International Journal of Academic Research in Business and Social Sciences
June 2012, Vol. 2, No. 6
ISSN: 2222-6990

Improvement of the Gupta and Thornton Scrambling Model through Double Use of Randomization Device

Zawar Hussain
Department of Statistics, King Abdulaziz University 80230, Jeddah 21589, Kingdom of Saudi Arabia
e-mail: zhlangah@yahoo.com

Abstract

To eliminate or reduce the extent of effect of falsified reporting, randomized response models are now being used as a survey tool in almost every field of science. To contribute in the literature on randomized response models, we propose an improved randomized response model utilizing two responses from each respondent. The proposed model provides an unbiased estimator and may be thought of as more shielding in term of privacy protection. The proposed estimator is a weighted estimator having the minimum possible sampling variance. The responses are obtained through the Gupta and Thornton (2002) model. The suggested weighted estimator is unconditionally more efficient than all of the scrambled response estimators suggested until now.

Keywords: Randomized response models, sensitive surveys and weighted estimator.

1. Introduction

While conducting a field survey on sensitive items, direct questioning method is not recommended because the respondents do not provide truthful answers. Stigma carried by these sensitive items is an obvious cause of falsification of the responses. The social desirability of over-reporting on a socially desirable characteristics and under-reporting on socially undesirable characteristics can be collectively called miss-reporting. Due to this misrepresentation the estimates turn out to be biased. The bias introduced into the estimates due to misrepresentation is called social desirability bias. In such surveys, collection of a trustworthy data becomes an important and intricate issue. To reduce the extent of falsification of responses, Warner (1965) presented his pioneering work to estimate the unknown population prevalence of a sensitive characteristic. Warner (1965) presented his idea on the premise of randomizing the response. Later on, Greenberg et al. (1971) extended his work to the estimation of the mean of sensitive quantitative variable. Since then, a lot of articles have been appeared in different reputed journals on psychological, business, educational, behavioral, marketing, medical, social, and environmental sciences. These articles are on either assessing the physical applications of some existing randomization techniques or suggesting new randomization techniques. No article has appeared yet suggesting a technique that may be taken as the best.

International Journal of Academic Research in Business and Social Sciences
June 2012, Vol. 2, No. 6
ISSN: 2222-6990

Our purpose of doing this study is to provide a quantitative randomized response estimator having the minimum possible variance. Giving a literature review on quantitative randomized response model is not required here since all the proposed models have the variance of the form of equation (1) given below. Eichhorn and Hayre (1983); Gupta et al. (2002); Gupta and Thornton (2002); Ryu et al. (2005-2006) and Hussain and Shabbir (2007) are some of the randomized response models to mention. Our motivation of writing this paper is to present a model which is better than all the randomized response models presented so far. In this paper, we propose an unbiased estimator by improving Gupta and Thornton (2002) optional randomized response model. It has been observed that all the randomized response models, proposed so far, provide estimator, say $\hat{\mu}_R$, having two sources of variation, namely, the sampling design and the randomization model. Let $\hat{\mu}_T$ be the estimator based on the true responses then variance of the estimator $\hat{\mu}_R$ is always given by

$$V\left(\hat{\mu}_R\right) = V\left(\hat{\mu}_T\right) + \text{extra variance due to randomization device} \qquad (1)$$

All the research on randomized response modeling has been done focusing on reducing the randomization variance. Gupta and Thornton (2002) made use of optional additive scrambling and presented an unbiased estimator which is superior to a large number of quantitative randomized response models. Gupta and Thornton (2002) considered splitting the total sample into two subsamples and getting an unbiased estimator using data obtained through each of the two subsamples. Contrary to Gupta and Thornton (2002), we consider a single sample of size n but obtaining two responses from each respondent making use of additive and subtractive scrambling of the true responses. Simple random sampling with replacement is assumed in this study together with the assumption that respondents are truthful in their reporting. Actual generation of the data through proposed scheme is assumed to be unknown to the enumerator. We organize this paper as follows. In the next section, we present Gupta and Thornton (2002) model and introduce all the notations. We present the proposed estimation in Section 3 followed by efficiency comparison study in Section 4. Section 5 concludes the study.

2. Gupta and Thornton (2002) RRM

Let X be the sensitive variable of interest and Y be an unrelated non-sensitive variable. The population mean μ_X is the parameter of interest. The distribution of variable Y, say $f\left(Y\right)$, is completely known mean $\mu_Y\ \left(-\infty < \mu_Y < \infty\right)$ and variance $\sigma_Y^2\ \left(>0\right)$. To estimate the mean μ_X, Gupta and Thornton (2002) described a partial quantitative randomized response model. In their technique, some known proportion of respondents responds truthfully while the remaining proportion of respondents' reports scrambled responses. The scrambling is done in an additive way. The i^{th} respondents is first requested to generate a value Y_i from $f\left(Y\right)$ and then provided a randomization device consisting of two statements: (i) Report your true response on sensitive variable X, and (ii) Report the scrambled response as $\left(X_i + Y_i\right)$,

represented with probabilities T and $(1-T)$, respectively. Let Z_{1i} be the reported response of the i^{th} respondent then it can be written as

$$Z_{1i} = \alpha_i X_i + (1-\alpha_i)(X_i + Y_i), \quad (2)$$

where α_i is a Bernoulli random variable with mean$T$? The expected response from the i^{th} respondent is given by

$$\begin{aligned} E(Z_{1i}) &= E(\alpha_i)E(X_i) + E(1-\alpha_i)E(X_i + Y_i) \\ &= T\mu_X + (1-T)(\mu_X + \mu_Y) \\ E(Z_{1i}) &= \mu_X + (1-T)\mu_Y . \end{aligned} \quad (3)$$

An unbiased estimator of μ_X is given by

$$\hat{\mu}_{1X} = \bar{Z}_1 - (1-T)\mu_Y . \quad (4)$$

It can be shown that the variance of the estimator $\hat{\mu}_{1X}$ has the variance given by

$$V(\hat{\mu}_{1X}) = V(\bar{Z}_1) = \frac{\sigma_X^2}{n} + \frac{(1-T)(\sigma_Y^2 + T\mu_Y^2)}{n} \quad (5)$$

Same like the structure of equation (1), the second term in the above equation (5) is the cost, in terms of variance, one has to pay for randomizing the responses.

3. Proposed Estimation

The model proposed by Gupta and Thornton (2002) is improved by taking two responses from each respondent and defining two dependent estimators with equal variances. To obtain the second response, subtractive scrambling is used. In this way, the two responses from each respondent are correlated. Then, taking advantage of the equal variances, a weighted estimator is defined with minimum variance. Let Z_{2i} be the second response from i^{th} respondent taken as

$$Z_{2i} = \alpha_i X_i + (1-\alpha_i)(X_i - Y_i), \quad (6)$$

Where α_i is a Bernoulli random variable defined as above. Now, the second expected response from the i^{th} respondent is given by

$$E(Z_{2i}) = E(\alpha_i)E(X_i) + E(1-\alpha_i)E(X_i - Y_i)$$

$$= T\,\mu_X + (1-T)(\mu_X - \mu_Y)$$
$$E(Z_{2i}) = \mu_X - (1-T)\,\mu_Y \qquad (7)$$

This suggests defining another unbiased estimator, based on the second set of responses, of μ_X as

$$\hat{\mu}_{2X} = \bar{Z}_2 + (1-T)\,\mu_Y\ . \qquad (8)$$

Now, we find variance of the estimator defined in (8). By definition, variance of the estimator, $\hat{\mu}_{2X}$, is given by

$$V(\hat{\mu}_{2X}) = V\left\{\bar{Z}_2 + (1-T)\,\mu_Y\right\} = V(\bar{Z}_2) = \frac{1}{n}V(Z_{2i}) \qquad (9)$$

Consider

$$V(Z_{2i}) = E(Z_{2i}^2) - \left\{E(Z_{2i})\right\}^2$$
$$V(Z_{2i}) = E\left\{\alpha_i^2 X_i^{\,2} + (1-\alpha_i)^2 (X_i - Y_i)^2 + 2\alpha_i(1-\alpha_i)X_i(X_i - Y_i)\right\} - \left\{E(Z_{2i})\right\}^2$$
$$V(Z_{2i}) = E\left\{T(\mu_X^2 + \sigma_X^2) + (1-T)(\mu_X^2 + \sigma_X^2 + \mu_Y^2 + \sigma_Y^2 - 2\mu_X\mu_Y)\right\} - \left\{\mu_X - (1-T)\,\mu_Y\right\}^2$$
$$V(Z_{2i}) = \left\{(\mu_X^2 + \sigma_X^2) + (1-T)(\mu_Y^2 + \sigma_Y^2 - 2\mu_X\mu_Y)\right\} - \left\{\mu_X^2 + (1-T)^2\mu_Y^2 - 2(1-T)\mu_X\mu_Y\right\}$$
$$V(Z_{2i}) = \left\{\sigma_X^2 + (1-T)(T\mu_Y^2 + \sigma_Y^2)\right\} \qquad (10)$$

On substituting (10) in (9), we get

$$V(\hat{\mu}_{2X}) = \frac{\sigma_X^2}{n} + \frac{(1-T)(T\mu_Y^2 + \sigma_Y^2)}{n} \qquad (11)$$

From (11) and (5), it is clear that $V(\hat{\mu}_{2X}) = V(\hat{\mu}_{1X})$. Taking the advantage of equal variances and utilizing the full information, we define a new estimator of μ_X as

$$\hat{\mu}_{3X} = W\,\hat{\mu}_{1X} + (1-W)\,\hat{\mu}_{2X}, \quad (0 < W \le 1). \qquad (12)$$

Its variance is given by

$$V(\hat{\mu}_{3X}) = W^2 V(\hat{\mu}_{1X}) + (1-W)^2 V(\hat{\mu}_{2X}) + 2W(1-W)C(\hat{\mu}_{1X}, \hat{\mu}_{2X}), \qquad (13)$$

International Journal of Academic Research in Business and Social Sciences
June 2012, Vol. 2, No. 6
ISSN: 2222-6990

Where $C\left(\hat{\mu}_{1X},\hat{\mu}_{2X}\right)$ is the covariance of the two estimators $\hat{\mu}_{1X}$ and $\hat{\mu}_{2X}$. Using the first order derivative condition of optimization, It is straight forward to verify that the optimum value of $W=\frac{1}{2}$. 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

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

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

$$C\left(\hat{\mu}_{1X},\hat{\mu}_{2X}\right)=C\left(\bar{Z}_1-\left(1-T\right)\mu_Y,\bar{Z}_2+\left(1-T\right)\mu_Y\right)$$
$$C\left(\hat{\mu}_{1X},\hat{\mu}_{2X}\right)=C\left(\bar{Z}_1,\bar{Z}_2\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C\left(Z_{1i},Z_{2j}\right)=\frac{1}{n^2}\sum_{i=1}^{n}C\left(Z_{1i},Z_{2i}\right), \tag{16}$$

Since $C\left(Z_{1i},Z_{2j}\right)=0\ \ \forall\ i\neq j$. Now, the covariance of Z_{1i} and Z_{2i} is given by

$$C\left(Z_{1i},Z_{2i}\right)=E\left(Z_{1i}Z_{2i}\right)-E\left(Z_{1i}\right)E\left(Z_{2i}\right)$$
$$C\left(Z_{1i},Z_{2i}\right)=\sigma_X^2-\left(1-T\right)\left(T\,\mu_Y^2+\sigma_Y^2\right). \tag{17}$$

On substituting (17) in (16), we get

$$C\left(\hat{\mu}_{1X},\hat{\mu}_{2X}\right)=\frac{\sigma_X^2}{n}-\frac{\left(1-T\right)\left(T\,\mu_Y^2+\sigma_Y^2\right)}{n}. \tag{18}$$

On substituting (18) in (15), we get the variance of the weighted estimator $\hat{\mu}_{3X}$, given by

$$V\left(\hat{\mu}_3^X\right)=\frac{1}{2}\left(\frac{\sigma_X^2}{n}+\frac{\left(1-T\right)\left(T\,\mu_Y^2+\sigma_Y^2\right)}{n}\right)+\frac{1}{2}\left(\frac{\sigma_X^2}{n}-\frac{\left(1-T\right)\left(T\,\mu_Y^2+\sigma_Y^2\right)}{n}\right)$$
$$V\left(\hat{\mu}_{3X}\right)=\frac{1}{2n}\left(2\sigma_X^2\right)=\frac{\sigma_X^2}{n}, \tag{19}$$

This is the lower bound on the variance of an estimator based on simple random sampling with replacement and utilizing randomized responses. Scrambling variance is eliminated and no further reduction of scrambling is possible. It is obvious now that scrambling effect is removed

International Journal of Academic Research in Business and Social Sciences
June 2012, Vol. 2, No. 6
ISSN: 2222-6990

by taking two responses from each respondent and using additive and subtractive scrambling simultaneously.

1. Efficiency comparison

The proposed estimator $\hat{\mu}_{3X}$ is relatively more efficient than Gupta and Thornton (2002) estimator if

$$\begin{aligned} & V\left(\hat{\mu}_X\right) - V\left(\hat{\mu}_{3X}\right) > 0 \\ & n^{-1}\left\{\sigma_X^2 + (1-T)\left(T\,\mu_Y^2 + \sigma_Y^2\right)\right\} - n^{-1}\left(\sigma_X^2\right) > 0 \\ & (1-T)\left(T\,\mu_Y^2 + \sigma_Y^2\right) > 0 \\ & 1 - T > 0 \\ & T < 1. \end{aligned} \tag{20}$$

The above inequality (20) always holds true since $0 < T < 1$. Thus, there is no need of computing the relative efficiency of the proposed estimator numerically. Also, there is no need of comparing it to other existing estimators.

2. Conclusion

Utilizing the idea of obtaining two responses from each respondent, new estimator $\hat{\mu}_{3X}$ has been proposed. The proposed estimator is actually a weighted estimator and unconditionally more efficient than that of Gupta and Thornton (2002) estimator. The variance of the proposed estimator is equal to the lower bound on the variance of an unbiased estimator. Hence the proposed estimator is a uniformly minimum variance unbiased estimator. The proposed estimator is uniformly better than any other existing estimators proposed so far. If we compare the variance of the proposed estimator to the expression (1), we see that it has no scrambling variance. It may be argued that privacy of the respondents would be at stake when we obtain the same responses from a respondent. This will be the case only when a respondent randomly chooses statement (i) to answer. To avoid this situation and keep the privacy of the respondents intact, we propose collecting the two responses from the respondents by requesting them to write their scrambled responses on separate cards and put them into different boxes without disclosing their identities. In this way, one box contains the first set of responses and the other contains the set of second responses. Then, any of the response in each box cannot be attributed to any of the respondents and complete privacy protection to the respondents is guaranteed.

Acknowledgement

The author highly appreciates the excellent research facilities provided by the King Abdulaziz University.

International Journal of Academic Research in Business and Social Sciences
June 2012, Vol. 2, No. 6
ISSN: 2222-6990

References

Eichhorn, B. H. & Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference, 7, 307-316

Greenberg, B. G., Kubler, R. R. & Horvitz, D. G. (1971). Applications of RR technique in obtaining quantitative data. Journal of the American Statistical Association, 66, 243-250.

Gupta, S. & Thornton, B. (2002). Circumventing social desirability response bias in personal interview surveys. American Journal of Mathematical and Management Sciences, 22, 369-383.

Gupta, S., Gupta, B. & Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. Journal of Statistical Planning and Inference, 100, 239-247.

Hussain, Z. & Shabbir, J. (2007). Estimation of mean of a sensitive quantitative variable. Journal of Statistical research, 41(2), 83-92.

Ryu, J. B., Kim, J. M., Heo, T. Y. & Park, C. G. (2005-2006). On stratified Randomized response sampling. Model Assisted Statistics and Applications, 1(1), 31-36.

Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.