J. Ind. Soc. Agril. Statist.
58(3), 2004 : 269-278

An Alternative Method of Optimum Stratification for Scrambled Response

P.K. Mahajan and M.R. Verma
University of Horticulture and Forestry, Nauni-Solan 173 230
(Received : May, 2002)

SUMMARY

Mahajan *et al.* (1994, 1997) considered the problem of finding optimum strata boundaries when the samples from different strata are selected with simple random sampling with replacement and the data are collected by scrambled randomized response technique on the sensitive character. The rule developed by Mahajan *et al.* (1994) for stratification variable x is suitable for the cases where correlation between c(x) and estimation variable y is small. This paper deals with the most ideal situation where correlation between c(x) and y is high. A limiting expression for the variance of the estimator of population mean and approximate expression for $[n_h]$ have also been suggested. The paper concludes with a numerical illustration.

Key words : Scrambling variable, Minimal equations, Strata boundaries, Limiting variance.

1. Introduction

Mahajan *et al.* (1994, 1997) have considered the problem of optimum stratification on an auxiliary variable x when the samples from different strata are selected with simple random sampling and with replacement (SRSWR) and the data on sensitive character are collected by scrambled randomized response technique proposed by Eichhorm and Hayre (1983). The scrambled randomized method involves the respondent multiplying his sensitive answer Y by a random number S from a known distribution and giving the scrambled response $Z = YS$ to the interviewer, who does not know the particular values of the random number S. Let the population of N units be stratified into L strata and the samples from each stratum be selected with SRSWR. Further, for the hth stratum let Y_h denote the value of the sensitive character under study, and S_h be a scrambling random variable independent of Y_h and with finite mean and variance. The respondent generates S_h using some specified method and multiplies the variable value Y_h by S_h. The particular values of S_h are unknown to the interviewer, but its distribution is known. In this way, the respondents privacy is not violated.

Let $E(S_h) = \theta_h$, $V(S_h) = r_h$, $E(Y_h) = \mu_{hy}$, $V(Y_h) = \sigma^2_{hy}$ and $C_h = \sqrt{r_h/\theta_h}$ where θ_h and r_h are known to the interviewer but μ_{hy} and σ^2_{hy} are unknown. Mahajan *et al.* (1994) showed that, an unbiased estimator of population mean μ is

$$\hat{\mu}_{St} = \sum_{h=1}^{L} W_h \hat{\mu}_{hy} \text{ with a variance}$$

$$V(\hat{\mu}_{St}) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \{\sigma^2_{hy}(1 + C_h^2) + \mu^2_{hy} C_h^2\} \tag{1.1}$$

where $\hat{\mu}_{hy}$ is an unbiased estimator for μ_{hy} and W_h is the proportion of units falling in the hth stratum.

Let the regression of y on x in the population be of the form

$$y = c(x) + e \tag{1.2}$$

where $c(x)$ is a function of x and e is the error term such that $E(e/x) = 0$ and $v(e/x) = \phi(x) \geq 0 \; \forall \; x$ in the range (a, b) of x with $(b - a) < \infty$. If f(x) denotes the marginal density of x, then we have

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx; \; \mu_{hy} = \mu_{hc} = W_h^{-1} \int_{x_{h-1}}^{x_h} c(x)f(x)dx$$

and $$\sigma^2_{hy} = \sigma^2_{hc} + \mu_{h\phi} \tag{1.3}$$

where (x_{h-1}, x_h) are the boundaries of the hth stratum; $\mu_{h\phi}$ is the expected value of $\phi(x)$ and σ^2_{hc} is the variance of $c(x)$ in the hth stratum.

Using the relations (1.3), the variance expression (1.1) under Neyman Allocation reduces to

$$nV(\hat{\mu}_{St.})_N = \left[\sum_{h=1}^{L} W_h \sqrt{(\sigma^2_{hc} + \mu_{h\phi})(1 + C_h^2) + \mu^2_{hc} C_h^2}\right]^2 \tag{1.4}$$

The relative magnitude of the two terms in $(\sigma^2_{hc} + \mu_{h\phi})$ depends on the correlation coefficient (ρ) between $c(x)$ and y. The cases in which ρ is high, σ^2_{hc} is much larger than $\mu_{h\phi}$ and if it is low σ^2_{hc} is much smaller than $\mu_{h\phi}$. In

the rule developed by Mahajan *et al.* (1994) for stratification of x, $\left|\frac{\sigma_{hc}^2}{\mu_{h\phi}}\right|$ has been taken to be less than one. This rule, therefore, is suitable for the cases where correlation between c(x) and y is small. There is therefore, a need to develop new rules for stratification on x which considers an ideal case where ρ is large i.e. when $\left|\frac{\mu_{h\phi}}{\sigma_{hc}^2}\right| < 1$. This is what we propose to do in this paper.

2. *The Variance* $V(\hat{\mu}_{St})$

Using relations (3.2) and (3.3) from Section 3, the approximation to $\mu_{h\phi}$ for hth stratum becomes

$$\mu_{h\phi} = \frac{12\sigma_{h\psi}^2}{K_h^2}[1 + O(K_h^2)] \tag{2.1}$$

where, $K_h = x_h - x_{h-1}$ and the function $\psi(x)$ is such that $\psi'(x) = \sqrt{\phi(x)}$

Now when the number of strata is large then the terms of $O(K_h^2)$ can be neglected in comparison to 1, the variance expression in (1.4), using (2.1) and taking $k_h = (b-a)/L$ for $h = 1, 2, \ldots, L$ reduces to

$$nV(\hat{\mu}_{St.})_N = \left[\sum_{h=1}^{L} W_h \sqrt{(\sigma_{hc}^2 + \theta\sigma_{h\psi}^2)(1 + C_h^2) + \mu_{hc}^2 C_h^2}\right]^2 \tag{2.2}$$

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

In cases where the correlation between y and c(x) is high the magnitude of $\mu_{h\phi}$ is quite small in comparison to σ_{hc}^2. The term $(\sigma_{hc}^2 + \mu_{h\phi})$ is, therefore, mainly dominated by σ_{hc}^2. Thus the magnitude of error, when $\mu_{h\phi}$ is approximated by $\theta\sigma_{h\psi}^2$ will not be much in comparison to σ_{hc}^2. It may be noted that Serfling (1968) while proposing the use of cum. $\sqrt{f}$ method for finding the approximately optimum strata boundaries (AOSB) has altogether neglected $\mu_{h\phi}$ in comparison to σ_{hc}^2.

3. Minimal Equations and their Approximate Solutions

For finding the minimal equations, solutions to which give the optimum points of stratification $[x_h]$, we shall consider the variance of the estimate $\hat{\mu}_{St.}$ as given in (2.2). The minimization of the variance in (2.2) is equivalent to the minimization of $\sum_{h=1}^{L} W_h \sqrt{(\sigma_{hc}^2 + \theta\sigma_{h\psi}^2)(1+C_h^2) + \mu_{hc}^2 C_h^2}$. On equating to zero the partial derivative of this function with respect to x_h, the minimal equations become

$$\frac{(1+C_h^2)[(C(x_h)-\mu_{hc})^2 + \theta(\psi(x_h)-\mu_{h\psi})^2 + \sigma_{hc}^2 + \theta_{h\psi}^2] + 2C_h^2\mu_{hc}C(x_h)}{\sqrt{(\sigma_{hc}^2 + \theta\sigma_{h\psi}^2)(1+C_h^2) + C_h^2\mu_{hc}^2}}$$

$$= \frac{(1+C_i^2)[(C(x_h)-\mu_{ic})^2 + \theta(\psi(x_h)-\mu_{ic})^2 + \sigma_{ic}^2 + \theta_{i\psi}^2] + 2C_i^2\mu_{ic}C(X_h)}{\sqrt{(\sigma_{ic}^2 + \theta\sigma_{i\psi}^2)(1+C_i^2) + C_h^2\mu_{ic}^2}}$$

$$i = h+1;\ h = 1, 2, \ldots, L-1. \qquad (3.1)$$

The system of equations (3.1) involves strata parameters which are themselves functions of the solutions of these equations. Due to this implicity, it is not possible to obtain exact solutions. We shall, therefore, proceed to find the method of solving them at least approximately.

To find approximate solutions to the minimal equations, we obtain the series expansions of the minimal equations about the point x_h, the common boundary point of hth and (h + 1)th strata. The existence of the various functions and their derivatives occurring in these expansions will be assumed for all x in (a, b).

Using Taylor's theorem, it has been shown by Singh and Sukhatme (1969) that 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

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

$$= \phi\left[1 + \frac{\phi'}{2\phi}K + \frac{\phi' f' + 2f\phi''}{12f\phi}K^2 + \frac{(ff''\phi' + ff'\phi'' + f^2\phi''' - \phi f'^2)}{24f^2\phi} + O(K^4)\right] \qquad (3.2)$$

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

$$\sigma_\phi^2(y, x) = \frac{K^2}{12}\phi'^2\left[1 + \frac{\phi''}{\phi'}K + 0(K^2)\right] \tag{3.3}$$

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

$$\left[\int_y^x \sqrt[\lambda]{f(t)}dt\right]^3 = K^\lambda f(y)\left[1 + \frac{K}{2}.\frac{f'(y)}{f(y)} + 0(K^2)\right]$$

$$= K^{\lambda-1}\int_y^x f(t)\,dt[1 + 0(K^2)] \tag{3.4}$$

In order to obtain the series expansions of the minimal equations in (3.1), these relations are to be used with (y, x) being replaced by (x_{h-1}, x_h). The expansions for $\mu_{hc}, \mu_{h\psi}, \sigma_{hc}^2$ etc. are obtained by substituting $c(x), \psi(x)$, respectively for $\phi(x)$.

Using these expansions, the system of minimal equations in (3.1) reduces to

$$K_h^2\left[1 - \frac{K_h}{3}.\frac{t'(x_h)}{t(x_h)} + 0(K_h^2)\right] = K_i^2\left[1 + \frac{K_i}{3}.\frac{t'(x_h)}{t(x_h)} + 0(K_i^2)\right]$$

where, $t(x_h) = \dfrac{f(c'^2 + \theta\psi'^2)(1 + C_h^2)}{c^{3/2}}$; $i = h + 1$

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

$$K_h^3\left[1 - \frac{K_h}{2}.\frac{t'(x_h)}{t(x_h)} + 0(K_h^2)\right] = K_i^3\left[1 + \frac{K_i}{2}.\frac{t'(x_h)}{t(x_h)} + 0(K_i^2)\right] \tag{3.5}$$

On comparing it with (3.4), with $\lambda = 3$, the system of equations (3.5) can be written in the form

$$K_h^2\int_{x_{h-1}}^{x_h} t(x)\,dx\,[1 + 0(K_n^2)] = K_i^2\int_{x_h}^{x_{h-1}} t(x)\,dx\,[1 + 0(K_h^2)] \tag{3.6}$$

If the terms of order $0(m^5)$ where $m = \underset{(a,b)}{\text{sub.}}(K_h)$ are neglected then the system of equations (3.1) or equivalently (3.6) can be approximated by

$$K_h^2 \int_{x_{h-1}}^{x_h} t(x)dx = K_i^2 \int_{x_h}^{x_{h+1}} t(x)dx = \text{constant}$$

or $$\int_{x_{h-1}}^{x_h} \sqrt[3]{t(x)}dx = \frac{1}{L}\int_a^b \sqrt[3]{t(x)}dx \qquad \text{[By virtue of (3.4)]}$$

where $$t(x) = G(x)\,f(x) = \left\{\frac{[c'^2(x) + \theta\psi'^2(x)]\,(1+C_h^2)}{[c(x)]^{3/2}}\right\} f(x) \qquad (3.7)$$

This gives us the following rule for finding the approximately optimum strata boundaries (AOSB) on the auxiliary variable x.

Cum. $\sqrt[3]{f(x)\,G(x)}$ Rule: If the function $p(x) = f(x)\,G(x)$ is bounded and possesses first two derivatives for all x in (a, b), then for a given value of L, taking equal intervals on the cum. $\sqrt[3]{p(x)}$ yields AOSB.

4. Expression for Limiting Variance $V(\hat{\mu}_{St.})$

Limiting expression for the variance is particularly important in optimum stratification as it gives an insight into the manner in the which the variance of the estimator $\hat{\mu}_{St.}$ is reduced with the increase in the number of strata. We will express $V(\hat{\mu}_{St.})$ as given in (1.4) in terms of the number of strata L and some other constants which do not depend on strata boundaries. For this purpose, we prove the following lemma.

Lemma 4.1 : If (x_{h-1}, x_h) are the boundaries of the hth stratum and $K_h = x_h - x_{h-1}$, then

$$W_h\sqrt{(\sigma_{hc}^2 + \theta\sigma_{h\psi}^2)\,(1+C_h^2) + \mu_{hc}^2 C_h^2} - \int_{x_{h-1}}^{x_h} c(x)\,f(x)\,C_h d_x$$

$$= \frac{1}{24}\left[\int_{x_{h-1}}^{x_h} \sqrt[3]{G(x)\,f(x)}\;dx\right]^3 [1 + 0(K_h^2)] \qquad (4.1)$$

Proof : Utilizing the series expansions in powers of interval width $K_h = x_h - x_{h-1}$ from the relations (3.2) and (3.3) and on using the Taylor's theorem for the expansion of the integrand about point $x = x_h$, the L.H.S. of (4.1) reduces to

$$\frac{K_h^2}{24}\left[f(x)\,G(x).\,K_h - \frac{1}{2}\frac{d}{dx_h}[f(x).G(x)]K_h^2 + 0(K_h^3)\right]$$

$$= \frac{K_h^2}{24}\int_{x_{h-1}}^{x_h} f(x)\,G(x)\,dx.[1 + 0(K_h^2)] \tag{4.2}$$

$$= \frac{1}{24}\left[\int_{x_{h-1}}^{x_h}\sqrt[3]{f(x)\,G(x)}dx\right]^3 [1 + 0(K_h^2)] \qquad \text{[By virtue of (3.4)]}$$

Hence the lemma.

Using (4.1) in (1.4), we get limiting expression for the variance as

$$V(\hat{\mu}_{St}) = \frac{1}{n}\left(\alpha + \frac{\beta}{L^2}\right)^2 \tag{4.3}$$

where

$$\alpha = \int_a^b c(x)\,f(x)\,C_h dx \tag{4.4}$$

$$\beta = \frac{1}{24}\int_a^b [\sqrt[3]{G(x)f(x)}dx]^3 \tag{4.5}$$

5. Approximate Expression for $[n_h]$

If for a given value of L, the variance expression in (1.4) is minimized w.r.t. $[n_h]$ and subject to the condition $n = \sum_h n_h$, the optimum value of $[n_h]$ are given by

$$n_h = \frac{n}{\Sigma W_h K_{hy}} W_h K_{hy} \tag{5.1}$$

where, $K_{hy} = \sqrt{(\sigma_{hc}^2 + \theta \mu_{hy}^2)(1 + C_h^2) + \mu_{hc}^2 C_h^2}$

Some times, it may be tedious to determine $[n_h]$, using (5.1) because of the integrations involved. Therefore, an approximate expression for obtaining the sample size $[n_h]$, which is free from integration, has been obtained below.

Using (4.2), when the terms of order $0(m^5)$ are neglected, the expression (5.1) for sample size n_h in the hth stratum reduces to

$$n_h = \frac{n}{\alpha + \beta/L^2}\left(\int_{x_{h-1}}^{x_h} c(x)\, f(x)\, C_h dx + \frac{K_h^2}{24}\int_{x_{h-1}}^{x_h} G(x)\, f(x)\, dx\right) \tag{5.2}$$

If $\bar{x}_h = \frac{x_h + x_{h-1}}{2}$, then (5.2) is approximately given by

$$n_h = \frac{n}{\alpha + \beta/L^2}\left(c(\bar{x}_h)\, C_h + \frac{G(\bar{x}_h)}{24} K_h^2\right) W_h \tag{5.3}$$

Given the frequency table, the following analogous expressions can be used for calculating the value of α and β

$$\alpha = \Sigma c(\bar{x}_i)\, C_h W_h$$

$$\beta = \left[\Sigma \{K_i^2\, G(\bar{x}_i) W_i\}^{1/3}\right]^3 / 24$$

where for the ith class $\bar{x}_i$ is the mid value, K_i is the width and W_i is the relative frequency and summation is carried out over all the classes.

7. Numerical Illustration

In actual practice, the frequency distribution of the auxiliary variable x is known. Suppose the distribution of x is exponential with p.d.f. $f(x) = e^{-x+1}$, $1 \le x < \infty$. In order to have the finite range for the variable x, the distribution was truncated at x = 6 so that the probability for x to take values beyond the truncation point is extremely small.

In order to find the approximate optimum strata boundaries (AOSB), the range of the distribution was divided into 10 classes of equal width. The relative frequencies given in column 2 of Table 1 were the areas corresponding to these classes. Let us assume that from a priori information $c(x) = a + x$, $\phi(x) = \lambda x$ and $C_h = 0.2$ where a and λ were determined in such a way that 90% of the total variation was accounted for by the regression. The function G(x) defined in (3.7) was evaluated at mid points of the class intervals and then multiplied by f.

The cube roots of this product $fG(x)$ were then found for each of the 10 classes. These cube roots were cumulated and using linear interpolation by taking equal intervals on the cumulative totals, AOSB were obtained. These AOSB have been presented in column I of Table 2.

Table 1. Relative frequency distribution for the variable x

Class interval	f	G(x)
1.0-1.5	0.39347	1.28331
1.5-2.0	0.23865	0.90490
2.0-2.5	0.14475	0.71000
2.5-3.0	0.00779	0.59154
3.0-3.5	0.05326	0.51187
3.5-4.0	0.03229	0.45449
4.0-4.5	0.01959	0.41109
4.5-5.0	0.01188	0.37704
5.0-5.5	0.00721	0.34953
5.5-6.0	0.00437	0.32681

Table 2. Approximate sample size $[n_h]$

AOSB	W_h	$W_h c(\bar{x}_h) C_h$	$W_h \frac{G(\bar{x}_h)}{24} K_h^2$	(3 + 4)	n_h
1.00000-1.42666	0.34732	0.08428	0.03492	0.08777	24
1.42666-1.96896	0.27321	0.09277	0.03122	0.09589	23
1.96896-2.73299	0.20272	0.09532	0.03361	0.09868	21
2.73299-3.86618	0.11984	0.06908	0.03240	0.08232	20
3.86618-6.00000	0.05018	0.04951	0.03487	0.52299	12

In order to use the proposed expression (5.3) for obtaining optimum sample size to be allocated to different strata, the results of column 2 of Table 2 were obtained by reconstructing the frequency table with AOSB given in column 1. Columns 3 and 4 show the values of $W_h c(\bar{x}_h) C_h$ and $W_h G(\bar{x}_h) K_h^2 / 24$, respectively. These values were worked out at the mid points $(\bar{x}_h)$ of their respective classes. Using approximations (4.4) and (4.5), we find that $\alpha = 0.39467$ and $\beta = 0.40853$. Finally, the required values of n_h obtained by using the expression (5.3) are recorded in column 6.

ACKNOWLEDGEMENT

The authors are thankful to Late Professor Ravindra Singh, PAU, Ludhiana, India for suggesting the problem and are also grateful to the referee for constructive suggestions.

REFERENCES

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

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

Mahajan, P.K. and Gupta, J.P. (1997). Optimum stratification for scrambled response. *J. Statist. Res.*, **31(1)**, 131-136.

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