MINIMUM VARIANCE STRATIFICATION FOR COMPROMISE ALLOCATION

Med Ram VERMA[1]
Jadish Prasad Singh JOOREL[2]
Rajeev Kumar AGNIHOTRI[3]

▪ *ABSTRACT: In this paper we has considered the problem of optimum stratification for two sensitive quantitative variables when the data from different strata have been collected by scrambled response technique and an auxiliary variable have been used as a stratification variable. In this paper we have obtained the limiting expression for the trace of generalized variance covariance matrix, expression for number of strata and expression for approximate sample size* $[n_h]$. *The paper concluded with a numerical illustration.*

▪ *KEYWORDS: Sensitive variable; scrambled response; optimum stratification; AOSB (Approximately Optimum Strata Boundaries).*

1 Introduction

In stratified sampling the main aim is to get estimators of the population parameters for the character under study with maximum precision at minimum cost. The precision of an estimator of population mean/total depends not only on the sample size and sampling fraction but also on the variability or heterogeneity among the units of the population. Apart from increasing the sample size one possible way to estimate the population mean with maximum precision is to divide the population into certain number of groups, which are more homogeneous within, and then selecting samples from each of the group independently. These groups are called strata and the whole procedure is called stratified sampling. Stratification in sample survey is well known device to increase the precision of the estimators. In stratified sampling efficiency of the estimator of population parameters depends on several factors such as choice of stratification variable, number of strata, determination of strata boundaries and allocation of sample sizes to the different strata. Once it is decided about the total number of strata and the procedure of allocating sample sizes to different strata, the problem of optimum stratification may be considered to consist of determination of optimum strata boundaries.

[1]Indian Veterinary Research Institute, Division of Livestock Economics, Statistics and Information Technology, Izatnagar (Bareilly) U.P., India. E-mail address: *mrverma19@yahoo.co.in / medramverma@rediffmail.com*
[2]University of Jammu, Department of Statistics, Jammu (J&K), Pin 180006, India.
[3]St. John's College Agra (U.P.), Department of Statistics, Pin 282002, India.

The pioneering work in this field was done by Dalenius (1950), Dalenius and Gurney (1951), Dalenius and Hodges (1957), Singh and Sukhatme (1969) and several other research workers also considered the problem of optimum stratification with respect to an auxiliary variable closely related to the study 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. It is unrealistic to assume the distribution of study variable is known in advance. 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 case of optimum stratification for two study variables in case of compromise method of allocation.

The randomized response technique is used to procure trustworthy data for estimating the proportion of people with a sensitive characteristic. Several research workers extended the technique since its introduction by Warner (1965). Greenberg *et al.* (1971) extended the method to the case when the responses to the sensitive questions were quantitative rather than qualitative. Eichhorn and Hayre (1983) introduced a scrambled randomized response technique, which does not contain the difficulties of Greenberg *et al.* (1971) unrelated question method. 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. Mahajan *et al.* (1994) developed the theory for determination of optimum strata boundaries for a sensitive variable using scrambled response technique. Verma *et al.* (2007) considered the problem of optimum stratification for two sensitive variables using equal allocation method.

In the present paper we have considered the problem of optimum stratification when two sensitive variables are present in the survey and data is collected by scrambled randomized response technique for the compromise method of allocation and auxiliary variable is taken as stratification variable. 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.

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 $\sum_{h=1}^{L} N_h = N$ (h = 1, 2, …, L). For drawing a stratified SRSWR (Simple Random Sampling With Replacement) sample of size n, the sample of sizesn_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_{hj} denote the value of sensitive variable Y_j for the h-th stratum. Suppose S_j be 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 value Y_{hj} by S_{hj} . The interviewer thus receives the scrambled answer $Z_{hj} = Y_{hj} \, . \, S_{hj}$. 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.

Let $E(S_{hj}) = \theta_{hj}$ and $V\left(S_{hj}\right) = \gamma_{hj}^2$;

$E(Y_{hj}) = \mu_{hy_j}$ and $V(Y_{hj}) = \sigma_{hy_j}^2$;

where θ_{hj} and γ_{hj}^2 are known to the interviewer but μ_{hy_j} and $\sigma_{hy_j}^2$ are unknown. Since Y_j and S_j are independent, we have:

$$E\left(Z_{hj}\right) = \mu_{hy_j} \, . \theta_{hj} \, . \tag{1}$$

$$V\left(Z_{hj}\right) = \sigma_{hy_j}^2 \left(\theta_{hj}^2 + \gamma_{hj}^2\right) + \mu_{hy_j}^2 \, . \gamma_{hj}^2 \, . \tag{2}$$

If Z_{hj} denote 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 $\hat{\mu}_{hy_j} = \dfrac{\overline{Z}_{hj}}{\theta_{hj}}$, where:

$$\overline{Z_{hj}} = n_h^{-1} \sum_{i=1}^{n_h} Z_{hij} \, ; \tag{3}$$

Z_{hij} = Scrambled response for the j-th sensitive variable for i-th element in h-th stratum.

Hence unbiased estimator of population mean $\overline{Y}_j$ under scrambled response is given as:

$$\overline{y}_{jst} = \sum_{h=1}^{L} W_h \hat{\mu}_{hy_j} \, ;$$

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

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

$$V\left(\overline{y}_{jst}\right) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \left[\sigma_{hy_j}^2 (1 + C_{hj}^2) + (\mu_{hy_j})^2 C_{hj}^2\right] \tag{4}$$

where $C_{hj} = \dfrac{\gamma_{hj}}{\theta_{hj}}$ is the coefficient of variation of the j-th scrambling variable S_j in *h*-th stratum.

2 Allocation in stratified sampling

Optimum allocation: The guiding principle in determination of n_h is to choose them in such away so as to minimize the cost for a desired precision or maximum precision for a given cost, thus making the most effective use of the resources available. The allocation of the sample to the different strata accordance with this principle is called the principle of optimum allocation.

A special case of this allocation is the Neyman optimum allocation where it is assumed that the investigating costs per element are the same for all the strata. The number of units allocated to *h*-th stratum for this allocation is given by:

$$n_h = n.\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}.$$

This method provides the best allocation an dissuitable in cases where the stratum variance differ much.

Compromise allocation: In multivariate stratified sampling where more than one characteristics 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 as '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, 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) $\sum_{j=1}^{2} V\left(\overline{y}_{jst}\right)$ is minimized where $V(\overline{y}_{jst})$ 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 is given by (5).

$$V\left(\overline{y}_{jst}\right) = \sum_{h=1}^{L} n_h^{-1} W_h^2 \left[\sigma_{hy_j}^2 (1 + C_{hj}^2) + \mu_{hy_j}^2 C_{hj}^2\right]. \quad (5)$$

We have to minimize:

$$\sum_{j=1}^{2} V\left(\overline{y}_{jst}\right) = V(\overline{y}_{1st}) + V(\overline{y}_{2st}). \quad (6)$$

Now minimizing (6) subject to the condition $\sum_{h=1}^{L} n_h = n$ the optimum value of n_h is given by:

$$n_h = n\,\frac{W_h\sqrt{K_{hy_1}^2 + K_{hy_2}^2}}{\sum_{h=1}^{L} W_h\sqrt{K_{hy_1}^2 + K_{hy_2}^2}}\,; \tag{7}$$

where: $K_{hy_j}^2 = \sigma_{hy_j}^2\left(1 + C_{hj}^2\right) + \mu_{hy_j}^2 C_{hj}^2\,(j = 1, 2).$

Using this value of n_h we have obtained the variance expression for compromise allocation. Under compromise method of allocation, the optimal variance of the estimated population mean of the sensitive variables Y_j is given:

$$V\left(\overline{y}_{jst}\right) = \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h K_{hy_j}^2}{\sqrt{K_{hy_1}^2 + K_{hy_2}^2}}\ \sum_{h=1}^{L} W_h\sqrt{K_{hy_1}^2 + K_{hy_2}^2}\right]\ (j = 1, 2). \tag{8}$$

3 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:

$$Y_j = c_j\left(X\right) + e_j\,; \tag{9}$$

where: $c_j\left(X\right)$ 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 (9) it can be easily seen that:

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

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

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

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

$$\sigma^2_{hy_j} = \sigma^2_{hc_j} + \mu_{h\varphi_j}.$$

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 function of the j-th sensitive variable.

The variance expression of $\overline{y}_{jst}$ for compromise allocation under super population model (9) are given by:

$$\sigma_1^2 = V(\overline{y}_{1st}) = \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h K^2_{hc_1}}{\sqrt{K^2_{hc_1} + K^2_{hc_2}}} \sum_{h=1}^{L} W_h \sqrt{K^2_{hc_1} + K^2_{hc_2}}\right]; \tag{10}$$

$$\sigma_2^2 = V(\overline{y}_{2st}) = \frac{1}{n}\sum_{h=1}^{L}\left[\frac{W_h K^2_{hc_2}}{\sqrt{K^2_{hc_1} + K^2_{hc_2}}} \sum_{h=1}^{L} W_h \sqrt{K^2_{hc_1} + K^2_{hc_2}}\right]; \tag{11}$$

where: $K^2_{hc_j} = (\sigma^2_{hc_j} + \mu_{h\varphi_j})(1 + C^2_{hj}) + \mu^2_{hc_j} C^2_{hj} \, (j = 1, 2).$

We assumed 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. Now generalized variance *G* is given by:

$$G = \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 \tag{12}$$

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:

$$tr(G) = \sigma_1^2 + \sigma_2^2 . \qquad (12)$$

Using (10) and (11) in (12) $tr(G)$ can be expressed as:

$$tr(G) = \frac{1}{n}\left[\sum_{h=1}^{L} W_h \sqrt{K_{hc_1}^2 + K_{hc_2}^2}\right]^2 \qquad (13)$$

Verma *et al.* (2003) obtained the solution of the minimal equations obtained by minimizing the trace of variance covariance matrix (13). They proposed following cumulative cube root rule for obtaining approximate optimum strata boundaries.

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

If the function $M(x) = P(x) f(x)$ 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 $M(x)$ will give approximately optimum strata boundaries (AOSB), where:

$$P(x) = \frac{4\left(\phi_1^*(x) + \phi_2^*(x)\right)\left(c_1'^2(x) + c_2'^2(x)\right) + \left(\phi_{11}(x) + \phi_{21}(x)\right)^2}{\left(\phi_1^*(x) + \phi_2^*(x)\right)^{3/2}}; \qquad (14)$$

and: $\phi_1^* = \phi_1 + \phi_1 C_{h1}^2 + c_1^2 C_{h1}^2$, $\qquad \phi_{11} = \phi_1' + \phi_1' C_{h1}^2 + 2c_1 c_1' C_{h1}^2$,

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

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

4 Limiting form of the trace of the variance covariance matrix

For obtaining the limiting expression for the trace of variance covariance matrix $tr(G)$ as defined in (13), 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). For this purpose we impose certain regularity conditions on the functions $f(x)$, $\phi_j{}^*(x)$ and $c_j(x)$ as given below.

A function $\omega(x)$ belongs to the class of functions Ω if it satisfies the following regularity conditions:

(i) $0 < \omega(x) < \infty$

(ii) $\omega(x)$, $\omega'(x)$ and $\omega''(x)$ exist and are continuous for all $x \in (a,b)$, where $(b-a) < \infty$.

Lemma 1: Under regularity conditions, for h-th stratum we have

$$W_h\sqrt{K_{hc_1}^2+K_{hc_2}^2} - \int_{x_{h-1}}^{x_h}\sqrt{\phi_1^*(x)+\phi_2^*(x)}\; f(x)dx \;=\; \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P(x)f(x)dx[1+O(k_h^2)];$$

where $P(x)$ is defined in (14).

Theorem 1: If the AOSB are obtained by using cumulative cube root rule $\sqrt[3]{M(x)}$ then limiting expression for the trace of variance covariance matrix $tr(G)$ is given by:

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

Where $\alpha=\int_a^b\sqrt{[\phi_1^*(x)+\phi_2^*(x)]}\; f(x)dx$ and $\beta=\frac{1}{96}\left[\int_a^b\sqrt[3]{P(x)f(x)}\,dx\,\right]^3$

Proof: Now making use of the Lemma 4.1 in the expression (13), we have:

$$tr(G)=\frac{1}{n}\left[\int_a^b\sqrt{[\phi_1^*(x)+\phi_2^*(x)]}\; f(x)dx+\sum_{h=1}^{L}\frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P(x)f(x)dx[1+O(k_h^2)]\right]^2. \qquad (15)$$

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

$$tr(G)=\frac{1}{n}\left[\int_a^b\sqrt{\left[\phi_1^*(x)+\phi_2^*(x)\right]}\; f(x)dx+\frac{1}{96}\sum_{h=1}^{L}\left\{\int_{x_{h-1}}^{x_h}\sqrt[3]{P(x)f(x)}\;\; dx\,\}^3\right\}\right]^2. \qquad (16)$$

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

$$\int_{x_{h-1}}^{x_h}\sqrt[3]{P(x)f(x)}\,dx \;=\frac{1}{L}\int_a^b\sqrt[3]{P(x)f(x)}\,dx. \qquad (17)$$

Therefore, equation (16) reduces to:

$$tr(G)=\frac{1}{n}\left[\alpha+\frac{\beta}{L^2}\right]^2; \qquad (18)$$

where:

$$\alpha = \int_a^b \sqrt{\left[\phi_1^*(x) + \phi_2^*(x)\right]}\, f(x)dx \text{ and } \beta = \frac{1}{96}\left[\int_a^b \sqrt[3]{P(x) f(x)}\, dx]\right]^3.$$

Now taking limit as $L \to \infty$ on both sides of (18) we get:

$$\underset{L\to\infty}{Lim}\ tr(G) = \frac{\alpha^2}{n}. \qquad (19)$$

From the above relation it may be concluded that with an increase in the number of strata *L*, the trace of generalized variance covariance matrix decreases and as the number of strata becomes large enough, $tr(G)$ tends to α^2/n. However if number of strata *L* goes to infinity then the sample size n goes more faster to infinity, because we have to select minimum one unit from each stratum hence $tr(G) \to 0$. But in general practice number of strata are always finite hence the equation (18) is useful for determining the trace of variance covariance matrix.

5 Optimum number of strata

The trace of the variance- covariance matrix of the estimator $\overline{y}_{jst}$ as given in (18) 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 for *L*, the number of strata to be constructed. The variance (18) is only the function of *L* as α and β 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 $tr(G)$ as given in (18) with respect to *L* we get.

$$\alpha L^2 + \beta = 0. \qquad (20)$$

6 Approximate expression for sample size $[n_h]$

After the strata boundaries have been obtained by cumulative $\sqrt[3]{M(x)}$ rule of Verma *et al.* (2003) for the number of strata *L* satisfying (20), the sample size $[n_h]$ allocated to the *h*-th stratum is given by (7). Since the functions *f(x), C(x)*, and $\phi(x)$ are known a priori, the parameter W_h, μ_{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 tedious to determine $[n_h]$ from (7) because of integrations involved in it. We now obtained the approximate expression for the sample size $[n_h]$.

Theorem 2: If the AOSB are obtained by using cumulative cube root rule $\sqrt[3]{M(x)}$ then the approximate expression for the sample size $[n_h]$ allocated to h-th stratum is given by:

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

Proof: For obtaining the approximate expression for the sample size $[n_h]$ allocated to h-th stratum we used Lemma 4.1.

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

$$n_h = \frac{n}{\left(\alpha + \frac{\beta}{L^2}\right)}\left[\int_{x_{h-1}}^{x_h}\sqrt{\left[\phi_1^*(x) + \phi_2^*(x)\right]}\, f(x)\,dx + \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P(x)\, f(x)\,dx\right]; \qquad (21)$$

Where:

$$\sum_{h=1}^{L} W_h \sqrt{\left[K_{hc_1}^2 + K_{hc_2}^2\right]} = \left(\alpha + \frac{\beta}{L^2}\right).$$

Then (21) is approximately given by

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

Where:

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

$$P(\bar{x}_h) = \frac{4\left(\phi_1^*(\bar{x}_h) + \phi_2^*(\bar{x}_h)\right)\left(c_1'^2(\bar{x}_h) + c_2'^2(\bar{x}_h)\right) + \left(\phi_{11}(\bar{x}_h) + \phi_{21}(\bar{x}_h)\right)^2}{\left(\phi_1^*(\bar{x}_h) + \phi_2^*(\bar{x}_h)\right)^{3/2}}.$$

If optimum points of stratification $\{x_h\}$ are obtained by using the cumulative cube root rule $\sqrt[3]{M(x)}$ then the equation (22) can be used for determination of optimum sample size n_h.

7 Imperial stud

To determine approximately optimum strata boundaries (AOSB) by the use of cumulative cube root rule we consider that stratification variable x follows following probability density function:

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 distributions are finite whereas range of exponential distribution is infinite. We have considered that sensitive study variables Y_j are related with the stratification variable x as $Y_1 = a_1 + x + e_1$, $Y_2 = a_2 + 2x + e_2$. 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 values of $g_1 = 1$ and $g_2 = 2$. 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.

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

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 ranges of uniform, right triangular and exponential distributions were divided into 10 classes of equal width. The Approximately optimum strata boundaries (AOSB) obtained by the use of cumulative cube root rule $\sqrt[3]{M(x)}$ as given in equation (14) along with the relative efficiency of stratification with no stratification. These strata boundaries can further be used for determination of trace of variance- covariance matrix and optimum sample sizes. From Table 1 it is clearly evident that for the uniform distribution as the number of strata increases relative efficiency increases from 173.07% (L=2) to 222.11% (L=6). From Table 2 we observed that for the right triangular distribution relative efficiency increases from 139.45% (L=2) to 160.41% (L=6). From Table 3 it is indicated that in case of exponential distribution relative efficiency increases from 179.92% (L=2) to 245.72% (L=6). Table 1 to Table 3 can further be used for finding the generalized variance and the approximate sample sizes.

Table 1 - AOSB for uniform distribution

No. of Strata L	Approximately optimum strata boundaries (AOSB)					$n\,tr(G)$	Percent Relative Efficiency
1						0.75723	100.00
2	1.47204					0.43752	173.07
3	1.30795	1.64224				0.37732	200.69
4	1.22856	1.47204	1.72955			0.35612	212.63
5	1.18156	1.37291	1.57340	1.78261		0.34629	218.67
6	1.15074	1.30795	1.47204	1.64224	1.81829	0.34093	222.11

Table 2 - AOSB for Right triangular distribution

No. of Strata L	Approximately optimum strata boundaries (AOSB)					$n\,tr(G)$	Percent Relative Efficiency
1						0.749912	100.00
2	1.38335					0.53778	139.45
3	1.244413	1.540356				0.494651	151.60
4	1.179356	1.38335	1.627781			0.478974	156.57
5	1.1421	1.297641	1.474946	1.684174		0.471569	159.02
6	1.117264	1.244413	1.38335	1.540356	1.724831	0.467498	160.41

Table 3 - AOSB for Exponential distribution

No. of Strata L	Approximately optimum strata boundaries (AOSB)					$n\,tr(G)$	Percent Relative Efficiency
1						7.045131	100.00
2	2.310008					3.915675	179.92
3	1.775334	3.017928				3.267838	215.59
4	1.542033	2.310008	3.488245			3.035232	232.11
5	1.425557	1.961975	2.712787	3.834811		2.926433	240.74
6	1.354631	1.775334	2.310008	3.017928	4.083356	2.867181	245.72

Conclusion

In the present paper we have considered the problem of optimum stratification when two sensitive variables are present in the survey and data are collected by scrambled randomized response technique for the compromise method of allocation and an auxiliary variable is taken as stratification variable. This strategy is important when we are dealing with stigmatized quantitative variables. For instance letY_1be the “Income understated in income tax return” and Y_2be the “Expenditure”. These two variables can be stratified by using an auxiliary variable X (Eye estimated value of property) as the stratification variable. Then optimum points of stratification can be found out. For example in our numerical example we have considered that the stratification variable followed uniform, right triangular and exponential distributions. Further optimum strata boundaries can be used to find the generalized variance and approximate sample sizes. In general if the distribution of the auxiliary variable is known and information on the two sensitive variables is collected by scrambled response then in that case present methodology can be used for finding the generalized variance and approximate sample sizes.

Acknowledgement

The authors would like to thank two learned referees and the Editor for their valuable suggestion to bring the paper in the present form.

VERMA, M. R.; JOOREL, J. P. S; AGNIHOTRI, R. K. Estratificação de variância mínima para alocação de tarefas. *Rev. Bras. Biom.*, São Paulo, v.30, n.2, p.278-291, 2012.

- RESUMO: Neste artigo estudou-se o problema da estratificação ótima para duas variáveis quantitativas sensitivas com dados obtidos a partir de diferentes camadas pela técnica de resposta embaralhada e uma variável auxiliar usada como variável de estratificação.Foi obtida a expressão limitante para o traço da matriz variância covariância generalizada, as expressões para o número de estratos e para o tamanho aproximado da amostra.O artigo apresenta uma ilustração numérica.

- PALAVRAS-CHAVE: Variável sensitiva; respostas embaralhadas; estratificação ótima; AOSB (Limites Estratos Aproximadamente Ótimos).

References

DALENIUS, T. The problem of optimum stratification. *Skand. Akt.*, Oslo, v.33, p.203-213, 1950.

DALENIUS, T.; GURNEY, M. The problem of optimum stratification II. *Skand. Akt.*, Oslo, v.34, p.133-148, 1951.

DALENIUS, T.; HODGES JR., J. L. Minimum variance stratification. *J. Am. Stat. Assoc.*, Baltimore, v.54, n.1, p.88–101, 1957.

EICHHORN, B.H.; HAYRE, L. S. Scrambled randomized response method for obtaining sensitive quantitative data. *J. Stat. Plan. Inference*, Amsterdam,v.7, n.4, p.307-316, 1983.

GHOSH, S. P. Optimum stratification with two characters. *Ann. Math. Statist.*, Bethesda, v.34, n.3, p.866-872, 1963.

GREENBERGE, B. G.; KUBLER, R. R.; ABERNATHY, J.R.; HORVITZ, D. G. Applications of randomized response technique in obtaining quantitative data. *J. Am. Stat. Assoc.*, New York, v.66, n.334, p.243-250, 1971.

NEYMAN, J. On the two different aspects of representative methods: The method of stratified sampling and method of purposive selection. *J. R. Stat. Soc.*, Chichester, v.97, n.4, p.558-606, 1934.

MAHAJAN, P.K.; GUPTA, J.P.; SINGH, R. Determination of optimum strata boundaries for scrambled response. *Statistica*, Rome, v.54, n.3, p.375-381, 1994.

RIZVI, S. E. H.; GUPTA, J.P.; SINGH, R. Approximately optimum stratification for two study variables using auxiliary information. *J. Indian Soc. Agric. Stat.*, New Delhi, v.53, n.3, p.287-298, 2000.

RIZVI, S.E. H.; GUPTA, J.P.; BHARGAVA, M. Optimum stratification based on auxiliary variable for compromise allocation. *Metron*, Rome, v.50, n.3-4, p.201-215, 2002.

SINGH, R.; SUKHATME, B. V. Optimum stratification. *Ann. Inst. Stat. Math.*, Heidelberg, v.21, p.515-528, 1969.

SUKHATME, P. V.; SUKHATME, B.V.; SUKHATME, S.; ASOK, C. *Sampling theory with applications*. 3.ed. New Delhi: Iowa, Indian Society of Agricultural Statistics, 1984. 526p.

VERMA, M. R.; SINGH JOOREL, J. P.; AGNIHOTRI, R. K. Approximately optimum stratification for two sensitive variables in case of compromise method of allocation. *Sri Lankan J. Appl. Stat.*, Abingdon, v.4, p.1-13, 2003.

VERMA, M. R.; SINGH JOOREL, J. P.; AGNIHOTRI, R. K. Optimum stratification for two sensitive quantitative variables using equal allocation method. *Braz. J. Math. Stat.*, São Paulo, v.25, n.2, p.73-85, 2007.

WARNER, S.L. Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.*, New York, v.60, p.63-69, 1965.

Received in 10.06.2012

Approved after revised in 17.10.2012