# ENHANCING PRECISION IN STRATIFIED SAMPLING USING MATHEMATICAL PROGRAMMING APPROACH

Mushtaq A. Lone[1], S. A. Mir[2], Kaisar Ahmad[3], Aafaq A. Rather[4,*], Danish Qayoom[5], S. Ramki[6]

•

[1, 2]SKUAST-Kashmir, India
[3]Department of Statistics, University of Kashmir, Srinagar, J&k, India
[4*, 5]Symbiosis Statistical Institute, Symbiosis International (Deemed University), Pune-411004, India
[6]Department of Community Medicine, Dhanalakshmi Srinivasan Institute of Medical Sciences and Hospital, Thuraiyur road, Perambalur-621212, India
[1]lonemushtaq11@gmail.com, [2]mir_98@msn.com, [3]ahmadkaisar31@gmail.com, [4*]aafaq7741@gmail.com, [5]danishqayoom11@gmail.com, [6]ramki.stat24@gmail.com

**Abstract**

*This article addresses the challenges of determining the optimal allocation of sample sizes in stratified sampling design to minimize the cost function. Researchers employed the iterative procedure of Rosen's Gradient projection method and obtained optimal allocation of non-linear programming problem through manual calculation, which are often susceptible to human errors, such as rounding or arithmetic mistakes especially for complex nonlinear programming problems. R software performs calculations with high precision and consistency. In this paper, we demonstrate how to solve the non-linear programming problem by using iterative based procedure of Rosen's Gradient projection method through R software.*

**Keywords:** Stratified random sampling**,** Optimal allocation, Gradient project method, Nonlinear programming problems

## 1. Introduction

Stratified sampling is widely utilized statistical method across various fields of scientific research, aimed enhancing the accuracy of estimates by reducing heterogeneity among population units. This is accomplished through a process known as stratification, where the entire population is segmented into distinct sub populations referred as strata. These strata are typically formed based on factors like administrative classifications, geographic locations and additional characters, ensuring they are non-overlapping and collectively they encompass the entire population. These strata are made to be homogeneous within and heterogeneous between. Once the strata established, samples are independently drawn from each stratum. The key challenge in stratified sampling is the determination of optimal allocation of sample sizes within each stratum, which can either aim to minimize the variance while adhering the cost or minimize cost while maintaining the variance. Thus, the problem of optimally selecting these the sample sizes is known as the optimal allocation problem, first addressed by Neyman [19] with further contributions by Cochran

[5], Sukhatme et al. [22] and Thompson [23]. The allocation problem of distribution becomes more difficult in in many studies, because an allocation optimal for one characteristic may not be optimal/suitable for others. Various researchers, including Wywial [24], Bethel [4], kreienbrock [13], Khan et al. [14, 15], Kozak [16], Ghosh [10], Yates [25], Aoyama [1], Hartley [12], Folks and Antle [9], Gren [11], Chatterjee [6], Ansari et al. [2], Chromy [7], have explored compromise allocations that suit multiple characteristics. The optimal allocation is characterized as a non-mathematical programming problem, the objective function being the variance subject to a cost constraint, or vice versa. This problem is solved using the Lagrange multiplier method, see Sukhatmeh et al [22] or the Cauchy-Schwarz inequality, see Cochran [5] for univariate case and Arthanari and Dodge [3] for multivariate one, both from deterministic point of view.

Dalenius [8] proposed a graphical solution for the problem involving two characteristics. Kokan and khan [17] demonstrated the existence and uniqueness of the solution and have given the optimal solution through iterative procedure. Chatterjee [6] developed an algorithm to solve the problem. In 1960, Rosen [20] developed the Gradient Projection method for linear constraints and later Rosen [21] 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. In this paper, objective to determine the optimal allocation of sample sizes using Rosen's [20, 21] Gradient projection method through R software instead of using manual calculations. Lone et al [18] employed the same iterative procedure of Rosen's [20, 21] Gradient projection method and obtained optimal allocation of non-linear programming problem through manual calculation. Manual procedure might rely on a simplified or less robust version of an optimization algorithm. Performing a sufficient number of iterations manually to reach the optimal allocation is a challenging task due to time constraints or computational limits which may sometime provide less accurate results and involves iterative calculations in case of complex problems. R software provides a more reliable and accurate approach to solving complex optimization problems, explaining the difference in optimal allocation compared to manual procedures with high precision and consistency.

## 2. Formulation of the problem

Assume that there are $p$ characteristics under study, with $Y_j$ being the $j$th characteristic considered.

$$\bar{y}_{ij} = \frac{1}{n_i}\sum_{h=1}^{n_i} y_{ijh} \quad \text{for all } i = 1, 2, 3, ..., L \text{ and } j = 1, 2, 3, ..., p \tag{1}$$

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

Then $\bar{y}_j(st) = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_{ij}$ is an unbiased estimate of population mean $\bar{Y}$.

$$V(\bar{y}(st)) = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_{ij} = \frac{1}{N^2}\sum_{i=1}^{L} N_i{}^2 V(\bar{y}_{ij}) = \sum_{i=1}^{L} W_i{}^2 S_{ij}^2 x_i \tag{2}$$

Where

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

Let $a_{ij} = W_i S_{ij}^2$. Also, let $C_i$ 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 $a_{ij}, C_{i.} > 0$ for all $i = 1, 2, 3, ..., L$ and $j = 1, 2, 3, ..., p$.

In this context, the challenge of deriving statistical information about the population characteristics using sample data, this can be framed as an optimization problem, where we aim to determine optimum allocation of sample size $n_i$ for all $i = 1, 2, 3, ..., L$ to minimize the survey cost is minimized. The multivariate sample design and its optimization are approached as a mathematical programming problem as discussed by Arthanari and Dodge [3]. Therefore, the allocation problem is defined accordingly, following the work of Sukhatme*et al*. [22] and Arthanari and Dodge [3].

$$\text{Minimize} \sum_{i=1}^{L} C_i n_i \text{ subject to } \sum_{i=1}^{L} a_{ij} x_i \le v_j \text{ and } 0 \le x_i \le 1 - \frac{1}{N_i} \tag{3}$$

Where $v_j$ is the allowable error in the estimate of the $j^{th}$ characteristics. The problem (3) can be equivalently written as

$$\text{Minimize} \sum_{i=1}^{L} \frac{C_i}{X_i} \text{ subject to } \sum_{i=1}^{L} a_{ij} X_i \le v_j \text{ and } \frac{1}{N_i} \le X_i \le 1 \tag{4}$$

Since $N_i$, are given, it is sufficient to minimize $\sum_{i=1}^{L} \frac{a_{ij}}{n_i}$. Where $X_i = \frac{1}{n_i'}$ and $\frac{C}{X_i}$ 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 exit. 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.
The formulated non- linear programming model has been taken from Lone *et al.* [18].

$$\text{Minimize} = \frac{3}{X_1} + \frac{4}{X_2} \tag{5}$$

Subject to

$$0.36 X_1 + 3.24 X_2 \le 0.30$$
$$0.81 X_1 + 8.12 X_2 \le 0.60$$
$$0.09 X_1 + 9.92 X_2 \le 0.50$$
$$\frac{1}{180} \le X_1 \le 1$$
$$\frac{1}{270} \le X_2 \le 1$$

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

$$V_1 \le 0.30,\ V_2 \le 0.60,\ \textit{and}\ V_3 \le 0.50.$$

The solution of the above NLLP is obtained from R software through Rosen's Gradient Projection method.

## 3. Results

The optimal allocation obtained through R software using Gradient Projection method is $X_1 = 7$ and $X_2 = 20$ and optimal Value of the objective function is $103$. The optimal solution using the same method through manual calculation is $X_1 = 4.0$ and $X_2 = 21$ and Value of the objective function is $96$.

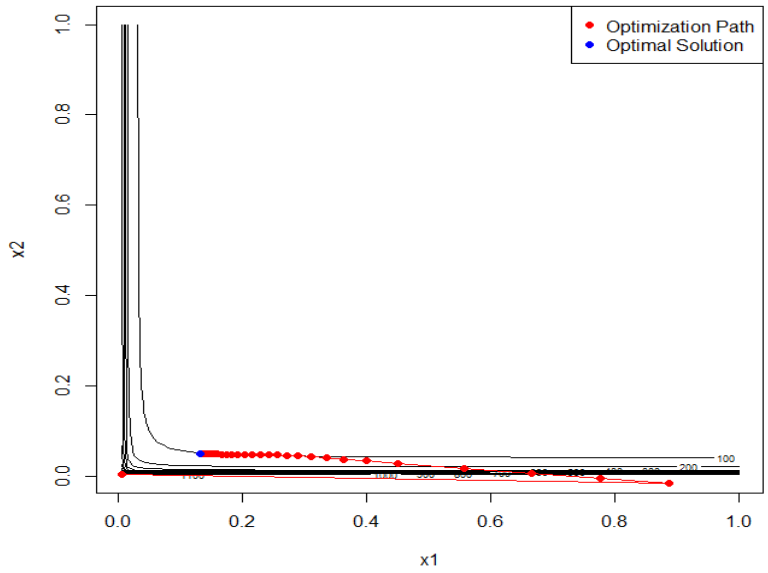


**Figure 1:** *Contour plot with optimization path*

## 4. Conclusion

This article highlights the complexities and potential for human error by calculation manually the optimal allocation of sample sizes using Gradient Project method in stratified sampling design. This study successfully demonstrates the use of R software to solve the NLPP for optimal allocation which shows significant improvements in precision and efficiency compared to manual calculation.

## References

[1] Aoyama, H. (1963). Stratified random sampling with optimum allocation for multivariate populations. *Annals of the Institute of Statistical Mathematics*, 14, 251–258.

References

[1] Aoyama, H. (1963). Stratified random sampling with optimum allocation for multivariate populations. *Annals of the Institute of Statistical Mathematics*, 14, 251–258.

[2] Ansari, A. H., Najmussehar, and Ahsan, M. J., (2009). On multiple response stratified random sampling design.*International Journal of Statistical Sciences*, 1(1), 45-54.

[3] Arthanari, T. S., and Dodge, Y. (1981). Mathematical programming in statistics. *New York, NY: John Wiley.*

[4] Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 40-57.

[5] Cochran, W. G. (1977). Sampling techniques (3rd ed.). *New York, NY: John Wiley & Sons.*

[6] Chatterjee, S. (1968). Multivariate stratified surveys. *Journal of the American Statistical Association*, 63, 530-535.

[7] Chromy, J. R. (1987). Design optimization with multiple objectives. *In Proceedings of the*

*Survey Research Methods Section, 194–199, Alexandria, VA: American Statistical Association.*

[8] Dalenius, T. (1957). Sampling in Sweden: Contributions to the methods and theories of sample survey practice. *Stockholm, Sweden: Almqvist & Wicksell.*

[9] Folks, J. K., and Antle, C. E. (1965). Optimum allocation of sampling units to the strata when there are R responses of interest. *Journal of the American Statistical Association,* 60, 225-233.

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

[11] Gren, J. (1966). Some application of non-linear programming in sampling methods. *Przeglad Statystyczny, 13, 203–217 (in Polish).*

[12] Hartley, H. O. (1965). Multiple purpose optimum allocation in stratified sampling. *In Proceedings of the American Statistical Association, Social Statistics Section, 258-261, Alexandria, VA: American Statistical Association.*

[13] Kreienbrock, L. (1993). Generalized measures of dispersion to solve the allocation problem in multivariate stratified random sampling. *Communications in Statistics: Theory and Methods,* 22(1), 219-239.

[14] Khan, M. G. M., Ahsan, M. J., and Jahan, N. (1997). Compromise allocation in multivariate stratified sampling: An integer solution. *Naval Research Logistics*, 44, 69-79.

[15] Khan, M. G. M., Ahsan, M. J., and Jahan, N. (2003). An optimal multivariate stratified sampling using dynamic programming. *Australian & New Zealand Journal of Statistics,* 45(1), 107-113.

[16] Kozok, M. (2006). On sample allocation in multivariate surveys. *Communications in Statistics - Simulation and Computation,* 35, 901-910.

[17] Kokan, A. R., and Khan, S. U. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B,* 29, 115-125.

[18] Lone, M.A., Mir, S. A., Khan, I., and Wani, M. S. (2017). Optimal allocation of stratified sampling design using Gradient Projection method. *Oriental Journal of Computer Science and Technology,* 10(1), 11-17.

[19] Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society,* 97, 558-625.

[20] Rosen, J. B. (1960). The gradient projection method for nonlinear programming, part I, linear constraints. *SIAM Journal of Applied Mathematics,* 8, 181-219.

[21] Rosen, J. B. (1961). The gradient projection method for nonlinear programming, part II, nonlinear constraints. *SIAM Journal of Applied Mathematics,* 9, 514-553.

[22] Sukhatme, P. V., Sukhatme, B. V., & Sukhatme, C. (1984). Sampling theory of surveys with applications.*Ames, IA: Iowa State University Press.*

[23] Thompson, M. E. (1997). Theory of sample surveys. *London, England: Chapman & Hall.*

[24] Wywial, J. (1988). Minimizing the spectral radius of means vector from sample variance-covariance matrix sample allocation between strata. *Prace Naukowe Akademii Ekonomicznej we Wroclawiu,*404, 223–235 (in Polish).

[25] Yates, F. (1960). Sampling methods for censuses and surveys (3rd ed.). *London, England: Charles Griffin.*