Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(3) 2012 401-412

Optimum Stratification for Sensitive Quantitative Variables using Auxiliary Information

Med Ram Verma[1*], Sarjinder Singh[2] and Rajiv Pandey[3]

[1]*Indian Veterinary Research Institute, Izatnagar, U.P.*

[2]*Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX 78363, USA*

[3]*Forest Research Institute, Dehradun, Uttrakhand, India*

Received 10 November 2010; Revised 06 August 2012; Accepted 08 August 2012

SUMMARY

The paper considers the problem of optimum stratification for two sensitive quantitative variables when data on sensitive variables are collected by scrambled randomized response technique and an auxiliary variable is taken as stratification variable. We have proposed a cumulative cube root rule for determination of optimum strata boundaries for ratio and regression method of estimation under compromise method of allocation. A limiting expression for the trace of variance covariance matrix and an approximate expression for sample size also have been suggested. The paper concludes with numerical illustration.

Keywords: Sensitive variable, Scrambled response, Optimum stratification.

1. INTRODUCTION

The randomized response technique is used to procure trustworthy data for estimating the proportion of people with a sensitive characteristic. Several research workers have extended the technique since its introduction by Warner (1965). Eichhorn and Hayre (1983) introduced a scrambled randomized response technique, which does not contain the difficulties of unrelated question method of Greenberg *et al.* (1971). The scrambled randomized response technique involves the respondent multiplying his sensitive answer Y by a random number S from the known distribution and giving the scrambled response $Z = Y S$ to the interviewer who does not know the particular values of the random number S.

Singh and Sukhatme (1973) considered the problem of optimum stratification for univariate case in case of ratio and regression method of estimation. Mahajan *et al.* (1994) developed the theory for determination of optimum strata boundaries for a sensitive character using scrambled response technique. Mahajan (2006) considered the problem of optimum stratification for ratio and regression methods using scrambled randomized response technique for a sensitive variable.

Ghosh (1963) considered the problem of optimum stratification with two characters under proportional method of allocation assuming stratification variable as identical to the estimation variable under consideration. Gupta and Seth (1979) considered the problem of optimum stratification for the study of more than one characters under proportional allocation. Rizvi *et al.* (2002) considered the case of optimum stratification for two study variables in case of compromise method of allocation for simple random sampling with

[*]*Corresponding author* : Med Ram Verma
E-mail address : mrverma19@yahoo.co.in, medramverma@rediffmail.com

replacement scheme using auxiliary information. Verma and Rizvi (2007) considered 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.

In many cases it is important to obtain information about two characteristics of sensitive nature for the same population at the same time. It is possible that the information on one of the two characteristics may be used for weighting (post-stratification) purposes (Christofieds 2005). So we have considered the problem of determining AOSB (Approximately Optimum Strata Boundaries) for two sensitive quantitative variables by using compromise allocation for ratio and regression method of estimation. This strategy is important when we are dealing with stigmatized quantitative variables. For instance, let Y_1 be the "income understated in income tax return" and Y_2 be the "expenditure". These two variables can be stratified by using an auxiliary variable X (Eye estimated value of property) as the stratification variable.

2. OPTIMUM STRATIFICATION FOR RATIO AND REGRESSION METHODS OF ESTIMATION

If information on the auxiliary variable, which is highly correlated with the variable under study, is available in that case the population mean can be estimated by using ratio and regression method of estimation as compared to simple random sampling. The information on the auxiliary variable can further be used to increase the precision of estimators using the technique of optimum stratification.

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 such that $\sum_{h=1}^{L} N_h = N$. For drawing a stratified SRSWR (Simple Random Sampling with replacement) sample of size n, the sample of sizes n_1, n_2, ... n_L are to be drawn from respective stratum so that $\sum_{h=1}^{L} n_h = n$. Let Y_j (j = 1, 2) be two sensitive quantitative variables. Let Y_j denote the value of sensitive characters for the j-th sensitive variable and S_j be the scrambling random variable independent of Y_j and with finite mean and variance. The respondent generates S_j using some specified methods and multiplies the sensitive variable values Y_j by S_j. Hence the interviewer receives the scrambled answer $Z_j = Y_j S_j$. The particular values of S_j are unknown to the interviewer but its distribution is known. In this way the privacy of the respondents is not violated. Now using results of Eichhorn and Hayre (1983) and Mahajan *et al.* (1994) for j-th sensitive variable, we get

Let $E_R(S_{hj}) = \theta_{hj}$, and $V_R(S_{hj}) = \gamma_{hj}^2$

$E_R(Y_{hj}) = \mu_{hy_j}$ and $V_R(Y_{hj}) = \sigma_{hy_j}^2$

where S_{hj} is the value of the j-th scrambling variable in h-th stratum and Y_{hj} is the value of the j-th sensitive variable in the h-th stratum. θ_{hj} and γ_{hj}^2 are known to the interviewer but μ_{hy_j} and $\sigma_{hy_j}^2$ are unknown. Since Y_{hj} and S_{hj} are independent, we have

$$E_R(Z_{hj}) = \mu_{hy_j}.\theta_{hj} \tag{2.1}$$

$$V_R(Z_{hj}) = \sigma_{hy_j}^2(\theta_{hj}^2 + \gamma_{hj}^2) + \mu_{hyj}^2.\gamma_{hj}^2 \tag{2.2}$$

If Z_{hj} denotes the value of scrambled response for j-th sensitive variable in the h-th stratum and sampling within each stratum is SRSWR, then unbiased estimator of μ_{hy_j} is

$$\hat{\mu}_{hy_j} = \frac{\overline{Z}_{hj}}{\theta_{hj}} \quad \text{where } \overline{Z}_{hj} = n_h^{-1}\sum_{i=1}^{n_h} Z_{hij} \tag{2.3}$$

and Z_{hij} is the scrambled response for the j-th sensitive variable for i-th element in h-th stratum.

When information on an auxiliary variable is available, we may obtain more efficient estimators of population mean/total, than the usual simple random sampling scheme estimators by making use of ratio and regression method of estimations.

Separate Ratio Estimator

The separate ratio estimator for sample mean of the j-th sensitive variable in stratified sampling is given by

$$\bar{y}_{j.R1} = \sum_{h=1}^{L} W_h \hat{\mu}_{hRy_j}$$

where $\hat{\mu}_{hRy_j} = \frac{\hat{\mu}_{hy_j}}{\bar{x}_h}.\bar{X}_h$

where

W_h = Proportion of units in the h-th stratum

$\bar{y}_j$ = Sample mean for j-th sensitive variable in the h-th stratum

$\bar{x}_h$ = Sample mean for auxiliary variable X in the h-th stratum

$\bar{X}_h$ = Population mean for auxiliary variable X in the h-th stratum

Variance of estimator $\bar{y}_j$ under ratio method of estimation is given by

$$V(\bar{y}_{j.R1}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}[\sigma_{hy_j}^2(1+C_{hj}^2) + \mu_{hy_j}^2 C_{hj}^2 + R_{jh}^2\sigma_{hx}^2 - 2R_{jh}\sigma_{hxy_j}] \quad (2.4)$$

where $C_{hj} = \frac{\gamma_{hj}}{\theta_{hj}}$ is the coefficient of variation of the j-th scrambling variable S_j in h-th stratum and $R_{jh} = \frac{E(\hat{\mu}_{hy_j})}{E(\bar{x}_h)}$.

Separate Regression Estimator

It is often found that even when the regression lines of the study variable are linear, the regression line does not pass through the origin. Under such conditions, ratio estimators are not efficient hence it seems more appropriate to use regression type of estimators. For estimating the population mean of Y of the certain characteristics difference method of estimation is well known. The difference estimator was proposed by Hansen *et al*. (1953), which is given by the following model.

$$\bar{y}_d = \bar{y} + \beta(\bar{X} - \bar{x}) \quad (2.5)$$

Now using this estimator in stratified sampling for sensitive quantitative variable, the separate regression estimators for estimating the population mean of the j-th sensitive quantitative variable is given by

$$\bar{y}_{j.R} = \sum_{h=1}^{L} W_h \hat{\mu}_{hDy_j}$$

where $\hat{\mu}_{hDy_j} = \hat{\mu}_{hy_j} + \beta_{jh}(\bar{X}_h - x_h)$

$$V(\bar{y}_{j.R}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}[\sigma_{hy_j}^2(1+C_{hj}^2) + \mu_{hy_j}^2 C_{hj}^2 + \beta_{jh}^2\sigma_{hx}^2 - 2\beta_{jh}\sigma_{hxy_j}] \quad (2.6)$$

We observe that the variance expression under separate ratio and regression method of estimation (2.4) and (2.6) are same except they differ only in constants R_j and β_j. Hence we will consider only regression estimator with variance given by (2.6). The variance expression is clearly a function of the strata boundaries.

The variance of the estimator of population mean under ratio and regression methods is a function of h-th stratum weight W_h, within stratum variance σ_{hy}^2, the sample size in the h-th stratum n_h, regression coefficient β_{jh}. The precision of the estimator of population mean will, therefore, change with the change in any one of these parameters. Once the decision about the total number of strata and method of allocation is taken, the variance depends on W_h, μ_{hy_j} and $\sigma_{hy_j}^2$ which in turn depends on the way the strata are constructed. Thus, a wrong choice of strata boundaries may result in considerable increase in the variance. The proper choice of strata boundaries is therefore, an important consideration in stratified sampling scheme.

3. COMPROMISE ALLOCATION IN STRATIFIED SAMPLING

In multivariate stratified sampling where more than one characteristic are to be estimated, an allocation which is optimum for one characteristic may or may not be optimum for other characteristics. In such situation a compromise is needed to work out a usable allocation, which is optimum in some sense for all characteristics. Such an allocation may be called a

'Compromise Allocation' because it is based on some compromise criterion.

The problem of allocation to strata with several characteristics was first considered by Neyman (1934). 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 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 is determined in such a way so that for a given total sample size (which amounts to fixed total cost where the cost per unit in each stratum is same) $\sum_{j=1}^{2} V(\bar{y}_{j.R})$ is minimized where $V(\bar{y}_{j.R})$ is the variance for j-th sensitive variable. If finite population correction factor can be neglected then the variance expression for j-th sensitive variable under regression method of estimation is given by the equation (2.6).

Now we have to minimize

$$\sum_{j=1}^{2} V(\bar{y}_{j.R}) \tag{3.1}$$

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

$$n_h = n \frac{W_h \sqrt{R_{hy_1}^2 + R_{hy_2}^2}}{\sum_{h=1}^{L} W_h \sqrt{R_{hy_1}^2 + R_{hy_2}^2}} \tag{3.2}$$

where

$$R_{hy_j}^2 = \sigma_{hy_j}^2 (1 + C_{hj}^2) + \mu_{hy_j}^2 C_{hj}^2 + \beta_{jh}^2 \sigma_{hx}^2 - 2\beta_{jh}\sigma_{hxy_j}$$

Using this value of n_h we can obtain the variance expression for compromise allocation. Under compromise method of allocation, the optimal variances of the estimated population means of the sensitive variables Y_j are given by

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

$$(j = 1, 2) \tag{3.3}$$

4. 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 sensitive 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

$$Y_j = \beta_{jh} X + e_j \tag{4.1}$$

where $\beta_{jh} X = c_j(X)$ is a real valued function of X, e_j is error component such that

$$E(e_j | X) = 0,\ E(e_j e_j' | X, X') = 0, \text{ for } x \neq x'$$

and $V(e_j | X) = \phi_j > 0$ 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 (4.1) it can be easily seen that

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

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

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

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

$$\sigma_{hxy_j}^2 = \beta_{jh}^2 \sigma_{hx}^2 + \mu_{h\phi_j}$$

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 sensitive variable.

The variance expression for compromise allocation under super population model (4.1) is therefore given by

$$\sigma_j^2 = V(\bar{y}_{j.R}) = \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h R_{hcj}^2}{\sqrt{R_{hc_1}^2 + R_{hc_2}^2}}\sum_{h=1}^{L} W_h\sqrt{R_{hc_1}^2 + R_{hc_2}^2}\right] \quad (j = 1, 2) \qquad (4.2)$$

where $R_{hcj}^2 = \mu_{h\phi_j} + C_{hj}^2(\mu_{h\phi_j} + \sigma_{hcj}^2 + \mu_{hcj}^2)$

5. 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 points of stratification is minimum. These $\{x_h\}$ are the solutions of the minimal equations. Now generalized variance G is given by

$$G = \begin{vmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{vmatrix} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \qquad (5.1)$$

It is cumbersome to obtain even approximate solution to the minimal equations obtained through minimization of G under compromise method of allocation. Sukhatme *et al.* (1984) had shown numerically that all the compromise allocations 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 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

$$tr(G) = \sigma_1^2 + \sigma_2^2 \qquad (5.2)$$

Using (4.2) in (5.2) tr(G) can be expressed as

$$tr(G) = \frac{1}{n}\left[\sum_{h=1}^{L} W_h\sqrt{(R_{hc_1}^2 + R_{hc_2}^2)}\right]^2 \qquad (5.3)$$

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

$$\sum_{h=1}^{L} W_h\sqrt{(R_{hc_1}^2 + R_{hc_2}^2)} \quad \text{which gives}$$

$$W_h\frac{\partial}{\partial x_h}\sqrt{(h)} + \sqrt{(h)}\frac{\partial}{\partial x_h}W_h + W_i\frac{\partial}{\partial x_h}\sqrt{(i)} + \sqrt{(i)}\frac{\partial}{\partial x_h}W_i = 0 \qquad (5.4)$$

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

$$(h) = R_{hc_1}^2 + R_{hc_2}^2$$

$$(i) = R_{ic_1}^2 + R_{ic_2}^2$$

The expressions of the partial derivative terms involved in (5.4) can be easily obtained on the lines of Singh and Sukhatme (1969). Now inserting the values of the required partial derivatives in the equation (5.4) and solving we have the required minimal equations as

$$\left[\frac{\begin{array}{l}[(1+C_{h1}^2)[\phi_1(x_h) + \mu_{h\phi_1}] + C_{h1}^2[\{\mu_{hc_1} - c_1(x_h)\}^2\\ +\sigma_{hc_1}^2 + 2\mu_{hc_1}c_1(x_h)] + (1+C_{h2}^2)[\phi_2(x_h) + \mu_{h\phi_2}]\\ +C_{h2}^2[\{\mu_{hc_2} - c_2(x_h)\}^2 + \sigma_{hc_2}^2 + 2\mu_{hc_2}c_2(x_h)]]\end{array}}{\sqrt{R_{hc_1}^2 + R_{hc_2}^2}}\right]$$

$$= \left[\frac{\begin{array}{l}[(1+C_{i1}^2)[\phi_1(x_h) + \mu_{i\phi_1}] + C_{i1}^2[\{\mu_{ic_1} - c_1(x_h)\}^2\\ +\sigma_{ic_1}^2 + 2\mu_{ic_1}c_1(x_h)] + (1+C_{i2}^2)[\phi_2(x_h) + \mu_{i\phi_2}]\\ +C_{i2}^2[\{\mu_{ic_2} - c_2(x_h)\}^2 + \sigma_{ic_2}^2 + 2\mu_{ic_2}c_2(x_h)]]\end{array}}{\sqrt{R_{hc_1}^2 + R_{hc_2}^2}}\right] \qquad (5.5)$$

Solution to these minimal equations (5.5) will give set of optimum points of stratification. This system of equations is the function of parameter values, which themselves are the function 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.

6. APPROXIMATE SOLUTION TO THE MINIMAL EQUATIONS

To obtain the approximate solutions to the minimal equations (5.5) 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} and $\sigma^2_{hc_j}$ 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

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

$$= \phi\left[1+\frac{\phi'}{2\phi}k+\frac{\phi'f'+2f\phi''}{12f\phi}k^2 + \frac{(ff''\phi'+ff'\phi''+f^2\phi'''-f'^2\phi')}{24f^2\phi}k^3+O(k^4)\right] \tag{6.1}$$

In order to obtain the series expansions of the minimal equations in (5.5), 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 (5.5) are obtained by using (6.1) as given below

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

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

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

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

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

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

and $\sigma^2_\phi(y,\,x)$, the conditional variance of $\phi(t)$ in the interval $(y,\,x)$, about the point $t = y$ is given by

$$\sigma^2_\phi(y,x) = \frac{k^2}{12}\phi'^2\left[1+\frac{\phi''}{\phi'}k+O(k^2)\right] \tag{6.2}$$

The expansions for σ^2_{hc} and σ^2_{ic} are obtained by using (6.2)

$$\sigma^2_{hc} = \frac{k_h^2}{12}c'^2\left[1-\frac{c'''}{c'}k_h+O(k_h^2)\right]$$

$$\sigma^2_{ic} = \frac{k_i^2}{12}c'^2\left[1+\frac{c'''}{c'}k_i+O(k_i^2)\right]$$

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

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

$$= k^{\lambda-1}\int_y^x f(t)dt[1+O(k^2)] \tag{6.3}$$

Now using the various expressions in minimal equation (5.5) we get on simplification

$$2\sqrt{\phi_1^*+\phi_2^*}\left[1+A_2k_h^2+A_3k_h^3+O(k_h{}^4)\right] = 2\sqrt{\phi_1^*+\phi_2^*}\left[1+A_2k_i^2+A_3k_i^3+O(k_i{}^4)\right] \tag{6.4}$$

where $k_h = x_h - x_{h-1}$, $k_i = x_{h+1} - x_h$ are the stratum widths, and

$$A_2 = \frac{(\phi_{11} + \phi_{21})^2}{32(\phi_1^* + \phi_2^*)^2}$$

$$A_3 = \frac{1}{96 f \sqrt{\phi_1^* + \phi_2^*}} \frac{d}{dx_h}\left[\frac{f(\phi_{11} + \phi_{21})^2}{(\phi_1^* + \phi_2^*)^{3/2}}\right]$$

$$\phi_1^* = \phi_1 + \phi_1 C_{h1}^2 + c_1^2 C_{h1}^2 \quad \phi_{11} = \phi_1' + \phi_1' C_{h1}^2 + 2c_1' c_1 C_{h1}^2$$

$$\phi_2^* = \phi_2 + \phi_2 C_{h2}^2 + c_2^2 C_{h2}^2 \quad \phi_{21} = \phi_2' + \phi_2' C_{h2}^2 + 2c_2 c_2' C_{h2}^2$$

where ϕ_{11} is the first order derivative of ϕ_1^* and ϕ_{21} is the first order derivative of ϕ_2^*.

Now after canceling $2\sqrt{\phi_1^* + \phi_2^*}$ from both sides of the equation (6.4) and multiplying by $f(x_h)$ we get on simplification

$$\frac{k_h^2}{16}\left[R(x)f(x) - \frac{1}{3}\frac{d}{dx_h}[R(x)f(x)]k_h + O(k_h^2)\right]$$

$$= \frac{k_i^2}{16}\left[R(x)f(x) + \frac{1}{3}\frac{d}{dx_h}[R(x)f(x)]k_i + O(k_i^2)\right]$$

where $R(x) = \dfrac{(\phi_{11}(x) + \phi_{21}(x))^2}{(\phi_1^*(x) + \phi_2^*(x))^{3/2}}$

Using these expansions, the system of minimal equations in (5.5) reduces to

$$k_h^2\left[1 - \frac{k_h}{3}.\frac{[R(x)f(x)]'}{R(x)f(x)} + O(k_h^2)\right]$$

$$= k_i^2\left[1 + \frac{k_i}{3}.\frac{[R(x)f(x)]'}{R(x)f(x)} + O(k_i^2)\right] \qquad (6.5)$$

On raising both sides of the equation (6.5) to the power 3/2 and using binomial theorem (for any index), we get

$$k_h^3\left[1 - \frac{k_h}{2}.\frac{[R(x)f(x)]'}{R(x)f(x)} + O(k_h^2)\right]$$

$$= k_i^3\left[1 + \frac{k_i}{2}.\frac{[R(x)f(x)]'}{R(x)f(x)} + O(k_i^2)\right] \qquad (6.6)$$

On comparing it with (6.3), with $\lambda = 3$, the system of equations (5.5) can be written in the form

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} R(x)f(x)dx[1 + O(k_h{}^2)]\right]$$

$$= \left[k_i^2 \int_{x_h}^{x_{h+1}} R(x)f(x)dx[1 + O(k_i{}^2)]\right] \qquad (6.7)$$

The functions ϕ_j^*, c_j and their derivatives are evaluated at the point x_h, and we assume that the function $R(x)f(x) \in \Omega$ 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 (5.5) can approximately be given as

$$\left[k_h^2 \int_{x_{h-1}}^{x_h} R(x)f(x)dx\right] = \left[k_i^2 \int_{x_h}^{x_{h+1}} R(x)f(x)dx\right] \qquad (6.8)$$

Or equivalently by

$$k_h^2 \int_{x_{h-1}}^{x_h} R(x)f(x)dx = \text{constant } h = 1, 2, ..., L \qquad (6.9)$$

where terms of $O(m^4), m = \sup_{(a,b)}(k_h)$ have been neglected on both sides of equation (6.8). Since $a \le x \le b$ and the points of demarcation forming L strata are $x_1, x_2, \ldots x_L$ 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

$$k_h^2 \int_{x_{h-1}}^{x_h} R(x)f(x)dx = Q(x_{h-1}, x_h)[1 + O(k_h^2)] \qquad (6.10)$$

Then the system of equations (5.5) can approximately be put as

$$Q(x_{h-1}, x_h) = \text{Constant}, h = 1, 2, ..., L \qquad (6.11)$$

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.

Various methods of finding approximate solutions to the minimal equations can be established through the system of equations (6.11). 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 cumulative $\sqrt[3]{M_4(x)}$ Rule according to which the approximately optimum strata boundaries (AOSB) are the solutions of the system of equation (5.5). Proceeding on the same lines, one such form of function $Q(x_{h-1}, x_h)$ can also be obtained as follows

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{M_4(x)}dx = \int_a^b \sqrt[3]{R(x)f(x)}\,dx \Big/ L \qquad (6.12)$$

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

Cumulative $\sqrt[3]{M_4(x)}$ Rule

If the function $M_4(x) = R(x)f(x)$ 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 $M_4(x)$ will give approximately optimum strata boundaries (AOSB).

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

For obtaining the limiting expression of the trace of variance-covariance matrix $tr(G)$ as defined in (5.2), 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).

Lemma 7.1

Under regularity conditions for h-th stratum we have

$$\sum_{h=1}^{L} W_h \sqrt{[R_{hc_1}^2 + R_{hc_2}^2]} - \int_{x_{h-1}}^{x_h} \sqrt{[\phi_1^*(x) + \phi_2^*(x)]}f(x)dx$$

$$= \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} R(x)f(x)dx[1 + O(k_h^2)]$$

where $R(x)$ is defined in (6.5)

Now making use of the lemma 7.1 in the expression (5.3), we have

$$tr(G) = \frac{1}{n}\left[\int_a^b \sqrt{[\phi_1^*(x) + \phi_2^*(x)]}f(x)dx + \sum_{h=1}^{L}\frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} R(x)f(x)dx[1 + O(k_h^2)]\right] \qquad (7.1)$$

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

$$tr(G) = \frac{1}{n}\left[\int_a^b \sqrt{[\phi_1^*(x) + \phi_2^*(x)]}f(x)dx + \frac{1}{96}\sum_{h=1}^{L}\{\int_{x_{h-1}}^{x_h} \sqrt[3]{R(x)f(x)}dx\}^3\right]^2 \qquad (7.2)$$

Now, if the strata boundaries are determined by making use of proposed cumulative cube root rule $\sqrt[3]{M_4(x)}$ then for h = 1, 2, ... L, we have

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{R(x)f(x)}dx = \frac{1}{L}\int_a^b \sqrt[3]{R(x)f(x)}dx \qquad (7.3)$$

Therefore, equation (6.5) reduces to

$$tr(G) = \frac{1}{n}\left(\alpha + \frac{\beta}{L^2}\right)^2 \qquad (7.4)$$

where

$$\alpha = \int_a^b \sqrt{[\phi_1^*(x) + \phi_2^*(x)]}f(x)dx$$

$$\beta = \frac{1}{96}\left[\int_a^b \sqrt[3]{[R(x)f(x)}dx\right]^3$$

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

$$\lim_{L\to\infty} tr(G) = \frac{\alpha^2}{n} \qquad (7.5)$$

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, $tr(G)$ tends to a^2/n.

8. APPROXIMATE EXPRESSION FOR $[n_h]$

After the strata boundaries have been obtained by solving the system of equations (5.5) for the number of strata L, the sample size $[n_h]$ allocated to the h-th stratum is given by (3.3). Since the functions $f(x)$, $c(x)$, and $\phi(x)$ are known a priori, the parameters W_h, $\mu_{hc_j}, \sigma^2_{hc_j}$ 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 be 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 7.1.

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

$$n_h = \frac{n}{\left(a + \frac{\beta}{L^2}\right)}\left[\int_{x_{h-1}}^{x_h} \sqrt{[\phi_1^*(x) + \phi_2^*(x)]} f(x)dx + \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} R(x) f(x)dx\right] \quad (8.1)$$

where

$$\sum_{h=1}^{L} W_h \sqrt{[K^2_{hy_1} + K^2_{hy_2}]} = \left(a + \frac{\beta}{L^2}\right)$$

$$\bar{x}_h = \frac{x_h + x_{h-1}}{2}$$

Then (8.1) is approximately given by

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

where $R(\bar{x}_h) = \dfrac{\left(\phi_{11}(\bar{x}_h) + \phi_{21}(\bar{x}_h)\right)^2}{\left(\phi_1^*(\bar{x}_h) + \phi_2^*(\bar{x}_h)\right)^{3/2}}$

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. NUMERICAL STUDY

To determine approximately optimum strata boundaries (AOSB) by the use of proposed cumulative cube root rule $\sqrt[3]{M_4(x)}$ we consider that stratification variable x follows the following distributions with probability density functions.

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

The ranges of both uniform and right triangular distributions are finite where as range of exponential distribution is infinite. We have considered that sensitive 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 to be 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 and ρ_1, ρ_2 by using the following formulae, where ρ_1 and ρ_2 are the correlation coefficients between the sensitive study variables Y_1 and Y_2 with stratification variable x.

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

σ_x^2 is the variance of the stratification variable x. For the purpose of numerical illustration we have assumed $\rho_1^2 = 0.9$, $\rho_2^2 = 0.7$, $C_{h1} = 0.2$ and $C_{h2} = 0.1$. For finding out the approximately optimum strata boundaries (AOSB), the range of uniform and rectangular distribution was divided into 10 classes of equal width. The function R(x) was evaluated at the

Table 9.1 Percent relative efficiency of stratification for uniform distribution

No of Strata L	Strata bounaries					*n tr*(*G*)	Percent Relative Efficiency
	$g_1 = 2$ and $g_2 = 1$						
1						0.39483	100.00
2	1.47439					0.39698	102.03
3	1.30986	1.64431				0.38551	102.42
4	1.23039	1.47439	1.73128			0.38499	102.55
5	1.18313	1.37518	1.57564	1.78406		0.38475	102.62
6	1.15214	1.30986	1.47439	1.64431	1.81956	0.38462	102.65
	$g_1 = 1$ and $g_2 = 2$						
1						0.25057	100.00
2	1.47214					0.24144	103.78
3	1.30813	1.64224				0.23973	104.52
4	1.22879	1.47214	1.72951			0.23912	104.79
5	1.18181	1.37307	1.57344	1.78254		0.23884	104.91
6	1.15100	1.30813	1.47214	1.64224	1.81823	0.23869	104.97
	$g_1 = 2$ and $g_2 = 2$						
1						0.25060	100.00
2	1.47172					0.24127	103.87
3	1.30774	1.64188				0.23952	104.63
4	1.22846	1.47172	1.72920			0.23890	104.90
5	1.18153	1.37266	1.57304	1.78228		0.23862	105.02
6	1.15076	1.30774	1.47172	1.64188	1.81800	0.23846	105.09

Table 9. 2 Percent relative efficiency of stratification for right triangular distribution

No of Strata L	Strata bounaries					*n tr*(*G*)	Percent Relative Efficiency
	$g_1 = 2$ and $g_2 = 1$						
1						0.28325	100.00
2	1.38282					0.27842	101.73
3	1.24394	1.53984				0.27744	102.09
4	1.17897	1.38282	1.62731			0.27708	102.23
5	1.14177	1.29714	1.47442	1.68378		0.27690	102.29
6	1.11697	1.24394	1.38282	1.53984	1.72445	0.27681	102.32
	$g_1 = 1$ and $g_2 = 2$						
1						0.18705	100.00
2	1.38334					0.18148	103.07
3	1.24441	1.54035				0.18035	103.72
4	1.17935	1.38334	1.62777			0.17994	103.96
5	1.14210	1.29763	1.47494	1.68417		0.17974	104.07
6	1.11726	1.24441	1.38334	1.54035	1.72482	0.17963	104.13
	$g_1 = 2$ and $g_2 = 2$						
1						0.18705	100.00
2	1.38305					0.18137	103.13
3	1.24416	1.54005				0.18022	103.79
4	1.17915	1.38305	1.62750			0.17980	104.04
5	1.14193	1.29737	1.47464	1.68393		0.17960	104.15
6	1.11600	1.24416	1.38305	1.54005	1.72459	0.17949	104.22

Table 9.3 Percent relative effciency of stratification for exponential distribution

No of Strata L	Strata bounaries $g_1 = 2$ and $g_2 = 1$					$n\ tr(G)$	Percent Relative Efficiency
1						3.18814	100.00
2	2.28329					3.03856	104.92
3	1.75331	2.99130				3.00593	106.06
4	1.52094	2.28329	3.46831			2.99420	106.48
5	1.41238	1.93921	2.68475	1.81664		2.98875	106.67
6	1.34365	1.75331	2.28329	2.99130	4.06619	2.98581	106.78
	$g_1 = 1$ and $g_2 = 2$						
1						2.76582	100.00
2	2.30964					2.56022	108.03
3	1.77505	3.01748				2.52774	109.42
4	1.54176	2.30964	3.48789			2.52160	109.69
5	1.42538	1.96167	2.71238	3.83445		2.52077	109.72
6	1.35449	1.77505	2.30964	3.01748	4.08298	2.52126	109.70
	$g_1 = 2$ and $g_2 = 2$						
1						2.75268	100.00
2	2.30353					2.54233	108.27
3	1.77060	3.00931				2.50999	109.67
4	1.53775	2.30353	3.48119			2.50431	109.92
5	1.42285	1.95687	2.70521	3.82751		2.50376	109.94
6	1.35238	1.77060	2.30353	3.00931	4.07561	2.50449	109.91

middle point of the class intervals and cumulative $\sqrt[3]{M_4(x)}$ 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 cube root rule $\sqrt[3]{M_4(x)}$ are given in Table 9.1 to Table 9.3 along with relative efficiency of stratification with no stratification.

ACKNOWLEDGEMENTS

The authors would like to thank the learned referee and Coordinating Editor for the valuable suggestions.

REFERENCES

Christofied, T.C. (2005). Randomized response technique for two sensitive characteristics at the same time. *Metrika,* **62**, 53-62.

Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response method for obtaining sensitive quantitative data. *J. Statist. Plann. Inf.*, **7**, 307-316.

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

Greenberg, B.G. Kubler, R.R, Aberanathy, J.R. and Horvitz, D.G. (1971). Applications of randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.*, **66**, 243-250.

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

Hansen, M.H., Hurvitz, W.N. and Midow, W.G. (1953). *Sample Surveys Methods and Theory*, Volume 2. John Wiley and Sons, Incl., New York.

Mahajan, P.K. (2006). Optimum stratification for scrambled response with ratio and regression methods of estimation. *Model Assist. Statist. Appl.*, **1(1)**, 17-22.

Mahajan, P.K., Gupta, J.P. and Singh, R. (1994). Determination of optimum strata boundaries for scrambled response. *Statistica*, **54(3)**, 375-381.

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. (2002). Optimum stratificaton based on anuxiliary vaiable for compromise allocation. *Metron*, **60(3-4)**, 201-215.

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

Singh, R. and Sukhatme, B.V. (1973). Optimum stratification with ratio and regression methods of estimation. *Ann. Inst. Stat. Math*., **25**, 627-633.

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

Verma, M.R. and Rizvi, S.E.H. (2007). Optimum stratification for PPS sampling using auxiliary information. *J. Ind. Soc. Agril. Statist.*, **61(2)**, 66-76.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60**, 63-69.