An Estimation of Sensitive Proportion Utilizing Higher Order Moments of Auxiliary Variable

Zawar Hussain[*] and Javid Shabbir[*]
*Department of Statistics, Quaid-i-Azam University, Pakistan.
Emails: zhlangah@yahoo.com, jsqau@yahoo.com

Abstract

To estimate the population proportion of a stigmatizing attribute use of auxiliary information in randomized response studies is rarely seen except a few studies including Zaizai (2006) and Diana and Perri (2009, 2010). Further, the use of higher order moments of auxiliary variable has not been made in the estimation of proportion of a stigmatizing attribute. To the best of our information there exists only one study by Singh and Chen (2009) concerning the estimation of proportion of a sensitive attribute through the use of higher order moments of scrambling variable but not the auxiliary (supplementary) information. With this thought of using auxiliary information with higher order moment a general class of estimators of population proportion has been anticipated which works outstandingly well than the most recent class recommended by Diana and Perri (2009).

Keywords: Estimation of proportion, higher order moments, auxiliary information, dichotomous population and sensitive attribute.

1. Introduction

In medical, sociological, psychological or human behavior studies surveys are now becoming popular to have timely and reliable estimates of proportion of individuals in the population possessing a sensitive attribute. In social, medical or economic surveys on sensitive traits obtaining reliable data has emerged as a challenging issue especially in socio-economic and behavioral studies since making reliable and valid inferences mainly depends upon the reliability of the data. Warner (1965), for the first time studied this issue and proposed an ingenious method, called randomized response procedure, to procure honest data for estimating proportion of a sensitive attribute. The initiative of randomizing the response was further enhanced by many researchers to the estimation of mean/proportion of a sensitive quantitative/qualitative variable. To date there is a wide range of randomized response models to estimate the mean / proportion of a sensitive variable / attribute. Indeed, the use of Randomized Response (RR) models is made as a device to cut the ambiguous answer bias and, of course, to impart privacy protection to the respondents in order to get them ready to reveal their response truthfully.

Some important latest randomized response models to estimate the proportion of a sensitive variable are Greenberg et al. (1969), Kuk (1990), Mangat and Singh (1990), Mangat (1994), Mangat et al. (1995), Tracy and Mangat (1996), Mahmood et al. (1998), Bhargava and Singh (2000), Singh et al. (2000), Singh and Mathur (2002), Kim and Warde (2004), Chaudhuri (2004), Huang (2005), Shabbir and Gupta (2005), Zaizai (2006), Diana and Perri (2009,2010) and many others. For a more precise and useful understanding of RR models the interested readers may be referred to Fox and Tracy (1986), Chaudhuri and Mukerjee (1998) and Hedayat and Sinha (1991) among many others.

It has been discussed by several authors that the auxiliary information can be used to acquire estimators with improved precision (see Cochran (1977) and Singh (2003)). The auxiliary information can be used in two ways. As a first use it is brought into study at design stage of a survey and secondly it may be explored at the estimation step of a survey. As far as the Randomized Response Models (RRMs) are concerned, there exists a small amount of articles utilizing the auxiliary information at estimation stage. For instance, Zaizai (2006), Van den Hout et al. (2007), Diana and Perri (2009) and Diana and Perri (2010). Also the studies by Chaudhuri and Mukerjee (1988), Allen and Singh (2001) and Grewal et al. (2005-2006) are some of studies which are based on utilizing auxiliary information at the design stage. In spite of using first order moments of auxiliary information in randomized response study, to the best of our knowledge, however, the idea of making use of second or higher order moments of the auxiliary (supplementary) information is to be taken in consideration. Though a broad review of articles on randomized response techniques we could find only one article by Singh and Chen (2009) which is based on making use of second order moments of scrambling variable in estimation of proportion of a sensitive attribute.

Further, again, to our knowledge Diana and Perri (2009) proposed class of estimators remains the best one among the class of estimators utilizing auxiliary information. In fact, Diana and Perri (2009) proposed a regression estimator in order to improve Zaizai (**2006**) study which suggests the use of a ratio estimator. Diana and Perri (2009) concluded that with single auxiliary variable their estimator is best in the class so further proliferation is not needed. However they indicated that more efficient estimator might be worked out using more than one auxiliary variable and/or using the information on the variability or shape of the auxiliary variables. Motivated by Singh and Chen (2009) and borrowing the idea of higher order moments of auxiliary variable from Diana and Perri (2009) we plan to study a general class of unbiased estimators in terms of its precision. Prior to formally exploring this idea, it is reasonable to discuss the Diana and Perri (2009) class of estimators in the Section to follow. Then we will present proposed class of estimators in the Section 3. In Section 4, we will showcase the comparisons and conclusions of our study.

2 Diana and Perri Estimator

Diana and Perri (2010) proposed a very motivating and efficient class of estimators making use of the auxiliary information. Their work is actually the extension of the proposal by Zaizai (2006). The Class of estimators proposed by Diana and Perri (2009) may be briefly outlined as given below.

Let $I=\{i_1,i_2,...,i_N\}$ be a finite population of individuals who can either be classified into a sensitive group of individuals possessing a sensitive attribute S or to its complementary group of individuals not possessing the attribute S. Let Y be a binary variable assuming value 0 if a particular individual does not possesses the sensitive attribute S and the value 1, otherwise. The problem of interest is to estimate the population proportion $\pi = N^{-1}\sum_{j=1}^{N} Y_j$. Let X be a non-sensitive auxiliary variable (qualitative or quantitative) with known mean μ_X and variance σ_X^2 respectively. To estimate the proportion (π) a sample of size n is drawn from the population using simple random sampling with replacement and each selected individual is provided a randomization device through which a randomized answer on sensitive attribute S and a true value on the auxiliary variable X are obtained. Any randomization device with probability of a *yes* answer given by

$$\theta = g\pi + f \ , \qquad (1)$$

may be used for this purpose. Here f and g are the known constants. For instance, if Warner (1965) randomization device consisting of the two statements: (i) "I possess the sensitive attribute S" and (ii) "I do not possess the sensitive attribute S", with probabilities P and $(1-P)$, is used then the probability of a *yes* answer is given by $\theta=(2P-1)\pi+(1-P)$ with $g=(2P-1)$ and $f=(1-P)$. This is the randomization device actually used by Zaizai (2006). For the values of the constants g and f under different randomization devices one can be referred to Diana and Perri (2009). Let Z_j be the binary response of the j^{th} respondent, obtained through a given randomization device, taking value yes (1) or no (0) with probability θ and $(1-\theta)$, respectively, and $(z_1,x_1),(z_2,x_2),...,(z_n,x_n)$ be the pairs of responses. Then Diana and Perri (2009) class of estimators is given by

$$\hat{\mu}_D = \frac{\overline{z}_d - f}{g}, g \neq 0, \qquad (2)$$

where $\overline{z}_d = \overline{z} + b(\mu_X - \overline{x})$, $\overline{x} = n^{-1}\sum_{j=1}^{n} x_j$, $\overline{z} = n^{-1}\sum_{j=1}^{n} z_j$ such that $E(\overline{z})=\theta$ and $E(\overline{x})=\mu_X$; g and f are constants depending the *RR* model used and b is linked to the efficient use of auxiliary variable. It is to be noted that if $b=0$ then there will be no use of auxiliary information.

The variance of the estimator in (2) is given by

$$Var(\hat{\mu}_D) = \frac{Var(\overline{z}_d)}{g^2} = \frac{1}{ng^2}\left(\sigma_z^2 - 2b\sigma_{XZ} + b^2\sigma_X^2\right). \qquad (3)$$

Diana and Perri (2009) reported that the $Var\left(\hat{\mu}_D\right)$ is minimum for $b=\frac{\sigma_{XZ}}{\sigma_X^2}$ with minimum variance given by

$$Var\left(\hat{\mu}_D\right)_{\min}=\frac{1}{ng^2}\sigma_Z^2\left(1-\rho_{ZX}^2\right). \tag{4}$$

An ordinary least squares estimate of b is suggested to be used as $\hat{b}=\frac{s_{ZX}}{s_X^2}$ when the parameter b is unknown. The result for the variance of $\hat{\mu}_D$ given in (4) will still be valid to the first order of approximation.

As mentioned earlier, this study is about the thought of making use of higher order moments of the auxiliary variable; we now present the planned class of estimators in the next section.

3. Proposed Class of estimators

Letting $\mu_2'=\mu_X^2+\sigma_X^2$ be the second order raw moment of the auxiliary variable X the new proposed class of estimators is given by

$$\hat{\mu}_N=\frac{\overline{z}_g-f}{g}, g\neq 0,$$

where $\overline{z}_g=\overline{z}+\lambda_1\left(\mu_X-\overline{x}\right)+\lambda_2\left(\mu_2'-m_2'\right)$, λ_1 and λ_2 are regression coefficients and $m_2'=n^{-1}\sum_{j=1}^{n}x_j^2$. Let $x_i^2=q_i$ then $m_2'=n^{-1}\sum_{j=1}^{n}q_j$. It is obvious that the regression coefficients λ_1 and λ_2 may be either known or unknown. We consider both the situations when λ_1 and λ_2 are known and unknown one by one.

(i) Case of known λ_1 and λ_2

Under the condition $f+g\pi=\theta$, the expectation and variance of $\hat{\mu}_N$ are given by

$$E\left(\hat{\mu}_N\right)=\frac{\theta-f}{g}=\pi\text{, and}$$

$$Var\left(\hat{\mu}_N\right)=\frac{1}{ng^2}\left(\sigma_Z^2+\lambda_1^2\sigma_X^2+\lambda_2^2\sigma_q^2+2\lambda_1\sigma_{ZX}+2\lambda_2\sigma_{ZQ}+2\lambda_1\lambda_2\sigma_{XQ}\right), \tag{5}$$

which attains its minimum when

$$\lambda_1=\frac{\sigma_Q^2\sigma_{ZX}^2-\sigma_{ZQ}^2\sigma_{XQ}^2}{\sigma_X^2\sigma_Q^2-\sigma_{XQ}^2}, \lambda_2=\frac{\sigma_X^2\sigma_{ZQ}^2-\sigma_{ZX}^2\sigma_{XQ}^2}{\sigma_X^2\sigma_Q^2-\sigma_{XQ}^2}. \tag{6}$$

The minimum variance of $\hat{\mu}_N$ now is given by

$$Var\left(\hat{\mu}_N\right)_{\min}=\frac{\sigma_Z^2}{ng^2}-\left(\frac{\sigma_Q^2\sigma_{XZ}^2+\sigma_X^2\sigma_{ZQ}^2-2\sigma_{ZX}\sigma_{ZQ}\sigma_{XQ}}{\sigma_Q^2\sigma_X^2-\sigma_{XQ}^2}\right). \tag{7}$$

To have a more simplified expression of $Var\left(\hat{\mu}_N\right)_{\min}$, we define

$$V_{200}=\frac{\sigma_z^2}{n},V_{020}=\frac{\sigma_x^2}{n},V_{002}=\frac{\sigma_q^2}{n},V_{110}=\frac{\sigma_{zx}}{n},V_{101}=\frac{\sigma_{zq}}{n},V_{011}=\frac{\sigma_{xq}}{n},$$

$$\phi_{abc}=\frac{V_{abc}}{V_{200}^{a/2}V_{020}^{(b+2c)/2}}\text{, and }\psi^2=\frac{\left(\phi_{101}-\rho_{zv}\phi_{011}\right)^2}{\phi_{002}-\phi_{011}^2}.$$

Then

$$\phi_{110}=\frac{V_{110}}{V_{200}^{1/2}V_{020}^{1/2}}=\frac{\sigma_{ZX}}{\sigma_Z\sigma_X}=\rho_{ZX}\text{ , }\phi_{101}=\frac{V_{101}}{V_{200}^{1/2}V_{020}}=\frac{\sigma_{ZQ}}{\sigma_Z\sigma_Q^2},\phi_{011}=\frac{V_{011}}{V_{200}^{0}V_{020}^{3/2}}=\frac{\sigma_{XQ}}{\sigma_X^3},$$

$$\phi_{002}=\frac{V_{002}}{V_{200}^{0}V_{002}^{2}}=\frac{\sigma_Q^2}{\sigma_X^4}\text{ and }\psi^2=\frac{\left(\rho_{ZQ}-\rho_{ZX}\rho_{XQ}\right)^2}{1-\rho_{XQ}^2}.$$

Now by substituting the above defined quantities in (7), we get

$$Var\left(\hat{\mu}_N\right)_{\min} = \frac{\sigma_Z^2}{ng^2}\left\{1-\left(\frac{\rho_{ZX}^2+\rho_{ZQ}^2-2\rho_{ZX}\rho_{ZQ}\rho_{XQ}}{1-\rho_{XQ}^2}\right)\right\}. \tag{8}$$

Now consider

$$\psi^2+\rho_{ZX}^2 = \frac{\left(\rho_{ZQ}-\rho_{ZX}\rho_{XQ}\right)^2}{1-\rho_{XQ}^2}+\rho_{ZX}^2 = \frac{\left(\rho_{ZQ}^2+\rho_{ZX}^2-2\rho_{ZQ}\rho_{ZX}\rho_{XQ}\right)}{1-\rho_{XQ}^2}. \tag{9}$$

Thus by (8) and (9) we get

$$Var\left(\hat{\mu}_N\right)_{\min} = \frac{\sigma_Z^2}{ng^2}\left\{1-\rho_{ZX}^2-\psi^2\right\}. \tag{10}$$

(ii) Case of unknown λ_1 and λ_2

As λ_1 and λ_2 are the regression coefficients of the variable X and Q in the regression of $Z=\lambda_0+\lambda_1 X+\lambda_2 Q+\varepsilon$, where λ_0 is constant and ε is the error term so the unbiased ordinary least squares estimators of λ_1 and λ_2 are given by $\hat{\lambda}_1=\frac{s_Q^2 s_{ZX}^2-s_{ZQ}^2 s_{XQ}^2}{s_X^2 s_Q^2-s_{XQ}^2}$ and $\hat{\lambda}_2=\frac{s_X^2 s_{ZQ}^2-s_{ZX}^2 s_{XQ}^2}{s_X^2 s_Q^2-s_{XQ}^2}$, which minimize the error sum of squares. Thus our class of estimators becomes

$$\tilde{\mu}_N = \frac{\hat{\bar{z}}_g-f}{g}, g\neq 0, \tag{11}$$

where $\hat{\bar{z}}_g=\bar{z}+\hat{\lambda}_1\left(\mu_X-\bar{x}\right)+\hat{\lambda}_2\left(\mu_2'-m_2'\right)$. Following the steps as in the case of known λ_1 and λ_2 it can be shown that to the first order of approximation through Taylor's series expansion the expectation and variance of $\tilde{\mu}_N$ are given by

$$E\left(\tilde{\mu}_N\right)=\frac{E\left(\hat{\bar{z}}_g\right)-f}{g}=\pi,$$

$$Var\left(\tilde{\mu}_N\right)\frac{\sigma_Z^2}{ng^2}\left\{1-\rho_{ZX}^2-\psi^2\right\}. \tag{12}$$

Now we give the comparisons of the proposed class of estimators with the Diana and Perri (2009) class of estimators. It is clear form (4), (10) and (12) that we do not need to consider both the cases of known and unknown regression coefficients separately since the comparison in one case remain valid in the other case. The proposed class of estimators will be more efficient than the Diana and Perri (2009) class if

$$Var\left(\hat{\mu}_D\right)_{\min}-Var\left(\tilde{\mu}_N\right)\geq 0$$

$$\frac{\sigma_Z^2}{ng^2}\left\{1-\rho_{ZX}^2\right\}-\frac{\sigma_Z^2}{ng^2}\left\{1-\rho_{ZX}^2-\psi^2\right\}\geq 0$$

$$\left\{1-\rho_{ZX}^2-1+\rho_{ZX}^2+\psi^2\right\}\geq 0$$

$$\psi^2\geq 0$$

which is always a non-negative quantity.

4. Conclusions

With the exception of the works by Chaudhuri and Mukerjee (1988), Allen and Singh (2001), Grewal et al. (2006), Zaizai (2006) and Diana and Perri (2009, 2010) it is hard to find any work improving the RRMs further in the situations where auxiliary information is utilized. However, in many social, medical, agricultural or economic studies some auxiliary variables may be found and measured with ease and no extra sampling expenditures.

For instance, in the study on taxable income of a large scale firm the number of workers and their salaries can be taken as auxiliary variables. To achieve the utmost gain from auxiliary variable it should be taken in a way that it is easy to collect, seemingly non-sensitive and correlated with the study variable, so that a rough guess by the interviewer himself may be applied as a test gauge on the collected answers.

With these settings a advantageous use of auxiliary variable, in terms of improved precision, can be made at the estimation stage of the study. Moving on this thought we proposed a general class of estimators utilizing the known higher order moments of the auxiliary variable. The best estimator in this class performs better than best estimator in Diana and Perri (2009) class. It is be noted that the proposed class of estimators has also an advantage of including the Diana and Perri (2009) class if $\lambda_2 = 0$. It may also be explored and proliferated further using some other function of auxiliary variables such as coefficients of variance, skewness and kurtosis or any other function of auxiliary variable for which an unbiased estimator is readily available.

References

1. Allen, J. and Singh, S. (2001). Response techniques to analyze various transformation and selection probabilities. *InterStat*. Available at http://interstat.statjournals.net/YEAR/2001/abstracts/0111002.php.
2. Bhargava, M. and Singh, R. (2000). A modified randomization device for Warner's model. Statistica, 60, 315-321.
3. Chaudhuri, A. (2004). Christofides' randomized response technique in complex survey. Metrika, 60, 223-228.
4. Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and Techniques. Marcel Dekker, Inc., New York.
5. Cochran, W. G. (1977). Sampling techniques. John Wiley & Sons, New York.
6. Diana, G. and Perri, P. F. (2009). Estimating a sensitive proportion through randomized response procedures based on auxiliary information. Statistical Papers (in press) DOI: 10.1007/s00362-009-0273-1.
7. Diana, G. and Perri, P. F. (2010). New Scrambled response models for estimating the mean of a sensitive character, Journal of Applied Statistics, 37(11), 1875-1890.
8. Fox, J. and Tracy, P. (1986). Randomized Response; A Method for sensitive Surveys. Sage, CA.
9. Greenberg, B. G., Abul-Ela Abdel-Latif, A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question RR model: Theoretical framework. Journal of the American Statistical Association, 64, 52-539.
10. Grewal, I. S., Bansal, M. L. and Sidhu, S. S. (2005-2006). Population mean corresponding to Horvitz-Thompson's estimator for multi-characteristics using randomized response technique. Model Assisted Statistics and Applications, 1, 215-220.
11. Huang, K. C. (2005). Estimation of sensitive data from a dichotomous population. Statistical Papers, 47, 149-156.
12. Hedayat, A. S. and Sinha, B. k. (1991). Design and Inference in finite population sampling. Wiley, New York.
13. Kim, J. M., Warde, D. W. (2004). A stratified Warner's randomized response model. Journal of Statistical Planning and Inference, 120/1-2 155-165.
14. Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. Biometrika, 77, 436-438.
15. Mahmood, M. Singh, S. and Horn, S. (1998). On the confidentiality guaranteed under randomized response sampling: a comparison with several new techniques. Biometrical Journal, 40, 237-242.
16. Mangat, N. S. (1994). An Improved randomized response strategy. Journal of Royal statistical Society, B. 56, 93-95.
17. Mangat, N. S., Singh, R. (1990). An alternative randomized response procedure. Biometrika, 77, 439-442.
18. Mangat, N. S., Singh, S. and Singh, R. (1995). On the use of a modified randomization device in Warner's model. Journal of Indian Society of Statistics and Operations Research, 16, 65-69.
19. Singh, H. and Mathur, N. (2002). On Mangat's improved randomized response strategy. Statistica LXII, 397-403.
20. Shabbir, J. and Gupta, S. (2005). On modified randomized device of Warner's model. Pakistan Journal of Statistics, 21, 123-129.
21. Singh, S. (2003).Advanced Sampling Theory with Application: Hoe Michael 'selected' Amy. Kluwer Academic Publishers, London.
22. Singh, S. and Chen, C. C. (2009). Utilization of higher order moments of scrambling variables in randomized response sampling. Journal of Statistical planning and Inference, 139, 3377-3380.
23. Singh, S., Singh, R. and Mangat, N. S. (2000). Some alternative strategies to Moor's model in randomized response sampling. Journal of Statistical Planning and Inference, 83, 243-255.
24. Tracy, D. and Mangat, N. (1996). Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical* Sciences, 4, 533-544.
25. Van den Hout, A., Van Heijden, P. G. M. and Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. Computational Statistics and data Analysis, 51, 6060-6069.
26. Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.
27. Zaizai, Y. (2006). Ratio method of estimation of population proportion using randomized device technique. *Model Assisted Statistics and Application*, 1, 125-130.