Engineering Sciences International Research Journal
ISSN (Print) : 2330 – 4338 ; Volume 1 Issue 1 ‖ February. 2013 ‖

Limiting Expression of Variance for Eliminating Evasive Answer Bias in Quantitative Sensitive Data

P. K. Mahajan[1]

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. Auxiliary information for quantitative sensitive data has received attention mainly reserved to regression model theory and stratification (see Singh et al. 1996; Chaudhuri and Roy 1997; Singh and Tracy 1999; Mahajan et al. 1994; Mahajan and Singh 2005; Mahajan 2005-2006). The optimum stratification with ratio method of estimation in randomized response methods for quantitative sensitive data has received less attention so far. Keeping in view its importance in socio-economic and industrial surveys, 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, regularity conditions, scrambled response.

1. INTRODUCTION

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 (1951), Singh and Sukhatme (1969, 1973) and several other workers considered the problem of optimum stratification by using an auxiliary variable closely related to study variable. Surveys on sensitive subjects such as tax evasion, illegal income, gambling, alcoholism, sexual abuse, abortions and many other, are generally affected by two serious problems: respondents may refuse to answer or deliberately give untruthful or misleading answers. This causes substantial bias in the estimation of population parameters. The randomized response technique (RRT), an ingenious interviewing procedures to reduce or eliminate evasive answer bias for eliciting information on sensitive character has attracted a lot of attention after the pioneering work of Warner (1965). For reference, see Diana and Perri (2011); Mahajan et al (2006). Eichhorn and Hayre (1983) introduced the scrambled randomized response method which did not contain the difficulties of the Greenberg *et al* (1971) unrelated question method. The scrambled randomized response 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 value of the random variable S. the theory related to the determination of optimum strata boundaries (OSB) for quantitative sensitive character has been discussed in Mahajan *et al.* (1994). In socio-economic survey and biometric research, we come across situations in which we have to use the aggregated data on auxiliary variable at the time of estimation of the sensitive characters, the ratio method of estimation has its own importance. When the knowledge of population mean $\overline{X}$ is known in advance. The optimum stratification with ratio method of estimation in RRT has received, to our knowledge, less attention so for keeping in view the importance of the aforementioned issues, we intend in this paper to obtain an expression for the limiting variance when the number of strata are large. This expression in particularly important as it 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. This expression in turn will establish the boundaries $[x_h]$ are approximately optimum.

2. SCRAMBLED RESPONSE IN STRATIFIED RANDOM SAMPLING

Let the population under consideration be divided into L strata and a stratified SRS of size n be drawn from it, the sample size in the h^th^ stratum being n_h so that $\sum_{h=1}^{L} n_h = n$. for h^th^ stratum, let Y_h 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 respondent generates S_h using some specified method and multiplies the variable value Y by S_h. the interviewer thus receives the scrambled answer $Z_h = YS_h$. 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.

If z_{hi} and x_{hi} respectively denote the value of the scrambled variable z and auxiliary variable x, for i^th^ unit of the sample in the h^th^ stratum, then define

$\bar{z}_h = n_h^{-1} \sum_{i=1}^{n_h} z_{hi}$, $\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi}$,

$s_{hxz}=(n_h - 1)^{-1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)(x_{hi} - \bar{x}_h)$

and $C_h^2 = \frac{\gamma_h^2}{\theta_h}$

Since sampling within each stratum is SRS and with replacement, therefore, an unbiased estimator for μ_{hy} is

$$\hat{\mu}_{hy} = \frac{\bar{z}_h}{\theta_h}$$

In practice, the ratio of population means of y and x differ from stratum to stratum and therefore, the use of separate ratio estimator is justified. The separate ratio estimator to estimate the population mean μ, of sensitive variable y assuming the knowledge of h^th^ stratum mean $\bar{x}_h$ for auxiliary variable x is defined as

$$\hat{\mu}_{st.R} = \sum_{h=1}^{L} W_h \hat{\mu}_{hRy}$$

where W_h is the proportion of units in the h^th^ stratum and

$$\hat{\mu}_{hRy} = \frac{\hat{\mu}_{hy}}{\bar{x}_h} \bar{X}_h \text{ where } \hat{\mu}_{hy} = \frac{\bar{z}_h}{\theta_h}$$

We now state the following Theorems 1 and 2, the proofs for which are straightforward (keeping in view that the scrambling variable is independent of both x and y).

Theorem 1. The variance of the estimator $\hat{\mu}_{St.R}$, upto the terms of order $0(n^{-1})$, is obtained as

$$V(\hat{\mu}_{St.R})=\sum_{h=1}^{L} W_h^2 n_h^{-1} \{\sigma_{hy}^2 (1 + C_h^2) + \mu_{hy}^2 C_h^2 + R_h^2 \sigma_{hx}^2 - 2R_h \sigma_{hxy}\} \tag{2.1}$$

where R_h is the stratum mean ratio.

Theorem 2. An unbiased estimator of variance $V(\hat{\mu}_{St.R})$ is given by

$$v(\hat{\mu}_{St.R}) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \{s_{hy}^2 (1 + C_h^2) + \mu_{hy}^2 C_h^2 + R_h^2 s_{hx}^2 - 2R_h s_{hxy}\}$$

where

s_{hy}^2 and s_{hxy} are as obtained above and

$$\hat{\mu}_{hy}^2 = \frac{(n_h^{-1} s_{hz}^2 - \bar{z}_h^2)}{\theta_h^2}$$

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

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

where $c(x) = R_h x$ 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

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

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

$$\sigma_{hy}^2 = \sigma_{hc}^2 + \mu_{h\phi}$$

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

The model in Equation (3.1) is appropriate for separate ratio estimator while for combined ratio estimator, we shall have $R_1 = R_2 = - - - = R_L = R$ and as such it is a particular case of separate ratio estimator.

Under model Equation (3.1), we have

$$\sigma_{hy}^2 = \sigma_{hc}^2 + \mu_{h\phi} = R_h^2 \sigma_{hx}^2 + \mu_{h\phi}$$

$$\sigma_{hxy} = \sigma_{hxc} = R_h^2 \sigma_{hx}^2 \text{ and } \mu_{hy} = \mu_{hc} = R_h \mu_{hx}$$

Thus variance expression given in (2.1) reduces to

$$V(\hat{\mu}_{St..R}) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \{\mu_{h\phi} + C_h^2 (\mu_{h\phi} + R_h \sigma_{hx}^2 - R_h^2 \mu_{hx}^2\} \tag{3.2}$$

Minimizing the variance in (3.2) with respect to $n_h$subject to given total sample size $\sum_{h=1}^{L} n_h$, the expression for minimum variance ($V(\hat{\mu}_{St.R})_{opt}$) reduces to

$$\frac{1}{n}\left(\sum_{h=1}^{L} W_h \sqrt{\mu_{h\phi} + C_h^2 (\mu_{h\phi} + \sigma_{hc}^2 + \mu_{hc}^2}\right)^2 \tag{3.3}$$

where $c(x) = R_h x$

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

4. 4. AN EXPRESSION FOR $V(\hat{\mu}_{St.R})$

The expression for the variance $V(\hat{\mu}_{St.R})$ that we shall obtain in this section is particularly important in obtaining approximately optimum stratification on the auxiliary variable. This expression gives an insight into the manner in which the variance of the estimator $(\hat{\mu}_{St.R})$ 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

$$W_h\sqrt{\mu_{h\phi} + C_h^2\left(\mu_{h\phi} + \sigma_{hc}^2 + \mu_{hc}^2\right)} - \int_{x_{h-1}}^{x_h} \sqrt{\phi * (x)}.f(x)dx$$

$$= \frac{1}{96}\left[\int_{x_{h-1}}^{x_h} \sqrt[3]{P_2(x)dx}\right]^3 \left(1 + 0(k_h^2)\right) \qquad (4.1)$$

where $P_2(x) = \frac{\phi_1^2(x)}{\left(\sqrt{\phi*(x)}\right)^3}.f(x)$ and $c(x) = R_h x$

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_{hc}^2 as

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

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

and

$$\sigma_{hc}^2 = \frac{c_{kh^2}'^2}{12}\left[1 - \frac{c''}{c'}k_h + 0(k_h^2)\right] \qquad (4.2)$$

Various functions and their derivatives in (4.2) are evaluated at the upper boundary x_h of the h^th^ stratum can be put as

$$W_h\sqrt{\mu_{h\phi} + c_h^2\left(\mu_{h\phi} + \sigma_{hc}^2 + \mu_{hc}^2\right)}$$

$$=K_h f\sqrt{\phi *}\,[1 - A_1K_h + A_2K_h^2 - A_3K_h^3 + 0(K_h^4)] \qquad (4.3)$$

where

$$A_1 = \frac{[f\phi_1 + 2f'\phi*]}{4f\phi*}$$

$$A_2 = \frac{1}{96\phi^{*2}}\left[4f'^{\phi_1\phi^*} + 8f\phi_2\phi^* + 4f\phi^*c'^2 + 16f\phi^*c'^2c_h^2 - 3f\phi_1^2 + 12f'^{\phi_1\phi^*} + 16f''\phi^{*2}\right]$$

and

$$A_3 = \frac{1}{384f^2\phi^{*3}}\left[24ff''\phi_1\phi^{*2} + 24ff'\phi_2\phi^{*2} + 48ff'\phi^{*2}c'^2C_h^2 - 10ff'\phi_1^2\phi^* + 8ff'\phi^{*2}c'^2 + 8f^2\phi_3\phi^{*2} + 16f^2\phi^{*2}c'c'' + 48f^2\phi^{*2}c'c''C_h^2 - 8f^2\phi_1\phi_2\phi^* - 4f^2\phi_1\phi^*c'^2 - 16f^2\phi_1\phi^*c'^2C_h^2 + 3f^2\phi_1^3 + 16ff'''^{\phi^{*3}}\right]$$

where $\phi_2 = \phi'' + \phi''C_h^2 + 2cc''C_h^2$ and $\phi_3 = \phi''' + \phi'''C_h^2 + 2cc''C_h^2$

Similarly we have on using the Taylor's theorem

$$\int_{x_{h-1}}^{x_h} \sqrt{\phi^*(x)}\,f(x)dx = k_h f\sqrt{\phi^*}\,[\,1 - (f\phi_1 + 2f'\phi^*)/4f\phi^*\,K_h + (2f\phi^*\phi_1' + 4f'\phi^*\phi_1 + 4f''\phi^{*2} - f\phi_1^2)/24f\phi^{*2}\,K_h^{\,2} - (4f\phi^{*2}\phi_1'' + 12f'\phi^{*2}\phi_1' - 6f'\phi^*\phi_1^2 + 12f''\phi^{*2}\phi_1 + 8f'''\phi^{*3} - 6f\phi^*\phi_1\phi_1' + 3f\phi_1^3)\,/192f\phi^{*3}\,K_h^{\,3}$$

By subtraction, we have

$$W_h\sqrt{\mu_{h\phi} + C_h^2\left(\mu_{h\phi} + \sigma_{hc}^2 + \mu_{hc}^2\right)}$$

$$- \quad \int_{x_{h-1}}^{x_h} \sqrt{\phi^*(x)}\,f(x)dx$$

$$= \frac{k_h}{96}\left[\frac{f\phi_1^2}{\phi^{*3/2}}k_h^2\,\frac{2f'\phi^*\phi_1^2 + 4f\phi^*\phi_1\phi_1' - 3f\phi_1^3}{4\phi^{*5/2}}k_h^3 + 0(k_h^4)\right]$$

$$= \frac{k_h}{96}\left[P_2(x)k_h^2 - \frac{1}{2}P_2'(x)k_h^3 + 0(k_h^4)\right]$$

$$= \frac{k_h^2}{96}\int_{x_{h-1}}^{x_h} P_2(x)dx[1 + 0(k_h^2)]$$

$$= \frac{1}{96}\left[\int_{x_{h-1}}^{x_h} \sqrt[3]{P_2(x)dx}\right]^3 \left(1 + 0(k_h^2)\right) \qquad (4.4)$$

This completes the proof of the lemma.

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

$V(\hat{\mu}_{St.R})$

$$=\frac{1}{n}\left[\sum_{h-1}^{L}\left(\left\{\int_{x_{h-1}}^{x_h}\sqrt{\emptyset^*(t)}f(t)dt+\right.\right.\right.$$
$$196xh-1xh3P2tdt32$$

$$=\frac{1}{n}\left[\int_a^b\sqrt{\emptyset^*(t)}f(t)dt+\frac{1}{96L^2}\left(\int_a^b\sqrt[3]{P_2}(t)dt\right)^3\right]^2 \qquad (4.6)$$

$$=\frac{1}{n}\left(A+\frac{B}{L^2}\right)^2$$

where $A=\int_a^b\sqrt{\emptyset^*(t)}f(t)dt$ and

$$B=\frac{1}{96}\left(\int_a^b\sqrt[3]{P_2}(t)dt\right)^3$$

In obtaining these expressions for the variance $V(\hat{\mu}_{St.R})$ the terms of order $0(m^4)$ have been neglected. Also,

$$\lim_{L\to\infty}V(\hat{\mu}_{St.R})=\lim_{L\to\infty}\frac{1}{n}\left(A+\frac{B}{L^2}\right)^2=\frac{A^2}{n}$$

It can be easily seen by proceeding on the lines of Mahajan (2006) that boundaries $[x_h]$ satisfying (4.6) are approximately optimum.

5. 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] Greenberg, B. G. Kuebler, R. R. Abernathy, J. R. and Horvitz, D. G. (1971): Application of randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.* 66, 243-250.

[8] Mahajan, P. K. (2005-2006): Optimum stratification for scrambled response with ratio and regression methods of estimation, *Model Assist Stat Appl.* 1: 17-22.

[9] Mahajan, P. K. (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, R. and Sukhatme, B. V. (1973): Optimum stratification with ratio and regression methods of estimation. *Ann. Inst. Statist. Math.* 25, 627-633.

[14] Singh, S., Joarder A. H and Kinh, M. L. (1996): Regression analysis using scrambled response. *Aust. N Z J Stat.38*, 201-211.

[15] Singh, S. and Tracy, D.S. (1999): Ridge regression analysis using scrambled responses. *Metron* LVII: 147-157.

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

* * *

[1]Dr. Y. S. Parmar University of Horticulture & Forestry, Solan 173 230, India
pawan_uhf@yahoo.com