Research Article

Variance Expression for Eliminating Evasive Answer Bias in Optimum Stratification

Ashu Chandel, P K Mahajan*

Department of Basic Sciences (Mathematics and Statistics Section), Dr Y. S. Parmar University of Horticulture and Forestry, Nauni, Solan 173230, INDIA.
Email: pawan_uhf@yahoo.com

Abstract

Randomized response methods for quantitative sensitive data are treated in a unified approach which includes the use of auxiliary information at the estimation stage (refer Singh et al. 1996; Chaudhuri and Roy 1997; Singh and Tracy 1999; Mahajan and Singh 2005; Mahajan 2006). Auxiliary information for quantitative sensitive data in optimum stratification has received less attention so far. Keeping in view its importance in personal interview surveys on sensitive issues, the limiting expression of variance for eliminating evasive answer bias in quantitative sensitive data has been proposed. This expression gives an insight into the manner in which the variance of the estimator of mean for the sensitive character under optimum allocation changes with the increase in the number of strata. The paper concludes that proposed limiting expression of variance in turn also establishes the strata boundaries $[x_h]$ which are approximately optimum.
Keywords: Auxiliary variable, Neyman allocation, optimum stratification, scrambled response.

***Address for Correspondence:**
Dr. P K Mahajan Professor and HOD, Department of Basic Sciences (Mathematics and Statistics Section), Dr Y. S. Parmar University of Horticulture and Forestry, Nauni, Solan 173230, INDIA.
Email: pawan_uhf@yahoo.com
Received Date: 22/03/2015 Accepted Date: 04/04/2015

Access this article online

Quick Response Code:

Website: www.statperson.com

DOI: 06 April 2015

INTRODUCTION

Sample surveys on human populations have established that the innocuous questions usually receive good response, whereas information on highly sensitive issues such as tax evasion, illegal income, and life expectancy of the HIV/AIDS affected persons etc. excites resistance. Direct questions about them often result in either refusal to respond or falsification of their answers. Several randomized response models have been developed by researchers for collecting data on both the qualitative and the quantitative variables since its introduction by Warner (1965). For details one can refer to Diana and Perri (2011), Mahajan and Singh (2005), Guerriero and Sandri (2007). In stratified random sampling, the efficiency of the estimator of population parameters mainly depends upon choice of stratification variables and optimum strata boundaries. The pioneering work in this field was done by Dalenius (1950), Dalenius and Gurney (1951), Dalenius and Hodges (1959), Singh and Sukhatme (1969, 1973). Because of the formidable difficulties involved in determining exactly the optimum strata boundaries (OSB), several attempts were made to obtain approximate solutions to this problem. For reference, see Singh and Sukhatme (1969). Keeping in view the importance of stratification in randomized response technique (RRT), Mahajan *et al.* (1994) proposed the Cum. $\sqrt[3]{P_1(x)}$ rule for obtaining approximate

How to site this article: Ashu Chandel, P K Mahajan. Variance Expression for Eliminating Evasive Answer Bias in Optimum Stratification. *International Journal of Statistika and Mathemtika* Feb. to Apr. 2015; 13(3): 91-95. http://www.statperson.com (accessed 10 April 2015).

optimum strata boundaries (AOSB). This rule is applicable in a situation in which the samples from different strata are selected with simple random sampling with replacement (SRSWR) and the data are collected by Eichhorn and Hayre (1983) method, which involves the respondent multiplying his sensitive answer Y and a random scrambling variable S which is drawn from some pre-assigned distribution, giving the scrambled response Z = YS to the interviewer, who does not know the particular values of the random number S. The present paper proposes an expression for the limiting variance, which is particularly important as it gives an insight into the manner in which the variance of the estimator of mean for the sensitive study variable under optimum allocation changes with the increase in the number of strata. This variance expression in turn will establish the strata boundaries $[x_h]$ which are obtained approximately optimum when the samples from different strata are selected with SRSWR and the data on sensitive character are collected by scrambled response additive model Z = Y + S.

2. SCRAMBLED RESPONSE IN STRATIFIED RANDOM SAMPLING

Let the population under consideration be divided into L strata and a stratified simple random sample (SRS) of size n be drawn from it, the sample size in the h^{th} stratum being n_h so that $\sum_{i=1}^{n_h} n_h = n$. For h^{th} stratum, let Y denote the value of the sensitive character and let S_h be a scrambling random variable independent of Y_h and with finite mean and variance. The interviewer receives the scrambled answer $Z_h = Y_h + S_h$. The particular values of S_h are unknown to the interviewer, but its distribution is known. In this way, the respondent's privacy is not violated.

For theoretical development, let $E(S_h) = \theta_h$, $V(S_h) = \gamma_h$, $E(Y_h) = \mu_{hy}$ and $V(Y_h) = \sigma_{hy}^2$ where γ_h and θ_h are known to the interviewer but μ_{hy} and σ_{hy}^2 are unknown.

Since Y_h and $S_h$are independent, we have

$E(Z_h) = \mu_{hy} + \theta_h$ and $V(Z_h) = \sigma_{hy}^2 + \gamma_h$

If z_{hi} denote the value of the scrambled response for i^{th} unit of the sensitive variable in the h^{th} stratum, and sampling within each stratum is SRSWR, then unbiased estimator of μ_{hy} is

$\hat{\mu}_{hy} = \bar{z}_h - \theta_h$ where $\bar{z}_h = n_h^{-1} \sum_{i=1}^{n_h} z_{hi}$

It follows that an unbiased estimator for population mean is

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

$$V(\hat{\mu}_{St}) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \{\sigma_{hy}^2 + \gamma_h\} \quad (2.1)$$

Where W_h is the proportion of units in the h^{th} stratum and

If the cost of observing any unit in the population is assumed to be the same, the variance in (2.1) is minimised by adopting Neyman method of allocating the sample to different strata i.e. minimising the variance in (2.1) with respect to $n_h$subject to given total sample size $\sum_{i=1}^{n_h} n_h = n$, the variance in (2.1) reduces to

$$V(\hat{\mu}_{St})_N = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sqrt{\sigma_{hy}^2 + \gamma_h}\right)^2 \quad (2.2)$$

3. MINIMUM VARIANCE UNDER A MODEL

In this section, we shall consider the question of optimum allocation with constant cost of observing a unit in each stratum. Suppose we have a sensitive study variable y (e.g. income understated in income tax return) and non-sensitive stratification variable x (e.g. eye estimated value of the property) be related as

$$y = \eta(x) + e \quad (3.1)$$

where $\eta(x)$ is a real valued function of x and e is the error term such that $E(e|x) = 0$ and $V(e|x) = \phi(x) > 0 \; \forall \; x \in (a, b)$ such as $(b - a) < \infty$. If f(x) is the marginal density function of x then define

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

$\mu_{hy} = \mu_{h\eta} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \eta(x) f(x) dx$ and

$$\sigma_{hy}^2 = \sigma_{h\eta}^2 + \mu_{h\phi} \quad (3.2)$$

where (x_{h-1}, x_h) are respectively the lower and upper boundaries of the h[th] stratum, $\mu_{h\phi}$and $\sigma^2_{h\eta}$ are respectively the expected value of $\phi(x)$ and the variance of $\eta(x)$ in the hth stratum. From (2.2) and (3.2) the variance of the estimate $\hat{\mu}_{St}$becomes

$$V(\hat{\mu}_{St})_N = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sqrt{\sigma^2_{h\eta} + \mu_{h\phi} + \gamma_h}\right)^2 \tag{3.3}$$

The variance expression (3.3) is clearly a function of the strata boundaries [x_h]. The variance can, therefore, be further reduced by using the optimum strata boundaries which corresponds to the minimum of $V(\hat{\mu}_{St})_N$ in (3.3).

4. LIMITING EXPRESSION FOR $V(\hat{\mu}_{St})_N$

The expression for the variance $V(\hat{\mu}_{St})_N$ that we shall obtain in this section is particularly important in approximately optimum stratification on the auxiliary variable. This expression gives an insight into the manner in which the variance of the estimator $(\hat{\mu}_{St})_N$ under optimum allocation is reduced with the increase in the number of the strata. For this purpose we first prove the following lemma.

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

$$W_h\sqrt{\sigma^2_{h\eta} + \mu_{h\phi} + \gamma_h} - \int_{x_{h-1}}^{x_h} \sqrt{\theta^*(x)}.f(x)dx = \frac{1}{96}\left[\int_{x_{h-1}}^{x_h} \sqrt[3]{G_1(x)f(x)}\,dx\right]^3 \left(1 + 0\left(k_h^2\right)\right) \tag{4.1}$$

where $G_1(x) = \frac{\theta_1^2(x) + 4\theta^*(x)\eta'^2(x)}{\left(\sqrt{\theta^*(x)}\right)^3}$, $\theta_1 = \frac{\partial}{\partial x_h}\phi(x)$ and $\theta^* = \phi + \gamma_h$

Proof: Assuming the existence of the various functions and their derivatives occurring in (4.1) for all x in open interval (a, b), Singh and Sukhatme (1969) have given the following series expansions for W_h, μ_{hc} and σ^2_{hc} as

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

$$\mu_{h\eta} = \eta\left[1 - \frac{\eta'}{2\eta}k_h + \frac{f'\eta' + 2f\eta''}{12f\eta}k_h^2 - \frac{ff''\eta' + ff'\eta'' + f^2\eta''' - f'^2c'}{24f\eta^2}k_h^3 + 0\left(k_h^4\right)\right]$$

and

$$\sigma^2_{h\eta} = \frac{\eta'^2 k_h^2}{12}\left[1 - \frac{\eta''}{\eta'}k_h + 0(k_h^2)\right] \tag{4.2}$$

Various functions and their derivatives in (4.2) are evaluated at the upper boundary x_h of the h[th] stratum. Using the relations given in (4.2), the first term on L.H.S of (4.1) after simplification can be put as can be put as

$$W_h\sqrt{\sigma^2_{h\eta} + \mu_{h\phi} + \gamma_h} = f\,K_h\sqrt{\theta^*}\left[1 - A_1K_h + A_2K_h^2 - A_3K_h^3 + 0\left(K_h^4\right)\right] \tag{4.3}$$

where $A_1 = \frac{1}{4f\theta^*}\left[f\theta_1 + 2f'\theta^*\right]$

$$A_2 = \frac{1}{96f\theta^{*2}}\left[4f'\theta^*\eta'^2 + 4f'\theta_1\theta^* + 8f\theta_2\theta^* - -3f\theta_1^2 + 12f'\theta_1\theta^* + 16f''\theta^{*2}\right]$$

and

$$A_3 = \frac{1}{384f^2\theta^{*3}}\left[8f^2\theta_3\theta^{*2} + 16f^2\theta^*\eta'\eta'' + 24ff''\theta_1\theta^{*2} + 24ff'\theta_2\phi^{*2} + 3f^2\theta_1^3 - 4f^2\theta_1\theta^*\eta'^2 - 10ff'\theta_1^2\theta^* - 8f^2\theta_1\theta_2\theta^* + 8ff'\theta^{*2}\eta'^2 + 16ff'''\theta^{*3}\right]$$

Similarly we have on using the Taylor's theorem

$$\int_{x_{h-1}}^{x_h} \sqrt{\theta^*(x)}.f(x)dx = K_hF\left[1 - \frac{F'}{2F}K_h + \frac{F''}{6F}K_h^2 - \frac{F'''}{24F}K_h^3 + 0\left(K_h^4\right)\right] \tag{4.4}$$

Where $F(x) = \sqrt{\theta^*(x)}\,f(x)$ and its derivatives are evaluated at $x = x_h$.

On subtracting (4.4) from (4.3) after substituting for the derivativesF',F'' and F''' in terms ot the functions f, θ, η and their derivatives at x = x_h, the L.H.S. of (4.1) becomes on simplification as

$$\frac{K_h}{96}\left[B_2K_h^2 - B_3K_h^3 + 0\left(K_h^4\right)\right] \tag{4.5}$$

Where

$B_2 = \frac{4f\theta^*\eta'^2+f\theta_1^2}{\theta^{*3/2}} = (f.G_1)_{x_h}$ and

$B_3 = \frac{16f\eta'\eta''\theta^{*2} - 3f\theta_1^3 - 4f\eta'^2\theta_1\theta^* + 2f'\theta_1^2\theta^* + 4f\theta_1\theta_2\theta^* + 8f'\theta^{*2}\eta'^2}{4\theta^{*5/2}} = \frac{1}{2}\frac{\partial}{\partial x_h}(f.G_1)_{x_h}$

Thus (4.5) can be rewritten as

$\frac{K_h}{96}[G_1(x)f(x)K_h^2 - \frac{1}{2}\frac{\partial}{\partial x_h}(G_1(x)f(x))\,K_h^3 + 0\left(K_h^4\right)]$

$= \frac{K_h^2}{96}\left[G_1(x)f(x)K_h - \frac{1}{2}\frac{\partial}{\partial x_h}(G_1(x)f(x))\,K_h^2 + 0\left(K_h^3\right)\right]$

It follows that if we have large number of strata so that strata widths K_h are small and their higher powers in the expansions can be neglected. Proceeding on the lines of Singh and Sukhatme (1969), we have

$W_h\sqrt{\sigma_{h\eta}^2 + \mu_{h\phi} + \gamma_h} - \int_{x_{h-1}}^{x_h}\sqrt{\theta^*(x)}.f(x)dx$

$= \frac{K_h^2}{96}\int_{x_{h-1}}^{x_h} G_1(x)f(x)\,dx\,[1 + 0\left(K_h^2\right)]$

$= \frac{1}{96}\left[\int_{x_{h-1}}^{x_h}\sqrt[3]{G_1(x)f(x)}\,dx\right]^3\left(1 + 0\left(k_h^2\right)\right)$

This completes the proof of the lemma.

From (3.3) and (4.1) we therefore, get

$V(\hat{\mu}_{St})_N = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\sqrt{\sigma_{h\eta}^2 + \mu_{h\phi} + \gamma_h}\right)^2$

$= \frac{1}{n}\left[\sum_{h=1}^{L}\left\{\int_{x_{h-1}}^{x_h}\sqrt{\theta^*(x)}\,f(x)dx + \frac{1}{96}\left(\int_{x_{h-1}}^{x_h}\sqrt[3]{G_1(x)\,f(x)}dx\right)^3\right\}\right]^2$

$= \frac{1}{n}\left[\int_a^b\sqrt{\theta^*(x)}\,f(x)dx + \frac{1}{96L^2}\left(\int_a^b\sqrt[3]{G_1(x)\,f(x)}dx\right)^3\right]^2$

$= \frac{1}{n}\left(\alpha + \frac{\beta}{L^2}\right)^2$ (4.6)

Where $\alpha = \int_a^b\sqrt{\theta^*(x)}f(x)dx$ and $\beta = \frac{1}{96}\left(\int_a^b\sqrt[3]{G_1(x)\,f(x)}dx\right)^3$

It can be easily seen that in obtaining the expression (4.6) for the variance of the estimator $\hat{\mu}_{St}$ under Neyman allocation, the terms of order $O(m^4)$ have been neglected. Thus the limiting expression of the variance becomes $\lim_{L\to\infty} V(\hat{\mu}_{St})_N = \lim_{L\to\infty}\left(\alpha + \frac{\beta}{L^2}\right)^2 = \frac{\alpha^2}{n}$. This relation gives the exact manner in which the variance $V(\hat{\mu}_{St})_N$ will approach $\frac{\alpha^2}{n}$ as the value of number of strata (L) is increased. It can be easily seen by proceeding on the lines of Mahajan *et.al* (1994) that boundaries $[x_h]$ satisfying $\int_{x_{h-1}}^{x_h}\sqrt[3]{G_1(x)f(x)}\,dx$ = constant = $\frac{1}{L}\int_a^b\sqrt[3]{G_1(x)f(x)}dx$ are approximately optimum.

REFERENCES

1. Chaudhuri A, Roy D (1997): Model assisted survey sampling strategies with randomized response. *J Stat Plan Inference* 60, 61-68.
2. Dalenius, T. (1950): The problem of optimum stratification. *Skand Akt*. 33, 203-213.
3. Dalenius, T. and Gurney, M. (1951): The problem of optimum stratification. H. Skandinavisk Aktuarietidskrift, 34, 133-148.
4. Dalenius, T. and Hodges, J.L. (1959): Minimum variance stratification. *J. Amer. Statist. Assoc.,* 54, 88-101.
5. Diana, G. and Perri, P. F. (2011): A class of estimators for quantitative sensitive data. *Stat Papers* 52, 633-650.
6. Eichhorn, B. H. and Hayre, L. S. (1983): Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Planning and Infer*. 7, 307-316.

7. Mahajan, P. K. (2006): Optimum stratification for scrambled response with ratio and regression methods of estimation, *Model Assist Stat Appl.* 1: 17-22.
8. Guerriero, M. and Sandri, M. F. (2007): A note on the comparison of some randomized response procedure. *J. Statist. Planning and Infer. 13*7, 2184-2190.
9. Mahajan, P. K. and Singh, R. (2005): Optimum stratification for scrambled response in pps sampling, *Metron LXIII*: 103-114.
10. Mahajan, P. K., Gupta, J. P. and Singh, R. (1994): Determination of optimum strata boundaries for scrambled randomized response, *Statistica, 54:* 375-381.
11. Singh, R. and Sukhatme, B. V. (1969): Optimum stratification. *Ann. Inst. Statist. Math.* 21, 515-528.
12. Singh, R. and Sukhatme, B. V. (1973): Optimum stratification with ratio and regression methods of estimation. *Ann. Inst. Statist. Math.* 25, 627-633.
13. Singh, S., Joarder A. H and Kinh, M. L. (1996): Regression analysis using scrambled response. *Aust. N Z J Stat.38*, 201-211.
14. Singh, S. and Tracy, D.S. (1999): Ridge regression analysis using scrambled responses. *Metron* LVII: 147-157.
15. Warner, S. L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* 60, 63-69.

Source of Support: None Declared
Conflict of Interest: None Declared