

On Estimation of Mean of a Sensitive Quantitative Variable

Zawar Hussain^{*†} and Javed Shabbir^{*}

^{*}Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan.

[†]Corresponding Author's email: zhlangah@yahoo.com

Abstract. To estimate the mean of a sensitive quantitative variable two Randomized Response Models similar to Ryu et al (2005) model are studied. It is observed that the estimators based on proposed models work better than Ryu et al (2005) estimator in terms of efficiency as well as the provision of privacy to the survey respondents.

Key words: randomized response technique, sensitive character, estimation of mean, anonymity, double randomization, and scrambled response.

1. Introduction

Collecting information on sensitive attributes, both qualitative and quantitative, is a tricky issue. The most serious problem in studying certain social problems that are sensitive in nature (e.g. induced abortion, drug usage, tax evasion, etc.) is the lack of a reliable measure of their incidence or prevalence. Social stigma and fear of reprisals usually result in lying by the respondents when approached with the conventional or direct-response survey method. An easy consequence of false reporting is seen to be an unavoidable estimation bias. Warner (1965) showed this evasive answer bias to prevail in the estimate obtained by direct questioning, and proposed a randomized response method to estimate proportion of prevalence of the sensitive character in the population. Greenberg et al. (1971) extended the Randomized Response model to the estimation of mean of a sensitive quantitative variable. The recent articles on the estimation of mean of

a sensitive variable include Singh, Horn, and Chowdhury (1998), Singh (1999), Singh, Mahmood, and Tracy (2001), Chang and Haung (2001), Gupta et al. (2002), Bar-Lev, Bobovitch, and Boukai (2004), Singh and Mathur (2004), Ryu et al. (2005) and many others. In this note we present an unbiased estimator of the mean and compare it with the estimator proposed by the Ryu et al. (2005). Ryu et al (2005) RRM is based on the use of the two Randomization Devices (RDs) but the use of second RD is conditioned on the outcome of first RD. Also, there are two ways to report actual response in Ryu et al (2005) RRM. And respondent may suspect that his/her actual response can be traced. In the proposed Model I we use the idea of increasing the number of ways the responses can be scrambled and in the proposed Model II we use the idea of double randomization but the use of a particular randomization device (RD) is known to the interviewer only. Similar idea has also been used by Hussain and Shabbir (2007) to estimate the proportion of individuals with sensitive attribute. Since we intend to compare proposed RRM with Ryu et al (2005) RRM, we briefly outline in section 2, the Ryu et al. (2005) estimation procedure. In section 3 we present our proposed model. Section 4 contains the efficiency comparison of the proposed procedure with the procedure of Ryu et al (2005) and numerical example for relative efficiency of the proposed procedure followed by a short discussion in Section 5.

2. Ryu et al. Proposed Model

Based on Mangat and Singh (1990) two-stage randomized response model, Ryu et al. (2005) 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_i which consists of two statements: (i) "Report the true response A of sensitive

question” and (ii) “Go to randomization device R_2 in the second stage” represented with probabilities P_0 and $1 - P_0$. The randomization device R_2 consists of two statements: (i) “Report the true response A of sensitive question” and (ii) “Report the scrambled response AB of sensitive question” represented with probabilities T_0 and $1 - T_0$ respectively. Using the assumption of known distribution of scrambling variable S such that $\mu_S = 1$ and $\sigma_S^2 = \psi^2$, the response of i^{th} respondent can be written as

$$Z_i = \alpha A_i + (1 - \alpha) [\beta A_i + (1 - \beta) A_i S_i], \quad (2.1)$$

where $\alpha = 1$, if a respondent choose a statement (i) in R_1 , and $\alpha = 0$, if a respondent chooses a statement (ii) in R_1 . Also $\beta = 1$, if a respondent choose a statement (i) in R_2 , and $\beta = 0$, if a respondent chooses a statement (ii) in R_2 .

The expected value of the observed response is

$$\begin{aligned} E(Z_i) &= E\left\{\alpha A_i + (1 - \alpha) [\beta A_i + (1 - \beta) A_i S_i]\right\} \\ &= E(\alpha) E(A_i) + E(1 - \alpha) E[\beta A_i + (1 - \beta) A_i S_i] \\ &= P_0 \mu_A + (1 - P_0) \{T_0 \mu_A + (1 - T_0) \mu_A \mu_S\} = \mu_A, \end{aligned} \quad (2.2)$$

where α and β are Bernoulli random variables with means P_0 , T_0 and variances $P_0(1 - P_0)$, $T_0(1 - T_0)$ respectively. The estimator based on the responses Z_i , $i = 1, 2, \dots, n$, Ryu et al. (2005) suggested an unbiased estimator of the mean μ_A as

$$\hat{\mu}_R = \frac{1}{n} \sum_{i=1}^n Z_i \quad (2.3)$$

with variance

$$V(\hat{\mu}_R) = \frac{1}{n} \left\{ (\mu_A^2 + \sigma_A^2) \left[P_0 + (1-P_0)T_0 + (1-P_0)(1-T_0)(1+\psi^2) \right] - \mu_A^2 \right\}$$

or

$$V(\hat{\mu}_R) = \frac{1}{n} \left[\sigma_A^2 + (\mu_A^2 + \sigma_A^2)(1-P_0)(1-T_0)\psi^2 \right] \quad (2.4)$$

3. Proposed RRM

Model I

In the proposed RRM a sample of size n is drawn from the population with SRSWR sampling scheme. Each individual in the sample is requested to use a randomization device R_1 which consists of the two statements:

- (i) “report your true response A_i of the sensitive question” and
- (ii) “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:

- (i) “report the scrambled response $A_i + bS_i$ ” and
- (ii) “report your scrambled response $A_i - aS_i$ ”,

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 (2005) model) and variance $\sigma_s^2 = \psi^2$. Let y_i be the response of the i^{th} respondent then it can be written as

$$y_i = \alpha A_i + (1 - \alpha) \{ \beta (A_i + b S_i) + (1 - \beta) (A_i - a S_i) \},$$

where α and β are the Bernoulli random variables as defined in section 2.

As A_i and S_i are independent random variables therefore

$$E(y_i) = E \left[\alpha A_i + (1 - \alpha) \{ \beta (A_i + b S_i) + (1 - \beta) (A_i - a S_i) \} \right]$$

$$E(y_i) = P \mu_A + (1 - P) \{ T (\mu_A + \beta \mu_S) + (1 - T) (\mu_A - \alpha \mu_S) \}$$

$$E(y_i) = P \mu_A + (1 - P) \mu_A = \mu_A.$$

An unbiased estimator of μ_A is given by

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n y_i. \quad (3.1)$$

The variance of the proposed estimator is given by

$$Var(\hat{\mu}_j) = \frac{1}{n} \left[\sigma_A^2 + (1 - P) ab (1 + \psi^2) \right]. \quad (3.2)$$

Model II

Each individual in the sample provided two randomization devices to R_1 and R_2 to use one of them with probabilities P_1 and $1 - P_1$ respectively. The randomization device R_1 consists of the two statements:

- (i) “report you scrambled response $A_i + b_1 S_i$ ” and
- (ii) “report you scrambled response $A_i - a_1 S_i$ ”

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:

- (i) “report you scrambled response $A_i + b_2 S_i$ ” and
- (ii) “report you scrambled response $A_i - a_2 S_i$ ”

represented with probabilities $T_2 = \frac{a_2}{a_2 + b_2}$ and $1 - T_2 = \frac{b_2}{a_2 + b_2}$ respectively. Let U_i be

the response of the i^{th} individual then it can be written as

$$U_i = \alpha \left[\beta (A_i + b_1 S_i) + (1 - \beta) (A_i - a_1 S_i) \right] + (1 - \alpha) \left[\gamma (A_i + b_2 S_i) + (1 - \gamma) (A_i - a_2 S_i) \right]$$

where

$$\alpha = \begin{cases} 1 & \text{if a respondent chooses randomization device } R_1 \\ 0 & \text{if a respondent chooses randomization device } R_2 \end{cases}$$

$$\beta = \begin{cases} 1 & \text{if a respondent chooses a statement (i) in } R_1 \\ 0 & \text{if a respondent chooses a statement (ii) in } R_1 \end{cases}$$

and
$$\gamma = \begin{cases} 1 & \text{if a respondent chooses a statement (i) in } R_2 \\ 0 & \text{if a respondent chooses a statement (ii) in } R_2 \end{cases} .$$

The expected value of the observed response is,

$$\begin{aligned} E(U_i) &= P_1 \left[T_1 (\mu_A + b_1 \mu_S) + (1 - T_1) (\mu_A - a_1 \mu_S) \right] \\ &\quad + (1 - P_1) \left[T_2 (\mu_A + b_2 \mu_S) + (1 - T_2) (\mu_A - a_2 \mu_S) \right] \\ &= \mu_A , \end{aligned}$$

where α is a Bernoulli random variable with $E(\alpha) = P_1$, $Var(\alpha) = P_1(1 - P_1)$, β is a Bernoulli random variable with $E(\beta) = T_1$, $Var(\beta) = T_1(1 - T_1)$ and γ is a Bernoulli random variable with $E(\gamma) = T_2$, $Var(\gamma) = T_2(1 - T_2)$.

An unbiased estimator of the population mean μ_A is given by,

$$\hat{\mu}_z = \frac{1}{n} \sum_{i=1}^n U_i. \quad (3.3)$$

The variance of the estimator $\hat{\mu}_z$ is given by,

$$Var(\hat{\mu}_z) = \frac{1}{n} \left[\sigma_A^2 + (1 + \psi^2) \{ (1 - P_1)a_2b_2 + P_1a_1b_1 \} \right]. \quad (3.4)$$

4. Efficiency comparison

(i) Model I versus Ryu et al (2005) Model

Our proposed estimator based on RRM I will be more efficient than that of Ryu et al (2005) estimator if

$$Var(\hat{\mu}_R) - Var(\hat{\mu}_J) \geq 0,$$

or if

$$\frac{1}{n} \left[\sigma_A^2 + (\mu_A^2 + \sigma_A^2)(1 - P)(1 - T)\psi^2 \right] - \frac{1}{n} \left[\sigma_A^2 + (1 - P)ab(1 + \psi^2) \right] \geq 0,$$

or if

$$(\mu_A^2 + \sigma_A^2)(1 - P)(1 - T)\psi^2 - (1 - P)ab(1 + \psi^2) \geq 0$$

Or if

$$(\mu_A^2 + \sigma_A^2)(1 - T)\psi^2 - ab(1 + \psi^2) \geq 0 \quad (4.1)$$

The condition (4.1) is very much likely to be true in most of the survey situations because a and b are controllable and can be chosen very small. e.g. $a = 0.0001$ and $b = 0.0003$.

(ii) Model II versus Ryu et al (2005) RRM

The estimator based on the RRM II will be more efficient than that of Ryu et al (2005)

estimator if $Var(\hat{\mu}_R) - Var(\hat{\mu}_Z) \geq 0$,

or if

$$\frac{1}{n} \left[\sigma_A^2 + (\mu_A^2 + \sigma_A^2)(1-P)(1-T)\psi^2 \right] - \frac{1}{n} \left[\sigma_A^2 + ((1-P_1)a_2b_2 + P_1a_1b_1)\psi^2 \right] \geq 0,$$

or if

$$(\mu_A^2 + \sigma_A^2)(1-P)(1-T) - ((1-P_1)a_2b_2 + P_1a_1b_1) \geq 0. \quad (4.2)$$

The inequality (4.2) can be made true by suitably setting the values of the a_1, b_1, a_2 , and b_2 .

If the restriction $a_1b_1 = a_2b_2$ is imposed then inequality (4.2) reduces to

$$(\mu_A^2 + \sigma_A^2)(1-P)(1-T) - (a_2b_2) \geq 0,$$

this can be made true more easily.

Table 1 : Relative efficiencies of Model I and Model 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 $n = 1000$.

		P_0					
μ_A	σ_A^2	T_0	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

5 Important notes and discussion

(1). From equations (3.2) and (3.4) we can see that the variance of the proposed estimator does not depend upon the unknown true value of the population mean μ_A .

(2). In both the proposed models we do not need to set $\mu_s = 1$ (as is the case with Ryu et al (2005), and Gupta et al (2002) RR models.) to make the estimator unbiased. Instead, we can select any larger value (negative or positive) of μ_s as far as the unbiased ness is concerned.

(3). In Model I, if $P_1 = 1$, all the respondents will be using R_1 . In this case estimator remains the same but its variance changes to

$$Var(\hat{\mu}_z) = \frac{1}{n} \left[\sigma_A^2 + (1 + \psi^2) a_1 b_1 \right]. \quad (3.5)$$

This change in variance is due to restricting the extent of the randomization up to one stage. Similarly if $P_1 = 0$, all the respondents will be using R_2 and again estimator remains same but its variance changes to

$$Var(\hat{\mu}_z) = \frac{1}{n} \left[\sigma_A^2 + (1 + \psi^2) a_2 b_2 \right]. \quad (3.6)$$

(4). We can reduce the variance of our proposed estimator $\hat{\mu}_z$ without cutting down the stages of randomization by putting a restriction that $a_1 b_1 = a_2 b_2$. With this restriction imposed, the variance of the proposed estimator reduces to

$$Var(\hat{\mu}_z) = \frac{1}{n} \left[\sigma_A^2 + (1 + \psi^2) a_2 b_2 \right]. \quad (3.7)$$

This strategy seems more logical because it provides more stages of randomization: one for selecting the randomization device itself and other for randomizing the responses. And even then the variance of the estimator remains the same as that of the variance obtained by the use of R_2 only. So this strategy provides more protection against the privacy of the respondents and in result the rate of response is increased. It is interested to

note that if we are imposing the restriction $a_1 b_1 = a_2 b_2$, then a different RRM can be developed. Respondents may be requested to report the two responses U_{1i} and U_{2i} . The two estimators of μ_A may then be defined with equal variances. Taking advantage of this equality in variances we can define a weighted estimator of μ_A as

$$\hat{\mu}_{AW} = \frac{1}{2} \frac{\sum_{i=1}^n U_{1i}}{n} + \frac{1}{2} \frac{\sum_{i=1}^n U_{2i}}{n}, \quad (3.8)$$

where 0.5 is the optimum weight attached with each estimator. The variance of the weighted estimator will then be

$$Var(\hat{\mu}_{AW}) = \frac{1}{2n} \left[\sigma_A^2 + (1 + \psi^2) a_2 b_2 \right]. \quad (3.9)$$

The estimator given by (3.8) is very much similar to the estimator proposed by Hussain et al (2007) but the RD used in it is different and, apparently, provides more privacy.

(5). As S_i could be chosen any real valued random variable, any of the responses $A_i + b_1 S_i$, $A_i - a_1 S_i$, $A_i + b_2 S_i$, and $A_i - a_2 S_i$ can not be traced back to the actual response of the respondents.

(6). Smaller values of a_1, b_1, a_2 , and b_2 can be easily and fruitfully set to have any desired values of the probabilities T_1 and T_2 . For example if in a particular survey a value of $T_2 = 0.333$ is desired we can choose $a_2 = 0.01$ and $b_2 = 0.02$ or even smaller as $a_2 = 0.001$ and $b_2 = 0.002$. This shows that variance can be reduced to the desired level. As a_i 's and b_i 's are small numbers, we must choose the random variable S which assumes larger values so that the reported responses would be very much different from the actual responses.

The proposed Models I and II are equally efficient compared to Ryu et al (2005) RRM for smaller values of a_i 's and b_i 's but Model II may be preferred because it provides more stages of Randomization and protection against privacy. That is, Model II preserves the anonymity of the survey respondents and therefore can be expected to result in a greater cooperation from the respondents.

Although the estimator given by (3.8) is unbiased and has as small variance as half of the variance of the estimator based on the use of the randomization device R_2 only, its application in field surveys may be problematic because the individuals in the samples may get annoyed / irritated of reporting again and again. (for that matter twice). Therefore respond preferably should not be asked to report multiple answers.

References

- Bar-Lev S K, Bobovitch E, and Boukai B (2004) A note on randomized response models. *Metrika* 60: 255-260.
- Chaudhuri A & R Mukerjee (1988) *Randomized response: theory and techniques*. Marcel Dekker, New York.
- Eichhorn B H and 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, Kuebler R R Jr, Abernathy J R, and Hovertz D G (1971) Application of the randomized response techniques in obtaining quantitative data. *Journal of the American Statistical Associations* 66: 243-250.
- Gjestvang C R & Singh S (2006) A new randomized response model. *Journal of Royal Statistical Society Series. B* 68: 523-530.

Greenberg G, Kuebler R R Jr, Abernathy R, and Hovertz D G (1969) The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Associations* 64: 520-539.

Gupta S, Gupta B, and Singh S (2002) Estimation of Sensitivity level of personal interview survey questions. *Journal of Statistical Planning and inference* 100: 239-247.

Hong K, Yum J, and Lee H (1994) A stratified randomized response technique. *Korean Journal of Applied Statistics* 7: 141-147.

Horvitz D G, Shah B V, Simmmons W R (1967) The unrelated question randomized response model. *Proceedings of Social Statistics Sec. American Statistical Associations*: 65-72.

Hussain Z, Shabbir J, & Gupta S (2007) An alternative to Ryu et al randomized response model. *journal of Statistics & Management Systems*.(accepted).

Hussain Z, and Shabbir J (2007) Randomized use of Warner's randomized response model. *InterStat*: April # 7.

Kuk A Y C (1990) Asking sensitive questions directly. *Biometrika* 77: 436-438.

Moors J J A (1971) Optimization of the unrelated question randomized response model. *Journal of the American Statistical Associations* 66: 627-629.

Mangat N S & Singh R (1990) An alternative randomized response procedure. *Biometrika* 77: 439-442.

Mangat N S (1994) An improved randomized response strategy. *Journal of Royal Statistical Society Series B* 56: 93-95.

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

Raghavarao D (1978) On an estimation problem in Warner's randomized response technique. *Biometrics* 34: 87-90.

Singh S, R Singh, & Mangat N S (2000) Some alternative strategies to Moor's model in randomized response sampling. *Journal of Statistical Planning and inference* 83: 243-255.

Singh S, Mahmood M, & 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.

Singh S (1999) An addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn. *Biometrical journal* 41(8): 955-966.

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