



Optimal Allocation of Stratified Sampling Design Using Gradient Projection Method

M. A. LONE*, S. A. MIR , IMRAN KHAN and M. S. WANI

Division of Agricultural Statistics, SKUAST-K, Kashmir, India.

*Corresponding author E-mail: mushtaqstat11@gmail.com

<http://dx.doi.org/10.13005/ojcs/10.01.02>

(Received: January 26, 2017; Accepted: February 08, 2017)

ABSTRACT

This article deals with the problem of finding an optimal allocation of sample sizes in stratified sampling design to minimize the cost function. In this paper the iterative procedure of Rosen's Gradient projection method is used to solve the Non linear programming problem (NLPP), when a non integer solution is obtained after solving the NLPP then Branch and Bound method provides an integer solution.

Keywords: Stratification, Optimal allocation, Nonlinear programming, Gradient projection method, Branch and Bound method and integer allocation.

INTRODUCTION

One of the areas of statistics that is most commonly used in all fields of scientific investigation is that of stratified sampling. In order to achieve greater precision of the estimate it is desirable to decrease the heterogeneity of the population units. This is achieved by a technique known as stratification, in which the entire population is divided into a number of sub population called strata. Generally, the stratification is done according to administrative grouping, geographic regions and on the basis of auxiliary characters. These sub-population are non-overlapping and together they comprise the whole of the population. These strata are so formed that they are homogeneous within

and heterogeneous between. When the strata have been determined, a sample is drawn from each stratum, the drawing being made independently in different strata. In stratified sampling the most important consideration is the allocation of sample sizes in each stratum either to minimize the variance subject to cost or minimize cost subject to variance. The problem of optimally choosing the sample sizes is known as the optimal allocation problem. The problem of optimal allocation for univariate stratified population was first considered by Neyman (1934). In univariate stratified random sampling for one characteristics, the sample size and its allocation is cited in Cochran (1977), Sukhatme *et al.* (1984) and Thompson (1997). The allocation problem becomes more complicated in multivariate surveys

because in univariate an allocation that is optimal for one characteristic may be far from optimal for other characteristics. Some of the author who addressed the problem of obtaining a compromise allocation that suits well to all the characteristics are Ghosh (1958), Yates (1960), Aoyama (1963), Hartley (1965), Folks and Antle (1965), Gren (1966), Chatterjee (1972), Chromy (1987), Wywial (1988), Bethel (1989), kreienbrock (1993), Khan *et al.* (1997,2003). Ahsan *et al.* (2005), Kozak (2006), Ansari *et al.* (2009), Optimum allocation has been stated as a non-linear mathematical programming problem in which the objective function is the variance subject to a cost restriction or vice versa. This problem has been solved by using Lagrange's multiplier method, see Sukhatme *et al.* (1984) or the Cauchy-Schwarz inequality, see Cochran(1977) for univariate case and Arthanari and Dodge (1981) for multivariate one, both from deterministic point of view. In this paper objective is to use Rosen's Gradient projection method to determine the optimal allocation of sample sizes in stratified sampling design.

Formulation of the problem

Let us assume that there are p characteristics under study and Y_j be the j th characteristics to be considered. Suppose we have L strata, and N be the total population, N_i units in the i th stratum such that $\sum_{i=1}^L N_i = N$. Also, we assume that n_i samples are drawn independently from each stratum and y_{ij} is an unbiased estimate of Y_j such that

$$\bar{y}_{ij} = \frac{1}{n_i} \sum_{h=1}^{n_i} y_{ijh} \text{ for all } (i=1,2,3,\dots,L, j=1,2,3,\dots,p)$$

Where Y_{ijh} is the observed value for Y_j in the i th stratum for the h th sample unit.

Then $\bar{y}_j(st)$, given by

$$\bar{y}_j(st) = \frac{1}{N} \sum_{i=1}^L N_i \bar{y}_{ij}$$

is an unbiased estimate of population mean. The precision of this estimate is measured by the variance of the sample estimate of the population characteristics.

$$\begin{aligned} V(\bar{y}_j(st)) &= \frac{1}{N} \sum_{i=1}^L N_i \bar{y}_{ij} \\ &= \frac{1}{N^2} \sum_{i=1}^L N_i^2 V(\bar{y}_{ij}) \\ &= \frac{1}{N^2} \sum_{i=1}^L N_i^2 S_{ij}^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) \\ &= \sum_{i=1}^L W_i^2 S_{ij}^2 x_i \end{aligned}$$

Where,

$$W_i = \frac{N_i}{N}, S_{ij}^2 = \frac{1}{N_i - 1} \sum_{h=1}^{N_i} (y_{ijh} - \bar{Y}_{ij})^2 \text{ and } x_i = \frac{1}{n_i} - \frac{1}{N_i}$$

Let $a_{ij} = W_i^2 S_{ij}^2$.

Also, let be the cost of sampling all the p characteristics on a single unit in the i th stratum. The total variable cost of the survey assuming linearity is

$$C = \sum_{i=1}^L C_i n_i$$

Assume that

$$a_{ij}, C_i > 0, \text{ for all } (i=1,2,3,\dots,L, j=1,2,3,\dots,p)$$

Here we consider the problem of deriving statistical information on population characteristics, based on sample data, and can be formulated as an optimization problem, in which we determine optimum allocation of sample size n_i , ($i=1,2,3,\dots,L$), such that cost of the survey is minimized. The multivariate sample design and optimization has been treated as a mathematical programming problem (Arthanari and Dodge (1981)). Thus the problem of allocation can be stated as (see Sukhatme *et al.* (1984) and Arthanari and Dodge (1981)).

$$\left. \begin{aligned} \text{Minimize } & \sum_{i=1}^L C_i n_i && \dots(i) \\ \text{Subjected to } & \sum_{i=1}^L a_{ij} x_i \leq v_j && \dots(ii) \\ & 0 \leq x_i \leq 1 - \frac{1}{N_i} && \dots(iii) \end{aligned} \right\} \text{ (P1)}$$

Where V_j is the allowable error in the estimate of the j th characteristic. Here objective function as well as set of constraints is linear but the restriction (3) is non linear. The problem (P1) can also be written as

$$\begin{aligned} & \text{Minimize } \sum_{i=1}^L a_{ij}x_i \leq v_j \\ & \text{Subjected to } 0 \leq x_i \leq 1 - \frac{1}{N_i} \\ & 0 \leq x_i \leq 1 - \frac{1}{N_i} \end{aligned} \quad (P2)$$

The above problem (P1) can be equivalently written as

$$\begin{aligned} & \text{Minimize } \sum_{i=1}^L \frac{c_i}{X_i} \\ & \text{Subject to } \sum_{i=1}^L a_{ij}X_i \leq v_j \\ & \frac{1}{N_i} \leq X_i \leq 1 \end{aligned} \quad (P3)$$

Since $\frac{c_i}{X_i}$ are given, it is sufficient to minimize $\sum_{i=1}^L \frac{a_{ij}}{n_i}$.

Where, $X_i = \frac{1}{n_i}$ since $\frac{c_i}{X_i}$ is strictly convex for $c_i > 0$ because of this objective function is strictly convex and the set of constraints provides a bounded convex feasible region and an optimal solution will also exist. Dalenius (1957) proposed a graphical solution to the problem for two characteristics. Kokan and Khan (1967) have proved the existence and uniqueness of the solution and have given the optimal solution through iterative procedure. Chatterjee (1968) gives an algorithm to solve the problem. In 1960, Rosen developed the Gradient Projection method for linear constraints and later in 1961, generalized it for nonlinear constraints. It uses the projection of the negative gradient in such a way that improves the objection function and maintains feasibility. Although the method has been described by Rosen for a general non linear

programming problem, its effectiveness is confined primarily to problems in which the constraints are all linear. The procedure involved in the application of the gradient projection method can be described in the following Algorithm.

Algorithm of Gradient Projection method

Step 1: Read in $X^{(0)}$, $X^{(0)}$ is an initial feasible point.

Step 2: set the iteration as $K=0$.

Step 3: compute $S^{(k)} = -\frac{P_k \nabla f(X^{(0)})}{\|P_k \nabla f(X^{(0)})\|}$,

where $P_k = I - N_p(N_p^T N_p)^{-1} N_p^T$.

Step 4: Repeat

Step 5: if $S^{(k)} \neq 0$, then

Step 6: Make non active constraints Active,

$$g_j(X^k) + \lambda a_j^T S^{(k)} = 0$$

where $a_j = (a_{1j}, \dots, a_{nj})$

$$\lambda_M = \min(\lambda_k) = \min_j \left[\frac{-g_j(X^{(k)})}{a_j^T S^{(k)}} \right], \lambda_k > 0.$$

Step 7: compute maximum step size λ^* that minimizes

$$\phi(\lambda) = f(X^{(k)} + \lambda S^{(k)}) \quad \lambda \in [0, \alpha_M]$$

Step 8: $X^{(k+1)} = X^{(k)} + \lambda^* S^{(k)}$

Step 9: else

$$\text{Step 10: } \lambda = -(N_p^T N_p)^{-1} N_p^T \nabla f(n^{(k)})$$

Step 11: if all the components of λ are non negative then

Step 12: Optimal solution $X_{opt} = X^{(k)}$.

Step 13: else

Step 14: delete the most negative value of λ from P_k .

Step 15: end if

Step 16: end if

Step 17: until all components of are λ non-negative.

Step 18: optimal solution X^* at $X^{(k)}$.

Numerical illustration

The given data has been taken from Mohd. Vaseem Ismail *et al* (2010). The population contains 450 units, the stratum weights and stratum variance of a population which has been divided into two strata with three characteristics under study is given below in the table 1.

Assume that C (available budget) =400 units including C^o and $C^o = 300$ units (overhead cost). Therefore the total amount available for the survey is 100 units. Also we assume that the cost of measurement C_i in various strata are $C_1= 3$ and $C_2= 4$ for the cost function

$$C = C^o + \sum_{i=1}^L c_i n_i \text{ or } \sum_{i=1}^L c_i n_i = C - C^o$$

Substituting these values of the parameters into problem (P3), the allocation problem may be stated as:

$$\text{Minimize} = \frac{3}{X_1} + \frac{4}{X_2}$$

Subject to

$$\left. \begin{aligned} 0.36X_1 + 3.24X_2 &\leq 0.30 \\ 0.81X_1 + 8.12X_2 &\leq 0.60 \\ 0.09X_1 + 9.92X_2 &\leq 0.50 \\ \frac{1}{180} &\leq X_1 \leq 1 \\ \frac{1}{270} &\leq X_2 \leq 1 \end{aligned} \right\} \text{ (P4)}$$

It is also assume that, the variance of the estimate for each character can not be greater than the specified limit. i. e.

$$V_1 \leq 0.30, V_2 \leq 0.60 \text{ and } V_3 \leq 0.50$$

Solution procedure

Step 1: let the initial feasible point be

$$X^{(0)} = \begin{bmatrix} 0.38 \\ 0.036 \end{bmatrix}$$

Step 2: set the iteration number as k=0.

Step 3: Since $g_2(X^{(0)}) = 0$ for j=2, we have

$$N_1 = [\nabla g_2(X^{(0)})] = \begin{bmatrix} 0.81 \\ 8.12 \end{bmatrix}, \text{ the projection matrix is given by}$$

$$P_1 = I - N_1(N_1^T N_1)^{-1} N_1^T = \begin{bmatrix} 0.990147 & -0.098771 \\ -0.098771 & 0.009853 \end{bmatrix}$$

The search direction $S^{(0)}$ is given by

$$S^{(0)} = -\frac{P_1 \nabla f(X^{(1)})}{\|P_1 \nabla f(X^{(1)})\|} = \begin{bmatrix} -284.277845 \\ 28.358465 \end{bmatrix}$$

$$\text{As } \nabla f(X^{(0)}) = \begin{bmatrix} -20.775623 \\ 3086.419753 \end{bmatrix}$$

The normalized search direction can be obtained as

$$S^{(0)} = \frac{1}{\sqrt{(-284.277842)^2 + (28.358465)^2}} = -\begin{bmatrix} -284.277845 \\ 28.358465 \end{bmatrix} = \begin{bmatrix} -0.995061 \\ 0.099263 \end{bmatrix} \neq 0.$$

Table 1

i	N_i	$W_i = \frac{N_i}{N}$	S_{i1}^2	S_{i2}^2	S_{i3}^2	a_{i1}	a_{i2}	a_{i3}
1	180	0.40	1.5	2.25	0.75	0.36	0.81	0.09
2	270	0.60	3.0	4.75	5.25	3.24	8.12	9.92

Step 4: since $S^{(0)} \neq 0$.

Step 5: find the step length λ_M we get

$$X^{(1)} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = X^{(0)} + \lambda S^{(0)} \\ = \begin{bmatrix} 0.38 - 0.995061\lambda \\ 0.036 + 0.099263\lambda \end{bmatrix}$$

For $j=1$, $g_1(X^{(1)}) = 0, \Rightarrow \lambda_1 = -1.3300$

For $j=3$, $g_3(X^{(1)}) = 0, \Rightarrow \lambda_3 = 0.1214$

Step 6: therefore $\lambda_M = \min(-1.3300, 0.1214)$
 $= \lambda_4 = 0.1214, \lambda_k > 0$

$$f(n) = \frac{3}{X_1} + \frac{4}{X_2}$$

$$f(X^{(1)}) = \frac{3}{(0.38 - 0.995061\lambda)} + \frac{4}{(0.036 + 0.099263\lambda)}$$

$$\frac{\partial f}{\partial \lambda} = (-3) \times -0.99506(0.38 - 0.99506)^{-2} \\ + (-4) \times 0.099263(0.036 + 0.099263\lambda)^{-2}$$

$$\frac{\partial f}{\partial \lambda}(\lambda_M) = \frac{3}{(0.38 - 0.120812)^2} - \frac{4}{(0.036 + 0.048052)^2} \\ = 44.4356 - 171.9584 < 0$$

Thus some new constraints will be active at $X^{(0+1)}$ iteration.

Step 1: We obtain a new point as $x^{(1)} = x^{(1)} + \lambda S^{(0)}$

$$X^{(1)} = \begin{bmatrix} 0.2565 \\ 0.0481 \end{bmatrix}$$

Step 2 : $k+0=1$

Step 3: since $g_3(X^{(1)})$ for $j=3$, we have

$$N_1 = [\nabla g_3(X^{(1)})] = \begin{bmatrix} 0.09 \\ 9.92 \end{bmatrix}, \text{ the projection matrix}$$

is given with one more active constraints g_3 .

$$P_2 = I - N_2(N_2^T N_2)^{-1} N_2^T \\ = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The search direction $S^{(1)}$ is given by

$$S^{(1)} = -\frac{P_2 \nabla f(X^{(1)})}{\|P_2 \nabla f(X^{(1)})\|} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\text{Where } \nabla f(X^{(1)}) = \begin{bmatrix} -45.5981 \\ -1728.8999 \end{bmatrix}$$

Step 5: $S^{(1)} = 0$.

Step 6: compute the vector λ at $X^{(1)}$ as

$$\lambda = -(N_2^T N_2)^{-1} N_2^T \nabla f(X^{(1)})$$

$$\lambda = \begin{bmatrix} 40.7127 \\ 1410285 \end{bmatrix} > 0,$$

The non-negative value of λ indicates that we have reached the optimum point and the optimum value is $\text{Min } Z = 94.85$.

Step 11: The optimum allocation is

$$X_1 = 0.2565 \text{ and } X_2 = 0.0481$$

$$\Rightarrow n_1 = \frac{1}{X_1} = 3.90, n_2 = \frac{1}{X_2} = 20.79$$

Since n_1 and n_2 are required to be the integers. For practical purpose if the solution is non integer then the NLPP is solved using Branch and Bound method instead of rounding the non integer sample sizes to the nearest integral values. However in some situation for small samples the rounded off allocation may become infeasible and non optimal. In order to get the integer value we use Branch and bound method as given below for problem (p1). Using LINGO-13.0, we obtain the integer optimum allocation as $n_1 = 4$ and $n_2 = 21$ and the optimal value is 96.

CONCLUSION

This paper concludes that when a non integer solution exists after solving the NLPP then Branch and Bound method is used to obtain the integer solution of the NLPP.

REFERENCES

1. Ahsan, M. J., Najmussehar and Khan ,M. G.M. Mixed allocation in stratified sampling. *Aligrah Journal of Statistics sciences* **25**,87-97, (2005).
2. Ansari A.H., Najmussehar and Ahsan M.J. On Multiple Response Stratified Random Sampling Design, *International Journal of Statistical Sciences*, **1**, 45-54, (2009).
3. Aoyama H. Stratified Random Sampling with Optimum Allocation for Multivariate Populations, *Annals of the institute of statistical Mathematics*. **14**, 251–258, (1963).
4. Arthanari ,T.S., and Dodge, Y. Mathematical programming in statistics. New York: John Willy (1981).
5. Bethel J. Sample Allocation in Multivariate Surveys. *Survey Methodology*, **15**, 40-57, (1989) .
6. Chatterjee, S. A study of optimum allocation in multivariate stratified surveys. *Skandinavisk Actuarietidskrift* **55**,73-80, (1972).
7. Chatterjee, S. Multivariate stratified surveys. *Journal of American statistics Association*. **63**,530-535, (1968).
8. Chromy, J. R. Design Optimization with Multiple Objectives. In: Proceedings of the Survey Research Methods section. *Alexandria, VA: American Statistical Association*, (1987)194–199 .
9. Cochran W.G. Sampling Techniques, *3rd edition. John Wileyand Sons, New York* (1977).
10. Dalenius. Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice. Stockholm: Almqvist and Wicksell (1957).
11. Folks, J.K., Antle, C.E. Optimum allocation of samplingunits to the strata when there are R responses of interest. *Journal of American Statistical Association* **60**, 225-233, (1965).
12. Ghosh,S.P. A note on stratified random sampling with multiple characters. *Calcutta Statistical Bulletin* **8**, 81-89, (1958).
13. Gren , J. Some Application of Non-linear Programming in Sampling Methods, *Przegląd Statystyczny*, **13**,203–217, (1966) (in Polish).
14. Hartley H.O. Multiple purpose optimum allocation in stratified sampling, in : *Proceedings of the American Statistical Association, Social Statistics Section. Alexandria, VA: American Statistical Association*, (1965) 258-261.
15. Khan. M.G.M., and Ahsan, M. J., and Jahan, N. An optimal multivariate stratified sampling using dynamic programming .*Australian & New Zealand of statistics* **45**, 107-113, (2003).
16. Khan. M.G.M., and Ahsan, M. J., and Jahan, N. Compromise allocation in multivariate stratified sampling : An integer solution. *Naval Research Logistics* **44**, 69-79, (1997).
17. Kokan, A. R. and Khan, S. U. Optimum allocation in multivariate surveys: An analytical solution. *Journal of Royal Statistical Society, Ser. B*, **29**, 115-125, (1967).
18. Kozok M. On sample allocation in multivariate surveys, *Communication in Statistics-Simulation and Computation*, **35**, 901-910, (2006).
19. Kreienbrock, L . Generalized measures of dispersion to solve the allocation problem in multivariate stratified random sampling. *Communication in Statistics : theory and methods* **22**, 219-239, (1993).
20. Lingo 13.0, LINDO inc.ltd.
21. Mohd.Vaseem Ismail *et al.* Multi-objective convex programming problem arising in multivariate sampling. *International Journal of Engineering Science and Technology* **2**, 291-296, (2010).
22. Neyman, J. On the two Different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of Royal Statistical Society* **97**, 558-625, (1934).
23. P. V. Sukhatme, B. V. Sukhatme, Sukhatme, C. Asok. Sampling theory of Surveys with Applications. Iowa State University Press, Iowa, U. S. A. and Indian Society of Agricultural Statistics, New Delhi, India (1984).
24. Rosen, J.B. The gradient projection method For Nonlinear Programming, Part i, linear constraints. *SIAM journal of Applied*

- Mathematics* **8**, 181-219, (1960).
25. Rosen, J.B. The gradient projection method For Nonlinear Programming, Part ii, Nonlinear constraints. *SIAM journal of Applied Mathematics* **9**, 514-553, (1961).
 26. Thompson, M.E. Theory of Sample Surveys. Chapman & Hall (1997).
 27. Wywiał J. Minimizing the Spectral Radius of Means Vector from Sample Variance-covariance Matrix Sample Allocation between Strata, *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*. Wrocław, Poland, **404**, 223–235, (1988) (in Polish).
 28. Yates, F. Sampling Methods for Censuses and Surveys. 3rd ed. London: Charles Griffin (1960).