ProbStat Forum, Volume 08, January 2015, Pages 19–23
ISSN 0974-3235

ProbStat Forum is an e-journal.
For details please visit www.probstat.org.in

Optimum Allocation in Stratified Two Stage Design by Using Double Sampling for Multivariate Surveys

Monika Saini, Ashish Kumar

Department of Mathematics, School of Basic Sciences and Humanities, Manipal University Jaipur Jaipur 303007, Rajasthan, India.

Abstract. When more than one characteristics are under study it is not possible for one reason or the other to use the individual optimum allocation of first stage and second stage sampling units to each stage and to various strata while stratified two stage sampling designs when auxiliary information is estimated by using double sampling. In such situations some criterion is needed to work out an acceptable allocation which is optimum for all characteristics in some sense. In this paper the problems of the optimum allocation in multivariate stratified two stage sampling by using double sampling are formulated as Nonlinear Programming Problems (NLPP). The NLPPs are then solved by using Lagrange multiplier technique and explicit formulas are obtained for the optimum allocation of the first stage and second stage sampling units.

1. Introduction

In many surveys the use of two stage sampling design often specifies two stages of selection: clusters or primary sampling units (PSUs) at first stage, and subsamples from PSUs at second stage as a secondary units (SSUs). For the large-scale surveys, stratification may precede selection of the sample at any stage. Analysis of two stage designs are well documented when a single variable is measures and the method to obtain the optimum allocations of sampling units to each stage are readily available (Neyman (1934); Dalenius (1957); Ghosh (1958); Kokan and Khan (1967); Cochran (1977); Arnold (1986); Sadooghi-Alvandi (1986); Valliant and Gentle (1977); Clark and Steel (2000); Dever et al. (2001)) and many others. However, when more than one characteristic are under study the procedures for determining optimum allocations are not well defined. The traditional approach is to estimate optimum sample size for each characteristic individually and then chose the final sampling design from among the individual solutions. In practice it is not possible to use this approach of individual optimum allocations, because an allocation, which is optimum for one characteristic, may not be optimum for other characteristics. Moreover, in absence of a strong positive correlation between the characteristics under study the individual optimum allocation may differ a lot and there may be no obvious compromise. In such situations some criterion is needed to work out an acceptable sampling design which is optimum, in some sense, for all characteristics. Several authors have studied various criteria for obtaining a compromise allocation. Among them are Prekopa (1995), Garcia and Tapia (2007), Javad et al. (2009), Bakhshi et al. (2010) and many others.

Keywords. Double sampling, Stratified two-stage design, Optimum allocation, Nonlinear programming problem
Received: 21 June 2014; Accepted: 21 January 2015
Email addresses: drmnksaini4@gmail.com (Monika Saini), ashishbarak2020@gmail.com (Ashish Kumar)

In this paper a method of optimum allocation for multivariate stratified two-stage sampling designs by using double sampling is developed. The problems of determining the optimum allocations are formulated as Nonlinear Programming problems (NLPP) in which each NLPP has a convex objective function and a single linear cost constraint. Several techniques are available for solving these NLPPs, better known as Convex Programming Problems (CPP). We used Lagrange multiplier technique to solve the formulated NLPPs and explicit formula for the optimum allocation of PSUs and the optimum size of SSUs or the subsamples to various strata are obtained. The Kuhn and Tucker (1951) necessary conditions, which are also sufficient, for this problem, are verified at the optimum solutions.

2. Formulation of the Problem in Stratified Two Stage Design by using Double Sampling

The most common design in surveys is stratified two-stage design. The population of FSU is divided into strata within each stratum a simple random sample without replacement of FSUs is selected and each of the FSUs is further sub sampled. If information is not known for strata, the technique of double sampling can be used which consists of selecting a preliminary sample of n' units from N FSUs distinct and identify units without replacement, to collect information for constructing strata then classify them into strata $n'_1, n'_2, n'_3, \ldots, n'_L$ respectively, where $n' = \sum_h^L n'_h$ and then further selecting a sub-sample of n units with n_i units from the ith stratum such that $n = \sum_h n_h$ Also let M_{hi} be the number of SSUs in the ith FSU and $M_{h0} = \sum_{i=1}^{N} M_{hi}$ be the total number of SSUs in the hth stratum. A random sample of m_{hi} i.e. number of ssu's to be selected from each sampled first stage units out of M_{hi} in hth stratum. In a multivariate stratified two-stage sampling, where p characteristics are under study, let $y_{k,hij}$ denote the value of kth characteristic on the jth SSU of ith FSU of hth stratum.

Let $\overline{\overline{y}}_{k,std} = \sum_{h=1}^{L} w'_h \overline{\overline{y}}_{k,h}$ denote the overall sample mean per SSU for k^{th} characteristic in h^{th} stratum where $\overline{\overline{y}}_{k,h} = \frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} \overline{y}_{k,hi}$ and $\overline{y}_{k,hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{k,hij}$. Note that $w'_h = \frac{n'_h}{n'}$ is an unbiased estimator of strata weights $W_h = \frac{N_h}{N}$. Throughout we assume that $n'$is large enough so that pr $(n' = 0) = 0$ for all h. It could be shown that $\overline{\overline{y}}_{k,std}$ is conditionally unbiased estimate of the overall population mean $\overline{\overline{Y}}_k$ of kth characteristic with conditional variance

$$
\begin{aligned}
V(\overline{\overline{y}}_{k,std}) = {} & \left(\frac{1}{n'} - \frac{1}{N}\right) S^2_{k,y} + \frac{1}{n'} \sum_h W_h \left(\frac{1}{v_h} - 1\right) S^2_{k,yh} \\
& + \frac{1}{n'} \sum_h \frac{W_h}{n_h} \sum_i M^2_{hi} \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right) S^2_{k,yhi} \qquad (1)
\end{aligned}
$$

where $S^2_{k,y}$ and $S^2_{k,yh}$ is the variance among primary unit means. $S^2_{k,yhi}$ is the variance among subunits within primary units for kth characteristic respectively.

Assume that the total cost of the survey consist of two components depending upon the numbers of PSUs using double sampling and number of SSUs in the sample. The PSUs using double sampling so, the cost of PSUs also consist of two components depending upon the number of first phase and second phase in the PSUs. Let c_1 denote the cost per unit of first phase of PSU for measuring auxiliary variate, c_{1h} denote the cost per unit of second phase of PSU and $c_{2h} = \sum_{k=1}^{p} c_{2kh}$ denote the cost of measurement all the p characteristics per SSUs in hth stratum, respectively where c_{2kh} are the per unit costs of measuring the kth characteristic of a SSU. Thus the total cost of the survey may be expressed as a function of first stage sample size using double samplingn', n_h and second stage sample size m_{hi} as:

$$
c_0 + c_1 n' + \sum_{h=1}^{L} \left[c_{1h} n_h + c_{2h} \sum_{i=1}^{n_h} m_{hi} \right]
$$

where c_0 is the overhead cost of the survey. The second component in (1) varies from sample to sample. It is, therefore the expected cost function could be considered as:

$$c_0 + c_1 n' + \sum_{h=1}^{L} \left[c_{1h} n_h + c_{2h} \frac{n_h}{N_h} \sum_{i=1}^{n_h} m_{hi} \right] \tag{2}$$

If the total amount available for a multivariate stratified two stage sampling is predetermined, a compromise allocation of n', n_h and m_{hi} may be one that minimizes the weighted sum of the sampling variances of the estimates of various characteristics, that is

$$\sum_{k=1}^{p} a_k V(\overline{\overline{y}}_{k,std}) \tag{3}$$

where a_k is the weights assigned to the kth characteristic in proportion to its importance as compared to other characteristics and $V(\overline{\overline{y}}_{k,std})$ as given in (1). For the minimization, the term independent of n', n_h and m_{hi} in (3) is ignored. Also letting

$$A = \sum_{k=1}^{p} a_k S_{k,y}^2, \;\; A_h = \sum_{k=1}^{p} a_k [n_h' S_{k,hy}^2 - \sum_{i=1}^{N_h} M_{hi} S_{k,hiy}^2 \text{ and } B_{hiy}^2 = \sum_{k=1}^{p} a_k S_{k,hiy}^2 \tag{4}$$

the problem of finding the compromise allocation of n', n_h and m_{hi} for a fixed cost C_0 may be given as the following NLPP:

$$\min Z = \frac{1}{n'} \left[A + \sum_{h=1}^{L} \frac{W_h}{n_h} \left\{ A_h + \sum_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}} \right\} \right]$$

such that

$$c_0 + c_1 n' + \sum_{h=1}^{L} \left[c_{1h} n_h + c_{2h} \frac{n_h}{N_h} \sum_{i=1}^{n_h} m_{hi} \right] \leq C_0 \tag{5}$$

and n', n_h and $m_{hi} \geq 0$

$$(i = 1, 2, \ldots, N_h;\; h = 1, 2, \ldots, L)$$

where $C_0 = C - c_0$.

3. Solution

The objective function Z of the NLPP given in (5) will be minimum when the values of n', n_h and m_{hi} are large as permitted by the cost constraint. Therefore, this problem also suggest that at the optimum point the cost constraint will be active and one can use Lagrange multipliers technique to determine the optimum values of n'^*, n_h^* and m_{hi}^* considering the cost constraint as an equation and ignoring the non-negative restrictions on the variables.

The Lagrangian function ϕ is defined as

$$\phi(n', n_h, m_{hi}, \lambda) = \frac{1}{n'} \left[A + \sum_{h=1}^{L} \frac{W_h}{n_h} \left\{ A_h + \sum_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}} \right\} \right] + \lambda \left[c_1 n' + \sum_{h=1}^{L} \left[c_{1h} n_h + c_{2h} \frac{n_h}{N_h} \sum_{i=1}^{n_h} m_{hi} \right] - C_0 \right] \tag{6}$$

where λ is Lagrange multiplier.

The necessary conditions for the solution of the problem are

$$\frac{\delta\phi}{\delta n'} = -\frac{1}{n'^2}\left[A + \sum_{h=1}^{L}\frac{W_h}{n_h}\left\{A_h + \sum_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}}\right\}\right] + \lambda c_1 \tag{7}$$

$$\frac{\delta\phi}{\delta n_h} = -\frac{1}{n'}\left[\frac{W_h}{n_h^2}\left\{A_h + \sum_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}}\right\}\right] + \lambda\left\{c_{1h} + c_{2h}\frac{1}{N_h}\sum_{i=1}^{N_h} m_{hi}\right\} \tag{8}$$

$$\frac{\delta\phi}{\delta m_{hi}} = -\frac{1}{n'}\frac{W_h}{n_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}} + \lambda c_{2h}\frac{n_h}{N_h} \tag{9}$$

and

$$\frac{\delta\phi}{\delta\lambda} = \left[c_1 n' + \sum_{h=1}^{L}\left[c_{1h}n_h + c_{2h}\frac{n_h}{N_h}\sum_{i=1}^{n_h} m_{hi}\right] - C_0\right] \tag{10}$$

Multiplying by $\dfrac{m_{hi}}{n_h}$ and summing over i $(i = 1, 2, \ldots, N_h)$, (9) reduces to

$$-\frac{1}{n'}\frac{W_h}{n_h^2}\sum_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}} + \lambda c_{2h}\frac{1}{N_h}\sum_{i=1}^{N_h} m_{hi} \tag{11}$$

(8) and (11) give

$$n_h = \frac{\sqrt{W_h}}{\sqrt{\lambda}}\frac{\sqrt{A_h}}{\sqrt{c_{1h}}} \text{ provided } A_h > 0 \tag{12}$$

Substituting the values of n_h from (12) in (9), the optimum values are obtained

$$m_{hi}^* = M_{hi}B_{hiy}\sqrt{\frac{C_{ih}N_h}{C_{2h}N_h}} \tag{13}$$

For $(i = 1, 2, \ldots, N_h;\ h = 1, 2, \ldots, L)$ from equation (7)

$$n' = \frac{\sqrt{A + \sum\limits_{h=1}^{L}\dfrac{W_h}{n_h}\left\{A_h + \sum\limits_{i=1}^{N_h} M_{hi}^2 \dfrac{B_{hiy}^2}{m_{hi}}\right\}}}{\sqrt{\lambda}\sqrt{c_1}} \tag{14}$$

Substituting the value of n', n_h and $m_{h_i}^*$ from (12), (13) and (14) respectively, (10) gives

$$\frac{1}{\sqrt{\lambda}} = \frac{C_0 - c_{2h}\dfrac{n_h}{N_h}\sum\limits_{i=1}^{N_h} M_{hi}B_{hiy}^2\sqrt{\dfrac{c_{1h}N_h}{c_{2h}A_h}}}{\sqrt{c_1\left[A + \sum\limits_{h=1}^{L}\dfrac{W_h}{n_h}\left\{A_h + \sum\limits_{i=1}^{N_h} M_{h_i}^2 \dfrac{B_{hiy}^2}{m_{hi}}\right\}\right]} + \sum\limits_{h=1}^{L}\sqrt{W_h A_h c_{1h}}} \tag{15}$$

From (12) and (15) the optimum value of n_h^* is

$$n_h^* = \frac{\sqrt{W_h A_h}\left[C_0 - c_{2h}\dfrac{n_h}{N_h}\sum\limits_{i=1}^{N_h} M_{hi}B_{hiy}^2\sqrt{\dfrac{c_{1h}N_h}{c_{2h}A_h}}\right]}{\sqrt{c_{1h}}\left[\sqrt{c_1\left[A + \sum\limits_{h=1}^{L}\dfrac{W_h}{n_h}\left\{A_h + \sum\limits_{i=1}^{N_h} M_{h_i}^2 \dfrac{B_{hiy}^2}{m_{hi}}\right\}\right]} + \sum\limits_{h=1}^{L}\sqrt{W_h A_h c_{1h}}\right]} \tag{16}$$

From (14) and (15) the optimum value of n'^* is

$$n'^* = \frac{\left[\sqrt{A + \sum\limits_{h=1}^{L} \frac{W_h}{n_h}\left\{A_h + \sum\limits_{i=1}^{N_h} M_{hi}^2 \frac{B_{hiy}^2}{m_{hi}}\right\}}\right]\left[C_0 - c_{2h}\frac{n_h}{N_h}\sum\limits_{i=1}^{N_h} M_{hi}B_{hiy}^2\sqrt{\frac{c_{1h}N_h}{c_{2h}A_h}}\right]}{\sqrt{c_1}\left[\sqrt{c_1\left[A + \sum\limits_{h=1}^{L}\frac{W_h}{n_h}\left\{A_h + \sum\limits_{i=1}^{N_h} M_{h_i}^2\frac{B_{hiy}^2}{m_{hi}}\right\}\right]} + \sum\limits_{h=1}^{L}\sqrt{W_h A_h c_{1h}}\right]} \tag{17}$$

As the objective function of (4) is convex for

$$A_h = \sum_{k=1}^{p} a_k \left[n'_h S_{k,hy}^2 - \sum_{i=1}^{N_h} M_{hi} S_{k,hiy}^2\right] > 0$$

and the constraint is linear, the (K-T) necessary conditions of the NLPP (10) are sufficient also. It can be easily verified that the K-T conditions hold at the point n'^*, n_h^* and m_{hi}^* are given by (13), (16) and (17). Hence, n'^*, n_h^* and m_{hi}^* is optimum for NLPP (5).

4. Conclusion

In Section 2 we formulate the NLPP for optimum allocation in stratified two stage design by using double sampling under certain conditions. Further in Section 3 we determine the optimum values of n'^*, n_h^* and m_{hi}^* for NLPP.

References

[1] Arnold, B.F. (1986). *Procedure to determine optimum two-stage sampling plans by attributes, Metrika*, **33**, 93–109.

[2] Bakhshi, Z.H., Khan, M.F. and Ahmed, Q.S. (2010). Optimal sample numbers in multivariate stratified sampling with a probabilistic cost constraint, *International Journal of Mathematics and Applied Statistics*, 1 (2), 111–120.

[3] Clark, R. G. and Steel, D. G. (2000). Optimum allocation of sample to strata and stages with simple additional constraints, *Journal of the Royal Statistical Society*, Series D: The Statist. **49**, 197–207.

[4] Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition, New York: John Wiley and Sons, Inc., 1977.

[5] Dalenius, T. (1957). *Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice*, Alm-Qvist Och Wiksell, Stockholm, 1957.

[6] Dever, J. A; Jun, L; Iannacchione, V. G; and Kendrick, D. E. (2001). An Optimum Allocation Method for Two-stage Sampling Designs with Stratification at Second Stage. *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association (Alexandria, VA), 2001.

[7] Garcia, J. A. D. and Tapia, M. M. G. (2007). Optimum Allocation in stratified surveys: stochastics programing, *Computational Statistics and Data Analysis*, **51**(6), 3016–3026.

[8] Ghosh, S. P. (1958). A note on stratified random sampling with multiple characters, *Calcutta Statistical Association Bulletin*, **8**, 81–89.

[9] Javed, S., Bakhshi, Z. N. and Khalid, M. M. (2009). Optimum allocation in stratified sampling with random costs, *International Review of Pure and Applied Mathematics*, **5** (2), 363–370.

[10] Kokan, A. R. and Khan, S. U.. Optimum allocation in multivariate surveys: an analytical solution, *Journal of the Royal Statistical Society, Series B*, 29, 115–125.

[11] Kuhn, H. W. and Tucker, A. W. (1951). *Nonlinear Programming. Proceeding of the Second Bakley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkley, 481–492.

[12] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and methods of purposive selection, *Journal of Royal Statistical Society*, **97**(4).

[13] Prekopa, A. (1995). *Stochastic Programming, Series Mathematics and its Applications*, Kluwer Academic Publishers, Berlin.

[14] Sadooghi-Alvandi, M. (1986). The choice of subsample size in two-stage sampling, *Journal of American Statistical Assocation*, **81**, 555–558.

[15] Valliant, R. and Gentle, J. E. (1997). An application of mathematical programming to sample allocation, *Computational Statistics & Data Analysis*, **25**, 337–360.