

**ON ESTIMATION OF MEAN OF A SENSITIVE QUANTITATIVE
VARIABLE IN COMPLEX SURVEYS**

Zawar Hussain¹ and Javid Shabbir²

Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan

Email: ¹zhlangah@yahoo.com; ²jsqau@yahoo.com

ABSTRACT

Estimation of mean of a stigmatized quantitative variable is considered here for complex survey situations. The estimators proposed by Hussain and Shabbir (2007) and that of Ryu et al. (2006) are studied. It is observed that the estimators proposed by Hussain and Shabbir work better than Ryu et al. (2006) estimator in terms of efficiency as well as the provision of privacy to the survey respondents, in complex survey situations.

KEY WORDS

Randomized response technique; sensitive characteristic; estimation of mean; anonymity; double randomization; and scrambled response.

1. INTRODUCTION

Direct questioning about a sensitive characteristic (for example: about abortion, sexual abuse, or taking drugs), generally, results in falsified responses or refusal to respond at all. Having a reliable measure of the prevalence of the sensitive variables is a serious issue in most of the social surveys. As a consequence of false reporting, an unavoidable estimation bias creeps into the estimators. Warner (1965) was the first to show this evasive answer bias, prevailing in the estimators based on direct questioning. He proposed a randomized response method to estimate proportion of prevalence of the sensitive characteristic in the population. Greenberg et al. (1971) extended the idea of the randomized response model (RRM) to the estimation of mean of a sensitive quantitative variable. The recent articles on the estimation of mean of a sensitive variable include Eichhorn and Hayre (1983), Singh (1999), Singh et al. (2001), Gupta et al. (2002), Bar-Lev et al. (2004), Ryu et al. (2006), Hussain et al. (2007) and many others. In all the research cited above simple random sampling with replacement (SRSWR) sampling is assumed. There are a few articles, which assume unequal probability sampling. Saha (2006) and Arnab and Dorffner (2006) are some of the papers to be cited. In this paper, we have studied Hussain and Shabbir (2007) RRTs and compared them with the estimator proposed by the Ryu et al. (2006) in complex survey situations using unequal probability sampling. In Ryu et al. (2006) RRM a respondent may report his actual value on sensitive variable in two ways. Moreover, he/she may be suspicious about being traced back to his/her actual response. Assuming SRSWR sampling design, Hussain and Shabbir (2007) used the idea of increasing the number of ways a response can be scrambled and an idea of double randomization to estimate the mean of the sensitive quantitative variable. Since we intend to compare proposed RRMs with Ryu et al. (2006)

RRM, we briefly outline in Section 2, the Ryu et al. (2006), and Hussain and Shabbir (2007) RRM. In Section 3, efficiency conditions are derived and numerical results are given in Table 3 (see Appendix). In Section 4, results of a simulation study are discussed.

2. SOME RANDOMIZED RESPONSE MODELS

Let $U = \{u_1, \dots, u_N\}$ be a finite population of N units and z_i be the value of the i th unit of sensitive variable z under study. The objective is to estimate the population mean $\mu_z = \sum_{i=1}^N z_i / N$ on the basis of a sample s selected with $p(s)$ according to design p . Here the value z_i cannot be obtained directly from the respondent. So a randomized response y_i is obtained from the i th ($i \in s$) unit by using suitable RR technique R (say). Let r_i be a function of y_i such that $E_R(r_i) = z_i$ and $V_R(r_i) = \sigma_i^2(R)$ where $E_R(V_R)$ denotes expectation (variance) operators with respect to RR technique. In this paper we will consider the following RR techniques:

2.1 RYU et al. Model

Based on Mangat and Singh (1990) two-stage randomized response model, Ryu et al. (2006) proposed a model to estimate the mean of the sensitive quantitative variable. The i^{th} respondent selected in the sample of size n is requested to use the randomization device R_1 , which consists of two statements: (i) "Report the true response Z_i of sensitive question" and (ii) "Go to randomization device R_2 in the second stage", represented with probabilities P and $1-P$ respectively. The randomization device R_2 consists of two statements: (i) "Report the true response Z_i of sensitive question", and (ii) "Report the scrambled response $Z_i S_i$ of sensitive question", represented with probabilities T and $1-T$ respectively. The possible outcomes from Ryu et al. (2006) RRM are presented graphically in Figure 1 (see Appendix). Using the assumption of known distribution of scrambling variable S such that $\mu_S = 1$ and $\sigma_S^2 = \psi^2$, the variance of the estimator r_i of Z_i is given by

$$V_R(r_i) = (1-P)(1-T)y_i^2\psi^2 = V_{li}, \quad (2.1)$$

2.2 Hussain and Shabbir Model I

Each individual in the sample is requested to use a randomization device R_1 , which consists of the two statements:

- a) report your true response Z_i of the sensitive question", and
- b) go to the randomization device R_2 ",

represented with the probabilities P and $1-P$ respectively.

The randomization device R_2 consists of the two statements:

- a) "report the scrambled response $Z_i + bS_i$ ", and
- b) "report your scrambled response $Z_i - aS_i$ ",

represented with probabilities $T = \frac{a}{a+b}$ and $1-T = \frac{b}{a+b}$ respectively, where a and b are any positive real numbers, S is a scrambling variable with mean $\mu_S = 1$ (any value of mean of scrambling variable can be set, not particularly 1 as in Ryu et al. (2006) model as far as the unbiasedness is concerned) and variance $\sigma_S^2 = \psi^2$. The graphical representation of the outcomes of Model I is given in Figure 2 (See Appendix). The variance of the estimator r_i of Z_i is given by

$$V_R(r_i) = (1-P)(1+\psi^2)ab = V_{2i}. \quad (2.2)$$

2.3 Hussain and Shabbir Model II

Each individual in the sample is provided a randomization device R with two outcomes: ‘Use randomization devices R_1 ’ and ‘Use the randomization device R_2 ’, with probabilities P and $1-P$ respectively. The randomization device R_1 consists of the two statements:

- a) “report you scrambled response $Z_i + b_1 S_i$ ”, and
- b) “report you scrambled response $Z_i - a_1 S_i$ ”,

represented with probabilities $T_1 = \frac{a_1}{a_1+b_1}$ and $1-T_1 = \frac{b_1}{a_1+b_1}$ respectively.

The randomization device R_2 consists of the two statements:

- a) “report you scrambled response $Z_i + b_2 S_i$ ”, and
- b) “report you scrambled response $Z_i - a_2 S_i$ ”,

represented with probabilities $T_2 = \frac{a_2}{a_2+b_2}$ and $1-T_2 = \frac{b_2}{a_2+b_2}$ respectively. The outcomes of the Model II are represented graphically in Figure 3 (see Appendix). The variance of the estimator r_i of Z_i is given by

$$V_R(r_i) = (1+\psi^2)(Pa_1b_1 + (1-P)a_2b_2) = V_{3i}. \quad (2.3)$$

3. EFFICIENCY COMPARISON

In case z_i is available from the respondents, one could set the following homogeneous unbiased estimator of the population mean

$$t(z, s) = \sum_{i \in s} b_{si} z_i, \quad (3.1)$$

where b_{si} ’s are known constants satisfying $\sum_{s \ni i} b_{si} p(s) = 1/N$.

In the present situation z_i ’s are not available from the respondents, so, we replace z_i by y_i in (3.1) to find an unbiased estimator for the RR survey as follows.

$$t_R = t(z, y) = \sum_{i \in S} b_{si} y_i . \quad (3.2)$$

It is clear that

$$Et(z, y) = E_p \sum_{i \in S} b_{si} E_R y_i = \mu_z ,$$

and

$$V(t_R) = V_p \left(\sum_{i \in S} b_{si} z_i \right) + E_p \left(\sum_{i \in S} b_{si}^2 \sigma_i^2(R) \right) = Q + W(R),$$

where $Q = V_p \left(\sum_{i \in S} b_{si} z_i \right)$, $W(R) = \sum_{i \in U} \sigma_i^2(R) \alpha_i$ and $\alpha_i = \sum_{s \supset i} b_{si}^2 p(s)$. Let D_1 and D_2 be two different RR techniques. Then we say D_1 is more efficient than D_2 if $V(t_{D_1}) \leq V(t_{D_2})$ i.e. $W(D_1) \leq W(D_2)$. A sufficient condition of superiority of D_1 over D_2 is $\sigma_i^2(D_1) \leq \sigma_i^2(D_2) \quad \forall i \in U$.

3.1 Model I versus Ryu et al. (2006) Model

The estimator based on RR Model I will be more efficient than that of Ryu et al. (2006) estimator if

$$V_{1i} - V_{2i} \geq 0 ,$$

or

$$(1-P)(1-T) y_i^2 \psi^2 - (1-P)(1+\psi^2) ab \geq 0 .$$

Above condition is most likely to be achieved in most of the survey situations because a and b are controllable and can be chosen very small. For example, one can set $a = 0.0001$ and $b = 0.0003$.

3.2 Model II versus Ryu et al. (2006) RRM

The estimator based on RR Model II will be more efficient than the Ryu et al. (2006) RRT if

$$V_{1i} - V_{3i} \geq 0 ,$$

or

$$(1-P)(1-T) y_i^2 \psi^2 - (1+\psi^2) (P a_1 b_1 + (1-P) a_2 b_2) \geq 0 . \quad (3.3)$$

The inequality (3.3) can be made true by suitably setting the values of the a_1, b_1, a_2 and b_2 . If the restriction $a_1 b_1 = a_2 b_2$ is imposed then inequality (3.3) reduces to

$$(1-P)(1-T) y_i^2 \psi^2 - (1+\psi^2) a_2 b_2 \geq 0 . \quad (3.4)$$

The inequality (3.3) can be made true more easily simply by choosing the constants a_2 and b_2 smaller irrespective of the value of P .

4. NUMERICAL RESULTS AND SIMULATIONS STUDY

The numerical results for the both the Models I and II are given in Table 3 (see Appendix) for suitable values of $a_i, b_i, i=1,2, \psi^2$, and σ_A^2 , assuming unequal probability sampling using Lahiri (1951) method. We have presented the relative efficiency of our proposed Models I and II in a Table because for small values of $a_i, b_i, i=1,2$ Models I and II are equally efficient when compared to RRM by Ryu et al. (2006). To study the performance of the estimators based on Models I and II as compared to the Ryu et al. (2006) estimator, we have performed a simulation study. A sample of size 100 was taken from the population. We assumed that the study variable follows a Gamma (1, 2) distribution so that mean of the sensitive variable is 2. It was also assumed that the scrambling variable follows a normal distribution with mean 1 and variance 0.5. We have performed 5000 iterations and calculated the simulated means and standard errors of the three estimators. All the estimators are unbiased but compared to Ryu et al. (2006) model, both the estimators based on Model I and Model II perform well in terms of the sampling variance. In Tables 1 and 2 (see Appendix), we have given the simulated means and standard deviation of the three estimators discussed in this article.

REFERENCES

1. Arnab, R. and Dorffner, G. (2006). Randomized response technique for complex survey designs. *Statistical Papers*, 48, 131-141.
2. Bar-Lev, S.K., Bobovitch, E., and Boukai, B. (2004). A note on randomized response models. *Metrika*, 60, 255-260.
3. Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Plann. and Infer.* 7, 307-316.
4. Greenberg, B.G., Kuebler, R.R. Jr., Abernathy, J.R. and Hovertz, D.G. (1971). Application of the randomized response techniques in obtaining quantitative data. *J. Amer. Statist. Assoc.*, 66, 243-250.
5. Gupta S., Gupta, B. and Singh, S. (2002). Estimation of Sensitivity level of personal interview survey questions. *J. Statist. Plann. and Infer.*, 100, 239-247.
6. Hussain, Z., Shabbir, J. and Gupta, S. (2007). An alternative to Ryu et al. randomized response model. *J. Statist. and Management Systems*, 10(4), 511-517.
7. Hansen and Hurwitz (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, 14, 333-362.
8. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47, 663-685.
9. Hussain, Z. and Shabbir, J. (2007). On estimation of mean of a sensitive quantitative variable. *Inter Stat*: July # 6. <http://www.interstat.statjournals.net/YEAR/2007/abstracts/0707006.php>.
10. Lahiri, D.B. (1951). A method for sample selection providing unbiased ratio estimates. *Bull. Ins. Statist. Inst.*, 33(2), 133-140.
11. Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
12. Ryu, J.B., Kim, J.M., Heo, T.Y. and Park, C.G. (2006). On stratified Randomized response sampling. *Model Assisted Statist. and App.*, (1), 31-36.
13. Raj, D. (1968). *Sampling Theory*. Mc Graw Hill, N.Y.

14. Rao, J.N.K. (1975). Unbiased variance estimation for multistage designs. *Sankhya*, C 37, 133-139.
15. Singh, S. (1999). An addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn. *Biometrical J.*, 41(8), 955-966.
16. Singh, S., Mahmood, M. and Tracy, D.S. (2001): Estimation of mean and variance of stigmatized quantitative variable using distinct units in randomized response sampling. *Statistical Papers*, 42, 403-411.
17. Saha, A. (2005). Kim and Warde mixed Randomized response technique for complex survey. *J. Modern App. Statist. Methods*, 4(2), 538-544.
18. Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, 60, 63-69.

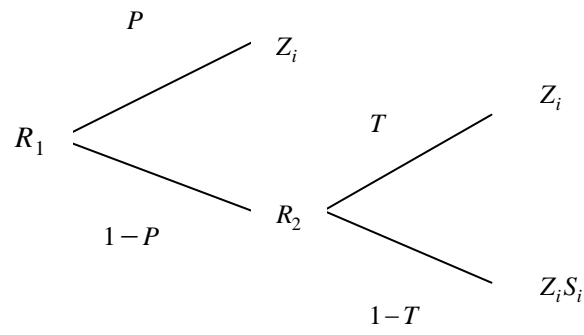


Fig. 1: Graphical representation of Ryu et al. (2005) RRM.

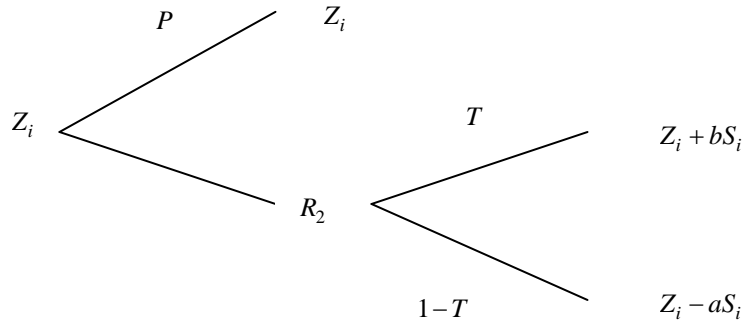


Fig. 2: Graphical representation of the outcomes of Model I

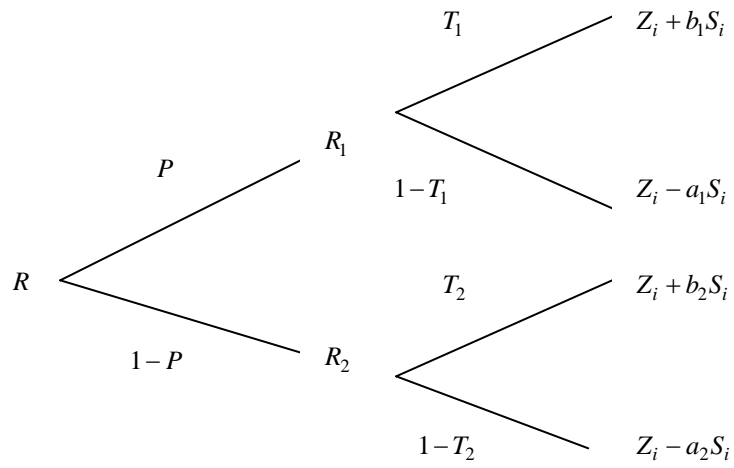


Fig. 3: Graphical representation of the outcomes of Model II

Table 1:
Simulated mean and standard error (Stdev) of the estimator based on model I and Ryu et al. (2005) estimator for $a = 0.01$, $b = 0.02$ and $n = 1000$.

| P | Mean (\bar{Z}_2) | Stdev (\bar{Z}_2) | Mean (\bar{Z}_1) | Stdev (\bar{Z}_1) |
|-----|----------------------|-----------------------|----------------------|-----------------------|
| 0.1 | 1.9967 | 0.2030 | 1.9967 | 0.2197 |
| 0.3 | 2.0019 | 0.2033 | 2.0026 | 0.2165 |
| 0.5 | 1.9979 | 0.2017 | 1.9987 | 0.2124 |
| 0.7 | 2.0050 | 0.2051 | 2.0051 | 0.2130 |
| 0.9 | 1.9992 | 0.2016 | 1.9992 | 0.2078 |

Table 2:
Simulated mean and standard error (Stdev) of the estimators based on model II and Ryu et al. (2005) estimator for $a_1 = a_2 = 0.01$, $b_1 = b_2 = 0.02$ and $n = 1000$.

| P | Mean (\bar{Z}_3) | Stdev (\bar{Z}_3) | Mean (\bar{Z}_1) | Stdev (\bar{Z}_1) |
|-----|----------------------|-----------------------|----------------------|-----------------------|
| 0.1 | 2.0021 | 0.1982 | 2.0018 | 0.2159 |
| 0.3 | 1.9983 | 0.1994 | 1.9977 | 0.2126 |
| 0.5 | 1.9991 | 0.1991 | 1.9990 | 0.2099 |
| 0.7 | 1.9974 | 0.2007 | 1.9990 | 0.2091 |
| 0.9 | 2.0032 | 0.2002 | 2.0033 | 0.2063 |

Table 3:
Relative efficiencies of Models I and II relative to Ryu et al. (2005) RRM for $\psi^2 = 0.5$, $\sigma_A^2 = 1$, $a_1 = a_2 = b_1 = 0.01$, $b_2 = 0.02$, and $n = 1000$

| μ_A | σ_A^2 | T | P | | | | |
|---------|--------------|-----|-------|-------|------|------|------|
| | | | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
| 2 | 1 | 0.1 | 2.01 | 1.78 | 1.56 | 1.33 | 1.11 |
| | | 0.3 | 1.78 | 1.61 | 1.43 | 1.26 | 1.08 |
| | | 0.5 | 1.56 | 1.43 | 1.31 | 1.18 | 1.06 |
| | | 0.7 | 1.33 | 1.26 | 1.18 | 1.11 | 1.03 |
| | | 0.9 | 1.11 | 1.08 | 1.06 | 1.03 | 1.01 |
| 4 | 1 | 0.1 | 4.44 | 3.67 | 2.91 | 2.14 | 1.38 |
| | | 0.3 | 3.67 | 3.08 | 2.48 | 1.89 | 1.29 |
| | | 0.5 | 2.91 | 2.48 | 2.06 | 1.63 | 1.21 |
| | | 0.7 | 2.14 | 1.89 | 1.63 | 1.38 | 1.12 |
| | | 0.9 | 1.38 | 1.29 | 1.21 | 1.12 | 1.04 |
| 6 | 1 | 0.1 | 8.49 | 6.82 | 5.16 | 3.49 | 1.83 |
| | | 0.3 | 6.82 | 5.53 | 4.23 | 2.94 | 1.64 |
| | | 0.5 | 5.16 | 4.23 | 3.31 | 2.38 | 1.46 |
| | | 0.7 | 3.49 | 2.94 | 2.38 | 1.83 | 1.27 |
| | | 0.9 | 1.83 | 1.64 | 1.46 | 1.27 | 1.09 |
| 8 | 1 | 0.1 | 14.15 | 11.23 | 8.31 | 5.38 | 2.46 |
| | | 0.3 | 11.23 | 8.96 | 6.68 | 4.41 | 2.13 |
| | | 0.5 | 8.31 | 6.68 | 5.06 | 3.43 | 1.81 |
| | | 0.7 | 5.38 | 4.41 | 3.43 | 2.46 | 1.48 |
| | | 0.9 | 2.46 | 2.13 | 1.81 | 1.48 | 1.16 |