Danish, Faizan

Article

A MATHEMATICAL PROGRAMMING APPROACH FOR OBTAINING OPTIMUM STRATA BOUNDARIES USING TWO AUXILIARY VARIABLES UNDER PROPORTIONAL ALLOCATION

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Danish, Faizan (2018) : A MATHEMATICAL PROGRAMMING APPROACH FOR OBTAINING OPTIMUM STRATA BOUNDARIES USING TWO AUXILIARY VARIABLES UNDER PROPORTIONAL ALLOCATION, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 19, Iss. 3, pp. 507-526, https://doi.org/10.21307/stattrans-2018-028

This Version is available at:
https://hdl.handle.net/10419/207911

STATISTICS IN TRANSITION new series, September 2018
Vol. 19, No. 3, pp. 507–526, DOI 10.21307/stattrans-2018-028

A MATHEMATICAL PROGRAMMING APPROACH FOR OBTAINING OPTIMUM STRATA BOUNDARIES USING TWO AUXILIARY VARIABLES UNDER PROPORTIONAL ALLOCATION

Faizan Danish[1]

ABSTRACT

Optimum stratification is the method of choosing the best boundaries that make the strata internally homogenous. Many authors have attempted to determine the optimum strata boundaries (OSB) when a study variable is itself a stratification variable. However, in many practical situations fetching information regarding the study variable is either difficult or sometimes not available. In such situations we find help in the variable (s) closely related to the study variable. Using auxiliary information many authors have formulated the problem as a MPP by redefining the problem as the problem of optimum strata width, and developed a solution procedure using dynamic programming technique. By using many distributions they worked out the optimum strata boundary points for the population under different allocation. In this paper, under proportional allocation OSBs are determined for the study variable using two auxiliary variables as the basis of stratification with uniform, right-triangular, exponential and lognormal frequency distribution by formulating the problems which are executed by using dynamic programming. Empirical studies are presented to illustrate the computation details of the solution procedure and its comparison with the existing literature.

Key words: optimum stratification, multistage decision problem, mathematical programming problem.

Mathematical Classification: 62D05

1. Introduction

Stratified random sampling is the most commonly used sampling technique for estimating population parameters with greater precision in sample surveys. In order to use the stratified random sampling the sample needs to choose the best boundary points such that the strata internally homogenous and the variance of the estimator within the strata be as small as possible. However, when the single characteristic is under study and its frequency distribution is known, one could use this information effectively to achieve the best boundary strata boundaries.

[1] Division of Statistics and Computer Science, Faculty of Basic Sciences, SKUAST-Jammu, Main campus, Chatha-180009 (J&K), India. E-mail: danishstat@gmail.com.

If in many situations the frequency distribution of the study variable is unknown, it may be approximated from the past experience or using some prior knowledge obtained in a recent study. This problem was pioneered by Dalenius (1950) and he obtained a set of minimal equations that could be solved for obtaining the optimum stratification points. However, the equations so obtained could not be solved provided the number of strata is small. Since then, several steps have been made for obtaining the stratification points such as Dalenius and Gurney (1951), Mahalanobis (1952), Aoyama (1954), Dalenius and Hodges (1959), Singh and Sukhatme (1969, 1973), Singh (1977), etc. Most of the authors suggested different approaches and obtained the calculus equations in terms of stratum mean and stratum variance for determining the strata boundaries.

Buhler and Deutler (1975) formulated the problem of optimum strata boundaries (OSB) as an optimization problem and developed a computational technique to solve the problem using dynamic programming. Khan *et al*. (2002, 2008) applied their procedure to determine OSB to the population various distributions. Danish et al. (2017a) made an attempt to present all the developed methods introduced for construction of stratification points using mathematical programming technique. Also, Danish *et al*. (2017b) proposed a method for determining OSB for single study variable with one auxiliary variable when the cost of every unit varies in the whole strata.

In this study, a procedure has been produced for constructing stratification points under proportional allocation for two auxiliary variables with uniform, exponential, right triangular and lognormal distributions.

2. Formulation of problem

Let us assume we have a population consisting of 'N' units stratified into L×M strata on the basis of two auxiliary variables 'X' and 'Z' when the estimation of the mean of the study variable 'Y' is of interest. We divide the whole population into the L×M (say) number of strata, such that each stratum is homogenous within itself and heterogeneous between strata with respect to the character under study such that the number of units in the (h, k)th stratum is N_{hk}, so that $\sum_{h=1}^{L}\sum_{k=1}^{M} N_{hk} = N$.

A sample of size n_{hk} (h=1,2,...,L; k=1,2,...,M) is to be drawn from each such that $\sum_{h}\sum_{k} n_{hk} = n$. The population unit in the (h , k)th stratum can be expressed as $Y = \sum_{h}\sum_{k}\sum_{i} y_{hki}$. We know, under stratified random sampling, the unbiased estimator of the population mean $\overline{Y}_N$ is

$$\overline{y}_{st} = \sum_{h}\sum_{k} W_{hk}\overline{y}_{hk}$$

where $W_{hk} = \frac{N_{hk}}{N}$ denotes the weight of the (h, k)th stratum and $\bar{y}_{hk} = \frac{1}{n_{hk}} \sum_i y_{hki}$

However, for an unbiased estimator $\bar{y}_{st}$ we have

$$V\left(\bar{y}_{st}\right) = \sum_h \sum_k \left(\frac{1}{n} - \frac{1}{N}\right) W_{hk}^2 \sigma_{hky}^2$$

where σ_{hky}^2 is the variance for the $\left(h,k\right)^{th}$ stratum (h = 1, 2, ..., L; k = 1,2,..., M). If finite population is ignored (fpc), we have

$$V\left(\bar{y}_{st}\right) = \sum_h \sum_k \frac{W_{hk}^2 \sigma_{hky}^2}{n}$$

Since 'n' is constant, thus it is sufficient to minimize

$$V\left(\bar{y}_{st}\right) = \sum_h \sum_k W_{hk}^2 \sigma_{hky}^2 \tag{2.1}$$

Let us assume the regression model of the study variable on auxiliary variables is of the form as:

$$Y = \lambda\left(x,z\right) + \varepsilon \tag{2.2}$$

where $\lambda\left(x,z\right)$ is a linear or non-linear function of 'X' and 'Z' and 'ε' denotes the error term such that its conditional expectation is zero and variance is finite and equal to $\phi\left(x,z\right)$ for all x and z.

For (h,k)th stratum the mean 'μ_{hky}' and the stratum variance 'σ_{hky}^2' can be written as

$$\mu_{hky} = \mu_{hk\lambda}$$

and

$$\sigma_{hky}^2 = \sigma_{hk\lambda}^2 + \mu_{hk\phi} \tag{2.3}$$

where $\mu_{hk\phi}$ are the expected values of $\phi\left(x,z\right)$ and $\mu_{hk\lambda}$ & $\sigma_{hk\lambda}^2$ denote the mean variance of $\lambda(x,z)$ in the (h, k)th stratum.

If 'λ' and 'ε'are uncorrelated, then in the model (2.2) then 'σ^2_{hky}' can be expressed as

$$\sigma^2_{hky} = \sigma^2_{hk\lambda} + \sigma^2_{hk\varepsilon}$$

where $\sigma^2_{hk\varepsilon}$ is the variance of error term in (h, k)th stratum.

Let the joint density function of (Y, X, Z) in the super population be f(y, x, z) and let f(x,z) be the joint function of X and Z, and f(x) & f(z) be the frequency function of the auxiliary variables X and Z, respectively, defined in the interval [a, b] and [c, d].

For determining the strata boundaries is to cut up the ranges $d_x = b - a$ and $t_z = d - c$, at (L-1) and (M-1) intermediate points as $a = x_0 \le x_1 \le ... \le x_{L-1} \le x_L = b$ and $c = z_0 \le z_1 \le ... \le z_{M-1} \le z_M = d$, respectively, such that the equation (2.1) is minimum.

Thus, while using (2.3), we have

$$\sum_h \sum_k W^2_{hk} \left(\sigma^2_{hk\lambda} + \mu_{hk\phi} \right) \tag{2.4}$$

$W_{hk}, \sigma^2_{hk\lambda}$ and $\mu_{hk\phi}$ can be obtained as a function of boundary points $\left(x_{h-1}, x_h, z_{k-1}, z_k \right)$ if $f\left(x,z\right), \lambda\left(x,z\right)$ and $\phi\left(x,z\right)$ are known and also integrable. Then, by using the following expression

$$W_{hk} = \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} f\left(x,z\right) \partial x \partial z \tag{2.5}$$

$$\sigma^2_{hk\lambda} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} \lambda^2\left(x,z\right) f\left(x,z\right) \partial x \partial z - \mu^2_{hk\lambda} \tag{2.6}$$

and $$\mu_{hk\phi} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} \phi\left(x,z\right) f\left(x,z\right) \partial x \partial z$$

where $\mu_{hk\lambda} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} \lambda\left(x,z\right) f\left(x,z\right) \partial x \partial z$ and $\left(x_h, x_{h-1}\right)$ & $\left(z_k, z_{k-1}\right)$

Thus, the objective function (2.4) could be expressed as the function of boundary points $\left(x_{h-1}, x_h, z_{k-1}, z_k \right)$ only.

Let $$\phi_{hk}\left(x_h, x_{h-1}, z_k, z_{k-1}\right) = W^2_{hk}\left(\sigma^2_{hk\lambda} + \mu_{hk\phi}\right) \tag{2.7}$$

and the ranges as:

$$d_x = b - a = x_L - x_0 \tag{2.8}$$

$$t_z = d - c = z_M - z_0 \tag{2.9}$$

Then, in the bivariate stratification the problem of determining the strata boundaries (x_h, z_k) is to break up the ranges of (2.8) and (2.9) at intermediate points. Then, the reasonable criterion for determining optimum strata boundaries (OSB) (x_h, z_k) is to minimize

Minimize $\sum_h \sum_k \phi_{hk}(x_h, x_{h-1}, z_k, z_{k-1})$

Subject to

$$a = x_0 \le x_1 \le ... \le x_{L-1} \le x_L = b$$
$$c = z_0 \le z_1 \le ... \le z_{M-1} \le z_M = d \tag{2.10}$$

and

$$\sum_h \sum_k n_{hk} = n$$

Let $V_h = x_h - x_{h-1}$ and $U_k = z_k - z_{k-1}$ denote the total length or width of the (h, k)[th] stratum for rectangular stratification. Then, using (2.8) and (2.9), the ranges can be expressed as

$$\sum_h V_h = d_x \tag{2.11}$$

$$\sum_k U_k = t_z \tag{2.12}$$

The objective function in (2.3) suggests that, for determining two way stratification, a two-dimensional dynamic programming approach should be used, employing the general concept of dynamic programming with the state and decision variables by the pairs (h, k). Then, the problem of two-way optimum stratification can be expressed as to

Minimize $\sum_h \sum_k \phi_{hk}(x_h, x_{h-1}, z_k, z_{k-1})$

Subject to (2.13)

$$(x_h, z_k) = (x_{h-1} + V_h, z_{k-1} + U_k)$$
$$(x_h, z_k) \in [a, d] \times [c, d]$$
$$(V_h, U_k) \in B_h(x_{h-1}) \times B_k(z_{k-1})$$
$$= [0, b - x_{h-1}] \times [0, d - z_{k-1}]$$
$$(x_0, z_0) = [a, c]$$
$$h = 1, 2, ..., L \quad and \quad k = 1, 2, ..., M$$

We propose a simple approach which permits a solution to the problem (2.13) using the unidimensional dynamic programming iteratively. Before the first iteration, some trail values, say x_0 and z_0, such that $a = x_0 \le x_1 \le ... \le x_{L-1} \le x_L = b$ and $c = z_0 \le z_1 \le ... \le z_{M-1} \le z_M = d$ are chosen for the initial points of the stratification. Then, for the i[th] iteration (i=1, 2, ...) the points of stratification z^{i-1} are first considered as fixed. Note that the points of stratification x^{i-1}could also be chosen instead of z^{i-1}. Fixing the values of z^{i-1} has in fact the effect of reducing the problem exactly to the one of two-way optimum stratification with one categorical stratification variable. This can be seen by comparing the formulation (2.13) to the one which is defined on univariate auxiliary variable used as stratification variable with the values of the points of stratification Z taken as constant in (2.13).

Let $\phi^*_{x_h}\left(x_{h-1}, z^{i-1}\right)$ be the optimal value for the objective function (2.10) for the strata (h, k) to (L, k) for all k = 1, 2, ..., M given that the lower bound for the strata (h, k) for k = 1, 2, ..., M is x_{h-1}.The functional equation of Bellman with respect to the first part of the i[th] iteration is then given by

$$\phi^*_{x_h}\left(x_{h-1}, z^{i-1}\right) = \underset{V_h \in B_h\left(x_{h-1}\right)}{Minimize} \left\{ \sum_{k=1}^{M} \phi\left(x_{h-1}, x_h, z^{i-1}_{k-1}, z^{i-1}_{k}\right) + \phi^*_{x_{h+1}}\left(x_h, z^{i-1}\right) \middle| x_h = x_{h-1} + V_h \right\}$$

where $B_h\left(x_{h-1}\right)$ is defined in (5.1.17).

Restating the problem of determining OSB as the problem of determining optimum points $\left(V_h, U_k\right)$, adding equation (2.11) and (2.12) as a constraint, the problem (2.10) can be treated as an equation problem of determining Optimum Strata Width (OSW), $V_1, V_2, ..., V_L \quad and \quad U_1, U_2, ..., U_M$, and expressed as the following Mathematical Programming Problem (MPP):

Minimize $\sum_h \sum_k \phi_{hk}\left(x_h, x_{h-1}, z_k, z_{k-1}\right)$

Subject to (2.14)

$$\sum_h V_h = d_x$$

$$\sum_k U_k = t_z \text{ , h = 1, 2, ..., L and k = 1, 2, ..., M}$$

and

$$V_h \ge 0 \quad and \quad U_k \ge 0$$

Therefore, the first term $\phi_{11}(x_1, x_0, z_1, z_0)$ in the objective function (2.14) is the function of (V_1, U_1) alone as (x_0, z_0) are initially known, once the (V_1, U_1) is known. The second term $\phi_{22}(x_2, x_1, z_2, z_1)$ will be the function of (V_2, U_2) alone, and so on. Due to the special nature of function, MPP (2.14) may be treated as the function of (V_h, U_k) and can be expressed as

Minimize $\sum_h \sum_k \phi_{hk}(V_h, U_k)$

Subject to (2.15)

$$\sum_h V_h = d_x$$

$$\sum_k U_k = t_z, \text{ h=1, 2, ..., L and k=1,2,...,M}$$

and $V_h \geq 0 \quad and \quad U_k \geq 0$

3. Proportional Allocation

Proportional allocation was originally proposed by Bowley (1926), which is very common in practice because of its simplicity, when no other information other than N_{hk}, which denotes the total number of units in the $(h,k)^{th}$ stratum, is available, the allocation of a given sample size 'n' to different strata is done in proportion to their sizes, i.e. in the $(h,k)^{th}$ stratum

$$n_{hk} = \frac{n}{N} N_{hk}$$

This means that the sampling fraction is the same in all strata. It gives a self-weighing sample by which numerous estimates can be made with greater speed and a higher degree of precision.

Under proportional allocation the variance is given by

$$V(\bar{y}_{st}) = \frac{(1-f)}{n} \sum_h \sum_k W_{hk} \sigma^2_{hky}$$

where $f = \frac{n}{N}$ is sampling fraction. If the finite population correction is ignored, we get

$$V(\bar{y}_{st}) = \frac{1}{n} \sum_h \sum_k W_{hk} \sigma^2_{hky}$$

Minimizing this function is equivalent to minimizing

$$\sum_h \sum_k W_{hk} \sigma^2_{hky} \tag{3.1}$$

Using the same procedure as discussed in the case of general and equal allocation, we need to replace the equation the objective function by $\sum_h \sum_k W_{hk} \sigma^2_{hky}$, Thus, MPP that we have to minimize is

Minimize $\sum_h \sum_k W_{hk} \sigma^2_{hky}$

Subject to

$$\begin{aligned} &\sum_h V_h = d_x \\ &\sum_k U_k = t_z \\ &\forall V_h \geq 0, U_k \geq 0 \quad , \quad \begin{matrix} h = 1,2,...,L \\ k = 1,2,...,M \end{matrix} \end{aligned} \tag{3.2}$$

4. The solution procedure

The problem (2.15) is a problem of multistage decision in which the objective function and the constraints are separable functions of (V_h, U_k), which allows us to use a dynamic programming technique, and a dynamic programming model is generally a recursive equation. These recursive equation links to different stages of the problem.

Consider the following sub-problem of equation (2.15) for first $(L_1 \times M_1)$ strata, where $(L_1 \times M_1) \leq (L \times M)$, i.e. $L_1 < L, M_1 < M$

Minimize $\sum_{h=1}^{L_1} \sum_{k=1}^{M_1} \phi_{hk}(x_{h-1}, x_h, z_{k-1}, z_k)$

Subject to (4.1)

$$\sum_{h=1}^{L-1} V_h = d_{L_1}$$

$$\sum_{k=1}^{M-1} U_k = t_{M_1} \qquad \text{, h=1, 2, ..., } L_1 \text{ and k =1, 2, ..., } M_1$$

and $V_h \geq 0 \quad and \quad U_k \geq 0$

where $d_{L_1} < d_x, t_{M_1} < t_z$

Note: If $d_{L_1} = d_x$ and $t_{M_1} = t_z$ then $(L_1 \times M_1) = (L \times M)$

The transformation functions are given by

$$d_{L_1} = V_1 + V_2 + ... + V_{L_1}$$

$$d_{L_1-1} = V_1 + V_2 + ... + V_{L_1-1} = d_{L_1} - V_{L_1}$$

$$\vdots$$

$$d_1 = V_1 = d_2 - V_2$$

Similarly, we have

$$t_{M_1} = U_1 + U_2 + ... + U_{M_1}$$

$$t_{M_1-1} = U_1 + U_2 + ... + U_{M_1-1} = t_{M_1} - U_{M_1}$$

$$\vdots$$

$$t_1 = U_1 = t_2 - U_2$$

Let the minimum value of the objective function of the equation (4.1) be denoted as

$$\phi_{L_1 \times M_1}\left(d_{L_1}, t_{M_1}\right) = Min\left[\sum_{h=1}^{L_1-1}\sum_{k=1}^{M_1-1} \phi_{hk}\left(V_h, U_k\right) \middle| \sum_{h=1}^{L_1-1} V_h = d_{L_1-1}, \sum_{k=1}^{M_1-1} U_k = t_{M_1-1}\right] = A_1$$

$$and \qquad V_h \geq 0, U_k \geq 0; h = 1,2,3,...,L_1 \quad ; \quad k = 1,2,3,...,M_1$$

with the above definition of $\phi_{L_1\times M_1}\left(V_{L_1},U_{M_1}\right)$, MPP (2.15) is equivalent to finding $\phi_{L\times M}\left(d_x,t_z\right)$ recursively by defining $\phi_{L_1\times M_1}\left(V_{L_1},U_{M_1}\right)$ for $L_1=1,2,...,L$ and $M_1=1,2,...,M$; $0\le d_{L_1}\le V, 0\le t_{M_1}\le U$.

$$\phi_{L_1\times M_1}\left(d_{L_1},t_{M_1}\right) = Min \left[A_1+\left[\sum_{h=1}^{L_1-1}\sum_{k=1}^{M_1-1}\phi_{hk}\left(V_h,U_k\right)\middle|\sum_{h=1}^{L_1-1}V_h=d_{L_1}-V_{L_1},\ \sum_{k=1}^{M_1-1}U_k=t_{M_1}-U_{M_1}\right]\right] \quad (4.2)$$

$$and \quad V_h\ge 0, U_k\ge 0; h=1,2,3,...,L_1 \quad and \quad k=1,2,3,...,M_1$$

For fixed value of $\left(V_{L_1},U_{M_1}\right)$, $0\le d_{L_1}\le V$, $0\le t_{M_1}\le U$.

$$\phi_{L_1\times M_1}\left(d_{L_1},t_{M_1}\right)=A_1+Min\left[\sum_{h=1}^{L_1-1}\sum_{k=1}^{M_1-1}\phi_{hk}\left(V_h,U_k\right)\middle|\sum_{h=1}^{L_1-1}V_h=d_{L_1}-V_{L_1},\ \sum_{k=1}^{M_1-1}U_k=t_{M_1}-U_{M_1}\right]$$

and

$$V_h\ge 0 \quad ,h=1,2,...,L_1, U_k\ge 0 \quad ,k=1,2,...,M_1, 1\le L_1\le L \quad ,1\le M_1\le M$$

Using the same procedure to write the forward recursive equation of the dynamic programming technique and could obtain OSB.

Let the estimation variable and the stratification variables take the regression model defined in (2.2) be of the form as

$$Y=\alpha+\beta x+\gamma z+\varepsilon \quad (4.3)$$

then $$\sigma_{hky}^2=\beta^2\sigma_{hkx}^2+\gamma^2\sigma_{hkz}^2$$

The weight and variance of the (h, k)th stratum having auxiliary variables as 'X' and 'Z'.

$$W_{hk}=\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k} f\left(x,z\right)\partial x\partial z \quad (4.4)$$

$$\sigma_{hkx}^2=\frac{1}{W_{hk}}\int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} x^2 f\left(x\right)\partial x\partial z-\mu_{hkx}^2 \quad (4.5)$$

$$\sigma_{hkz}^2=\frac{1}{W_{hk}}\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k} z^2 f\left(z\right)\partial z\partial x-\mu_{hkz}^2 \quad (4.6)$$

where $\mu_{hkx} = \frac{1}{W_{hk}} \int_{z_{k-1}}^{z_k} \int_{x_{h-1}}^{x_h} x f(x) \partial x \partial z$,

$$\mu_{hkz} = \frac{1}{W_{hk}} \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} z f(z) \partial z \partial x$$

Thus, under proportional allocation with the model of the form given in (4.3), MPP will take the form as

Minimize $\sum_h \sum_k W_{hk} \left(\beta^2 \sigma_{hkx}^2 + \gamma^2 \sigma_{hkz}^2 \right)$

Subject to

$$\begin{aligned} &\sum_h V_h = d_x \\ &\sum_k U_k = t_z \\ &\forall V_h \geq 0, U_k \geq 0 \quad , \quad \begin{matrix} h = 1,2,...,L \\ k = 1,2,...,M \end{matrix} \end{aligned} \tag{4.7}$$

5. Empirical study

I:Let the variable X follow a distribution with pdf as

$$f(x) = \begin{cases} 2(2-x) & ;1 \leq x \leq 2 \\ 0 & ;otherwise \end{cases} \tag{5.1}$$

and the other auxiliary variable Z follow truncated exponential distribution with pdf

$$f(z) = \begin{cases} e^{-z+1} & ; \quad 1 \leq z \leq 6 \\ 0 & ;otherwise \end{cases} \tag{5.2}$$

In order to obtain OSB under proportional allocation with the pdf's of the auxiliary variables defined in (5.1) and (5.2), we need to obtain the value of W_{hk} and σ_{hky}^2 , for which we have to substitute (5.1) and (5.2) in equations (4.4)-(4.6), and get

$$W_{hk} = V_h e^{-z_k+1} \left(e^{U_k} - 1 \right) \left(4 - V_h - 2x_{h-1} \right) \tag{5.3}$$

$$\sigma_{hkx}^2 = \frac{a_1 U_k e^{-z_k+1}}{\left(a_1 e^{-z_k+1}\right)^2}\left\{\frac{4}{3}\left(V_h^2 + 3x_{h-1}^2 + 3V_h x_{h-1}\right) - \frac{1}{2}\left[\begin{matrix}(V_h + 2x_{h-1}) \\ \left[V_h\left(V_h + 2x_{h-1}\right) + x_{h-1}\left(1 + x_{h-1}\right)\right]\end{matrix}\right]\right\} - 4U_k^2 \tag{5.4}$$

and $$\sigma_{hkz}^2 = \frac{a_1 a_2 - e^{U_k}\left(1 + z_{k-1}\right) - U_k - z_{k-1} - 1}{a_1^2} \tag{5.5}$$

where $$a_1 = \left(e^{U_k} - 1\right)\left(4 - V_h - 2x_{h-1}\right)$$

$$a_2 = z_{k-1}^2 e^{U_k} - U_k^2 - z_{k-1}^2 - U_k z_{k-1} + 2\left[e^{U_k}\left(1 + z_{k-1}\right) - U_k - z_{k-1} - 1\right]$$

Substituting the values obtained in (5.3)-(5.5) in equation (4.7), we have MPP as

Minimize

$$\sum_h \sum_k \left(Sqrt\left(a_1 V_h e^{-z_k+1}\right)\right)$$

$$\left\{\beta^2 \frac{a_1 U_k e^{-z_k+1}}{6\left(a_1 e^{-z_k+1}\right)^2}\left\{\left[8\left(V_h^2 + 3x_{h-1}^2 + 3V_h x_{h-1}\right) - 3\left[\left(V_h + 2x_{h-1}\right)\left[\begin{matrix}V_h\left(V_h + 2x_{h-1}\right) \\ + x_{h-1}\left(1 + x_{h-1}\right)\end{matrix}\right]\right]\right] - 4U_k^2\right\}\right.$$

$$\left. + \gamma^2 \frac{a_1 a_2 - e^{U_k}\left(1 + z_{k-1}\right) - U_k - z_{k-1} - 1}{a_1^2}\right\}$$

Subject to

$$\begin{aligned} &\sum_h V_h = d_x \\ &\sum_k U_k = t_z \end{aligned} \tag{5.6}$$

$$\forall V_h \geq 0, U_k \geq 0 \quad , \quad \begin{matrix} h = 1, 2, ..., L \\ k = 1, 2, ..., M \end{matrix}$$

By using the given pdf's a simulation has been done in R-software and the values of β = 0.576 and γ = 0.257 and have been obtained. Thus, using the

values of β and γ, d_x = 1 and t_z = 5 as given above, the defined interval for X and Z respectively for total 6 (2×3) strata. Thus (5.6) can be written as

Minimize

$$\sum_h \sum_k \left[Sqrt\left(a_1 V_h e^{-z_k+1} \right) \right]$$

$$\left\{ (0.055) \frac{a_1 U_k e^{-z_k+1}}{\left(a_1 e^{-z_k+1} \right)^2} \left\{ \left[8\left(V_h^2 + 3x_{h-1}^2 + 3V_h x_{h-1} \right) - 3\left[\left(V_h + 2x_{h-1} \right) \left[\begin{matrix} V_h \left(V_h + 2x_{h-1} \right) \\ + x_{h-1}\left(1 + x_{h-1} \right) \end{matrix} \right] \right] \right] - 4U_k^2 \right\} \right.$$

$$\left. + (0.666) \frac{a_1 a_2 - e^{U_k}\left(1 + z_{k-1} \right) - U_k - z_{k-1} - 1}{a_1^2} \right\}$$

Subject to

$$\begin{aligned} &\sum_h V_h = 1 \\ &\sum_k U_k = 5 \end{aligned} \qquad (5.7)$$

$$\forall V_h \geq 0, U_k \geq 0 \quad , \quad \begin{matrix} h = 1,2 \\ k = 1,2,3 \end{matrix}$$

Executing a computer programme for MPP (5.5.10) using LINGO software, we get OSB as given in tables below:

Table 5.1. OSB when the auxiliary variables X and Z are independent with right triangular and exponential distribution respectively

Table 5.2. OSB and Variance when the auxiliary variables X and Z are independent with right triangular and exponential distribution respectively

OSB (x_h, z_k)	Variance (Proposed method)	Variance (Thomson 1973)	% R.E.
(1.3956,1.6568)	0.000864	0.00412	476.85
(2.0000,1.6568)			
(1.3956,3.5986)			
(2.0000,3.5986)			
(1.3956,6.0000)			
(2.0000,6.0000)			

Thus, while making 2 strata along x-axis and 3 along z-axis when the auxiliary variables X and Z are having Right triangular and Exponential distribution respectively independently. The results obtained in Table 5.1 and 5.2 reveal that the variance obtained by the proposed method is much less than Thomson (1973), for which the percentage relative efficiency comes out to be 476.85. Thereby, it is revealed that the use of two auxiliary variables is better than using one auxiliary variable.

II: The log normal distribution is a positively skewed distribution. Surveyors may use the log normal distribution for a positive valued study variable, which might increase without limit, such as the value of securities in financial problem or the values of properties in real estate or the failure rate of electronic parts in the engineering problems.

Let us assume that one of the auxiliary variable, say X, follows log-normal distribution with pdf as

$$f(x)=\begin{cases}\dfrac{1}{\sigma x\sqrt{2\pi}}e^{-\dfrac{(\log x-\mu)^2}{2\sigma^2}} & ;x>0,\sigma>0\\ 0 & ;otherwise\end{cases} \tag{5.8}$$

and the other auxiliary variable Z with pdf as:

$$f(z)=\begin{cases}\dfrac{1}{b-a} & ,a\le z\le b\\ o & ,otherwise\end{cases} \tag{5.9}$$

Then, in order to estimate OSB we need to find the value of W_{hk} and σ^2_{hky}. Substituting the pdf's (5.8) and (5.9) in equations (4.4)-(4.6), we shall get

$$W_{hk} = \frac{U_k}{2(b-a)} E_1 \tag{5.10}$$

$$\sigma^2_{hkx} = \frac{(b-a)\left[e^{2\left(\sigma^2+\mu\right)}(E_2)(E_1)\right] - U_k^2(b-a)^2\left[e^{\frac{1}{2}\left(\sigma^2-2\mu\right)}(E_3)\right]^2}{E_1^2} \tag{5.11}$$

and $$\sigma^2_{hkz} = \frac{2E_1\left(U_k^2 + 3z_{k-1}^2 + 3U_k z_{k-1}\right) - 3V_h^2\left(U_k + 2z_{k-1}\right)^2}{3E_1^2} \tag{5.12}$$

where $$E_1 = erf\left(\frac{\log\left(V_h + x_{h-1}\right) - \mu}{\sqrt{2\sigma^2}}\right) - erf\left(\frac{\log\left(x_{h-1}\right) - \mu}{\sqrt{2\sigma^2}}\right)$$

$$E_2 = erf\left(\frac{\log\left(V_h + x_{h-1}\right) - \mu - 2\sigma^2}{\sqrt{2\sigma^2}}\right) - erf\left(\frac{\log\left(x_{h-1}\right) - \mu - 2\sigma^2}{\sqrt{2\sigma^2}}\right)$$

and $$E_3 = erf\left(\frac{\log\left(V_h + x_{h-1}\right) - \mu - \sigma^2}{\sqrt{2\sigma^2}}\right) - erf\left(\frac{\log\left(x_{h-1}\right) - \mu - \sigma^2}{\sqrt{2\sigma^2}}\right)$$

It is to be noted here that the function 'erf', which repeats many times in the above result, is an error function, which is used to counter the integration with log-normal density function. It is defined as

$$erf\left(\omega\right) = \frac{2}{\sqrt{\pi}} \int_0^{\omega} e^{-j^2} \partial j$$

and some of its properties that need to be noted are

$$erf\left(-\omega\right)=-erf\left(\omega\right)$$
$$erf\left(0\right)=0$$
$$erf\left(\infty\right)=1$$
$$erf\left(-\infty\right)=1$$

Substituting values (5.10) to (5.12) in (4.7), we have MPP as

Minimize

$$\sum_h\sum_k\left[Sqrt\left(\frac{U_k}{2(b-a)}E_1\right)\right]\left(\begin{array}{l}\beta^2\dfrac{(b-a)\left[e^{2\left(\sigma^2+\mu\right)}\left(E_2\right)\left(E_1\right)\right]-U_k^2(b-a)^2\left[e^{\frac{1}{2}\left(\sigma^2-2\mu\right)}\left(E_3\right)\right]^2}{E_1^2}\\ +\gamma^2\dfrac{2E_1\left(U_k^2+3z_{k-1}^2+3U_kz_{k-1}\right)-3V_h^2\left(U_k+2z_{k-1}\right)^2}{3E_1^2}\end{array}\right)$$

Subject to

$$\sum_h V_h=d_x$$
$$\sum_k U_k=t_z \qquad (5.13)$$
$$\forall V_h\geq 0,U_k\geq 0 \quad , \quad \begin{array}{l}h=1,2,...,L\\ k=1,2,...,M\end{array}$$

In this case let us assume that the log-normal distribution is to be standardized, i.e. μ=0,σ=1 $z\in\left[0,1\right]i,e.z_0=0,z_M=1$ and the other variable $x\in\left[0,10\right],i,e.x_0=0,x_L=10$. Further, let us assume that the total strata to be made are 3×2(L×M)=6 and by simulation in R-software the value of β = 0.82 and γ = 0.437. Then, to obtain OSB we need to solve MPP

Minimize

$$\sum_{h=1}^{3}\sum_{k=1}^{2}\left[Sqrt\left(\frac{U_k}{2}E_1\right)\right]$$

$$\left((0.0722)\frac{\left[(7.389)\left(E_2^{'}\right)\left(E_1^{'}\right)\right]-U_k^2\left[(1.648)\left(E_3^{'}\right)\right]^2}{E_1^2} + (0.1909)\frac{2E_1\left(U_k^2+3z_{k-1}^2+3U_k z_{k-1}\right)-3V_h^2\left(U_k+2z_{k-1}\right)^2}{3E_1^2}\right)$$

Subject to

$$\sum_{h=1}^{3} V_h = 10$$

$$\sum_{k=1}^{2} U_k = 1 \qquad (5.14)$$

$$\forall V_h \geq 0, U_k \geq 0 \quad , \quad \begin{matrix} h=1,2,3 \\ k=1,2 \end{matrix}$$

where $$E_1^{'} = erf\left(\frac{\log\left(V_h + x_{h-1}\right)}{1.141}\right) - erf\left(\frac{\log\left(x_{h-1}\right)}{1.141}\right)$$

$$E_2^{'} = erf\left(\frac{\log\left(V_h + x_{h-1}\right)-2}{1.141}\right) - erf\left(\frac{\log\left(x_{h-1}\right)-2}{1.141}\right)$$

and $$E_3^{'} = erf\left(\frac{\log\left(V_h + x_{h-1}\right)-1}{1.414}\right) - erf\left(\frac{\log\left(x_{h-1}\right)-1}{1.414}\right)$$

By executing the computer programme of (5.14) MPP in LINGO, we get OSB value presented in the following tables.

Table 5.3. OSB when the auxiliary variables X and Z follow log-normal and uniform distribution respectively

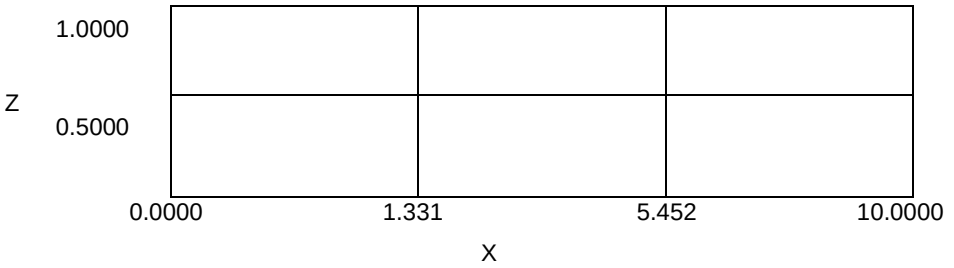

Table 5.4. OSB and Variance when the auxiliary variables X and Z follow log-normal and uniform distribution respectively

OSB (x_h, z_k)	Variance (Proposed method)	Variance (Khan *et al*. 2005)	% R.E.
(1.331,0.500) (5.452,0.500) (10.000,0.500) (1.331,1.000) (5.452,1.000) (10.000,1.000)	0.005916	0.014708	248.61

A perusal of Tab 5.4 indicates that the variance obtained by the proposed method is much less than not on Khan *et al*. (2005) and the percentage of relative efficiency comes out to be 248.61 of the proposed method over the other method in comparison. Thus, it may be concluded that using two auxiliary variables is better than using one auxiliary variable. In practice, the complete dataset of the study variable is unknown, which diminishes the uses of many stratification techniques. In such a situation, the proposed technique can be used as it requires only the values of parameters of the population, which can easily be available from the past studies

Conclusion

In this investigation, a scheme has been proposed to obtain the optimum strata boundaries (OSB) for two stratification variables highly related to the study variable. Numerical illustrations have been presented to explain the computational details of the application of the proposed method for two auxiliary variables. By using the frequency distribution the problem of constructing stratification points is formulated into the mathematical, programming problem, which results in a multistage decision problem, which is to be solved on a compromise distance.

In the empirical study I while comparing the proposed method with Thomson (1973), the percentage of relative efficiency comes out to be 476.85 when the auxiliary variables have right triangular and exponential distributions. Similarly, in study II the percentage of relative efficiency comes out to be 248.61 when comparing the proposed method with Khan *et al*. (2005). However, it is found that while obtaining the strata for two variables, when the frequency distributions are well-known, leads to the substantial gains in average relative efficiencies and have gains in precision of estimates. Thus, both the empirical studies suggest that the proposed method is more preferable than the existing methods.

REFERENCES

AOYAMA, H., (1954). A study of stratified random sampling. Annals of Institute of Statistical Mathematrica, Vol. 6, pp. 1–36.

BUHLER, W., DEUTLER, T., (1975). Optimal stratification and grouping by dynamic programming. Metrik, Vol. 22, pp. 161-175.

DALENIUS, T., (1950). The problem of optimum stratification. Skandinavisk Aktuarietidskrift, Vol. 33, pp. 203–213.

DALENIUS, T., HODGES, J. L., (1959). Minimum variance stratification. Journal of American Statistical Association, Vol. 54, pp. 88-101.

DALENIUS, T. AND GURNEY, M. (1951). The problem of optimum stratification II. Skandinavisk Aktuarietidskrift, Vol. 34, pp. 133–148.

DANISH, F., RIZVI, S. E. H., SHARMA, M. K., JEELANI, M. I., (2017a). Optimum Stratification Using Mathematical Programming Approach: A Review. Journal of Statistics Applications & Probability Letters, Vol. 4(3), pp. 123–129

DANISH, F., RIZVI, S. E. H., JEELANI, M. I., REASHI, J. A., (2017b). Obtaining Strata Boundaries under Proportional Allocation with Varying Cost of Every Unit. Pakistan Journal Of Statistics and Operations Research, Vol. 13(3), pp. 567–574

KHAN, E. A., KHAN, M. G. M., AHSAN, M. J., (2002). Optimum stratification: a mathematical programming approach. Calcutta Statistical Association bulletin, Vol. 52, pp. 323–333.

KHAN, M. G. M., NAND, N., AHMAD, N., (2008). Determining the optimum strata boundary points using dynamic programming. Survey Methodology, Vol. 34 (2), pp. 205–214.

KHAN, M. G. M., NAJMUSSEHAR, AHSAN, M. J., (2005). Optimum stratification for exponential study variable under Neyman allocation. Journal of the Indian Society of Agricultural Statistics, Vol. 59 (2), pp. 146–150.

MAHALANOBIS, P. C., (1952). Some aspect of design of sample surveys. Sankhya, Vol. 12, pp. 1–17.

SINGH, R., SUKHATME, B. V., (1969). Optimum stratification. Annals of Institute of Statistical Mathematrica, Vol. 21, pp. 515–528.

SINGH, R., SUKHATME, B. V., (1973). Optimum stratification with ratio and regression methods of estimation. Annals of Institute of Statistical Mathematrica, Vol. 25, pp. 627–633.

SINGH, R., (1977). A note on equal allocation with ratio and regression methods of estimation. Australian Journal of Statistics, Vol. 19, pp. 96–104.

THOMSEN, I., (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. Metrika, Vol. 23, pp. 15–25.