J. Ind. Soc. Agril. Statist.
62(3), 2008 : 266-275

A Method of Optimum Stratification for Two Study Variables

Med Ram Verma
ICAR Research Complex for NEH Region, Umiam (Barapani), Meghalaya 793 103

(Received: April 2008)

SUMMARY

The paper considers the problem of optimum stratification for two study variables when the units from different strata are selected with simple random sampling with replacement scheme. Under super population model a cumulative square root rule has been proposed to find approximately optimum strata boundaries for compromise allocation when correlation between auxiliary variable and study variables is high. A limiting expression for the trace of variance-covariance matrix has also been suggested. The paper concludes with a numerical illustration.

Key words: Auxiliary variable, Optimum stratification, AOSB (Approximately Optimum Strata Boundaries), Super population.

1. INTRODUCTION

Stratified sampling is the most popular among various sampling designs, which are extensively used in sample survey (Khan and Ahsan 2003). One reason for the stratification is that survey designer forms homogeneous strata, which is achieved if important study variables vary less within strata than in the entire unstratified population. When a stratified sampling is to be used a sampler has to deal with three basic problems such as (i) the problem of determining the number of strata, (ii) the problem of determination of optimum strata boundaries and (iii) the problem of optimum allocation of sample sizes to different strata. Once it is decided about the total number of strata and method of allocating the sample sizes to different strata, the problem of optimum stratification consist of determination of optimum strata boundaries so called construction of strata. These strata should be constructed in a proper way such that sampling variance is minimized. Optimum stratification is a technique of obtaining strata boundaries so that the variance of a particular estimator is minimized for the type of allocation envisaged.

Ghosh (1963) extended Dalenius (1950) theory for univariate stratification to more than one variates and theoretically solved the problem of optimum stratification with two characters and gave general theory under proportional method of allocation. Sadasivan and Aggarwal (1978) considered the problem of finding optimum strata boundaries with two study variables in case of Neyman allocation. They extended the exact equations given by Dalenius (1950) to the bivariate case by taking study variables as the basis for stratification. Gupta and Seth (1979) considered the problem of optimum stratification for the study of more than one characters on the basis of an auxiliary character under proportional method of allocation. Schneeberger and Pollot (1985) considered the problem of optimum stratification of two variates with proportional and optimum method of allocations of a bivariate normal distribution in dependence on the parameters. Rizvi *et al*. (2000) considered the problem of optimum stratification for two characters using proportional method of allocation by taking an auxiliary variable as stratification variable. Rizvi *et al.* (2002) considered the problem of optimum stratification for two characters under compromise method of allocation and proposed a cumulative cube root rule for determination of optimum points of stratification. Rizvi *et al.* (2004) considered the problem of optimum stratification for two study variables for varying probabilities under proportional allocation method.

For theoretical development, let us assume that there be a population of size N which is divided into L strata of N_1, N_2, $\ldots N_L$ units respectively so that

$\sum_{h=1}^{L} N_h = N$. For drawing a stratified SRSWR sample of size n, the sample of sizes n_1, n_2, ... n_L are to be drawn from respective stratum so that $\sum_{h=1}^{L} n_h = n$. It is assumed that the units from different strata are selected with simple random sampling with replacement scheme (SRSWR). Let Y_j (j = 1, 2) be the variable under study. Let Y_{hj} denote the value of the j-th study variable in the h-th stratum.

Hence unbiased estimator of population mean of $\overline{Y}_j$ is given by

$$\bar{y}_{jst} = \sum_{h=1}^{L} W_h \bar{y}_{hj} \qquad (j = 1, 2) \qquad (1.1)$$

where $W_h = \frac{N_h}{N} =$ Proportion of the elements present in the h-th stratum.

Variance of estimator $\bar{y}_{jst}$ is given by

$$V(\bar{y}_{jst}) = \sum_{h=1}^{L} \frac{W_h^2 \sigma_{hy_j}^2}{n_h} \qquad (1.2)$$

In multivariate stratified sampling where more than one characteristic is to be estimated, an allocation, which is optimum for one characteristic, may not be optimum for other characteristics. In such situations a compromise criterion is needed to work out a usable allocation, which is optimum for all characteristics. Such an allocation is called ‘compromise allocation’ because it is based on a compromise criterion. Various authors such as Neyman (1934), Chatterjee (1967), Khan *et al.* (1997) and Khan *et al.* (2003) discussed the compromise allocation.

Sukhatme *et al.* (1984) reviewed the problem of allocation with several characteristics as given by several research workers. They have shown numerically that all the compromise allocations as compared by them, are more efficient than proportional allocation. However, the compromise allocation based on the trace of the variance covariance matrix is most efficient. Hence, we have considered the case of compromise allocation based on minimization of trace of variance covariance matrix.

In the h-th stratum, the sample size n_h are determined in such a way so that for given total sample size (which amounts to fixed total cost where the cost per unit in each stratum is same) $\sum_{j=1}^{2} V(\bar{y}_{jst})$ is minimized where $V(\bar{y}_{jst})$ is the variance for j-th character.

If finite population correction factor can be neglected then the variance expression for j-th character is given by (1.2).

We have to minimize

$$\sum_{j=1}^{2} V(\bar{y}_{jst}) \qquad (1.3)$$

Now minimizing (1.3) subject to the condition $\sum_{h=1}^{L} n_h = n$ the optimum value of n_h is given by

$$n_h = n \frac{W_h \sqrt{\sigma_{hy_1}^2 + \sigma_{hy_2}^2}}{\sum_{h=1}^{L} W_h \sqrt{\sigma_{hy_1}^2 + \sigma_{hy_2}^2}} \qquad (1.4)$$

Using this value of n_h we shall obtain the variance expression for compromise allocation. Under compromise method of allocation, the optimal variance of the estimated population mean of the j-th study variable Y_j is by given

$$V(\bar{y}_{jst}) = \frac{1}{n} \sum_{h=1}^{L} \left[\frac{W_h \sigma_{hy_j}^2}{\sqrt{\sigma_{hy_1}^2 + \sigma_{hy_2}^2}} \sum_{h=1}^{L} W_h \sqrt{\sigma_{hy_1}^2 + \sigma_{hy_2}^2} \right]$$

(j = 1, 2) (1.5)

2. VARIANCE UNDER SUPER POPULATION MODEL AND MINIMAL EQUATIONS

Let us now assume that the population under consideration is a random sample from an infinite super population with same characteristics. Further, we assume that the study variables are linearly related with the auxiliary variable X so that the regression of Y_j on X is given by the linear model

$$Y_j = c_j(X) + e_j \qquad (j = 1, 2) \quad (2.1)$$

where $c_j(X)$ is a real valued function of X, e_j (X) is a error component which is normally distributed with mean zero and variance $\phi_j(x)$. Mathematically, we can express that $E(e_j \mid X) = 0$, $E(e_j e'_j \mid X, X') = 0$ for 0 $x \neq x'$ and $V(e_j \mid X) = \phi_j > 0$ for all $x \in (a, b)$ where $(b - a) < \infty$. It may be noted that $E(e_j(X)c_j(X)) = 0$ but $E(c_1(X)c_2(X)) \neq 0$ and $E(e_1(X)e_2(X)) \neq 0$.

If the joint density function of (X, Y_1, Y_2) in the super population is $f_s(x, y_1, y_2)$ and the marginal density function of X is f(x), then under model (2.1) it can be easily seen that

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx$$

$$\mu_{hy_j} = \mu_{hc_j} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} c_j(x)f(x)dx$$

$$\sigma^2_{hcj} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} c_j^2(x)f(x)dx - \mu^2_{hc_j}$$

$$\sigma_{hc_1c_2} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} c_1(x)c_2(x)f(x)dx - \mu_{hc_1}\mu_{hc_2}$$

$$\sigma^2_{hyj} = \sigma^2_{hc_j} + \mu_{h\phi_j}$$

where (x_{h-1}, x_h) are the boundaries of the h-th stratum, $\mu_{h\phi_j}$ is the expected value of the function $\phi_j(x)$ and $\phi_j(x)$ is the conditional variance of the j-th study variable.

The variance expression for compromise allocation under super population model (2.1) are therefore given by

$$\sigma_1^2 = V(\bar{y}_{1st})$$

$$= \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h(\sigma^2_{hc_1}+\mu_{h\phi_1})}{\sqrt{\sigma^2_{hc_1}+\mu_{h\phi_1}+\sigma^2_{hc_2}+\mu_{h\phi_2}}}\sum_{h=1}^{L}W_h\sqrt{\sigma^2_{hc_1}+\mu_{h\phi_1}+\sigma^2_{hc_2}+\mu_{h\phi_2}}\right] \quad (2.2)$$

$$\sigma_2^2 = V(\bar{y}_{2st})$$

$$= \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h(\sigma^2_{hc_2}+\mu_{h\phi_2})}{\sqrt{\sigma^2_{hc_1}+\mu_{h\phi_1}+\sigma^2_{hc_2}+\mu_{h\phi_2}}}\sum_{h=1}^{L}W_h\sqrt{\sigma^2_{hc_1}+\mu_{h\phi_1}+\sigma^2_{hc_2}+\mu_{h\phi_2}}\right] \quad (2.3)$$

We assume that stratification variable x is continuous with p.d.f. f(x), $a \leq x \leq b$ and the points of demarcation forming L strata are $x_1, x_2, \ldots x_L$. Let us denote the optimum points of stratification as $\{x_h\}$ then corresponding to these strata boundaries the generalized variance G, the determinant of variance covariance matrix, which is a function of point of stratification is minimum. Now determinant of generalized variance G is given by

$$|G| = \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \quad (2.4)$$

It is cumbersome to obtain even approximate solution obtained through minimization of G under compromise method of allocation; therefore, we have considered the minimization of trace of variance covariance matrix for the purpose of obtaining minimal equations and their solution.

Let us denote the trace of variance covariance matrix by tr(G) which is given by

$$\text{tr}(G) = \sigma_1^2 + \sigma_2^2 \quad (2.5)$$

Using (2.2) and (2.3) in (2.5) we get

$$\text{tr}(G) = \frac{1}{n}\left[\sum_{h=1}^{L}W_h\sqrt{\sigma^2_{hc_1}+\mu_{h\phi_1}+\sigma^2_{hc_2}+\mu_{h\phi_2}}\right]^2 \quad (2.6)$$

The relative magnitude of the two terms $(\sigma^2_{hc_j}+\mu_{h\phi_j})$ in (2.6) depends on the correlation coefficient ρ between auxiliary variable x and study variable Y_j. The cases in which ρ is high, $\sigma^2_{hc_j}$ is much larger than $\mu_{h\phi_j}$ and if it is low $\sigma^2_{hc_j}$ is much smaller than $\mu_{h\phi_j}$. In the rule developed by Rizvi *et al.* (2002)

for stratification of $x, \left|\frac{\sigma^2_{hc_j}}{\mu_{h\phi_j}}\right|$ has been taken to be less than one. This rule, therefore, is suitable for the cases where correlation between auxiliary variable x and study variable y is small. There is therefore, a need to develop new rule for stratification on x which considers an ideal case where ρ is large i.e. when $\left|\frac{\mu_{h\phi_j}}{\sigma^2_{hc_j}}\right| < 1$. Singh (1975) also discussed the similar case for univariate stratification and gave an alternative method of optimum stratification when the correlation between the study variable and auxiliary variable is high.

Using Singh and Sukhatme (1969) and Singh (1975), the approximation to $\mu_{h\phi_j}$ in h-th stratum is given by

$$\mu_{h\phi_j} = \frac{12\sigma^2_{h\Psi_j}}{k_h^2}[1+0\,(k_h^2)] \qquad (2.7)$$

where $k_h = x_h - x_{h-1}$ and the function ψ(x) is such that $\psi_j{}'(x) = \sqrt{\phi_j(x)}$.

Now when the number of strata is large then the terms of order $0(k_h^2)$ can be neglected in comparison to 1, the variance expression in (2.6) by using (2.7) and taking $k_h = (b-a)/L$ for h = 1, 2, ..., L reduces to

$$tr(G) = \frac{1}{n}\left[\sum_{h=1}^{L} W_h \sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta\left[\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}\right]}\right]^2 \qquad (2.8)$$

where $\theta = 12L^2/(b-a)^2$

In cases where the correlation ρ between auxiliary variable x and study variable Y_j is high the magnitude of $\mu_{h\phi_j}$ is quite small in comparison to $\sigma^2_{hc_j}$. The terms $(\sigma^2_{hc_j}+\mu_{h\phi_j})$ is, therefore, mainly dominated by $\sigma^2_{hc_j}$. Thus the magnitude of error, when $\mu_{h\phi_j}$ is approximated by $\theta\sigma^2_{h\psi_j}$ will not be much in comparison to $\sigma^2_{hc_j}$. It may be noted that Serfling (1968) while proposing the use of cumulative $\sqrt{f}$ method for finding the approximately optimum strata boundaries (AOSB) has altogether neglected $\mu_{h\phi_j}$ in comparison to $\sigma^2_{hc_j}$.

Now minimization of tr(G) is equivalent to the minimization of

$$\sum_{h=1}^{L} W_h \sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})} \text{ where gives}$$

$$W_h\frac{\partial}{\partial x_h}\sqrt{(h)}+\sqrt{(h)}\frac{\partial}{\partial x_h}W_h+W_i\frac{\partial}{\partial x_h}\sqrt{(i)}$$

$$+\sqrt{(i)}\frac{\partial}{\partial x_h}W_i = 0 \qquad (2.9)$$

where

$$(h) = \sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta\left(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}\right)$$

$$(i) = \sigma^2_{ic_1}+\sigma^2_{ic_2}+\theta(\sigma^2_{i\psi_1}+\sigma^2_{i\psi_2})$$

$$i = h+1 \text{ and } h = 1, 2, \ldots L$$

The expressions of the partial derivative terms involved in (2.9) can be easily obtained on the lines of Singh and Sukhatme (1969).

$$\begin{aligned}\frac{\partial(h)}{\partial x_h} = \frac{f(x_h)}{W_h} & [\{c_1(x_h)-\mu_{hc_1}\}^2+\{c_2(x_h)-\mu_{hc_2}\}^2 \\ & -(\sigma^2_{hc_1}+\sigma^2_{hc_2})+\theta[\{\psi_1(x_h)-\mu_{h\psi_1}\}^2 \\ & +\{\psi_2(x_h)-\mu_{h\psi_2}\}^2-(\sigma^2_{i\psi_1}+\sigma^2_{i\psi_2})]\end{aligned}$$

$$\begin{aligned}\frac{\partial(i)}{\partial x_h} = \frac{f(x_h)}{W_h} & [\{c_1(x_h)-\mu_{ic_1}\}^2+\{c_2(x_h)-\mu_{ic_2}\}^2 \\ & -(\sigma^2_{ic_1}+\sigma^2_{ic_2})+\theta[\{\psi_1(x_h)-\mu_{i\psi_1}\}^2 \\ & +\{\psi_2(x_h)-\mu_{i\psi_2}\}^2(\sigma^2_{i\psi_1}+\sigma^2_{i\psi_2})]\end{aligned}$$

After inserting the values of the required partial derivatives in the equation (2.9) and solving we have the required minimal equations as

$$\frac{\left(\begin{array}{l}\{c_1(x_h)-\mu_{hc_1}\}^2+\{c_2(x_h)-\mu_{hc_2}\}^2+\sigma^2_{hc_1}+\sigma^2_{hc_2} \\ \qquad +\theta[\{\psi_1(x_h)-\mu_{i\psi1}\}^2+\{\psi_2(x_h)-\mu_{h\psi_2}\}^2 \\ \qquad +\sigma^2{}_{h\psi_1}+\sigma^2_{h\psi_2}]\end{array}\right)}{\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta\{\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}\}}}$$

$$\frac{\left(\begin{array}{r}\{c_1(x_h)-\mu_{ic_1}\}^2+\{c_2(x_h)-\mu_{ic_2}\}^2+\sigma^2_{ic_1}+\sigma^2_{ic_2}\\ +\theta[\{\psi_1(x_h)-\mu_{i\psi_1}\}^2+\{\psi_2(x_h)-\mu_{i\psi_2}\}^2\\ +\sigma^2_{i\psi_1}+\sigma^2_{i\psi_2}]\end{array}\right)}{\sqrt{\sigma^2_{ic_1}+\sigma^2_{ic_2}+\theta\{\sigma^2_{i\psi_1}+\sigma^2_{i\psi_2}\}}} \tag{2.10}$$

Solution to these minimal equations (2.10) will give set of optimum points of stratification. These systems of equations are the functions of parameter values, which themselves are the function of the strata boundaries. Since, it is very difficult to obtain exact solutions of minimal equations, therefore we will try to find approximate solutions to these equations.

3. APPROXIMATE SOLUTION OF THE MINIMAL EQUATIONS

To find the approximate solutions to the minimal equations (2.10) we have to expand both sides of the minimal equations (2.10) about the point x_h, the common boundary point of the h-th and i-th strata. The series expansion for W_h, μ_{hc_j} and $\mu_{h\psi_j}$, can be obtained by using Taylor's theorem about both the upper and lower boundaries of h-th stratum on the lines of Singh and Sukhatme (1969).

The expansions for various terms used in minimal equations (2.10) are given below

$$W_h = f\,k_h\left[1-\frac{f'}{2f}K_h+\frac{f''}{6f}K_h^2-\frac{f'''}{24f}K_h^3+0(k_h^4)\right]$$

where f', f'' and f''' are respectively the first, second and third derivatives of the function f (x).

$$\mu_{h\phi_j} = \phi_j\left[1-\frac{\phi'_j}{2\phi_j}k_h+\frac{f'\phi'_j+2f\phi''_j}{12f\phi_j}k_h^2\right.$$

$$\left.-\frac{ff''\phi_j+ff'\phi''_j+f^2\phi'''_j-f'^2\phi'_j}{24f^2\phi_j}k_h^3+0(k_h)^4\right]$$

where ϕ'_j, ϕ_j'' and ϕ_j''' are respectively the first, second and third derivatives of the function $\phi_j(x)$.

$$\mu_{hc_j} = c_j\left[1-\frac{c_j'}{2c_j}k_h+\frac{f'c'_j+2fc''_j}{12fc_j}k_h^2\right.$$

$$\left.-\frac{ff''c'_j+f'f'c''_j+f^2c'''_j-f'^2c'_j}{24f^2c_j}k_h^3+0(k_h)^4\right]$$

where c'_j, c_j'' and c'''_j are respectively the first, second and third derivatives of the function $c_j(x)$.

$$\mu_{h\psi_j} = \psi_j\left[1-\frac{\psi'_j}{2\psi_j}k_h+\frac{f'\psi'_j+2f\psi''_j}{12f\psi_j}k_h^2\right.$$

$$\left.-\frac{ff''\psi'_j+ff'\psi''_j+f^2\psi'''_j-f'^2\psi'_j}{24f^2\psi_j}k_h^3+0(k_h)^4\right]$$

where ψ'_j, ψ''_j and ψ'''_j are respectively the first, second and third derivatives of the function $\psi_j(x)$.

The expansions for σ^2_{hc} and $\sigma^2_{h\psi_j}$ are obtained by

$$\sigma^2_{hc_j} = \frac{k_h^2}{12}c_j'^2\left[1-\frac{c'''_j}{c'_j}k_h+0(k_h^2)\right]$$

$$\sigma^2_{h\psi_j} = \frac{k_h^2}{12}\psi_j'^2\left[1-\frac{\psi'''_j}{\psi'_j}k_h+0(k_h^2)\right]$$

$$\{c_1(x_h)-\mu_{hc_1}\}^2$$

$$= \frac{k_h^2}{4}\left[c_1'^2+\frac{f'c_1'^2+2fc'^2+2fc'_1c''_1}{3f}k_h+0(k_h^2)\right]$$

$$\theta[\{\psi_1(x_h)-\mu_{h\psi_1}\}^2]$$

$$= \frac{k_h^2}{4}\theta\left[\psi_1'^2+\frac{f'\psi_1'^2+2f\psi'_1\psi''_1}{3f}k_h+0(k_h{}^2)\right]$$

Now putting the values of all the series expansions in the minimal equations (2.10) and on further simplification, the system of minimal (2.10) giving optimum points of stratification after simplification can, therefore, be written in the form

$$k_h[A_2-A_3k_h+0(k_h^2)] = k_i[A_2+A_3k_h+0(k_h^2)] \tag{3.1}$$

where

$$A_2 = \sqrt{f\sqrt{c_1'^2 + c_2'^2 + \theta[\psi_1'^2 + \psi_2'^2]}}$$

$$A_3 = \frac{\begin{pmatrix} f'\{c_1'^2 + c_2'^2 + \theta(\psi_1'^2 + \psi_2'^2)\} \\ +f[c_1'c_1''' + c_2'c_2''' + \theta\{\psi_1'\psi_1'' + \psi_2'\psi_2''\}] \end{pmatrix}}{\begin{pmatrix} \left(4\sqrt{f\sqrt{c_1'^2 + c_2'^2 + \theta(\psi_1'^2 + \psi_2'^2)}}\right) \times \\ \left(\sqrt{\{c_1'^2 + c_2'^2 + \theta(\psi_1'^2 + \psi_2'^2)\}}\right) \end{pmatrix}}$$

$$= \frac{1}{2}\frac{d}{dx_h}\left(\sqrt{f\sqrt{c_1'^2 + c_2'^2 + \theta[\psi_1'^2 + \psi_2'^2]}}\right)$$

The equations in (3.1) is equivalent to

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\psi_1'^2 + \psi_2'^2]}}\, dx[1 + 0(k_h^2)]\right]$$

$$= \left[k_i^2 \int_{x_h}^{x_{h+1}} \sqrt{f\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\psi_1'^2 + \psi_2'^2]}}\, dx[1 + 0(k_i^2)]\right] \tag{3.2}$$

Since $\psi_j'(x) = \sqrt{\phi_j(x)}$ hence equation (3.2) can be given by

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\, dx[1 + 0(k_h^2)\right]$$

$$= \left[k_i^2 \int_{x_h}^{x_{h+1}} \sqrt{f\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\, dx[1 + 0(k_i^2)\right] \tag{3.3}$$

Thus, if the number of strata is large so that the strata width k_h is small and the higher powers of k_h in the expansion can be neglected, then the system of minimal equations (2.10) can approximately be given as

$$\int_{x_{h-1}}^{x_h} \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\,dx = \text{constant} \tag{3.4}$$

where terms of $0(m^4), m = \sup_{(a,b)}(k_h)$ have been neglected on both sides of equation. Since $a \le x \le b$ and the points of demarcation forming L strata are x_1, x_2, ... x_L with x_1 = a and x_L = b.

While developing these equations, the higher order terms have been ignored and this is justified for a large number of strata so that the error that might have been introduced would hardly affect the solutions. Therefore, the solutions to the minimal equations, that is the set $\{x_h\}$ of approximately optimum strata boundaries (AOSB) obtained from approximate system of equations, shall be quite close to optimum values.

To calculate the sets of points $\{x_h\}$ that will satisfy the equation (3.3), the knowledge of the constant on the right hand side of (3.4) is very much essential. The value of this constant can be easily obtained by

$$\text{Contant} \quad = \frac{1}{L}\int_a^b \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\ dx \tag{3.5}$$

Then the solution of the system of minimal equations (2.10) or equivalently the system of equation (3.4) can be put as

$$\int_{x_{h-1}}^{x_h} \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\ dx$$

$$= \frac{1}{L}\int_a^b \sqrt{f(x)\sqrt{c_1'^2(x) + c_2'^2(x) + \theta[\phi_1(x) + \phi_2(x)]}}\ dx \tag{3.6}$$

We now propose the following rule for finding approximately optimum strata boundaries (AOSB) as given below.

Cumulative $\sqrt{M_7(x)}$ Rule

If the function

$$M_7(x) = f(x)\sqrt{c_1'^2(x) + c_2'^2(x)\theta[\phi_1(x)\phi_2(x)]}$$

is bounded and its first two derivatives exists for all x in (a, b) with (b – a) < ∞, then for a given value of L taking equal intervals on the cumulative of $\sqrt{M_7(x)}$ will give approximately optimum strata boundaries (AOSB).

Remarks : When $\phi_1(x) = \phi_2(x) = 0$ and $c_1(x) = c_2(x) = a + bx$ then this reduces to the cumulative $\sqrt{f}$ rule of Dalenius and Hodges (1957). Thus cumulative $\sqrt{f}$ rule becomes a particular case of the proposed rule. However, the proposed rule can be applied even more general situations when $\phi_j(x)$ is not constant.

4. LIMITING FORM OF THE TRACE OF THE VARIANCE-COVARIANCE MATRIX

In this section, we shall express the trace of variance-covariance matrix tr(G), as given in (2.8) in terms of number of strata L and some other constants which do not depend on the strata boundaries. This expression is particularly important in approximately optimum stratification on the auxiliary variable as it gives an insight into the manner in which the variance of the estimator $\bar{y}_{jst}$ is reduced with the increase in the number of strata.

Lemma 4.1. Under regularity conditions as given in Section 3, for h-th stratum we have

$$W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}$$

$$=\frac{1}{\sqrt{12}}\left[\int_{x_{h-1}}^{x_h}\sqrt{M_7(x)}\,dx\right]^2[1+0(k_h^2)]$$

where

$$M_7(x) = f(x)\sqrt{c_1'^2(x)+c_2'^2(x)+\theta[\phi_1(x)+\phi_2(x)]}$$

Proof. Using the various series expansion from Singh and Sukhatme (1969) we get

$$\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})$$

$$=\frac{K_h^2}{12}[c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)$$
$$-[c_1'c_1'''+c_2'c_2'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2''\})]k_h+0(k_h^2)] \tag{4.1}$$

Taking square root of the equation (4.1) we get

$$\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}$$

$$=\frac{K_h}{\sqrt{12}}\begin{array}{l}[\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}\\ -[c_1'c_1'''+c_2'c_2'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2''\}]k_h+0(k_h^2)]\end{array} \tag{4.2}$$

Equation (4.2) after further simplification can be written as

$$\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}$$

$$=\frac{K_h}{\sqrt{12}}\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}$$

$$\times\left[1-\frac{[c_1'c_1'''+c_2'c_2'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2''\}]}{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}k_h+0(k_h^2)\right] \tag{4.3}$$

Multiplying the equation (4.3) by W_h we get

$$W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}$$

$$=\frac{K_h}{\sqrt{12}}f\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}$$

$$\times\left[1-\frac{\begin{array}{l}[f'\{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)\}\\ +f[c_1'c_1'''+c_2'c_2'''+\theta\{\psi'\psi_1''+\psi_2'\psi_2''\}]\end{array}}{2f\{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)\}}k_h+0(k_h^2)\right] \tag{4.4}$$

Now taking the square root of the equation (4.4) we get

$$\left[W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}}\right]^{1/2}$$

$$=\frac{K_h}{\sqrt[4]{12}}\sqrt{f\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}}$$

$$\times\left[1-\frac{\begin{array}{l}[f'\{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)\}\\ +f[c_1'c_1'''+c_1'c_1'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2''\}]\end{array}}{2f\{c_1'^2+c_2'^2+(\psi_1'^2+\psi_2'^2)\}}k_h+0(k_h^2)\right]^{1/2} \tag{4.5}$$

$$\left[W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}}\right]^{1/2}$$

$$=\frac{K_h}{\sqrt[4]{12}}\sqrt{f\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}}$$

$$\times\left[1-\frac{\begin{array}{c}[f'\{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)\}\\ +f[c_1'c_1'''+c_2'c_2'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2''\}]\end{array}}{4f\{c_1'^2+c_2'^2+(\psi_1'^2+\psi_2'^2)\}}k_h+0(k_h^2)\right] \tag{4.6}$$

$$\left[W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}\right]^{1/2}$$

$$=\frac{K_h}{\sqrt[4]{12}}\sqrt{f\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}}$$

$$\frac{\begin{array}{c}[f'(c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)\\ +f[c_1'c_1'''+c_2'c_2'''+\theta\{\psi_1'\psi_1''+\psi_2'\psi_2'''\}]\end{array}}{\begin{array}{c}[\left(4\sqrt{f\sqrt{c_1'^2+c_2'^2+\theta(\psi_1'^2+\psi_2'^2)}}\right)\times\\ \left(\sqrt{\{c_1'^2+c_2'^2+(\psi_1'^2+\psi_2'^2)\}}\right)]\end{array}}k_h+0(k_h^2) \tag{4.7}$$

$$\left[W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2})}\right]^{1/2}$$

$$=\frac{1}{\sqrt[4]{12}}\int_{x_{h-1}}^{x_h}\sqrt{M_7(x)}\,dt[1+0(k_h^2)] \tag{4.8}$$

Now squaring the equation (4.8) we get

$$W_h\sqrt{\sigma^2_{hc_1}+\sigma^2_{hc_2}+\theta\left(\sigma^2_{h\psi_1}+\sigma^2_{h\psi_2}\right)}$$

$$=\frac{1}{\sqrt{12}}\left[\int_{x_{h-1}}^{x_h}\sqrt{M_7(x)dx}\right]^2[1+0(k_h^2)]$$

which completes the proof of the Lemma.

Using Lemma 4.1 in the expression (2.8), we have

$$\mathrm{tr}(G)=\frac{1}{n}\left[\sum_{h=1}^{L}\frac{1}{\sqrt{12}}\left(\int_{x_{h-1}}^{x_h}\sqrt{M_7(x)}\,dx\right)^2\right]^2 \tag{4.9}$$

Now if the strata boundaries are determined by making use of cumulative $\sqrt{M_7(x)}$ rule then for h = 1, 2, …. L we have

$$\int_{x_{h-1}}^{x_h}\sqrt{M_7(x)}\,dx = \frac{1}{L}\int_a^b\sqrt{M_7(x)}\,dx \tag{4.10}$$

Using (4.10) in (4.9) we have

$$\mathrm{tr}(G) = \frac{1}{n}\left[\frac{1}{\sqrt{12}L}\left(\int_a^b\sqrt{M_7(x)}\,dx\right)^2\right]^2 \tag{4.11}$$

which can be written as

$$\mathrm{tr}(G) = \frac{\lambda^2}{nL} \tag{4.12}$$

where $$\lambda = \frac{1}{\sqrt{12}}\left(\int_a^b\sqrt{M_7(x)}\,\,dx\right)^2$$

Now taking limit as L → ∞ on both sides of (4.12) we get

$$\lim_{L\to\infty}\mathrm{tr}(G) = 0 \tag{4.13}$$

From the above relation it may be concluded that with an increase in the number of strata L, the trace of generalized variance decreases and as the number of strata becomes large enough, tr(G) tends to zero.

5. EMPIRICAL STUDY

To determine approximately optimum strata boundaries (AOSB) by the use of proposed cumulative square root rule we consider that stratification variable x follows uniform, right triangular and exponential distributions with probability density functions given by

Uniform distribution	$f(x) = 1$	$1 \le x \le 2$
Right triangular distribution	$f(x) = 2(2-x)$	$1 \le x \le 2$
Exponential distribution	$f(x) = e^{-x+1}$	$1 \le x < \infty$

The ranges of both uniform and right triangular distribution are finite where as range of exponential distribution is infinite. For the purpose of numerical computation exponential distribution is truncated at x = 6 so that the probability beyond the truncation point is very small. We have considered that study variables Y_j are related with the stratification variable x as

$Y_1 = x + e_1$, $Y_2 = 2x + e_2$. The conditional variances of the error terms i.e. $V(e_1 / x)$ and $V(e_2 / x)$ are assumed to be of the forms $A_1x_1^g$ and $A_2x_2^g$ respectively where $A_1, A_2 > 0$, g_1 and g_2 being constants. Here we have taken $g_1 = 1$ and $g_2 = 2$. The values of A_1 and A_2 were determined for the values of g_1, g_2, ρ_1 and ρ_2 by using the following formulae.

$$A_1 = \frac{\beta_1\sigma_x^2(1-\rho_1^2)}{\rho_1^2 E(x^{g_1})} \text{ and } A_2 = \frac{\beta_2\sigma_x^2(1-\rho_2^2)}{\rho_2^2 E(x^{g_2})}$$

where ρ_1 and ρ_2 are the correlation coefficients between the study variables Y_1 and Y_2 with stratification variable x. σ_x^2 is the variance of the stratification variable x.

Table 5.1. Percent relative efficiency of stratification for uniform distribution

No. of Strata L	Strata boundaries					n tr(G)	Percent Relative Efficiency
1						0.568767	100.00
2	1.536961					0.256966	221.34
3	1.368413	1.697858				0.196882	288.89
4	1.280719	1.536961	1.775369			0.175356	324.35
5	1.226701	1.436872	1.634204	1.821288		0.165255	344.18
6	1.19035	1.368413	1.536961	1.697858	1.851486	0.159725	356.09

Table 5.2. Percent relative efficiency of stratification for right triangular distribution

No. of Strata L	Strata boundaries					n tr(G)	Percent Relative Efficiency
1						0.379189	100.00
2	1.400927					0.176394	214.97
3	1.262237	1.552118				0.134247	282.46
4	1.194956	1.400927	1.634553			0.118917	318.87
5	1.155597	1.316807	1.489507	1.686958		0.111696	339.48
6	1.129358	1.262237	1.400927	1.552118	1.725223	0.107729	351.98

Table 5.3. Percent relative efficiency of stratification for exponential distribution

No. of Strata L	Strata boundaries					n tr(G)	Percent Relative Efficiency
1						6.646229	100.00
2	2.594839					3.645847	182.30
3	1.990916	3.336816				2.952229	225.13
4	1.731112	2.594839	3.745069			2.701630	246.01
5	1.575230	2.226342	3.010215	4.104782		2.585356	257.07
6	1.473836	1.990916	2.594839	3.336816	4.340938	2.522812	263.45

For the purpose of numerical illustration we have assumed $\rho_1^2 = 0.9$, and $\rho_2^2 = 0.7$. For finding out the approximately optimum strata boundaries (AOSB), the ranges of uniform, right triangular and exponential distributions were divided into 10 classes of equal width. The function $\sqrt{M_7(x)}$ was evaluated at the middle point of the class intervals and $\sqrt{M_7(x)}$ was then found for each of 10 classes. These cube roots were cumulated and AOSB were obtained by taking equal intervals on the cumulative totals. Approximately optimum strata boundaries (AOSB) obtained by the use of proposed cumulative square root rule are given in Table 5.1 to Table 5.3 along with relative efficiency of stratification with no stratification.

REFERENCES

Chatterjee, S. (1967). A note on optimum allocation. *Scond. Actu. J.*, **50,** 40-44.

Dalenius, T. (1950). The problem of optimum stratification. *Skand. Akt*., **33,** 203-213.

Dalenius, T. and Hodges (jr.), J.L. (1957). Minimum variance stratification. *J. Amer. Statist . Assoc.*, **54,** 88 –101.

Ghosh, S.P. (1963). Optimum stratification with two characters. *Ann. Math. Statist.*, **34**, 866-872.

Gupta, P.C. and Seth, G.R. (1979). On stratification in sampling investigation involving more than one character. *J. Ind. Soc. Agril. Statist.*, **31(2)**, 1-15.

Khan, M.G.M. and Ahsan, M.J. (2003). A note on optimum allocation in multivariate stratified sampling. *South Pacific J. Nat. Sci.*, **21,** 91-95.

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.

Khan, M.G.M., Khan, E.A. and Ahsan, M.J. (2003). An optimal multivariate stratified sampling design using dynamic programming. *Austr. and Newzealand J. Statist.*, **45(1)**, 107-113..

Neyman, J. (1934). On the two different aspects of representative methods: The method of stratified sampling and method of purposive selection. *J. Roy. Statist. Soc.*, **97**, 558-606.

Rizvi, S.E.H., Gupta, J.P. and Bhargava, M. (2004). Effect of optimum stratification on sampling with varying probabilities under proportional allocation. *Statistica*, **64(4)**, 721-733.

Rizvi, S.E.H., Gupta, J.P. and Bhargava, M. (2002). Optimum stratification based on auxiliary variable for compromise allocation. *Metron*, **LX (3-4)**, 201-215.

Rizvi, S.E.H., Gupta, J.P. and Singh, R. (2000). Approximately optimum stratification for two study variables using auxiliary information. *J. Ind. Soc. Agril. Statist.*, **53(3)**, 287-298.

Sadasivan, G. and Aggarwal, R. (1978). Optimum points of stratification in bi-variate populations. *Sankhya*, **C40,** 84-97.

Schneeberger, H. and Pollot, J.P. (1985). Optimum stratification with two variates. *Statistische Hefte*, **26,** 97-113.

Serfling, R.J. (1968). Approximately optimal stratification. *J. Amer. Statist. Assoc.*, **63**, 1298-1309.

Singh, R. (1975). An alternative method of stratification on the auxiliary variable. *Sankhya*, **C37**, 100-108.

Singh, R. and Sukhatme, B.V. (1969). Optimum stratification. *Ann. Inst. Statist. Math.*, **21**, 515-528.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory with Applications*. Indian Society of Agricultural Statistics, New Delhi and IOWA State University Press, Ames, USA.