

# Obtaining Strata Boundaries under Proportional Allocation with Varying Cost of Every Unit

Faizan Danish

Division of Statistics and Computer Science

Faculty of Basic Sciences, SKUAST-Jammu, Main campus, Chatha-180009 (J&K), India

danishstat@gmail.com

S.E.H. Rizvi

Division of Statistics and Computer Science, Faculty of Basic Sciences

SKUAST-Jammu, Main campus, Chatha-180009 (J&K), India

sehrizvi\_stats@yahoo.co.in

M. Iqbal Jeelani

Division of Statistics and Computer Science, Faculty of Basic Sciences

SKUAST-Jammu, Main campus, Chatha-180009 (J&K), India

jeelani.miqbal@gmail.com

Javaid Ahmad Reashi

Department of Statistics, Govt. Degree College Anantnag

J&K Higher Education, India

reshijavid19@gmail.com

## Abstract

One of the main reasons for stratifying the population is to produce a gain in precision of the estimates, in the sample surveys. For achieving this, one of the problem is determination of optimum strata boundaries. The strata boundaries should be obtained in such a way, so that it can reasonably expect to reduce the cost of the survey as much as possible without sacrificing the accuracy or alternatively, reducing the margin of error to the greatest possible extent for the same expected cost. In this paper, we have discussed the way of obtaining optimum strata boundaries when the cost of every unit varies in the whole strata. The problem is formulated as non-linear programming problem which is solved by using Bellman's principle of optimality. For numerical illustration an example is presented for uniformly distributed study variable.

**Keywords:** Optimum strata boundaries, Non-linear programming problem, Bellman's principle of optimality.

## 1. Introduction

A stratified random sampling design is a sampling plan in which a population is partitioned into mutually exclusive strata and a sample is drawn from each stratum with simple random sampling. Dividing the population into sub-populations is known as stratification and the essential objective of this is to construct strata to allow for efficient estimation. While dealing with the stratified sampling design choosing of stratification variable, total number of strata, determination of points of stratification and selecting a sample from each stratum by using selected sampling design are some of the necessary points about whom care must be taken. The problem of determining optimum strata boundaries (OSB) is one of the famous research area in sampling literature when the probability distribution of the stratifying variable is known. The simple way involved in

determining the OSB is that the construction of strata should be done in such a way so that they are internally as homogeneous as possible. To achieve this, the distribution of the stratified variable is splitted at suitable points.

Dalenius (1950) first studied the problem of obtaining OSB and obtained minimal equations which were very difficult to solve them. Dalenius and Gurney (1951) showed that if stratification is not done in well defined way in some cases would lead us to the loss in precision. Subsequently the attempts for determining OSB have been made by several authors, such as Mahalanobis (1953), Hansen and Hurwitz (1953), Sethi (1963), Aoyama (1954), Ekman (1959), Rizvi *et al.* (2000) etc. Des R. (1964) studied the performance of equal size stratification with equal allocation by comparing it with optimum stratification when same allocation of samples in different strata is used. Yadava and Singh (1984) obtained the approximate OSB for allocations proportional to strata totals. Mehta *et al.* (1996) extended the above work for the use of ratio, regression and product estimators of population mean.

Buhler and Deutler (1975) formulated the problem of determining the OSB as an optimization problem that can be solved by dynamic programming technique. Khan *et al.* (2002) proposed the technique of determining OSB for the study variable using different frequency density functions which formulate problem as equivalent of determining optimum strata width (OSW), using non linear programming problem (NLPP) and finally solved it by dynamic programming approach. Khan *et al.* (2005) formulated the problem as mathematical programming problem (MPP) for obtaining OSB assuming exponential probability distribution of the main study variable. The stratum boundaries were obtained are optimum in the sense they minimize the sampling variance of the stratified sample under Neyman allocation. The formulated problem were separable with respect to the decision variables and is treated as a multistage decision problem and the solution procedure were developed using dynamic programming technique. Kosak and Verma (2006) proved that optimization approach used for obtaining OSB is better than classical approach. Khan *et al.* (2015) developed a technique for determining OSB when the population is skewed and compare the given results with the results obtained by using classical technique with the conclusion of superiority of mathematical programming approach over classical approach.

In the present paper a technique would be developed when the cost measurement per unit varies from stratum to stratum, that certainly influences the OSB that maximizes the precision of the estimate. The problem would be formulated ad non linear programming problem and would be solved by dynamic programming technique in which Bellman's principle of optimality would be used.

## 2. The problem of optimum strata boundaries

Let the target population be stratified into L strata based on the single study variable Y, which is of interest. If a simple random sample of size  $n_i$  is to be drawn from the  $i$ th stratum with the sample mean  $\bar{y}_i$  ( $i = 1, 2, \dots, L$ ), then by stratified sampling sample mean  $\bar{y}_{st}$  is an unbiased estimate of population mean  $\bar{Y}$ , which is equal to

$$\bar{y}_{st} = \sum_{i=1}^L W_i \bar{y}_i \quad (2.1)$$

where  $W_i = N_i / N$  is defined as the proportion of the population contained in the  $i$ th stratum.

Under proportional allocation the variance of the unbiased sample estimate is given below

$$V(\bar{y}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^L W_i S_i^2 \quad (2.2)$$

Since in this paper we are dealing with the cost, so the simplest cost function in stratified sampling can be written as

$$C = a + \sum_{i=1}^L n_i c_i$$

where 'a' represents the overhead cost,  $c_i$  represents the average cost of surveying one unit in the  $i$ th stratum and  $C$  denotes the total cost of the survey.

The problem of optimum stratification can be obtained as to find the intermediate stratum boundaries  $y_0 \leq y_1 \leq y_2 \leq \dots \leq y_{L-1} \leq y_L$ , such that the variance of the sample mean is minimum, which is

$$\begin{aligned} V(\bar{y}_{st}) &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^L W_i S_i^2 \\ &= \sum_{i=1}^L \frac{W_i S_i^2}{n_i} - \sum_{i=1}^L \frac{W_i S_i^2}{N_i} \end{aligned} \quad (2.2)$$

Subject to fixed cost factor

$$C = a + \sum_{i=1}^L n_i c_i$$

which gives an optimum allocation of size

$$n_i = \frac{C-a}{\sum_{i=1}^L W_i S_i \sqrt{C_i}} \times \frac{W_i S_i}{\sqrt{C_i}} \quad (2.3)$$

Since we are dealing with fixed cost for obtaining optimum strata boundaries (OSB). Thus, using value of  $n_i$  from (2.3) in (2.2), we get

$$V(\bar{y}_{st}) = \frac{(\sum_{i=1}^L S_i \sqrt{W_i C_i})^2}{C-a} - \sum_{i=1}^L \frac{W_i S_i^2}{N_i} \quad (2.4)$$

If the finite population correction (f.p.c) i.e  $1 - \frac{n}{N}$  is ignored, then minimizing (2.4) is equal to minimizing of

$$\sum_{i=1}^L S_i \sqrt{W_i C_i} \quad (2.5)$$

If  $y_0$  and  $y_L$  denote the smallest and largest values of stratification variable  $Y$ , then for determining the OSB in the range

$$y_L - y_0 = g \quad (2.6)$$

Let  $f(y)$  being the probability distribution of the study variable  $Y$  and  $(y_{i-1}, y_i)$  denotes the boundaries of the  $i$ th stratum. Then the weight  $W_i$ , Mean ( $\mu_i$ ) and Variance ( $S_i^2$ ) of the  $i$ th stratum can be obtained as

$$W_i = \int_{y_{i-1}}^{y_i} f(y) dy \quad (2.7)$$

$$S_i^2 = \frac{1}{W_i} \int_{y_{i-1}}^{y_i} y^2 f(y) dy - \mu_i^2 \quad (2.8)$$

$$\text{where } \mu_i = \frac{1}{W_i} \int_{y_{i-1}}^{y_i} y g(y) dy \quad (2.9)$$

When the frequency function  $f(y)$  is known (2.5) can be expressed as the function of boundary points of the  $i$ th stratum as

$$f_i(y_i, y_{i-1}) = S_i \sqrt{W_i C_i} \quad (2.10)$$

Define  $V_i$ , the width of the  $i^{\text{th}}$  stratum, as

$$V_i = y_i - y_{i-1}, \quad i = 1, 2, \dots, L \quad (2.11)$$

where  $V_i \geq 0$

Using (2.11) we can express (2.6) as

$$\begin{aligned} \sum_{i=1}^L V_i &= \sum_{i=1}^L (y_i - y_{i-1}) \\ &= y_L - y_0 = g \end{aligned} \quad (2.12)$$

Keeping in view the above results we can express the  $j$ th stratification point as

$$\begin{aligned} y_j &= y_0 + V_1 + \dots + V_j, \quad j = 1, 2, \dots, L-1 \\ &= y_{j-1} + V_j \end{aligned} \quad (2.13)$$

Thus, the problem of determining OSB can now be considered as the problem of determining optimum strata widths (OSW) as the following MPP

$$\begin{aligned} &\text{Minimize} && \sum_{i=1}^L f_i(y_i, y_{i-1}) \\ &\text{Subject to constraint} && \\ &\text{and} && V_i \geq 0, \quad i = 1, 2, \dots, L \end{aligned} \quad (2.14)$$

For  $i = 1$ , the term  $f_1(y_1, y_0)$  is the objective function of MPP (2.14) which is a function of  $V_1$  only as  $y_0$  is known. Similarly the second term  $f_2(y_2, y_1) = f_2(y_2, y_0 + V_1)$  is a function of  $V_2$  alone once  $V_1$  is known. Thus, stating the objective function as a function of  $V_i$  alone we may replace MPP (2.14) as

$$\begin{aligned} &\text{Minimize} && \sum_{i=1}^L f_i(V_i) \\ &\text{Subject to constraint} && \\ &\text{and} && V_i \geq 0 \\ &&& \text{where } i = 1, 2, \dots, L \end{aligned} \quad (2.15)$$

### 3. Determination of OSB under uniform study variable

Let the study variable  $Y$  follows uniform distribution with probability density function (pdf) given as

$$f(y) = \begin{cases} \frac{1}{b-a}, & a \leq y \leq b \\ 0, & \text{otherwise} \end{cases} \quad (3.1)$$

where 'a' represents location parameter and b-a denotes the scale parameter.

Note that here  $y_0 = a$  and  $y_L = b$

By using pdf (3.1), we can obtain the values of equations (2.7) and (2.8) as

$$W_i = \frac{V_i}{b-a} \quad (3.2)$$

$$\text{and } S_i^2 = \frac{V_i^2}{12} \quad (3.3)$$

Using (3.2) and (3.3), the problem of determining OSB when the probability distribution of the main study variable  $Y$  is given as (3.1), then the MPP would take the form as

$$\begin{aligned} &\text{Minimize} && \sum_{i=1}^L \sqrt{\frac{V_i^3 c_i}{12(b-a)}} \\ &\text{Subject to constraint} && \sum_{i=1}^L Z_i = g \\ &\text{and} && V_i \geq 0 \\ &&& \text{where } i = 1, 2, \dots, L \end{aligned} \quad (3.4)$$

#### 4. The solution procedure

Let us consider the following sub-problem of (2.15) for first  $k$  strata

$$\begin{aligned} &\text{Minimize} && \sum_{i=1}^k f_i(V_i) \\ &\text{Subject to constraint} && \sum_{i=1}^k V_i = g_k \\ &\text{and} && V_i \geq 0, \quad i = 1, 2, \dots, k \end{aligned} \quad (4.1)$$

where  $g_k < g$  is the total width available of the division into  $k$  strata.

Note that

$$g_k = g \text{ for } k = L$$

Also

$$\begin{aligned} g_k &= V_1 + V_2 + \dots + V_K \\ g_{k-1} &= V_1 + V_2 + \dots + V_{K-1} \\ &= g_k - V_k \\ &\vdots \\ g_2 &= V_1 + V_2 = g_3 - V_3 \\ g_1 &= V_1 \\ &= g_2 - V_2 \end{aligned}$$

If  $f(k, g_k)$  denotes the minimum value of the objective function (4.1), then

$$\begin{aligned} f(k, g_k) &= \text{Min} \left[ \sum_{i=1}^k f_i(V_i) / \sum_{i=1}^k V_i = g_k \right] \\ \text{and} & \quad V_i \geq 0, \quad i = 1, 2, \dots, k \end{aligned}$$

With this definition of  $f(k, g_k)$  the recurrence relations of the dynamic programming takes the form as

$$g(k, t_k) = \text{Min}_{0 \leq V_k \leq g_k} [f_k(V_k) + f(k-1, g_k - V_k)], \quad k \geq 2 \quad (4.2)$$

for  $k=1$ , using (4.1)

$$\begin{aligned} f_1(1, g_1) &= f_1(g_1) \\ V_1 &= g_1 \end{aligned} \quad (4.3)$$

From  $f(L, g)$  the optimum width of  $i$ th stratum,  $V_L$ , is obtained from  $f(L-1, g-V_L)$  the optimum width of  $(L-1)$ th stratum,  $V_{L-1}$ , is obtained and so on till we obtain  $V_1$ .

Using (4.2) and (4.3) the recurrence relation for MPP (3.4) are given as

for  $k = 1$

$$f(1, g_1) = \sqrt{\frac{g_1^3 C_1}{12(b-a)}} \quad \text{at } V_1 = g_1 \quad (4.4)$$

Similarly, for  $j$ th stage use (4.2) we would get the results.

## 5. Numerical Illustration

Solving the recurrence relations (4.2) and (4.4) to seek optimum strata widths  $V_j$  ( $j = 1, 2, \dots, L$ ) for uniform study variable with density function defined in (3.1) by taking  $[a, b] = [1, 2]$  and assuming  $g = 20$  and  $C_1 = 2, C_2 = 3, C_3 = 4, C_4 = 5, C_6 = 7$  table 1 gives optimum values of  $V_i, y_i$  and  $\sum_{i=1}^L f_i(V_i)$  below:

**Table I: Optimum Strata Width, Optimum Strata Boundaries and Optimum Value Of The Objective Function For Uniform Study Variable**

No. of strata	OSW	OSB	Optimum value of objective function $\sum_{i=1}^L S_i \sqrt{W_i S_i}$
L	$V_i^*$	$y_i^* = y_{i-1}^* + Z_h^*$	
2	$Z_1^* = 0.5000$ $Z_2^* = 0.5000$	$y_1^* = y_0 + Z_1^* = 0.5000$	1.063
3	$Z_1^* = 0.3333$ $Z_2^* = 0.3333$ $Z_3^* = 0.3333$	$y_1^* = y_0 + Z_1^* = 0.3333$ $y_2^* = y_1^* + Z_2^* = 0.6666$	1.017
4	$Z_1^* = 0.2500$ $Z_2^* = 0.2500$ $Z_3^* = 0.2500$ $Z_4^* = 0.2500$	$y_1^* = y_0 + Z_1^* = 0.2500$ $y_2^* = y_1^* + Z_2^* = 0.5000$ $y_3^* = y_2^* + Z_3^* = 0.7500$	1.006
5	$Z_1^* = 0.2000$ $Z_2^* = 0.2000$ $Z_3^* = 0.2000$ $Z_4^* = 0.2000$ $Z_5^* = 0.2000$	$y_1^* = y_0 + Z_1^* = 0.2000$ $y_2^* = y_1^* + Z_2^* = 0.4000$ $y_3^* = y_2^* + Z_3^* = 0.6000$ $y_4^* = y_3^* + Z_4^* = 0.8000$	1.004
6	$Z_1^* = 0.1666$ $Z_2^* = 0.1666$ $Z_3^* = 0.1666$ $Z_4^* = 0.1666$ $Z_5^* = 0.1666$ $Z_6^* = 0.1666$	$y_1^* = y_0 + Z_1^* = 0.1666$ $y_2^* = y_1^* + Z_2^* = 0.3333$ $y_3^* = y_2^* + Z_3^* = 0.4999$ $y_4^* = y_3^* + Z_4^* = 0.6665$ $y_5^* = y_4^* + Z_5^* = 0.8332$	1.003

## Conclusion

It is concluded from the above table that the optimum values goes on decreasing way on increasing the number of strata. Also available methods present in the literature such as Aoyama (1954), Ekman (1959), Sethi (1963) etc. are mostly classical methods by which

we obtain approximately optimum strata boundaries. Many authors have suggested iterative procedures, but there is no guarantee that they will give us a global minimum in the absence of initial solution and variance function have more than one local minimum. The advantage of developed method is that it gives global minimum and doesn't require any kind of initial approximate solution while dealing with varying cost of each stratum.

## **References**

1. Aoyama, H. (1954). A study of stratified random sampling. *Ann. Ins. Stat. Math.* 6, 1-36.
2. Bellman, R. (1957). *Dynamic programming*. Princeton University Press Princeton, New Jersey.
3. Buhler, W. and Deutler, T. (1975). Optimum stratification and grouping by dynamic programming. *Metrika*, 22, 121-175.
4. Cochran, W.G. (1977). *Sampling technique*, 3rd ed. New York: John Wiley and Sons, Inc.
5. Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*. 33, 203-213.
6. Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*. 34, 133-148.
7. Ekman, G. (1959). Approximate expression for the conditional mean and variance over small intervals of a continuous distribution. *Annals of Mathematical Statistics*, 30, 1131-1134.
8. Hansen, M. H., Hurwitz, W. N. and Madow, W.G. (1953). *Sample survey methods and theory*. Vol. I & II, John Wiley and Sons, Inc., New York.
9. Khan, E. A. Khan, M. G. M. and Ahsan, M. J. (2002). Optimum stratification: A mathematical programming approach. *Calcutta Statistical Association Bulletin*. 52, 323-333.
10. Khan, M.G.M., Najmussehar and Ahsan, M.J. (2005). Optimum stratification for exponential study variable under Neyman allocation. *Journal of the Indian Society of Agricultural Statistics*. 59(2), 146-150.
11. Khan, M.G.M., Reddy, K.G. and Rao, D.K. (2015). Designing stratified sampling in economic and business surveys. *Journal of applied statistics*. 42(10), 2080-209.
12. Kozak, M. and Verma, M. R. (2006). Geometric versus optimization stratification: a comparison of efficiency. *Survey Methodology*. 32(2), 157-163.
13. Mahalanobis, P.C. (1953). Some aspect of design of sample surveys. *Sankhya*, 12, 1-17.
14. Mehta, S., Singh, R. and Kishore (1996). On optimum stratification for allocation proportional to strata totals. *Journal of Indian Statistical Association*. 34, 9-19.
15. Rizvi, S.E.H., Gupta, J.P. and Singh, R. (2000). Approximately optimum stratification for two study variables using auxiliary information. *Journal of the Indian Society of the Agricultural Statistics*. 53 (3), 287-298.

16. Sethi, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*. 5, 20-23.
17. Yadava, S.S. and Singh, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communication in statistics-theory and methods*, 13(22), 2793-2806.