STATISTICS IN TRANSITION new series, December 2021
Vol. 22, No. 4 pp. 19–40, DOI 10.21307/stattrans-2021-036
Received – 23.01.2019; accepted – 16.03.2021

Approximately optimum strata boundaries for two concomitant stratification variables under proportional allocation

Faizan Danish[1], **S. E. H. Rizvi**[2]

ABSTRACT

The proper choice of strata boundaries is an important factor determining the efficiency of the estimator of the considered characteristics of a population. In this article, the Cum$\sqrt[3]{D_i(x,z)}$ Rule (i=3,4) for obtaining approximately optimum strata boundaries has been applied, taking into account a single-study variable along with two concomitant variables serving as the basis of the stratification variables. The relative efficiency of the proposed methods has been demonstrated theoretically and empirically by comparing them to a selection of already-existing methods in a simulation study with the use of the proportional allocation method.

Key words: stratification points, proportional allocation, minimal equation.

1. Introduction

Let there be a finite population consisting of N units, for which it is required to estimate the total or mean for the characteristic Y under study, using simple random sampling technique. In order to have this, we partition population L × M strata:

$$\sum_{h=1}^{L}\sum_{k=1}^{M} N_{hk} = N$$

where N_{hk} indicates the number of units in $(h\ ,k)^{th}$ stratum.

Let 'n' be the number of units to be drawn from the whole population and suppose that the allocation of sample size n_{hk} such that

$$\sum_{h=1}^{L}\sum_{k=1}^{M} n_{hk} = n$$

[1] Corresponding author. Division of Biostatistics, Department of Population Health, School of Medicine, New York University, USA. E-mail: Faizan.Danish@nyulangone.org. ORCID: https://orcid.org/0000-0002-0476-9744.
[2] Division of Statistics and Computer Sciences, SKUAST-J, J&K, India. E-mail: sehrizvi_stats@yahoo.co.in.

Let y_{hki} (i = 1 , 2 , 3,…, N_{hk}) be the population unit and then the population total is

$$Y = \sum_{h=1}^{L} \sum_{k=1}^{M} \sum_{i=1}^{N_{hk}} y_{hki}$$

For the study variable, the unbiased estimate of $\overline{Y}$, $\overline{y}_{st} = \sum_{h=1}^{L} \sum_{k=1}^{M} W_{hk} \overline{y}_{hk}$,where $\overline{y}_{hk} = \frac{1}{n_{hk}} \sum_{i=1}^{n_{hk}} y_{hki}$ and W_{hk} denotes the weight of the (h,k)th stratum.

For stratified simple random sampling, the sample estimate $\overline{y}_{st}$ is unbiased with sampling variance as below:

$$V\left(\overline{y}_{st}\right) = \sum_{h} \sum_{k} \left(1 - f_{hk}\right) \frac{W_{hk}^2 \sigma_{hky}^2}{n_{hk}}$$

where $f_{hk} = \frac{n_{hk}}{N_{hk}}$ and if f.p.c is ignored, we have

$$V\left(\overline{y}_{st}\right) = \sum_{h} \sum_{k} \frac{W_{hk}^2 \sigma_{hky}^2}{n_{hk}}$$

σ_{hky}^2 represents the population variance for the character Y and is defined as

$$\sigma_{hky}^2 = \frac{1}{N_{hk}} \sum_{i=1}^{N_{hk}} \left(y_{hki} - \overline{y}_{hk}\right)^2$$

$\overline{y}_{hk}$ being the population mean of all the N_{hk} units in the $\left(h, k\right)^{th}$ stratum.

Construction of stratification points was pioneered by Dalenius (1950), while minimizing variance set of equations as the functions of population parameters were obtained and due to their implicit nature, it becomes complicated to obtain solutions. Cochran (1961,1963) has also discussed the cases regarding the optimum boundaries. Yadav and Singh (1984), Rizvi *et al.* (2000), Danish *et al.* (2020), Khan *et al.* (2008), Khan *et al.* (2014), Danish *et al.* (2017), Danish *et al.* (2018), Danish and Rizvi (2018,2019) and Danish, F. (2018). Rizvi and Danish (2018) made an attempt to summarise the proposed contribution towards obtaining stratification points.

The allocation procedure in which a sample size is selected as per proportion of the stratum is known as proportional allocation. In such allocation, the sample size is selected as

$$n_{hk} = \frac{nN_{hk}}{N} = nW_{hk}$$

thus, $$n_{hk} \propto N_{hk} \text{ and } \sum_{h=1}^{L}\sum_{k=1}^{M} n_{hk} = n$$

Hence, under such allocation, variance is

$$V\left(\bar{y}_{st}\right)_p = \frac{1}{n}\sum_h \sum_k W_{hk}\sigma_{hky}^2 \tag{1.1}$$

In this paper, for obtaining stratification points using classical approach for two concomitant variables as the basis of stratification variables and a single study variable under the proportional allocation method by assuming different distributions of the concomitant variables and both dependent and independent cases have been discussed as well.

2. Variance expression

Let the regression model of the response variable Y and the two information variables X & Z be given as

Y = C (X, Z) + e

where 'e' is error term such that

$$E\left(e \mid x, z\right) = 0 \text{ and } V(e|x,z) = \eta(x,z) > 0, \forall\; x\varepsilon\left(a,b\right)\; z\varepsilon\left(c,d\right),$$

$(b – a) < \infty ,(c – d) < \infty$

If joint marginal of X and Z is f(x, z) and f(x) and f(z) denotes marginal densities of individual variables, respectively, then under above regression model,

we have $W_{hk} = \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} f\left(x,z\right)\partial x \partial z$ is weight of a stratum.

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

$\sigma_{hky}^2 = \sigma_{hkc}^2 + \mu_{hk\eta}$ denotes its mean and variation respectively,

where $\left(x_{h-1}, x_h, z_{k-1}, z_k\right)$ be the stratification points and $\mu_{hk\eta}$ is the average value of the function $\eta(x,z)$ and σ_{hkc}^2 as

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

Using these relations, the variance can be expressed in terms of the population parameters of the function of X and Y and $V(e|x,z)$. The variance expression for the case of proportional allocation is therefore given by

$$V\left(\bar{y}_{st}\right)_{prop} = \frac{\left[\sum_{h}\sum_{k} W_{hk}\left(\sigma_{hkc}^{2} + \mu_{hk\eta}\right)\right]}{n} \tag{2.1}$$

The expression for various terms can be in terms of Singh and Sukhatme (1969) and Danish *et al.* (2018).

3. Minimal equations for proportional allocation

Since $\sum_{h}\sum_{k} W_{hk}\mu_{hk\eta} = \mu_{\eta}$, which is the population parameter and therefore is a fixed constant. Hence, minimization of (2.1) is equivalent to minimization

$$V_p = \sum_{h}\sum_{k} W_{hk}\sigma_{hkc}^{2} \tag{3.1}$$

Thus, to obtain minimal equations, we minimize V_p by with respect of x_h and equate to zero, we get

$$\frac{\partial}{\partial x_h} V_p = \sum_{k}\left[W_{hk}\frac{\partial}{\partial x_h}\sigma_{hkc}^{2} + \sigma_{hkc}^{2}\frac{\partial}{\partial x_h} W_{hk} + W_{ik}\frac{\partial}{\partial x_h}\sigma_{ikc}^{2} + \sigma_{ikc}^{2}\frac{\partial}{\partial x_h} W_{ik}\right] = 0$$

After further simplification, we get

$$\begin{aligned}
&\sum_{k}\left[W_{hk}\int_{z_{k-1}}^{z_k} \frac{f(x_h,z)}{W_{hk}}\left\{\left[c(x_h,z) - \mu_{hkc}\right]^2 - \sigma_{hkc}^{2}\right\}\partial z\right] + \sigma_{hkc}^{2}\int_{z_{k-1}}^{z_k} f(x_h,z)\,\partial z \\
&= \sum_{k}\left[W_{ik}\int_{z_{k-1}}^{z_k} \frac{f(x_h,z)}{W_{ik}}\left\{\left[c(x_h,z) - \mu_{ikc}\right]^2 - \sigma_{ikc}^{2}\right\}\partial z\right] + \sigma_{ikc}^{2}\int_{z_{k-1}}^{z_k} f(x_h,z)\,\partial z
\end{aligned} \tag{3.2}$$

For obtaining minimal equations we also differentiate V_p partially w.r.t. z_k in a similar way, we get

$$\begin{aligned}
&\sum_{h}\left[W_{hk}\int_{x_{h-1}}^{x_h} \frac{f(x,z_k)}{W_{hk}}\left\{\left[c(x,z_k) - \mu_{hkc}\right]^2 - \sigma_{hkc}^{2}\right\}\partial x\right] + \sigma_{hkc}^{2}\int_{x_{h-1}}^{x_h} f(x,z_k)\,\partial x \\
&= \sum_{h}\left[W_{hj}\int_{x_{h-1}}^{x_h} \frac{f(x,z_k)}{W_{hj}}\left\{\left[c(x,z_k) - \mu_{hjc}\right]^2 - \sigma_{hjc}^{2}\right\}\partial x\right] + \sigma_{hjc}^{2}\int_{x_{h-1}}^{x_h} f(x,z_k)\,\partial x
\end{aligned} \tag{3.3}$$

However, for obtaining minimal equations we minimize V_p on equating the partial derivative of this expression with respect of x_h and z_k to zero, we get

$$W_{hk} f(x_h, z_k) \frac{\left\{ \left[c(x_h, z_k) - \mu_{hkc} \right]^2 - \sigma_{hkc}^2 \right\}}{W_{hk}} + f(x_h, z_k)\sigma_{hkc}^2$$
$$= W_{ij} f(x_h, z_k) \frac{\left\{ \left[c(x_h, z_k) - \mu_{ijc} \right]^2 - \sigma_{ijc}^2 \right\}}{W_{ij}} + f(x_h, z_k)\sigma_{ijc}^2$$

This gives the equation as

$$c(x_h, z_k) = \frac{(\mu_{hkc} + \mu_{ijc})^2}{2}, \quad \begin{array}{l} i = h+1, h = 1,2,\ldots,L-1 \\ j = k+1, k = 1,2,\ldots,M-1 \end{array} \tag{3.4}$$

On the condition that $\lambda(x,z) = c'(x,z) f(x,z)$ belongs to class Ω functions, solutions to the system of equation (3.4) give OSB in the sense of minimization of variance $V(\bar{y}_{st})_{prop}$. These equations are also very difficult to solve and, therefore, for these equations also we shall find methods of obtaining approximation to the exact solutions $[x_h, z_k]$. Further better approximation can be obtained by using some approximate iterative procedures.

4. Some miscellaneous results

In the case of complexities in the equations, let us impose few regularity conditions on $f(x,z), c(x,z)$ *and* $\eta(x,z)$. We state that $\zeta(x,z)$ belongs to class Ω if it satisfies

i) $0 < \zeta(x,z)$

ii) $\zeta(x,z) < \infty$

iii) $\zeta(x,z), \zeta'(x,z)$ *and* $\zeta''(x,z)$ exist and are continuous $\forall (x,z)$ in [(a, b), (c, d)] respectively such that $(b-a) < \infty$ *and* $(d-c) < \infty$.

Let us suppose $f(x,z)$ *and* $\eta(x,z)$ belong to class Ω and the function c(x, z) satisfies the conditions (ii) and (iii).

Before we proceed to prove the results, let us define the symbol 'O', which has been used in the present investigation.

For two functions $T_1(x,z)$ *and* $T_2(x,z)$, such that the ratio $T_1(x,z)/T_2(x,z)$ remains bounded as x and z tends to their limits, then we can write $T_1(x,z)=O(T_2(x,z))$.

Lemma 4.1: If the function $I_{ij}(x,z)$ is defined as

$$I_{ij}(x,z)=\int_{z_1}^{z_2}\int_{x_1}^{x_2}(t_1-x_1)^i(t_2-z_1)^j f(t_1,t_2)\partial t_1\partial t_2\ ,\ x_1<x_2\ \&\ z_1<z_2$$

then

$$I_{ij}(x,z)=\left[\begin{array}{l}\dfrac{k_1^{i+1}k_2^{j+1}}{(i+1)(j+1)}f+\dfrac{k_1^{i+1}k_2^{j+1}}{(i+1)(j+1)}f_x+\dfrac{k_1^{i+1}k_2^{j+2}}{(i+1)(j+2)}f_z\\ +\dfrac{1}{2!}\left[\dfrac{k_1^{i+3}k_2^{j+1}}{(i+3)(j+1)}f_{xx}+2\dfrac{k_1^{i+2}k_2^{j+2}}{(i+2)(j+2)}f_{xz}+\dfrac{k_1^{i+1}k_2^{j+3}}{(i+1)(j+3)}f_{zz}\right]+O\left(k^{i+j+5}\right)\end{array}\right] \tag{4.1}$$

where $f(t_1,t_2)=f,\dfrac{\partial f}{\partial t_1}=f_x,\dfrac{\partial f}{\partial t_2}=f_z,\dfrac{\partial^2 f}{\partial t_1^2}=f_{xx},\dfrac{\partial^2 f}{\partial t_2^2}=f_{zz},\dfrac{\partial^2 f}{\partial t_1\partial t_2}=f_{xz}$,

$k_1=x_2-x_1$ and $k_2=z_2-z_1$.

Lemma 4.2: Let $\mu_\eta(x,z)$ denote the conditional expectation of the function $\eta(t_1,t_2)$, so that

$$\mu_\eta(x,z)=\frac{\int_{z_1}^{z_2}\int_{x_1}^{x_2}\eta(t_1,t_2)\ f(t_1,t_2)\ \partial t_1\partial t_2}{\int_{z_1}^{z_2}\int_{x_1}^{x_2}f(t_1,t_2)\quad \partial t_1\partial t_2}$$

Then, the series expansion of $\mu_\eta(x,z)$ at point (t_1 , t_2) is given by

$$\mu_\eta(x,z)=\eta\left[\begin{array}{l}1+\dfrac{\eta'}{2\eta}(k_1+k_2)+\left(\dfrac{\eta'(f_x+f_z)+2f\eta''}{12f\eta}\right)(k_1+k_2)^2\\ +\left(\dfrac{f(f_{xx}+f_{zz}+f_{xz})\eta'+f(f_x+f_z)\eta''+f^2\eta'''-\eta'(f_x+f_z)^2}{12f^2\eta}\right)(k_1+k_2)^3\\ +O\left((k_1+k_2)^4\right)\end{array}\right] \tag{4.2}$$

Lemma 4.3: If $\sigma_{\eta}^{2}(x,z)$ denotes the conditional variance of the function $\eta(t_1,t_2)$ defined in the interval (x, z) , then

$$\sigma_{\eta}^{2}(x,z)=\frac{(k)^{2}}{12}\left(\eta^{'}\right)^{2}\left[1+\frac{\eta^{"}}{\eta^{'}}(k)^{1}+O(k)^{2}\right] \tag{4.3}$$

where $(k)^{1}$ *and* $(k)^{2}$ denote all k_i's with power '1' and '2' respectively.

Lemma 4.4:

$$(k_1k_2)^{\lambda-1}\int_{z_1}^{z_2}\int_{x_1}^{x_2} f(t_1,t_2)\partial t_1\partial t_2=\left[\int_{z_1}^{z_2}\int_{x_1}^{x_2}\sqrt[\lambda]{f(t_1,t_2)}\partial t_1\partial t_2\right]^{\lambda}\left[1+O(k^{2})\right] \tag{4.4}$$

Lemma 4.5: With $I_{00}(x,z)$ and $\sigma_{\eta}^{2}(x,z)$ defined as in Lemma 4.1 and Lemma 4.3 respectively, we have

$$I_{00}(x,z)\sigma_{\eta}^{2}(x,z)=\frac{k^2}{12}\int_{z_1}^{z_2}\int_{x_1}^{x_2}\eta'^{2}(t_1,t_2)\quad f(t_1,t_2)\partial t_1\partial t_2 \tag{4.5}$$

where k denotes any k_h or k_k .

5. Minimal equations and their approximate solutions

In this section we will obtain expansion of the series given in (3.4) about the points x_h and z_k the common boundary of $(h,k)^{th}$ *and* $(h+1,k+1)^{th}$ strata and obtain the approximate systems of equation, which will give approximately optimum points of stratification as their solutions. In doing so we shall make use of the Lemma's already 4.1-4.5.

The minimal equations for this method are given by

$$c(x_h,z_k)-\mu_{hkc}=\mu_{ijc}-c(x_h,z_k)$$

i= h +1,h = 1, 2, … L , j= k+1, k= 1,2,…,M

For R.H.S., the corresponding expansion for the L.H.S. may be obtained by changing the signs of coefficients of even powers of $\left(k_i, k_j\right)$ the width of $(i, j)^{th}$ stratum, where $k_i = x_{h+1} - x_h$, $k_j = z_{k+1} - z_k$. From (4.2), after replacing $\left(x_1, x_2\right)$ *by* $\left(x_h, x_{h+1}\right)$ and $\left(z_1, z_2\right)$ *by* $\left(z_k, z_{k+1}\right)$, we have

$$\mu_{ikc} = c\left[1 + \frac{c'}{2c}k_i + \left(\frac{c' f_x + 2 fc''}{12 fc}\right)k_i^2 + \left(\frac{ff_{xx}c' + ff_x c'' + f^2 c'''}{24 f^2 c}\right)k_i^3 + O\left(k_i^4\right)\right]$$

where $k_i = x_{h+1} - x_h$ and derivatives are evaluated at x_h.

However, when the same functions are differentiated w.r.t. z_k, we have

$$\mu_{hjc} = c\left[1 + \frac{c'}{2c}k_j + \left(\frac{c' f_z + 2 fc''}{12 fc}\right)k_j^2 + \left(\frac{ff_{zz}c' + ff_z c'' + f^2 c'''}{24 f^2 c}\right)k_j^3 + O\left(k_j^4\right)\right]$$

where $k_j = z_{k+1} - z_k$

Here, the derivatives are evaluated at both x_h *and* z_k, we get

$$\mu_{ijc} = c\left[1 + \frac{c'}{2c}\left(k_i k_j\right) + \left(\frac{c'\left(f_x + f_z\right) + 2fc''}{12 fc}\right)\left(k_i k_j\right)^2 + \left(f_4\right)\left(k_i k_j\right)^3 + O\left(k_i k_j\right)^4\right]$$

where

$$f_4 = \frac{f\left(f_{xx} + f_{zz} + f_{xz}\right)c' + f\left(f_x + f_z\right)c'' + f^2 c'''}{24 f^2 c}$$

Thus, we have

$$\mu_{ijc} - c\left(x_h, z_k\right) = \frac{\left(k_i k_j\right)}{2}\left[c' + \left(\frac{c'\left(f_x + f_z\right) + 2fc''}{12 f}\right)\left(k_i k_j\right) + O\left(k_i k_j\right)^2\right]$$

Similarly, we get

$$\mu_{hkc} - c\left(x_h, z_k\right) = \frac{\left(k_h k_k\right)}{2}\left[c' + \left(\frac{c'\left(f_x + f_z\right) + 2fc''}{12 f}\right)\left(k_h k_k\right) + O\left(k_h k_k\right)^2\right]$$

Evaluating derivatives at $x_h \quad and \quad z_k$. Therefore, equation (3.4) can be put as

$$\frac{(k_h k_k)}{2}\left[c' + \left(\frac{c'(f_x + f_z) + 2fc''}{12f}\right)(k_h k_k) + O(k_h k_k)^2\right]$$

$$= \frac{(k_i k_j)}{2}\left[c' + \left(\frac{c'(f_x + f_z) + 2fc''}{12f}\right)(k_i k_j) + O(k_i k_j)^2\right] \quad (5.1)$$

Now, let us consider an expansion of the function

$$B_{hk} = \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2$$

about the point $[x_h, z_k]$. Expanding the integral about $x_h \quad and \quad z_k$ with the help of the Taylor's expansion for two variables, we have

$$B_{hk} = fc'^2 k_h k_k\left[1 - \frac{(f_x + f_z)c' + 2fc''}{fc'}\frac{(k_h k_k)}{2} + O(k_h k_k)^2\right] \quad (5.2)$$

where in (5.2) also the function of f, η and their derivatives are evaluated at $x_h \quad and \quad z_k$. Thus, we find that

$$\frac{(k_h k_k)^2}{8}\frac{B_{hk}c'}{f} = \frac{(k_h k_k)^3 c'^3}{8}\left[1 - \left(\frac{(f_x + f_z)c' + 2fc''}{2fc'}\right)(k_h k_k) + O(k_h k_k)^2\right]$$

or

$$\left[\frac{(k_h k_k)^2}{8}\frac{B_{hk}c'}{f}\right]^{\frac{1}{3}} = \frac{(k_h k_k)c'}{2}\left[1 - \left(\frac{(f_x + f_z)c' + 2fc''}{2fc'}\right)(k_h k_k) + O(k_h k_k)^2\right]^{\frac{1}{3}}$$

$$= \frac{(k_h k_k)c'}{2}\left[1 - \left(\frac{(f_x + f_z)c' + 2fc''}{6fc'}\right)(k_h k_k) + O(k_h k_k)^2\right]$$

(5.3)

Similarly, we obtain

$$\left[\frac{(k_i k_j)^2}{8}\frac{B_{ij}c'}{f}\right]^{\frac{1}{3}} = \frac{(k_h k_k)c'}{2}\left[1 - \left(\frac{(f_x + f_z)c' + 2fc''}{6fc'}\right)(k_i k_j) + O(k_i k_j)^2\right]$$

Therefore, the minimal equations (5.1) can be put as

$$\left[\frac{(k_h k_k)^2}{8}\frac{B_{hk}c'}{f}\right]^{\frac{1}{3}}\left[1+O(k_h k_k)^2\right]=\left[\frac{(k_i k_j)^2}{8}\frac{B_{ij}c'}{f}\right]^{\frac{1}{3}}\left[1+O(k_i k_j)^2\right] \tag{5.4}$$

Hence, if the terms of order $O\left(\underset{((a,b),(c,d))}{Sup}(k_h k_j)\right)^3$ can be neglected, we can replace the minimal equations approximately by

$$\left[\frac{(k_h k_k)^2}{8}\frac{B_{hk}c'}{f}\right]^{\frac{1}{3}}=\left[\frac{(k_i k_j)^2}{8}\frac{B_{ij}c'}{f}\right]^{\frac{1}{3}}$$

or $(k_h k_k)^2 B_{hk}$ = Constant (5.5)

In the case when it is possible to find a function $Q_1'(x_{h-1},x_h,z_{k-1},z_k)$ such that

$$(k_h k_k)^2 B_{hk} = (k_h k_k)^2 \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2 (t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2$$
$$= Q_1'(x_{h-1},x_h,z_{k-1},z_k)\left[1+O(k_h k_k)^2\right] \tag{5.6}$$

Thus, the system of equations (5.5) to the same degree of accuracy can be put as

$$Q_1'(x_{h-1},x_h,z_{k-1},z_k) = \text{Constant} \tag{5.7}$$

The above results can be put in the form of a note as follows.

Remark 1: If the regression of the dependent variable Y and stratification variables X and Z in an unbounded population is given by

$$Y = C(X,Z)+e$$

where ‘e’ is the error component such that $E(e\,|\,x,z) = 0$ and $V(e\,|\,x,z)=\eta(x,z)>0\ ,\forall x\in(a,b)$ *and* $z\in(c,d)$ with finite deviation of the intervals, and in addition if $c'^2(x,z)\,f(x,z)$ belong to Ω, then the system of equations (3.4) giving strata boundaries $\left[x_h, z_k\right]$, which correspond to the minimum of $V\left(\bar{y}_{st}\right)_{prop}$, can be put as

$$\left\{(k_h k_k)^2 \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2 \left[1+O(k_h k_k)^2\right]\right\}^{\frac{1}{3}}$$

$$=\left\{(k_i k_j)^2 \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2 \left[1+O(k_i k_j)^2\right]\right\}^{\frac{1}{3}}.$$

If the terms of order $O\left(\underset{((a,b),(c,d))}{Sup}(k_h k_j)\right)^3$ can be neglected, these equations can be replaced by the approximate system of equations

$$(k_h k_k)^2 \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2 \left[1+O(k_h k_k)^2\right] = \text{Constant}$$

Or equivalently by

$$Q_1'(x_{h-1}, x_h, z_{k-1}, z_k) = \text{Constant}$$

Therefore,

$$\therefore \; Q_1'(x_{h-1}, x_h, z_{k-1}, z_k)\left[1+O(k_h k_k)^2\right] = (k_h k_k)^2 \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2 \left[1+O(k_h k_k)^2\right]$$

The same result can also be obtained by minimizing the function

$$\sum_h\sum_k \int_{z_{k-1}}^{z_k}\int_{x_{h-1}}^{x_h} c'^2(t_1,t_2) f(t_1,t_2)\partial t_1 \partial t_2 \left[1+O(k_h k_k)^2\right]$$

as in the light of Lemma 5, $12\sum_h\sum_k W_{hk}\sigma_{hkc}^2$ equals to this function

Thus, we find that if the function $c'^2(x,z)\, f(t_1,t_2)$ belongs to the class Ω the minimum value of $\sum_h\sum_k W_{hk}\sigma_{hkc}^2$ and therefore $V\left(\bar{y}_{st}\right)_{prop}$ exists and the set of strata boundaries $\left[x_h, z_k\right]$, corresponding to this minimum, is the solution of the systems of equations (3.4) or equivalently of (5.4). These equations are very difficult to solve exactly and it becomes essential to find some approximation to stratification points.

It may be precisely solved by substituting the exact minimal equations by other systems of equations which are comparatively easy to solve but are only asymptotically equivalent to the exact equations. The error is introduced because we neglect terms of higher powers of the strata widths which can be justified when the total stratum is large. The approximate systems of equations are obtained by neglecting terms of order $O\left(m^3\right)$ where $m = \underset{((a,b),(c,d))}{Sup}(k_h k_k)$, on both sides of (5.4). For large strata, the terms of order $O\left(m^3\right)$ are small and therefore the error involved in the approximate systems of equations is small, although this error is comparatively larger than the one involved in the case of optimum allocation. Here, we shall develop the approximate systems of equations given in (5.5) and (5.7).

6. Approximate systems of equations

I. If in the expansion of the minimal equations (3.4) we neglect all terms except the first on both sides of the equation, the solution is obtained by taking

$$x_h = \text{constant} = \frac{b-a}{L}, \ h=1,2,\ldots,L \text{ and } z_k = \text{constant} = \frac{d-c}{M}, \quad k = 1,2,\ldots,M. \tag{6.1}$$

Therefore

$$x_h = a + \left(\frac{(b-a)}{L}\right)h \quad with \quad x_0 = a \quad and \quad x_L = b$$

and $$z_k = c + \left(\frac{(d-c)}{M}\right)k \quad with \quad z_0 = c \quad and \quad z_M = d$$

It cannot be suspected that this set of approximations can give good solutions as these are simple to obtain. However, the method is not applicable in the case of infinite range.

II. An approximation to the optimum points of stratification is obtained by solving the systems of equations

$$\left(k_h k_k\right)^2 \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} c'^2\left(t_1,t_2\right) f\left(t_1,t_2\right) \partial t_1 \partial t_2 = C_1 \tag{6.2}$$

as shown in (5.5). The solutions of this system of equations and also of those that will now follow, are expected to be closer to the optimum stratification points as compared to the solutions obtained from (6.1).

III. From Lemma 4 and equation (6.2), we get a general class of approximate systems of equations as

$$\left[(k_h k_k)^{3\lambda-1}\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k}\left(c'^2(t_1,t_2)\,f(t_1,t_2)\right)^{\lambda}\partial t_1\partial t_2\right]^{\frac{1}{\lambda}} = \text{Constant}$$

However, for $\lambda=\frac{1}{2}$,we have

$$\left[(k_h k_k)\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k} c'^2(t_1,t_2)\sqrt{f(t_1,t_2)}\partial t_1\partial t_2\right]^2 = C_2$$

and for $\lambda=\frac{1}{3}$, we have a system of equations as

$$\left[\int_{x_{h-1}}^{x_h}\int_{z_{k-1}}^{z_k}\sqrt[3]{c'^2(t_1,t_2)\,f(t_1,t_2)}\partial t_1\partial t_2\right]^3 = C_3$$

giving approximations to stratification points $[x_h, z_k]$. As remarked in the case of the optimum allocation method, in some particular cases some of the approximate systems given in the above equations may be meaningless. Therefore, depending upon the situation, one should make the approximate choice of the systems of equations for obtaining the approximations to optimum points $[x_h, z_k]$.

7. Cum $\sqrt[3]{D_3(x,z)}$ Rule

If the function $D_3(x,z)=c'^2(x,z)\,f(x,z)$ is bounded and its first two derivatives exists $\forall\, x\in[a,b]$ & $z\in[c,d]$, then taking equal intervals with a given values of L and M on the cumulative cube root of $D_3(x,z)$ will give AOSB $[x_h, z_k]$.

Remarks:

I. If we take either $c(x,z)=\alpha+\beta x$ or $c(x,z)=\alpha+\gamma z$ in $D_3(x,z)$ it reduces to the method proposed by Singh and Sukhatme (1969).

II. If the function $c'^2(x,z)$ is constant, therefore the proposed method reduces to Cum $\sqrt[3]{f(x,z)}$ rule.

Further, for any distribution and given number of strata the set of AOSB will remain unchanged with respect to the form of conditional variance. However, the efficiency of the stratification will differ from stratified simple random sampling estimators as well as other estimators with the choice of various forms of conditional variance.

8. Empirical study

For the purpose of empirical study, the effectiveness of the methods of finding approximation to the optimum points of stratification, we have considered the system of minimal equations obtained for the case of proportional allocation. In this illustration we shall consider equal interval approximation and the system of approximations given in (6) article. The former approximation is specially considered due to its simplicity. From all the later approximations we have only chosen one suitable method. Since the order of approximation involved in all these methods is the same, this one approximation will give the idea about the effectiveness of all other approximations given in article (6). For the sake of simplicity, the linear regression line Y on X and Z have been taken as the form $y=\alpha+\beta x+\gamma z+e$. Here, it is considered that the two auxiliary variables used for stratification are dependent. From all the subsequent approximations we have only chosen one suitable method. Since the order of approximation involved in all these methods is the same, this one approximation will give the idea about the effectiveness of all other approximations. For obtaining the stratification points under proportional allocation let us assume $c(x,z)=\alpha+\beta x+\gamma z.$ Further, let us assume that the correlation coefficient between X and Z is denoted by ρ and is equal to 0.65. Let us consider the following examples:

Empirical study 1:
Suppose

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \qquad ,x\geq 0$$

and the variable Z has

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \qquad ,z\geq 0$$

In order to obtain the OSB when both the variables are standard normally distributed by assuming the value of regression coefficients $\beta=0.65$ and $\gamma=0.57$. For obtaining total 16 strata, 4 along X variable and 4 along Z variable using the proposed Cum $\sqrt[3]{D_3\left(x,z\right)}$ rule, by solving it in Mathematica Software assuming that the distribution of X and Z are truncated at x = 6 and z = 4, respectively, we get the stratification points as below:

Table 1. OSB and Variance, for standard normally distributed auxiliary variables

OSB $\left(x_h, z_k\right)$	**Variance** Cum $\sqrt[3]{D_3(x,z)}$ Rule	**Variance (Singh 1975)**	**% R.E.**
(0.3347,0.2673)			
(0.5779,0.2673)			
(1.9004,0.2673)			
(6.0000,0.2673)			
(0.3347,0.5284)			
(0.5779,0.5284)			
(1.9004,0.5284)			
(6.0000,0.5284)	0.06798628	0.182346	268.21
((0.3347,0.9865)			
(0.5779,0.9865)			
(1.9004,0.9865)			
(6.0000,0.9865)			
(0.3347,4.0000)			
(0.5779,4.0000)			
(1.9004,4.0000)			
(6.0000,4.0000)			

Empirical study 2:

Let

$$f(x) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} ; x > 0, \sigma > 0 \\ 0 \qquad\qquad , elsewhere \end{cases}$$

and

$$f(z) = \begin{cases} \dfrac{1}{b-a}, a \le z \le b \\ 0 \qquad , elsewhere \end{cases}$$

To obtain the OSB in proportional allocation using the proposed Cum $\sqrt[3]{D_3\left(x,z\right)}$ rule for uncorrelated auxiliary variables having densities as defined above. Standardised log-normal distribution is defined in the interval $x \in [0,10]$ and the other variable $z \in [0,1]$, and β=0.82 & γ=0.437. For 3×2 (L×M) = 6 strata, i.e. 3 along X variable and

2 along Z variable, the results obtained after solving the function using Mathematica Software are presented in the following tables as:

Table 2. OSB and Variance, for standard lognormal and uniform distributions

OSB (x_h, z_k)	**Variance** (Cum $\sqrt[3]{D_3(x,z)}$ Rule)
(3.4216,0.4759) (6.5319,0.4759) (10.0000,0.4759) (3.4216,1.0000) (6.5319,1.0000) (10.0000,1.0000)	0.035281796

9. For independent auxiliary variables under proportional allocation (ρ = 0)

In order to propose a technique under proportional allocation when the two auxiliary variables are independent to each other we need to proceed in the same way as proceeded in the case when they were dependent only with the difference that here in this case we have to take marginal densities rather than joint densities under consideration. We can write (5.3) as

$$\left[\frac{(k_h)^2}{8}\frac{B_h c'(x)}{f(x)}\right]^{\frac{1}{3}}\left[\frac{(k_h)^2}{8}\frac{B_h c'(z)}{f(z)}\right]^{\frac{1}{3}}$$

$$=\frac{(k_h)c'(x)}{2}\left[1-\left(\frac{(f_x)c'+2f(x)c''(x)}{6f(x)c'(x)}\right)(k_h)+O(k_h)^2\right]$$

$$\frac{(k_k)c'(z)}{2}\left[1-\left(\frac{(f_z)c'+2f(z)c''(z)}{6f(z)c'(z)}\right)(k_k)+O(k_k)^2\right]$$

In a similar way, we have

$$\left[\frac{(k_i)^2}{8}\frac{B_i c'(x)}{f(x)}\right]^{\frac{1}{3}}\left[\frac{(k_j)^2}{8}\frac{B_j c'(z)}{f(z)}\right]^{\frac{1}{3}}$$

$$=\frac{(k_i)c'(x)}{2}\left[1-\left(\frac{(f_x)c'+2f(x)c''(x)}{6f(x)c'(x)}\right)(k_i)+O(k_i)^2\right]$$

$$\frac{(k_j)c'(z)}{2}\left[1-\left(\frac{(f_z)c'+2f(z)c''(z)}{6f(z)c'(z)}\right)(k_j)+O(k_j)^2\right]$$

where $B_h = \int_{x_{h-1}}^{x_h} c'^2(t_1) f(t_1) dt_1 \quad and \quad B_k = \int_{z_{k-1}}^{z_k} c'^2(t_2) f(t_2) dt_2$

However, if the terms of order $O\left(\underset{a,b}{Sup}(k_h)\right)^3 \quad and \quad O\left(\underset{c,d}{Sup}(k_k)\right)^3$ can be neglected, we replace the minimal equations approximately by

$$\left[\frac{(k_h)^2}{8}\frac{B_h c'(x)}{f(x)}\right]^{\frac{1}{3}}\left[\frac{(k_h)^2}{8}\frac{B_h c'(z)}{f(z)}\right]^{\frac{1}{3}} = \left[\frac{(k_i)^2}{8}\frac{B_i c'(x)}{f(x)}\right]^{\frac{1}{3}}\left[\frac{(k_j)^2}{8}\frac{B_j c'(z)}{f(z)}\right]^{\frac{1}{3}}$$

or in other words we have that $k_h^2 B_h \quad and \quad k_k^2 B_k$ are constants. In the case when it is possible to find a function $Q_1'(x_{h-1}, x_h) \quad and \quad Q_1'(z_{k-1}, z_k)$ such that

$$\begin{aligned}(k_h)^2 B_h &= (k_h)^2 \int_{x_{h-1}}^{x_h} c'^2(t_1) f(t_1) dt_1 \\ &= Q_1'(x_{h-1}, x_h)\left[1 + O(k_h)^2\right]\end{aligned}$$

and

$$\begin{aligned}(k_k)^2 B_k &= (k_k)^2 \int_{z_{k-1}}^{z_k} c'^2(t_2) f(t_2) dt_2 \\ &= Q_1'(z_{k-1}, z_k)\left[1 + O(k_k)^2\right]\end{aligned}.$$

The Remark 1 can be proceeded in the case of independent variables too. Similarly, an approximate system of equations can be proposed in the same way as proposed in the case when auxiliary variables are dependent, and can be written as

$$(k_h k_k)^2 \int_{x_{h-1}}^{x_h} \int_{z_{k-1}}^{z_k} c'^2(t_1) c'^2(t_2) f(t_1) f(t_2) \partial t_1 \partial t_2 = C_1.$$

The solution of this system of equations and also those that will now follow are expected to be closer to the optimum points of stratification as compared to the strata obtained composed from $k_h = \frac{b-a}{L}$ = Constant and $k_k = \frac{d-c}{M}$ = Constant. Therefore, depending on the situation one should make the approximate choice of the system of equations for obtaining the approximation of optimum points of stratification.

10. Cum $\sqrt[3]{D_4(x,z)}$ Rule

For equal intervals with given values of L and M on the cumulative cube root of $D_4(x,z)$ will give AOSB if the function $D_4(x,z) = c'^2(x)c'^2(z)f(x)f(z)$ is bounded and its first derivative exists in all $x \in [a,b]$ and $z \in [c,d]$.

Remarks:

1. If the functions $c'^2(x)$ *and* $c'^2(z)$ are constants, then the proposed method is reduced to cum $\sqrt[3]{f(x)f(z)}$ rule.
2. If we take $c(x) = c(z)$ and $f(x) = f(z)$, then the proposed method is reduced to the Yadava and Singh (1984) method of Cum $\sqrt[3]{B_2(x)}$

   where $B_2 = \dfrac{f(x)x^2c'^2(x) + x\phi'(x) - \phi(x)}{x^3}$

11. Empirical study

We shall demonstrate empirically the efficiency of the given method obtaining approximately optimum strata boundaries (AOSB). For this purpose, we have considered the system of minimal equations obtained for the case of proportional allocation when the two auxiliary variables used for stratification are independent. From all the subsequent approximations we have only chosen one suitable method. Since the order of approximation involved in all these methods is the same, this one approximation will give the idea about the effectiveness of all other approximations. For obtaining the stratification points under proportional allocation let us assume $c(x,z) = \alpha + \beta x + \gamma z$. Let us consider the following examples:

Empirical study 3:

Let $f(x) = 2(2-x), 1 \le x \le 2$ and $f(z) = e^{-z+1}, 1 \le z \le 6$

In order to obtain stratification points when the auxiliary variable X follows right-triangular distribution defined in [1,2] and auxiliary variable Z follows exponential distribution defined in [1,6] we assume the values of $\beta = 0.567$ and $\gamma = 0.257$. While execution for obtaining OSB using Cum $\sqrt[3]{D_4(x,z)}$ Rule by solving the function using Mathematica Software for 6 strata, 2 along X variable and 3 along Z variable. The results obtained are presented in the following table.

Table 3. OSB and Variance, with right-triangular and exponential distribution

OSB (x_h, z_k)	Variance		% R.E.
	Cum $\sqrt[3]{D_4(x,z)}$ Rule	Yadava and Singh, 1984	
(1.5000,1.9474) (1.0000,1.9474) (1.5000,3.3368) (1.0000,3.3368) (1.5000,6.0000) (1.0000,6.0000)	0.089542	0.152122	169.89

Empirical study 4: Let us consider the distribution of X as right-triangular having

$$f(x) = 2(2-x), 1 \le x \le 4$$

and Z variable is having a uniformly distributed having

$$f(z) = \frac{1}{b-a}, 1 \le z \le 2$$

In order to find the OSB when one of the auxiliary variable is following right-triangular distribution and the other uniform distribution, we assume the value of $\beta = 0.56$ and $\gamma = 0.762$. The stratification points obtained for total 6 strata among that 3 along X variable and 2 along Z variable for the Cum $\sqrt[3]{D_4(x,z)}$ Rule using Mathematica Software for solving the function are presented in the following table.

Table 4. Uncorrelated variables having right-triangular and exponential distribution, OSB and Variance

OSB (x_h, z_k)	Variance		% R.E.
	Cum $\sqrt[3]{D_4(x,z)}$ Rule	Khan *et al.* (2008)	
(1.7880,1.5000) (2.6870,1.5000) (4.0000,1.5000) (1.7880,2.0000) (2.6870,2.0000) (4.0000,2.0000)	0.0354952	0.08293	233.64

12. Simulation Study

In this section, we conduct a simulation study to investigate the effectiveness of the proposed dynamic programming with the following methods (1-3) in stratification package in the R statistical software and 4 & 5 in LINGO:

1. Dalenius and Hodges [1959] cum *f* method, which is the most frequently used and better known method.
2. Gunning and Horgan [2004] method.
3. Lavallée -Hidiroglou [1988] method with Kozak's [2004] algorithm.
4. Khan et al. [2015] method
5. Proposed method.

In this study, a data set (with N = 5000) following the uniform and exponential distribution with a = 0, b=1, c= 0 and d = 2 was randomly generated by the R software. Then, the OSBs using the proposed method as discussed earlier are obtained for the three different number of strata, that is, (L,M)= (3 ,4). Then, the OSBs using the proposed method are obtained for (L,M)= (3 ,4). The OSBs are determined using cum f method, geometric method and the Lavallée-Hidiroglou (Kozak's) method using the stratification package with CV = 0.75 and Khan et al. [10] and the proposed method using LINGO.

Table 5. The variance of variables for different stratification methods

Stratification Method	**Variance (in e-09)**
Dalenius and Hodges [1959] cum $\sqrt{f}$ method	312.8371
Gunning and Horgan [2004] method	2891.916
Lavallee-Hidiroglou [1988] method using Kozak's [2004] method	728.3791
Khan et al. [2015]	589 .7021
Proposed method	203.107

From the table above, it is noted that the OSBs obtained by the cum f method and the proposed dynamic programming method are very close to each other, whereas the OBSs in the other methods, geometric, Lavallée-Hidiroglou method with Kozak's algorithm and Khan et al. (2015) differ widely from that of the proposed method. However, the table reveals that the proposed method yields the smallest variances of the estimate for all (L, M) = (3,4) as compared to all the other methods. Although the variances for the dynamic programming method are closed to the cum f method, the other two methods produce a greater variance than the dynamic programming technique. Thus, the study reveals that the proposed dynamic programming technique is more efficient than the other methods while stratifying a population with a uniform and exponential distributions.

13. Conclusion

The optimum stratification is defined as subdividing heterogeneous population into the best possible manner that makes the homogeneity within subpopulation and heterogeneity between them. Demarcation of strata boundaries is one of the main factors for efficient results in stratified random sampling. In this regard, we have proposed Cum $\sqrt[3]{D_i(x,z)}$ Rule (i = 3,4) for obtaining approximately OSB for two stratification variables having single study variables for both the dependent as well as independent cases for concomitant variables. Thus, comparing the proposed method Cum $\sqrt[3]{D_3(x,z)}$ Rule for standard normally distributed auxiliary variables with the Singh (1975), the %RE obtained is 268.21, which indicates the efficiency of the proposed method. Further, the %RE obtained while making comparisons between the proposed method (Cum $\sqrt[3]{D_4(x,z)}$ Rule) and the method given by Yadava and Singh (1984) results in 169.89 for right-triangular and exponential auxiliary variables. In the same case for right-triangular and uniform auxiliary variables the %RE comes out to be 233.64 as compared with Khan *et al.* (2008) under proportional allocation. Further, the simulation study also proved the superiority of the proposed methods with regard to the existing methods. Thus, it can be concluded that the use of two stratification variables gains efficiency over a single auxiliary variable and the proposed methods are more precise than the existing methods. The proposed strategy can be entirely applied to different distributions that describe the concomitant variables.

References

Cochran, W. G., (1961).Comparison of methods for determining strata boundaries. *Bulletin of the International Statistical Institute*, 38, pp. 345–358.

Cochran , W. G., (1963). *Sampling Technique*. John Wiley and Sons, Inc.

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

Danish, F., Rizvi, S. E .H., (2018). Optimum Stratification in Bivariate Auxiliary Variables under Neyman Allocation. *Journal of Modern Applied Statistical Methods*, 17(1), 2580. DOI: 10.22237/jmasm/1529418671.

Danish, F., Rizvi, S. E. H., (2019). Optimum Stratification by two Stratifying Variables using Mathematical Programming. *Pakistan Journal of Statistics*, 35(1), pp. 11–24.

Danish, F., Rizvi, S. E. H., Jeelani, M. I and Reashi, J.A., (2017). Obtaining Strata Boundaries under Proportional Allocation with Varying Cost of Every Unit. *Pakistan Journal of Statistics and Operations Research*, 13(3), pp. 567–574.

Danish, F., Rizvi, S. E. H., Sharma, M. K. and Jelani, M. I., (2018). Construction of Optimum Strata Boundaries for Uniform and Exponential Auxiliary Variable. *Journal of Applied Mathematics & Information Sciences Letters,* 6 (1), pp. 37–43.

Danish, F., (2018). A Mathematical Programming approach for obtaining optimum strata boundaries using two auxiliary variables under proportional allocation. *Transition in Statistics*, 19(3), pp. 507–526. DOI 10.21307/stattrans-2018-028.

Danish, F., Rizvi, S. E. H. and Bouza, C., (2020). On Approximately Optimum Strata Boundaries Using Two Auxiliary Variables. *Investigación Operacional,* 41 (3), pp. 445–461.

Khan, M. G. M., Nand, N. and Ahmad, N., (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34(2), pp. 205–214.

Khan, M. G. M., Prasad, D. K. and Rao, D. K., (2014). On optimum stratification. *International Journal of Mathematical, Computational, Natural and Physical Engineering*, 8(3), pp. 513–517.

Khan, M. G. M., Reddy, K. G. and D. K. Rao, (2015). Designing stratified sampling in economic and business surveys. *Journal of Applied Statistics*. DOI: 10.1080/02664763.2015.1018674.

Rizvi, S. E. H., Danish, F., (2018). Construction of Strata Boundaries: A Review. *Appl. Math. Inf. Sci. Lett.,* 6(1), pp. 27–36.

Singh, R., (1975). An alternative method of stratification on the auxiliary variable. *Sankhya C*, 37, pp. 100–108

Singh, R., Sukhatme, B. V., (1969). Optimum stratification. *Annals of Institute of Statistical Mathematics,* 21, pp. 515–528.

Yadava, S. S., Singh, R., (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics – Theory and Methods,* 13, pp. 2793–2806.