratum. When c_i is the same for all strata, say c, then the cost function is

$$C = cn \tag{2.4}$$

Considering cost function represented by equation (2.3) we determine the optimum values of n_i, which are given by

$$n_i = \frac{p_i \left(\beta^2 S_{i\infty}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N_i} X_{ij}^g\right)^{1/2}}{\sqrt{\mu c_i}}, \; i = 1, 2, \ldots, k \tag{2.5}$$

where μ is some constant.

Now considering the fixed cost C_0, which is the budget amount within which it is desired to estimate the mean with the maximum precision. We get

$$n_i = \frac{p_i \left(\beta^2 S_{i\infty}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N_i} X_{ij}^g \right)^{1/2} C_0}{\sqrt{c_i} \sum_{j=1}^{k} p_i \sqrt{c_i} \left(\beta^2 S_{i\infty}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N} X_{ij}^g \right)^{1/2}} \tag{2.6}$$

This is called the optimum allocation.

Now

$$\beta^2 S_{ix}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N_i} X_{ij}^g$$

$$= (\beta^2 + r)\, S_{ix}^2 + \frac{rN_i}{N_i - 1}\left[\bar{X}_{N_i}^2 - \frac{1}{N_i}\left(\sum_{j=1}^{N_i} X_{ij}^2 - \sum_{j=1}^{N_i} X_{ij}^g\right)\right]$$

Assuming $N_i/N_i - 1 \approx 1$, and if S_{ix}^2 are proportional to $\bar{X}_{N_i}^2 - \frac{1}{N_i}\left(\sum_{j=1}^{N_i} X_{ij}^2 - \sum_{j=1}^{N_i} X_{ij}^g \right)$, i.e. the square of the corrected coefficients of variation (c.c.v) of the X-characteristic defined by $\frac{S_{ix}^2}{\bar{X}_{Ni}^2 - \delta_i/N_i^2}$, where $\delta_i = N_i\left(\sum_{j=1}^{N_i} X_i^2 - \sum_{j=1}^{N_i} X_{ij}^g \right)$ are equal in all the strata, which is the case in many sample surveys encountered in practice. Under this condition the optimum allocation is given by

$$n_i = \frac{(T_i^2 - \delta_i)^{1/2}\, C_0}{\sqrt{c_i} \sum_{i=1}^{k} \sqrt{c_i}\, (T_i^2 - \delta_i)^{1/2}} \tag{2.7}$$

where

$$T_i = \sum_{j=1}^{N_i} X_{ij},$$

and the total sample size required for estimating the population mean with maximum precision for a fixed cost C_0 is given by

$$n = \frac{C_0 \sum_{i=1}^{k} \left[(T_i^2 - \delta_i)^{1/2}/\sqrt{c_i}\right]}{\sum_{i=1}^{k} \sqrt{c_i}\, (T_i^2 - \delta_i)^{1/2}} \tag{2.8}$$

Thus we have the following theorem

THEOREM 2.1. *Under the finite population model optimum allocation reduces to the allocation proportional to* $(T_i^2 - \delta_i)^{1/2}/\sqrt{c_i}$, *where* $T_i = \sum_{j=1}^{N_i} X_{ij}$ *is the total of the auxiliary variate* X *for the ith stratum,* c_i *the cost per unit in the ith stratum and* $\delta_i = N_i\left(\sum_{j=1}^{N_i} X_{ij}^2 - \sum_{j=1}^{N_i} X_{ij}^g\right)$.

when the corrected coefficients of variation for the X-characteristics are equal in all the strata. For $g = 2$, the above theorem reduces to

THEOREM 2.2. *For $g = 2$, under the finite population model, optimum allocation reduces to the allocation proportional to $T_i/\sqrt{c_i}$ where $T_i = \sum_{j=1}^{N_i} X_{ij}$, is the total of the auxiliary variate X—for the ith stratum and c_i the cost per unit in the ith stratum, when the coefficients of variation for the X-characteristic are equal in all the strata.*

Proof: Under the finite population model, when $g = 2$, we have

$$\delta_i = N_i \left(\sum_{j=1}^{N_i} X_{ij}^2 - \sum_{j=1}^{N_i} X_{ij} \right) = 0,$$

Hence

$$n_i = \frac{C_0 T_i/\sqrt{c_i}}{\sum_{i=1}^{k} T_i \sqrt{c_i}}. \tag{2.9}$$

Now when $c_i = c$, i.e. the cost per unit in the ith stratum is the same for all strata, then we have from equations (2.7) and (2.8)

$$n_i = n \frac{(T_i^2 - \delta_i)^{1/2}}{\sum_{i=1}^{k} (T_i^2 - \delta_i)^{1/2}}. \tag{2.10}$$

Thus we have the following theorem

THEOREM 2.3. *Under the finite population model; Neyman's optimum allocation reduces to allocation proportional to $(T_i^2 - \delta_i)^{1/2}$, where $T_i = \sum_{j=1}^{N_i} X_{ij}$ is the total of the auxiliary variate for the ith stratum and $\delta_i = N_i \left(\sum_{j=1}^{N_i} X_{ij}^2 - \sum_{j=1}^{N_i} X_{ij}^{o} \right)$, when the corrected coefficients of variation for the X-characteristic are equal in all the strata. For $g = 2$, the above theorem reduces to*

THEOREM 2.4. *For $g = 2$, under the finite population model, Neyman's optimum allocation reduces to allocation proportional to stratum totals of the auxiliary variate X, when the coefficients of variation of X-characteristic are equal in all the strata.*

Proof. Under finite population model, when $g = 2$ we have $\delta_i = 0$. Hence,

$$n_i = n \frac{T_i}{\sum_{i=1}^{k} T_i}.$$

Now when the population mean is to be estimated with a given variance V_0 at minimum cost. We have

$$\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \left(\beta^2 S_{ix}^2 + \frac{r}{N_i - 1} \sum_{j=1}^{N_i} X_{ij}^g \right) = V_0.$$

Substituting for n_i from the equation (2.5) and solving we get,

$$n_i = \frac{p_i\left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)^{1/2} \sum_{i=1}^{k} p_i \sqrt{c_i}\left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)^{1/2}}{\sqrt{c_i}\left[V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i\left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)\right]},$$

$$i = 1, 2, \ldots, k.$$

Thus the minimum sample size required for estimating the mean with fixed variance V_0, under optimum allocation is given by

$$n = \frac{\sum_{i=1}^{k} \left[p_i \left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)^{1/2} \Big/ \sqrt{c_i}\right] \sum_{i=1}^{k} p_i \sqrt{c_i}\left(\bar{X}_{N_i} - \frac{\delta_i}{N_i^2}\right)^{1/2}}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i \left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)} \tag{2.11}$$

Putting $c_i = c$, we find the minimum sample size required for estimating the mean with fixed variance V_0 under Neyman's optimum allocation is given by

$$n = \frac{\left[\sum_{i=1}^{k} p_i \left(X_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)^{1/2}\right]^2}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i\left(\bar{X}_{N_i}^2 - \frac{\delta_i}{N_i^2}\right)} \tag{2.12}$$

when $g = 2$, the above equation (2.12) reduces to

$$n = \frac{\left(\sum_{i=1}^{k} p_i \bar{X}_{N_i}\right)^2}{V_0 + \frac{1}{N} \sum_{i=1}^{N} p_i \bar{X}_{N_i}^2}. \tag{2.13}$$

3. Illustration

We illustrate the foregoing theory by considering the data furnished by Cochran [3] which gives the number of inhabitants (in thousands), of 64 large cities in the United States, in 1920, grouped into two strata.

TABLE I—SIZES OF 64, CITIES (IN THOUSANDS) IN 1920, SIZE (X_{ij})

Stratum 1	*Stratum* 2		
797	314	171	121
773	298	172	120
748	296	163	119
734	258	162	118
588	256	161	118
577	243	159	116
507	238	153	116
507	237	144	113
457	235	138	113
438	235	138	110
415	216	138	110
401	208	138	108
387	201	136	106
381	192	132	104
384	180	130	101
315	179	126	100

TABLE II

Stratum	*Stratum size* N_i	*Stratum totals* $T_i = \sum_j X_{ij}$	$\sum_{j=1}^{N_i} X_{ij}^2$	S_i	*Allocation proportional to* S_i	T_i
1	16	8349	4756619	163.3	.48*n*	.51*n*
2	48	7941	1474871	58.5	.51*n*	.49*n*

TABLE III

Stratum	*g*	δ_i	$\sqrt{T_i^2 - \delta_i}$	*c.c.v.*	*Allocation* $\sqrt{T_i^2 - \delta_i}$	*Proportional to* *c.c.v.*
1	1.7	64862736.0	2200.69	.1187	.57*n*	.52*n*
2	1.7	56201145.6	2618.84	.68*n*	.47*n*	
1	1.8	54855360.0	3853.62	.6780	.48*n*	.49*n*
2	1.8	46120881.6	4115.65	.6829	.51*n*	.50*n*
1	1.9	36976704.0	5720.93	.4567	.49*n*	.48*n*
2	1.9	29024880.0	5833.91	.4817	.50*n*	.51*n*

ACKNOWLEDGEMENT

The authors are grateful to the referee for many useful suggestions.

REFERENCES

[1] Avadhani, M. S. and Sukhatme, B. V. (1970) : A comparison of two sampling procedures with an application to successive sampling, *J. Roy. Statist. Soc.* (c), Applied Statistics, **19** : 251-259.

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

[3] Cochran, W. G. (1963) : *Sampling Techniques*, 2nd Ed. New York : John Wiley and Sons.

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

[5] 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.