J. Ind. Soc. Agril. Statist.
61(2), 2007 : 66-76

Optimum Stratification for PPS Sampling using Auxiliary Information

Med Ram Verma and S.E.H. Rizvi[1]
ICAR Research Complex for NEH Region, Umaiam (Barapani), Meghalaya
(Received : May, 2006)

SUMMARY

The paper considers the problem of optimum stratification for two study variables when the units from different strata are selected with probability proportional to size with replacement (PPSWR) sampling scheme. Minimal equations solution which give optimum points of stratification have been obtained by minimizing trace of variance-covariance matrix. A cumulative cube root rule $\sqrt[3]{M_6(x)}$ has been proposed to find approximate solution to the minimal equations. Limiting expression for the generalized variance-covariance matrix, the optimum numbers of strata and expression for the approximate sample sizes have also been obtained. The paper concludes with a numerical example.

Key words : Compromise allocation, Auxiliary variable, Optimum stratification, AOSB (Approximately Optimum Strata Boundaries).

1. INTRODUCTION

The problem of determining optimum strata boundaries, when both the estimation and the stratification variables are the same, was first considered by Dalenius (1950). The subsequent work in this direction is also well known. Regarding the optimum stratification on the auxiliary variable x, Dalenius and Gurney (1951) considered the case of optimum stratification for Neyman allocation while Taga (1967) considered the case of proportional allocation. Subsequently, Singh and Sukhatme (1969) have obtained the minimal equations giving optimum strata boundaries on the auxiliary variable for the case of Neyman allocation (minimizing the variance for fixed n) and also have suggested various methods of finding their approximate solutions, in more general form.

Ghosh (1963) extended Dalenius (1950) theory for univariate stratification to more than one variates. He 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 (OSB) 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 for two variables with proportional and optimum method of allocations for a bivariate normal distribution. Rizvi *et al.* (2000) considered the 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.

Singh and Sukhatme (1972) considered the problem of optimum stratification for univariate using varying probabilities under Neyman allocation. Recently, Mahajan and Singh (2005) discussed the

[1] *Sher-e-Kashmir University of Agricultural Sciences and Technology, Main Campus, Jammu*

problem of optimum stratification for a sensitive variable using PPSWR sampling scheme. Rizvi *et al.* (2004) considered the problem of optimum stratification for two study variables using varying probabilities under proportional allocation. In this paper we have considered the problem of optimum stratification for two study variables based on an auxiliary variable x when the samples from different strata are selected with probability proportional to size with replacement scheme under compromise method of allocation.

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, ...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 drawn with probability proportional to the values of the auxiliary variable x and with replacement scheme (PPSWR). Let Y_j $(j=1, 2)$ be two variables 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 $\bar{Y}_j$ is given by

$$\bar{y}_j = \frac{1}{N}\sum_{h=1}^{L} \hat{y}_{hy_j} \qquad (1.1)$$

where $\hat{y}_{hy_j}$ is the unbiased estimator of the h-th stratum total of j-th variable.

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

$$V(\bar{y}_j) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{\sigma_{hj}^2}{n_h} \qquad (1.2)$$

$$\sigma_{hj}^2 = X_h^2 \sum_{i=1}^{N_h} \frac{y_{hij}^2}{x_{hi}} - Y_{hj}^2$$

where y_{hij} is the value of i-th unit of j-th study variable in h-th stratum, x_{hi} is the value of the i-th unit of the auxiliary variable and X_h is the stratum total of the auxiliary variable x.

2. VARIANCE UNDER SUPER POPULATION MODEL

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

$$Y_j = c_j(X) + e_j \qquad (2.1)$$

where $c_j(X)$ is a real valued function of X, e_j is a error component such that

$E(e_j \mid X) = 0$, $E(e_j e_j' \mid X, X') = 0$ for $x \neq x'$ 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

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

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

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

$$\sigma_{hc_j}^2 = W_h^{-1} \int_{x_{h-1}}^{x_h} c_j^2(x)f(x)dx - \mu_{hc_j}^2$$

$$\sigma_{hy_j}^2 = \sigma_{hc_j}^2 + \mu_{h\phi_j}$$

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.

Under model (2.1), the expected value of the variance $V(\bar{y}_j)$ is given by

$$E\left(V(\bar{y}_j)\right) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{E(\sigma_{hj}^2)}{n_h} \qquad (2.2)$$

Now

$$E(\sigma_{hj}^2) = N^2 W_h^2 [\mu_{hx}\mu_{h\theta_j} - \mu_{hc_j}^2] \qquad (2.3)$$

where W_h is the proportion of the units in the h-th stratum and $\theta_j(x) = \frac{c_j^2(x) + \phi_j(x)}{x}$

Now using (2.3) in (2.2) we get the expected value of the variance $EV(\bar{y}_j)$ as given below.

$$E\left(V(\bar{y}_j)\right) = \sum_{h=1}^{L}\frac{W_h^2 [\mu_{hx}\mu_{h\theta_j} - \mu_{hc_j}^2]}{n_h} \quad (j = 1, 2) \qquad (2.4)$$

3. COMPROMISE ALLOCATION IN STRATIFIED SAMPLING

The problem of allocation to strata with several characteristics was first considered by Neyman (1934). Sukhatme *et al.* (1984) have 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) $\sum_{j=1}^{2} EV(\bar{y}_j)$ is minimized where $EV(\bar{y}_j)$ is the expected value of variance for j-th variable. If finite population correction factor can be neglected then the variance expression for j-th character is given by (2.4).

We have to minimize

$$\sum_{j=1}^{2} EV(\bar{y}_j) = \sum_{j=1}^{2}\sum_{h=1}^{L}\frac{W_h^2 [\mu_{hx}\mu_{h\theta_j} - \mu_{hc_j}^2]}{n_h} \qquad (3.1)$$

Now minimizing (3.1) subject to the condition $\sum_{h=1}^{L} n_h = n$ the optimum value of n_h is given by

$$n_h = n\frac{W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}}{\sum_{h=1}^{L} W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}} \qquad (3.2)$$

where

$$P_{hc_j}^2 = \mu_{hx}\mu_{h\theta_j} - \mu_{hc_j}^2$$

Using this value of n_h we have obtained the variance expression under compromise allocation. The optimal variance expression of the estimated population mean of the Y_j under super population model is given by

$$\sigma_j^2 = EV(\bar{y}_j)$$

$$= \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h P_{hc_j}^2}{\sqrt{P_{hc_1}^2 + P_{hc_2}^2}}\sum_{h=1}^{L} W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}\right]$$

$$(j = 1, 2) \qquad (3.3)$$

4. MINIMAL EQUATIONS

We assume that stratification variable is continuous with pdf f(x), $a \le x \le b$ 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. These $\{x_h\}$ are the solutions of the minimal equations. Now determinant of generalized variance G is given by

$$G = \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \qquad (4.1)$$

It is cumbersome to obtain even approximate solution to the minimal equations 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 G^T which is given by

$$G^T = \sigma_1^2 + \sigma_2^2 \tag{4.2}$$

Using (3.3) in (4.2) G^T can be expressed as

$$G^T = \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h P_{hc_1}^2}{\sqrt{P_{hc_1}^2 + P_{hc_2}^2}}\sum_{h=1}^{L} W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}\right]$$

$$+\frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h P_{hc_2}^2}{\sqrt{P_{hc_1}^2 + P_{hc_2}^2}}\sum_{h=1}^{L} W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}\right]$$

$$\Rightarrow G^T$$

$$= \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h(P_{hc_1}^2 + P_{hc_2}^2)}{\sqrt{P_{hc_1}^2 + P_{hc_2}^2}}\sum_{h=1}^{L} W_h\sqrt{P_{hc_1}^2 + P_{hc_2}^2}\right]$$

$$G^T = \frac{1}{n}\left[\sum_{h=1}^{L} Wh\sqrt{(P_{hc1}^2 + P_{hc2}^2}\right]^2 \tag{4.3}$$

Now minimization of G^T is equivalent to the minimization of

$$\sum_{h=1}^{L} W_h\sqrt{P_{hc1}^2 + P_{hc2}^2} \quad \text{which gives}$$

$$W_h\frac{\partial}{\partial x_h}\sqrt{(h)} + \sqrt{(h)}\frac{\partial}{\partial x_h}W_h + W_i\frac{\partial}{\partial x_h}\sqrt{(i)} + \sqrt{(i)}\frac{\partial}{\partial x_h}W_i = 0 \tag{4.4}$$

where $I = h + 1$ and $h = 1, 2, \ldots L$

$(h) = P_{hc1}^2 + P_{hc2}^2$

$(i) = P_{ic1}^2 + P_{ic2}^2$

The expressions of the partial derivative terms involved in (4.4) can be easily obtained on the lines of Singh and Sukhatme (1969). Now inserting the values of the required partial derivatives in the equation (4.4) and solving we have the required minimal equations as

$$\frac{\theta_1(x_h)\mu_{hx} + x_h\mu_{h\theta 1} - 2\mu_{hc1}c_1(x_h) + \theta_2(x_h)\mu_{hx} + x_h\mu_{h\theta 2} - 2\mu_{hc2}c_2(x_h)}{\sqrt{\mu_{h\theta_1}\mu_{hx} - \mu_{hc_1}^2 + \mu_{h\theta_2}\mu_{hx} - \mu_{hc_2}^2}}$$

$$= \frac{\theta_1(x_h)\mu_{ix} + x_h\mu_{h\theta_1} - 2\mu_{ic_1}c_1(x_h) + \theta_2(x_h)\mu_{ix} + x_h\mu_{h\theta_2} - 2\mu_{hc_2}c_2(x_h)}{\sqrt{\mu_{h\theta_1}\mu_{ix} - \mu_{ic_1}^2 + \mu_{h\theta_2}\mu_{ix} - \mu_{hc_2}^2}} \tag{4.5}$$

Solution to these minimal equations (4.5) will give set of optimum points of stratification. This system of equations is the functions of parameter values, which themselves are the function of points of 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.

5. APPROXIMATE SOLUTION OF MINIMAL EQUATIONS

To obtain the approximate solutions to the minimal equations (4.5) we have to expand both sides of the minimal equations 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}, μ_{hx} and $\mu_{h\theta_j}$ 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 series expansions of $\mu_\phi(y, x)$, the mean of the function $\phi(t)$ in the interval (y, x), about the point $t = y$ is given by

$$\mu_\phi(y, x) = \int_y^x \phi(t)f(t)dt \Big/ \int_y^x f(t)dt$$

$$= \phi\left[1 + \frac{\phi'}{2\phi}k + \frac{\phi' f' + 2f\phi''}{12f\phi}k^2 + \frac{(ff''\phi' + ff'\phi'' + f^2\phi''' - f'^2\phi')}{24f^2\phi}k^3 + O(k^4)\right] \tag{5.1}$$

In order to obtain the series expansions of the minimal equations (4.5), these relations are to be used with (y, x) being replaced by (x_{h-1}, x_h). The expansions for various terms used in minimal equations (4.5) are obtained by using (5.1) as given below

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

$$W_i = fk_i[1+\frac{f'}{2f}k_i+\frac{f''}{6f}k_i^2 + \frac{f'''}{24f}k_i^3+O(k_i^4)\ ]$$

$$\mu_{h\phi} = \phi[1-\frac{\phi'}{2\phi}k_h+\frac{f'\phi'+2f\phi''}{12f\phi}k_h^2 - \frac{ff''\phi'+ff'\phi''+f^2\phi'''-f'^2\phi'}{24f^2\phi}k_h^3+O(k_h^4)]$$

$$\mu_{i\phi} = \phi[1+\frac{\phi'}{2\phi}k_i+\frac{f'\phi'+2f\phi''}{12f\phi}k_i^2 + \frac{ff''\phi'+ff'\phi''+f^2\phi'''-f'^2\phi'}{24f^2\phi}k_i^3+O(k_i^4)]$$

$$\mu_{hc} = \phi[1-\frac{c'}{2c}k_h+\frac{f'c'+2fc''}{12fc}k_h^2 - \frac{ff''c'+ff'c''+f^2c'''-f'^2c'}{24f^2c}k_h^3+O(k_h^4)]$$

$$\mu_{ic} = \phi[1+\frac{c'}{2c}k_i+\frac{f'c'+2fc''}{12fc}k_i^2 + \frac{ff''c'+ff'c''+f^2c'''-f'^2c'}{24f^2c}k_i^3+O(k_i^4)]$$

where the functions ϕ, f and their derivatives are evaluated at t = y and k = x − y.

Similarly, expanding $\sqrt[\lambda]{f(t)}$ about the point t = y, we have

$$\left[\int_y^x \sqrt[\lambda]{f(t)}dt\right]^\lambda = k^\lambda f(y)\left[1+\frac{k}{2}.\frac{f'(y)}{f(y)}+O(k^2)\right] = k^{\lambda-1}\int_y^x f(t)\,dt[1+O(k^2)] \quad (5.2)$$

Now using the above various expressions the system of minimal (4.5) giving optimum points of stratification after simplification can, therefore, be written in the form

$$\begin{aligned}&2\sqrt{\phi_1+\phi_2}\left[1+A_2k_h^2-A_3k_h^3+O(k_h^4)\right]\\ &=2\sqrt{\phi_1+\phi_2}\left[1+A_2k_i^2+A_3k_i^3+O(k_i^4)\right]\end{aligned} \quad (5.3)$$

where $k_h = x_h - x_{h-1}$, $k_i = x_{h+1} - x_h$ are the stratum widths for h-th and (h+1)th strata and

$$A_2 = \frac{(\phi_1'+\phi_2')^2-4(\phi_1+\phi_2)(\lambda_1+\lambda_2)}{32f(\phi_1+\phi_2)^2}$$

$$A_3 = \frac{1}{96f\sqrt{\phi_1+\phi_2}}\frac{d}{dx_h}\left[\frac{f(\phi_1'+\phi_2')^2-4f(\phi_1+\phi_2)(\lambda_1+\lambda_2)}{(\phi_1+\phi_2)^{3/2}}\right]$$

$\lambda_1 = \theta_1' - c_1'^2$ and $\lambda_2 = \theta_2' - c_2'^2$

where ϕ_1' is the first order derivative of ϕ_1 and ϕ_2' is the first order derivative of ϕ_2.

Now after canceling $2\sqrt{\phi_1+\phi_2}$ from both sides of the equation (5.3) and multiplying by $f(x_h)$ we get on simplification

$$\begin{aligned}&\frac{k_h^2}{16}\left[P(t)f(t)-\frac{1}{3}\frac{d}{dx_h}[P(t)f(t)]\ k_h+O(k_h^2)\right]\\ &=\frac{k_i^2}{16}\left[P(t)f(t)+\frac{1}{3}\frac{d}{dx_h}[P(t)f(t)]k_i+O(k_i^2)\right]\end{aligned} \quad (5.4)$$

where

$$P(t) = \frac{(\phi_1'(t)+\phi_2'(t))^2-4(\phi_1(t)+\phi_2(t))\ (\lambda_1(t)+\lambda_2(t))}{(\phi_1(t)+\phi_2(t))^{3/2}}$$

Using these expansions, the system of minimal equations (4.5) reduces to

$$\begin{aligned}&k_h^2\left[1-\frac{k_h}{3}.\frac{[P(t)f(t)]'}{P(t)f(t)}+O(k_h^2)\right]\\ &=k_i^2\left[1+\frac{k_i}{3}.\frac{[P(t)f(t)]'}{P(t)f(t)}+O(k_i^2)\right]\end{aligned} \quad (5.5)$$

On raising both sides of the above equation (5.5) to the power 3/2 and using binomial theorem (for any index), we get

$$\begin{aligned}&k_h^3\left[1-\frac{k_h}{2}.\frac{[P(t)f(t)]'}{P(t)f(t)}+O(k_h^2)\right]\\ &=k_i^3\left[1+\frac{k_i}{2}.\frac{[P(t)f(t)]'}{P(t)f(t)}+O(k_i^2)\right]\end{aligned} \quad (5.6)$$

On comparing it with (6.2), with $\lambda = 3$, the system of equations (5.6) can be written in the form

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} P(t)f(t)dt[1+O(k_h^2)]\right] = \left[k_i^2 \int_{x_h}^{x_{h+1}} P(t)f(t)dt[1+O(k_i^2)]\right] \quad (5.7)$$

The functions ϕ_1, ϕ_2, λ_1, λ_2 and their derivatives are evaluated at the point x_h and we assume that the function $P(x)f(x) \in \Omega$ for all x in (a, b). 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 (4.5) can approximately be given as

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} P(t)f(t)dt\right] = \left[k_i^2 \int_{x_h}^{x_{h+1}} P(t)f(t)dt\right] \quad (5.8)$$

Or equivalently by

$$k_h^2 \int_{x_{h-1}}^{x_h} P(t)f(t)dt = \text{constant}, \quad h = 1,2,...L \quad (5.9)$$

where terms of $O(m^4)$, $m = \sup_{(a,b)}(k_h)$ 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 $x_L = b$.

Further, if we take a function $Q(x_{h-1}, x_h)$ of order $O(m^3)$ such that

$$k_h^2 \int_{x_{h-1}}^{x_h} P(t)f(t)dt = Q(x_{h-1}, x_h)\ [1+O(k_h^2)] \quad (5.10)$$

Then the system of equations (4.5) can approximately be put as

$$Q(x_{h-1}, x_h) = \text{constant}, \quad h = 1,2,...L \quad (5.11)$$

Various methods of finding approximate solutions to the minimal equations can be established through the system of equations (5.11). Singh and Sukhatme (1969) developed different forms of the function $Q(x_{h-1}, x_h)$ corresponding to univariate case under Neyman allocation. One such function gives cum. $\sqrt[3]{M_6(x)}$ rule according to which the approximately optimum strata boundaries (AOSB) are solutions of the system of equation (4.5). Proceeding on the same lines, one such form of function $Q(x_{h-1}, x_h)$ can also be obtained as follows

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{M_6(t)}\ dt = \int_a^b \sqrt[3]{P(t)f(t)}\ dt \Big/ L \quad (5.12)$$

Thus, we get the following cumulative cube root rule for finding AOSB on the non-sensitive auxiliary variable when the estimation variables are also non-sensitive.

Cumulative $\sqrt[3]{M_6(x)}$ Rule

If the function $M_6(x) = P(x)f(x)$ is bounded and its first two derivatives exists for all x in (a,b) with $(b-a) < \infty$, then for a given value of L taking equal intervals on the cumulative cube root of $M_6(x)$ will give approximately optimum strata boundaries (AOSB).

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

For obtaining the limiting expression of the trace of variance-covariance matrix G^T as defined in (4.3), we give the following lemma for bivariate case, which can be proved by using the series expansion of the various terms involved in it, exactly as for the univariate case discussed in Singh and Sukhatme (1969) and bivariate case of Rizvi *et al.* (2002).

Lemma 6.1

Under certain regularity conditions as given in Section **5**, for h-th stratum, we have

$$\sum_{h=1}^{L} W_h \sqrt{P_{hc_1}^2 + P_{hc_2}^2} - \int_{x_{h-1}}^{x_h} \sqrt{\phi_1(t)+\phi_2(t)}f(t)dt = \frac{k_h^2}{96} \int_{x_{h-1}}^{x_h} P(t)f(t)dt[1+O(k_h^2)]$$

where P(t) is defined in (5.4).

Now making use of the Lemma 6.1 in the expression (4.3), we have

$$G^T = \frac{1}{n}\left[\int_a^b \sqrt{[\phi_1(t)+\phi_2(t)]}\ f(t)dt + \sum_{h=1}^{L} \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P(t)f(t)dt[1+O(k_h^2)]\right]^2 \quad (6.1)$$

Now, using the result (3.8) of Singh and Sukhatme (1969), the equation (6.1) can be put as

$$G^T = \frac{1}{n}\left[\int_a^b \sqrt{[\phi_1(t)+\phi_2(t)]}\ f(t)dt + \frac{1}{96}\sum_{h=1}^{L}\left\{\int_{x_{h-1}}^{x_h} \sqrt[3]{P(t)f(t)}dt\right\}^3\right]^2 \quad (6.2)$$

Now, if the strata boundaries are determined by making use of cumulative cube root rule then for h=1,2, ...L, we have

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{P(t)f(t)}dt = \frac{1}{L}\int_a^b \sqrt[3]{P(t)f(t)}dt \quad (6.3)$$

Therefore, (6.2) reduces to

$$G^T = \frac{1}{n}[\delta + \frac{\psi}{L^2}]^2 \quad (6.4)$$

where

$$\delta = \int_a^b \sqrt{[\phi_1(t)+\phi_2(t)]}\ f(t)dt$$

$$\psi = \frac{1}{96}\left[\int_a^b \sqrt[3]{P(t)f(t)}dt\right]^3$$

Now taking limit as $L \to \infty$ on both sides of (6.4) we get

$$\lim_{L\to\infty} G^T = \frac{\delta^2}{n} \quad (6.5)$$

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, G^T tends to δ^2/n.

7. OPTIMUM NUMBER OF STRATA

The trace of the variance-covariance matrix of the estimator $\bar{y}_j$ as given in (6.4) has an approximately minimal value for the given number of strata and fixed total cost. Now to obtain approximately optimum stratification it remains to find an optimum value of L, the number of strata to be constructed. The variance (6.4) is only the function of L as δ and ψ are constants for a given population and for the given auxiliary variable x. Now equating to zero the partial derivative of the trace of the variance-covariance matrix G^T as given in (6.4) with respect to L we get

$$\delta L^2 + \psi = 0 \quad (7.1)$$

8. APPROXIMATE EXPRESSION FOR $[n_h]$

After the strata boundaries have been obtained by cumulative $\sqrt[3]{M_6(x)}$ rule for the number of strata L satisfying (7.1), the sample size $[n_h]$ allocated to the h-th stratum is given by (3.2). Since the functions f(x), c(x) and $\phi(x)$ are known a priori, the parameter W_h, μ_{hc_j} and $\mu_{h\phi_j}$ can be evaluated and the value n_h can be determined. The total sample size n is

$$\sum_{h=1}^{L} n_h = n$$

It may sometime tedious to determine $[n_h]$ from (3.2) because of integrations involved in it. We now obtain the approximate expressions for the sample size $[n_h]$. For this we use Lemma 6.1.

Therefore, if the terms of under $O(m^4)$ are neglected, the sample size n_h in the h-th stratum is given by

$$n_h = \frac{n}{(\alpha + \frac{\beta}{L^2})}\left[\int_{x_{h-1}}^{x_h} \sqrt{[\phi_1(t)+\phi_2(t)]}f(t)dt + \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P(t)f(t)dt\right] \quad (8.1)$$

where

$$\sum_{h=1}^{L} W_h \sqrt{[P_{hc_1}^2 + P_{hc_2}^2]} = (\delta + \frac{\psi}{L^2})$$

If $\bar{x}_h = \frac{x_h + x_{h+1}}{2}$ then (8.1) is approximately given by

$$n_h = \frac{n}{(\alpha + \frac{\beta}{L^2})} \left[\sqrt{(\phi_1(\bar{x}_h) + \phi_2(\bar{x}_h))} + \frac{k_h^2}{96} P(\bar{x}_h) \right] W_h \qquad (8.2)$$

If optimum points of stratification $\{x_h\}$ are obtained by using the proposed cumulative cube root rule then the equation (8.2) can be used for determination of optimum sample size n_h.

9. EMPIRICAL STUDY

To determine approximately optimum strata boundaries (AOSB) by the use of proposed cumulative cube root rule $\sqrt[3]{M_6(x)}$ we consider that stratification variable x follows the following distributions with probability density functions.

Uniform distribution $f(x) = 1 \quad 1 \le x \le 2$

Right triangular distribution $f(x) = 2(2-x) \quad 1 \le x \le 2$

Exponential distribution $f(x) = e^{-x+1} \quad 1 \le x < \infty$

The range of both uniform and right triangular distributions are finite whereas the range of exponential distribution is infinite. 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 $A_1 x^{g_1}$ and $A_2 x^{g_2}$ respectively where $A_1, A_2 > 0$, g_1 and g_2 being constants. Here we have taken different combinations of g_1 and g_2. The values of A_1 and A_2 were determined for the values 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 and σ_x^2 is the variance of the stratification variable x. 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 distribution were divided into 10 classes of equal width. The function $M_6(x)$ was evaluated at the middle point of the class intervals and cumulative $\sqrt[3]{M_6(x)}$ were 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 cube root rule $\sqrt[3]{M_6(x)}$ are given in Table 9.1 to Table 9.3 along with relative efficiency of stratification with no stratification. The variance corresponding to L = 1 is the variance of the usual PPSWR estimator with no stratification. From the Table 9.1 and 9.2 we find that for uniform and right-triangular distributions relative efficiency is only trivial but in case of exponential distribution there is gain in the efficiency. Similar observations were made by Singh and Sukhatme (1972) and Mahajan and Singh (2005). It can also be seen that increase in efficiency when number of strata is increased, is slow in comparison to optimum stratification for two variates using simple random sampling (Rizvi *et al.* (2002)). We further observe that the gain in the efficiency decreases when g increases and becomes zero for g = 2.

ACKNOWLEDGEMENT

Authors would like to thank the referee and Dr. V.K. Bhatia, Joint Secretary, Indian Society of Agricultural Statistics for their valuable suggestions to bring the paper in the present form. The first author would like to express his gratitude to Dr. P.K. Mahajan, Associate Professor (Statistics), Dr. Y.S. Parmar University of Horticulture and Forestry, Nauni (Solan) for his kind help in understanding and initiating the work on optimum stratification during his M.Sc. degree.

REFERENCES

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

Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, **34**, 133-148.

Dalenius, T. and Hodges, J.L. (1959). 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.

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.

Mahajan, P.K. and Singh, Ravindra. (2005). Optimum stratification for scrambled response in pps sampling. *Metron*, **63(1)**, 103-114.

Rizvi, S.E.H., Gupta, J.P. and Bhargava, M. (2002). Optimum stratification based on auxiliary variable for compromise allocation. *Metron*, **60(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.

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.

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.

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

Singh, R. and Sukhatme, B.V. (1972). Optimum stratification in sampling with varying probabilities. *Ann. Inst. Statist. Math.*, **24**, 485-494.

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

Taga, Y. (1967). On optimum stratification for objective variable using prior information. *Ann. Inst. Statist. Math.*, **19**, 101-130.

Table 9.1. Percent relative efficiency of stratification for uniform distribution

No. of Strata L	Strata boundaries					n G^T	Percent Relative Efficiency
	$g_1 = 2$ and $g_2 = 1$						
1						0.15180000	100.0000
2	1.465311					0.15085252	100.6281
3	1.301841	1.636224				0.15067365	100.7475
4	1.223534	1.465311	1.724451			0.15061079	100.7896
5	1.177391	1.366473	1.566938	1.778295		0.15058164	100.8091
6	1.147168	1.301841	1.465311	1.636224	1.814506	0.15056580	100.8197
	$g_1 = 1$ and $g_2 = 2$						
1						0.14700000	100.00000
2	1.472618					0.14699574	100.00290
3	1.308580	1.642666				0.14699489	100.00348
4	1.229163	1.472618	1.729860			0.14699458	100.00369
5	1.182124	1.373533	1.573895	1.782840		0.14699444	100.00378
6	1.151278	1.308580	1.472618	1.642666	1.818490	0.14699436	100.00384
	$g_1 = 2$ and $g_2 = 2$						
1						0.146700	100.00
2	1.471716					0.146700	100.00
3	1.307743	1.641877				0.146700	100.00
4	1.228460	1.471716	1.729200			0.146700	100.00
5	1.181528	1.372657	1.573040	1.782284		0.146700	100.00
6	1.150760	1.307743	1.471716	1.641877	1.818004	0.146700	100.00

Table 9.2. Percent relative efficiency of stratification for right triangular distribution

No. of Strata L	Strata boundaries					n G^T	Percent Relative Efficiency
	$g_1 = 2$ and $g_2 = 1$						
1						0.101224	100.0000
2	1.378139					0.100734	100.4864
3	1.240046	1.534977				0.100633	100.5875
4	1.175881	1.378139	1.622766			0.100596	100.6244
5	1.139144	1.292906	1.469597	1.679908		0.100578	100.6418
6	1.114652	1.240046	1.378139	1.534977	1.720608	0.100569	100.6514
	$g_1 = 1$ and $g_2 = 2$						
1						0.098525	100.0000
2	1.383461					0.098523	100.0021
3	1.244505	1.540469				0.098523	100.0026
4	1.179429	1.383461	1.627888			0.098523	100.0027
5	1.142163	1.297741	1.475059	1.684265		0.098523	100.0028
6	1.117319	1.244505	1.383461	1.540469	1.724921	0.098523	100.0029
	$g_1 = 2$ and $g_2 = 2$						
1						0.09833787	100.00
2	1.383052					0.09833787	100.00
3	1.244161	1.54005				0.09833787	100.00
4	1.179154	1.383052	1.627497			0.09833787	100.00
5	1.141928	1.297368	1.474641	1.683932		0.09833787	100.00
6	1.117112	1.244161	1.383052	1.54005	1.724591	0.09833787	100.00

Table 9.3. Percent relative efficiency of stratification for exponential distribution

No. of Strata L	Strata boundaries					n G^T	Percent Relative Efficiency
	$g_1 = 2$ and $g_2 = 1$						
1						1.523741	100.0000
2	2.303533					1.481007	102.8855
3	1.770596	3.009307				1.471108	103.5778
4	1.537751	2.303533	3.481186			1.467434	103.8371
5	1.422852	1.956873	2.705209	3.827506		1.465705	103.9596
6	1.352377	1.770596	2.303533	3.009307	4.075606	1.464755	104.027
	$g_1 = 1$ and $g_2 = 2$						
1						1.284662	100.0000
2	2.311065					1.284464	100.0154
3	1.776089	3.019367				1.284409	100.0197
4	1.542704	2.311065	3.489436			1.284388	100.0214
5	1.425981	1.962797	2.714041	3.836049		1.284378	100.0221
6	1.354984	1.776089	2.311065	3.019367	4.084675	1.284372	100.0226
	$g_1 = 2$ and $g_2 = 2$						
1						1.268173	100.00
2	2.303533					1.268173	100.00
3	1.770596	3.009307				1.268173	100.00
4	1.537751	2.303533	3.481186			1.268173	100.00
5	1.422852	1.956873	2.705209	3.827506		1.268173	100.00
6	1.352377	1.770596	2.303533	3.009307	4.075606	1.268173	100.00