World Applied Sciences Journal 32 (4): 567-571, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.32.04.2048

Estimation of Sensitive Mean: Unrelated Variable Sum Technique

[1]Zawar Hussain and [2]Khushnoor Khan

[1]Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan
[2]Department of Statistics, King Abdulaziz University 80203, Jeddah 21589,
Kingdom of Saudi Arabia

Abstract: This article focuses on providing the correct expression for the variance of a recent estimator and suggesting its further improvement. An estimator of the population mean of a sensitive variable is suggested by extending a recently proposed additive scrambling technique. The proposed estimator is developed using simple random sampling with replacement and requires obtaining two responses from each respondent. It is actually a generalized additive model and provides privacy protection. The proposed estimator is relatively more efficient than some of the recent scrambling techniques.

Key words: Sensitive variable · Evasive answers · Estimation of mean · Additive scrambling models · Parsimonious models · Two responses

INTRODUCTION

Sensitive or incriminating behaviors are often encountered in social surveys. Before collecting the data on these sensitive behaviors, Social Desirability Bias (SDB) needs to be considered. SDB is a bias which creeps into the estimators due to the falsified reporting by the respondents. Respondents falsify their answers because they desire to show that they have the socially desirable behavior. To cope with the issue of SDB, several methods have been developed in literature. These techniques include the Item Count Technique (Droitcour *et al*. [1], Hussain *et al*. [2]), the Three card method (Droitcour, *et al*. [3], Droitcour and Larson [4]), the Nominative technique (Miller [5]), the Randomized Response Technique (Warner, [6]) and many others.

Sensitive behaviors are often dealt in social surveys. The Randomized Response Techniques are frequently applied in surveys about sensitive behaviors. The applications of RRT can be found in many studies (cf. Liu & Chow [7]), Reinmuth & Geurts [8], Geurts [9], Larkins *et al*. [10]). Although these techniques have been applied successfully in social surveys their applications may be found in other fields of research like business, marketing, education, psychology, criminology, medicine ans public health, etc.

The RRTs are based on the idea of scrambling the response and thereby providing complete privacy and anonymity to the respondents. The key feature of RRT is that reported scrambled responses cannot be traced back to the true response (of a given respondent) on sensitive variable. While applying RRT, respondents are asked to randomize or scramble their responses to a sensitive question. The responses are obtained on the premise of chance. Recent RRTs can be classified as qualitative and quantitative RRTs. In this study, we focus on the quantitative RRTs using additive scrambling.

Warner [11] introduced the idea of additive scrambling in quantitative RRTs. It was further studied by Himmelfarb and Edgell [11]. Many authors have advocated the use of additive scrambling due to ease in its practical applications and efficiency (cf. Gjestevang and Singh [13], Gupta *et al*. [14], Huang [15, 16], etc). In this study, we plan to enhance the Warner [11] additive scrambling RRT using the idea of obtaining two responses from each respondent. Similar idea has been used by Hussain and Khan [17] but the expression for the variance of their proposed estimator is incorrect. The reason for choosing the Warner [11] RRT is twofold. Firstly, we plan to give the correct expression for variance of Hussain and Khan [17] mean estimator. Secondly, we plan to further improve the Warner [11] and Hussain and

Corresponding Author: Zawar Hussain, Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan. Mob: +923335223544.

Khan [17] models. Another reason for choosing Warner [11] model lies in its easy application to real life problem.

Problem: Let we have a population $U=(u_1,u_2,...u_N)$ of N individuals and a simple random sample $S=(s_1,s_2,...s_n)$ of size n is drawn from U with replacement. Let $\underline{X}=(X_1,X_2,...X_N)$ is unknown set of values on the sensitive variable X with unknown mean $E(X)=\mu_X$ and variance $V(X)=\sigma_x^2$. The interest of study lies in the estimation of μ_X. Application of a scrambling/randomized response model inflates the variance of the estimator. A part of the variance is attributed to use of scrambling variable. Our problem, here, is to reduce the variance due to using scrambling variable.

MATERIALS AND METHODS

Let Y be the scrambling variable with known (or unknown) distribution $f(Y),\ -\infty<Y<\infty$, known mean $E(Y)=\mu_Y,(-\infty<\mu_Y<\infty)$ and known variance $\sigma_Y^2=V(Y)$. We will obtain two responses from each respondent. Additive and subtractive scrambling will be used to obtain the two set of responses. The respondents will be selected using simple random sampling with replacement. Performance of the proposed estimator will be measured through relative efficiency.

We now present a brief description of Warner [11] and Hussain and Khan [17] scrambling models.

Warner (1971) Additive Model: The Warner [11] additive scrambling model may be explained as follows. Assuming a simple random sampling with replacement a sample of size n is drawn. Each respondent in the sample is provided a randomization device which randomly produces values of scrambling variable having known distribution $f(Y)$ with known mean μ_Y and variance σ_Y^2. The i^{th} 2respondent is requested to generate a random value Y_i using the randomization device and add it to his/her true response X_i. Let Z_i be the reported response of the i^{th} respondent then it can be written as

$$Z_i = X_i + Y_i. \tag{1}$$

Warner [11] suggested an unbiased estimator of the mean μ_X as given by

$$\hat{\mu}_X = \bar{Z} - \mu_Y, \tag{2}$$

with variance given by

$$V(\hat{\mu}_X) = V(\bar{Z}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n} \tag{3}$$

The second term in the above equation represents the variance due to the use of scrambling variable.

Hussain and Khan (2013) Model: Motivated by Gupta and Shabbir [18], Christofides [19], Hussain *et al*. [20] and Lee *et al*. [21], Hussain and Khan [17] proposed the following model. The i^{th} respondent is asked to draw two different random numbers Y_1 and Y_2 from $f(Y)$ with $\mu_Y = 0$ and report the two responses $Z_1 = X + Y_1$ and $Z_2 = X - Y_2$. The expected responses from the i^{th} respondent may be written as $E(Z_1)=\mu_X+\mu_Y=\mu_X$ and $E(Z_2)=\mu_X-\mu_Y=\mu_X$. Hussain and Khan [17] proposed the following estimator

$$\hat{\mu}_{X(HK)} = \frac{\bar{Z}_1 + \bar{Z}_2}{2} \tag{4}$$

Hussain and Khan [17] derived the $V\left(\hat{\mu}_{X(HK)}\right)$, incorrectly, as $V\left(\hat{\mu}_{X(HK)}\right)=n^{-1}\sigma_X^2$, whereas the correct expression for $V\left(\hat{\mu}_{X(HK)}\right)$ is given by:

$$V\left(\hat{\mu}_{X(HK)}\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{2n}. \tag{5}$$

Hussain and Khan [17] mentioned that $\hat{\mu}_{X(HK)}$ is the minimum variance unbiased estimator (*MVUE*) but actually it is not a *MVUE*, though, relatively more efficient than Warner [11].

Proposed Model: The i^{th} respondent is asked to draw $2G$ (where G is a positive integer) random numbers $Y_1,Y_2,...,Y_{2G}$ from $f(Y)$ and report the two responses R_{1i} and R_{2i} as

$$\left.\begin{aligned} R_{1i} &= X_i + G^{-1}\sum_{g=1}^{G} Y_g = X_i + \bar{Y}_1 \\ R_{2i} &= X_i - G^{-1}\sum_{g=G+1}^{2G} Y_g = X_i - \bar{Y}_2 \end{aligned}\right] \tag{6}$$

The expected responses from the i^{th} respondent may be written as

$$\left.\begin{aligned} E(R_{1i}) &= E(X_i) + E(\bar{Y}) = \mu_X + \mu_Y \\ E(R_{2i}) &= E(X_i) - E(\bar{Y}_2) = \mu_X - \mu_Y \end{aligned}\right] \tag{7}$$

From (7), we define two unbiased moment estimators of μ_X as

$$\hat{\mu}_{1X} = \bar{R}_1 - \mu_Y \tag{8}$$

$$\hat{\mu}_{2X} = \bar{R}_2 + \mu_Y \tag{9}$$

Now, we find the variances of the estimators defined in (8) and (9).

By definition,

$$V(\hat{\mu}_{1X}) = V(\bar{R}_1) = \frac{1}{n}V(R_{1i}) \tag{10}$$

Now,

$$\begin{aligned} V(R_{1i}) &= E(R_{1i}^2) - (E(R_{1i}))^2 \\ V(R_{1i}) &= E(X_i + \bar{Y}_1)^2 - (\mu_X + \mu_Y)^2 \\ V(R_{1i}) &= \sigma_X^2 + \frac{\sigma_Y^2}{G} \end{aligned} \tag{11}$$

On substituting (11) in (10) we get variance of the estimator $\hat{\mu}_{1X}$ as given below

$$V(\hat{\mu}_{1X}) = \frac{1}{n}\left(\sigma_X^2 + \frac{\sigma_Y^2}{G}\right) \tag{12}$$

Similarly, the variance of the estimator $\hat{\mu}_{2X}$ can be calculated to be as given in expression (12). So, we can write that $V(\hat{\mu}_{1X}) = V(\hat{\mu}_{2X})$. Taking the advantage of equal variances and utilizing the full information we define a new estimator of μ_X as

$$\hat{\mu}_{3X} = \lambda\hat{\mu}_{1X} + (1-\lambda)\hat{\mu}_{2X}, \quad (0 < \lambda \le 1) \tag{13}$$

Its variance is given by

$$\begin{aligned} V(\hat{\mu}_{3X}) &= \lambda^2 V(\hat{\mu}_{1X}) + (1-\lambda)^2 V \\ &(\hat{\mu}_{2X}) + 2\lambda(1-\lambda)C(\hat{\mu}_{1X}, \hat{\mu}_{2X}). \end{aligned}$$

It is straight forward to verify that the optimum value of $\lambda = \frac{1}{2}$. Hence the optimum estimator is given by

$$\hat{\mu}_{3X} = \frac{\hat{\mu}_{1X} + \hat{\mu}_{2X}}{2} \tag{14}$$

with optimum variance given by

$$V(\hat{\mu}_{3X}) = \frac{1}{2n}\left(\sigma_X^2 + G\sigma_Y^2\right) + \frac{1}{2}C(\hat{\mu}_{1X}, \hat{\mu}_{2X}) \tag{15}$$

The covariance of $\hat{\mu}_{1X}$ and $\hat{\mu}_{2X}$ is calculated as

$$\begin{aligned} &C(\hat{\mu}_{1X}, \hat{\mu}_{2X}) = C(\bar{R}_1 - \mu_Y, \bar{R}_2 + \mu_Y) \\ &C(\hat{\mu}_{1X}, \hat{\mu}_{2X}) = C(\bar{R}_1, \bar{R}_2) = \\ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C(R_{1i}, R_{2j}) = \frac{1}{n^2}\sum_{i=1}^{n} C(R_{1i}, R_{2i}) \end{aligned} \tag{16}$$

since $C(R_{1i}, R_{2j}) = 0 \ \forall\ i \ne j$. Now, the covariance of Z_{1i} and Z_{2i} is given by

$$\begin{aligned} C(R_{1i}, R_{2i}) &= E(R_{1i}R_{2i}) - E(R_{1i})E(R_{2i}) \\ C(R_{1i}, R_{2i}) &= \sigma_X^2 \end{aligned} \tag{17}$$

Substituting (17) in (16), we get

$$C(\hat{\mu}_{1X}, \hat{\mu}_{2X}) = \frac{1}{n}\left(\sigma_X^2\right) \tag{18}$$

Now using (17) in (15), we get the variance of the weighted estimator $\hat{\mu}_{3X}$, given by

$$V(\hat{\mu}_{3X}) = \frac{1}{2n}\left(\sigma_X^2 + \frac{\sigma_Y^2}{G}\right) + \frac{1}{2n}\left(\sigma_X^2\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{2nG} \tag{19}$$

Relative Efficiency Comparison:

(I) $\hat{\mu}_{3X}$ versus $\hat{\mu}_X$

The proposed estimator $\hat{\mu}_{3X}$ will be relatively more efficient than the Warner [11] estimator $\hat{\mu}_X$ if and only if The relative efficiency (RE_1) of the proposed estimator relative to is given by

$$V(\hat{\mu}_X) - V(\hat{\mu}_{3X}) > 0 \cdot$$

Using (3) and (19) in the above inequality, we get

$$\begin{aligned} &n^{-1}\left(\sigma_X^2 + \sigma_Y^2\right) > n^{-1}\sigma_X^2 + (2nG)^{-1}\sigma_Y^2 \\ &or \\ &G > \frac{1}{2} \end{aligned}$$

which is always true.

(ii) $\hat{\mu}_{3X}$ versus $\hat{\mu}_{X(HK)}$

The proposed estimator $\hat{\mu}_{3X}$ will be relatively more efficient than Hussain and Khan [17] estimator $\hat{\mu}_{X(HK)}$ if and only if

$$V(\hat{\mu}_X) - V(\hat{\mu}_{3X}) > 0 \cdot$$

Again, using (5) and (19), we get

$$n^{-1}\left(\sigma_X^2 + \frac{\sigma_Y^2}{2}\right) > n^{-1}\sigma_X^2 + (2nG)^{-1}\sigma_Y^2$$

or

$$G > 1,$$

which is always true since in our proposed model we set $G > 1$.

Remark: It is interesting to note that for $G = 1$, the proposed scrambling model reduces to Hussain and Khan [17] model. Thus, the proposed scrambling model is a generalization of Hussain and Khan [17] model. Moreover, the proposed scrambling model is more protective compared to Hussain and Khan [17] model in that the respondents are asked to report scrambled response using average of the G random numbers. If the proposed scrambling model is applied such that the i^{th} respondent is asked to report his/her j^{th} (j = 1,2)response without disclosing their identity then it will help reducing the evasive answering.

Summary: With the motivation of improving the Hussain and Khan [17] scrambling model a generalized additive scrambling model has been suggested. Further, a correct expression for the variance of Hussain and Khan [17] estimator has been provided. It has been shown that the proposed generalized scrambling model is relatively more efficient than the Warner [11] and Hussain and Khan [17] models.

It is anticipated that the proposed scrambled model is easier to apply in the field surveys. Therefore, we suggest using the proposed scrambling model in collecting the data on sensitive variables.

REFERENCES

1. Droitcour, J.A., R.A. Caspar, M.L. Hubbard, T.L. Parsley, W. Visscher and T.M. Ezzati, 1991. The item count technique as a method of indirect questioning: A review of its development and a case study application, in P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz & S. Sudman, (eds), 'Measurement Errors in Serveys', Wiley, New York.
2. Hussain, Z., E.A. Shah and J. Shabbir, 2012. An alternative item count technique in sensitive surveys. Revista Colombiana de Estadistica, 35(1): 39-54.
3. Droitcour, J.A., E.M. Larson and F.J. Scheuren, 2001. The three card method: Estimating sensitive survey items with permanent anonymity of response, in 'Proceedings of the Social Statistics Section', American Statistical Association, Alexandria, Virginia.
4. Droitcour, J.A. and E.M. Larson, 2002. An innovative technique for asking sensitive questions: The three card method. Sociological Methodological Bulletin, 75: 5-23.
5. Miller, J.D., 1985. The nominative technique: A new method of estimating heroin prevalence. NIDA Research Monograph, 54: 104-124.
6. Warner, S.L., 1965. Randomized response: a survey technique for eliminating evasive answer bias. Journal of the A merican Statistical Association, 60: 63-69.
7. Liu, P.T. and L.P. Chow, 1976. A new discrete quantitative randomized response model. Journal of the American Statistical Association, 71: 72-73.
8. Reinmuth, J.E. and M.D. Geurts, 1975. The collection of sensitive information using a two stage randomized response model. Journal of Marketing Research, 12: 402-407.
9. Geurts, M.D., 1980. Using a randomized response design to eliminate non-response and response biases in business research. Journal of the Academy of Marketing Science, 8(2): 83-91.
10. Larkins, E.R., E.C. Hume and B.S. Garcha, 1997. The validity of randomized response method in tax ethics research. Journal of the Applied Business Research, 13(3): 25-32.
11. Warner, S.L., 1971. The Linear randomized response model. Journal of the American Statistical Association, 66: 884-888.
12. Himmelfarb, S. and S.E. Edgell, 1980. Additive constant model: a randomized response technique for eliminating evasiveness to quantitative response questions. Psychological Bulletin, 87: 525-530.
13. Gjestvang, C. and S. Singh, 2009. An improved randomized response model: estimation of mean. Journal of Applied Statistics, 36(12): 1361-1367.
14. Gupta, S., J. Shabbir and S. Sehra, 2010. Mean and Sensitivity Estimation in optional randomized response models. Journal of Statistical Planning and Inference, 140(10): 2870-2874.
15. Huang, K.C., 2008. Estimation for the sensitive characteristic using optional randomized response technique. Quality and Quantity, 42: 679-686.

16. Huang, K.C., 2010. Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling. Metrika, 71: 341-352.
17. Hussain, Z. and K. Khan, 2013. On estimation of mean using scrambled data. World Applied Sciences Journal, 23(9): 1201-1206.
18. Gupta, S. and J. Shabbir, 2006. An alternative to Warner's randomized response Model. Journal of Modern Applied Statistical Methods, 5(2): 328-331.
19. Christofides, ,T.C. 2005. Randomized response technique for the two sensitive characteristics at the same time. Metrika, 62: 53-63.
20. Hussain, Z., J. Shabbir and S. Gupta, 2007. An alternative to Ryu *et al*. randomized response model. Journal of Statistics & Management Sciences, 10(4): 511-517.
21. Lee, C.S., S.A. Sedory and S. Singh, 2012. Estimating at least seven measures of qualitative variable from a single sample using randomized response technique. Statistics & Probability Letters, doi:10.1016/j.spl. 2012. 10.004.