J. Ind. Soc. Agril. Statist.
61(2), 2007 : 164-168

On the Estimation of Population Mean and Sensitivity in Two-Stage Optional Randomized Response Model

Sat Gupta and Javid Shabbir[1]
Department of Mathematics and Statistics, University of North Carolina, Greensboro, USA

SUMMARY

The paper discusses a two-stage optional randomized response model and presents estimators for the mean and sensitivity level of a sensitive question. A simulation study is used to assess the validity of the proposed estimators.

Key words : Two-stage randomized response models, Efficiency comparison, Optional scrambling, Sensitivity estimation, Simulation study, Unbiased estimation.

1. INTRODUCTION

Randomized response techniques (RRT) have been widely used for personal interview surveys ever since the pioneering work of Warner (1965). Umesh and Peterson (1991), Scheers (1992) and Hosseini and Armacost (1993), among others, have shown that RRT methods do, in fact, lead to more valid answers and prove effective in circumventing social desirability bias. Several randomized response models have been developed by researchers for collecting data on both the qualitative and quantitative variables. Greenberg *et al.* (1971) have proposed the unrelated question model for estimating the mean and the variance of a sensitive quantitative variable. Mangat and Singh (1990) have introduced a two-stage RRT model for estimating the prevalence of a sensitive trait in a binary population and showed that an improvement over the Warner (1965) model is possible. Eichhorn and Hayre (1983) discussed a multiplicative RRT model for quantitative responses. Gupta *et al.*(2002) modified the Eichhorn and Hayre (1983) model and introduced an optional RRT model for estimating simultaneously the mean as well as the sensitivity level of a sensitive variable. They defined sensitivity to be the proportion of subjects in a population who consider a given question to be of sensitive nature. The argument put forward by Gupta *et al.* (2002) was that a question may be sensitive to one person but may not be sensitive to another.

Ryu *et al.* (2006) have recently attempted to combine the models introduced by Mangat and Singh (1990) and by Gupta *et al.* (2002) to introduce an estimator of the mean of a quantitative sensitive variable and show that their estimator of the mean is more efficient than that of the estimator by Gupta *et al.*(2002). However, the model by Ryu *et. al.* (2006) is not an optional RRT model and does not estimate the sensitivity level of the sensitive question, unlike Gupta *et al.* (2002) whose model estimated simultaneously both mean and sensitivity of the sensitive variable.

The focus of this paper is on introducing a truly optional two-stage RRT model and on comparing it with the Ryu *et al.* (2006) model.

2. THEORETICAL FRAMEWORK

We first discuss a few relevant RRT models for quantitative response.

Eichhorn and Hayre Model

Let X be the true response and S be some scrambling variable, independent of X, with mean θ and standard deviation σ_S. The respondent is asked to report the response Z as given by

[1] *Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan*

$$Z = \frac{SX}{\theta}$$

Since E(Z) = E(X), an unbiased estimator of the mean μ_x of X is given by $\bar{Z}$, the sample mean of the reported responses. It is easy to verify that

$$V(\hat{\mu}_1) = \frac{1}{n}[\sigma_x^2 + (\frac{\sigma_s}{\theta})^2\{\sigma_x^2 + \mu_1^2\}] \qquad (2.1)$$

Gupta, Gupta and Singh Model

In this model, each respondent selects one of the following two options. The interviewer does not know which option has been chosen.

(a) Report the true response if the question is perceived as non-sensitive.

(b) Report a scrambled response Z = SX if the question is perceived as sensitive.

Again, X is the true response and S is some scrambling variable, independent of X, with mean of $\theta = 1$ and standard deviation σ_s. If W is the proportion of respondents who consider the question sensitive and choose to provide a scrambled response, then Z can be expressed as

$$Z = S^Y X, \text{ where } Y \sim \text{Bernoulli (W)}$$

Gupta *et al.* (2002) call W the sensitivity level of the underlying question and estimate both W and μ_x.

It can be verified easily that E(Z) = E(X), leading to an unbiased estimator of the population mean μ_x given by

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

The variance of the proposed estimator $\hat{\mu}_2$ is given by

$$V\left(\hat{\mu}_2\right) = \frac{1}{n}\left[\sigma_x^2 + W\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)\right] \qquad (2.2)$$

It is easy to note that $V(\hat{\mu}_2)$ increases as W increases from 0 to 1.

The relative efficiency of this estimator with respect to the estimator $\hat{\mu}_1$ of Eichhorn and Hayre (1983) is given by

$$RE = \frac{\sigma_x^2 + \gamma^2\left(\sigma_x^2 + \mu_x^2\right)}{\sigma_x^2 + W\gamma^2\left(\sigma_x^2 + \mu_x^2\right)} \qquad (2.3)$$

where $\gamma^2 = V(S)$

Note that RE ≥ 1 since $0 \le W \le 1$.

Gupta *et al.* (2002) also provided an estimator for W using a first order approximation. This is given by

$$\hat{W}_{G1} \approx \frac{\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i) - \ln\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)}{\delta} \qquad (2.4)$$

where $\delta = E[1N(S)]$ denotes the known expected value of the logarithm of the scrambling variable.

Ryu *et al.* (2006) Model

Ryu *et al.* (2006) introduce the following two-stage model.

Stage 1: A randomly selected, pre-determined proportion (T) of the respondents, respond truthfully.

Stage 2: Among rest of the respondents (proportion (1–T)), a known proportion W of the respondents provide randomized response SX and the rest provide a true response.

The key point to be noted here is that in this model, unlike Gupta *et al.* (2002) model, W is assumed to be known. Only μ_x is estimated. Ryu *et al.* (2006) show that E(Z) = E(X), where Z is the reported response. This leads to an unbiased estimator of μ_x given by

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

It can be verified that

$$V\left(\hat{\mu}_3\right) = \frac{1}{n}\left[\sigma_x^2 + (1-T)(W)\sigma_S^2\left(\sigma_x^2 + \mu_x^2\right)\right] \qquad (2.5)$$

Recall that the variance for the Gupta *et al.* (2002) model is given by

$$V\left(\hat{\mu}_2\right) = \frac{1}{n}\left[\sigma_x^2 + W\sigma_S^2\left(\sigma_x^2 + \mu_x^2\right)\right]$$

Clearly, $V(\hat{\mu}_3) \le V(\hat{\mu}_2)$, but one should note that Ryu *et al.* (2006) estimate only μ_x and do not

simultaneously estimate the sensitivity level W that was the main aspect of Gupta *et al.* (2002) model.

3. PROPOSED TWO-STAGE OPTIONAL RRT MODEL

Note that the response in the Ryu *et al.* (2006) model is given by

$$Z = \begin{cases} X & \text{with probability } T + (1-T)(1-W) \\ SX & \text{with probability } (1-T)W \end{cases} \quad (3.1)$$

But this is not different than the partial RRT model for quantitative responses, discussed by Gupta and Thornton (2002), if W is assumed known, except that the proportion of respondents providing truthful responses has been increased from T to T + (1–T)(1-W).

We now discuss a true two-stage optional RRT model.

Stage 1: A randomly selected, pre-determined proportion (T) of the respondents, respond truthfully. Other respondents are instructed to go to Stage 2.

Stage 2: These respondents are asked to provide a randomized response SX if they think the question to be sensitive. Otherwise they are asked to report the true response X.

The interviewer does not know in which stage and how the response is provided.

The reported response under this model is given by

$$Z = \{X^V\}\ \{X\ S^U\}^{1-V} \quad (3.2)$$

where U ~ Bernoulli(W), V ~ Bernoulli(T)

We assume that E(S) = 1 and X, S, U, V are mutually independent.

Taking expected value on both sides of (3.2), we get

$$\begin{aligned} E(Z) &= E(X.S^U).P(V=0) + E(X).P(V=1) \\ &= E(X).\{E(S)\,P(U=1) + P(U=0)\}P(V=0) \\ &\quad + E(X).P(V=1) \\ &= E(X)P(V=0) + E(X)P(V=1), \text{ Since } E(S)=1 \\ &= E(X) \end{aligned}$$

Hence, μ_X can be estimated by

$$\hat{\mu}_X = \frac{\sum Z_i}{n} \quad (3.3)$$

It can be verified, as in Ryu *et al.* (2006) that the variance of this estimator is given by

$$V(\hat{\mu}_X) = \frac{1}{n}\left[\sigma_X^2 + (1-T)(W)\sigma_S^2\left(\sigma_X^2 + \mu_X^2\right)\right] \quad (3.4)$$

As noted earlier, this is smaller than the variance in (2.2) corresponding to the one-stage optional RRT model given by Gupta *et al.* (2002). But this is clearly on expected lines since in a two-stage model, a greater proportion of respondents are asked to provide truthful responses. However, this gain will be offset, as shown below, in estimating the sensitivity level W.

The sensitivity level W can be estimated by proceeding as in Gupta *et al.* (2002), except that we will use a second order approximation here. Taking natural log of both sides of (3.2), we get

$$\begin{aligned} \ln(Z) &= V.\ln(X) + (1-V)\{\ln(X) + U.\ln(S)\} \\ &= \ln(X) + (1\text{-}V).U.\ln(S) \end{aligned} \quad (3.5)$$

Taking expected values on both sides of (3.5), we get

$$\begin{aligned} E[\ln(Z)] &= E[\ln(X)] + E(1-V).E(U).E[1N(S)] \\ &= E[\ln(X)] + (1-T)(W)\delta \end{aligned} \quad (3.6)$$

where $\delta = E[\ln(S)]$

Rewriting (3.6), we get

$$W = \frac{E[\ln(Z)] - E[\ln(X)]}{(1-T)\delta} \quad (3.7)$$

Now proceeding as in Gupta *et al.* (2002), E[In(Z)] can be approximated by $\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i)$.

For the second term in the numerator, E[In(X)], we use a second order Taylor's approximation and write

$$\ln(X) \approx \ln(\mu_X) + (X-\mu_X)\frac{1}{\mu_X} - \frac{(X-\mu_X)^2}{2\mu_X^2} \quad (3.8)$$

Taking expectation, we get

$$E[\ln(X)] \approx \ln(\mu_X) - \frac{1}{2}\frac{V(X)}{\mu_X^2} \quad (3.9)$$

In (3.9), we can use the approximation $\hat{\mu}_X = \frac{\sum Z_i}{n}$

As for $V(X)$, note that

$$V(Z) = E(Z^2) - \mu_x^2$$

From (3.1), it can be verified that

$$E(Z^2) = E(X^2)[1-(1-T)W\{1-E(S^2)\}]$$

The factor (1–T) W is expected to be small, being the product of two fractions, and can be made even smaller by choosing a large value of T. Hence, $E(Z^2) \approx E(X^2)$, and consequently, $V(Z) \approx V(X)$. With this, the approximation in (3.9) can be written as

$$E[\ln(X)] \approx \frac{1}{n}\sum_{i=1}^{n} \ln(Z_i) - \frac{1}{2}\frac{V(Z)}{\mu_z^2} \qquad (3.10)$$

Substituting this in (3.7), we can estimate W, in the two-stage model by

$$\hat{W}_{G2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i) - \ln\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) - \frac{\hat{V}(Z)}{2\hat{\mu}_z^2}}{(1-T)\delta} \qquad (3.11)$$

Note that the variance of the two-stage estimator of the sensitivity in (3.11) is likely to be larger than the variance for the one-stage estimator in (2.4) because of the term (1-T) in the denominator of (3.11). Thus the gain in estimation of the mean in the two-stage model is somewhat neutralized in estimating the sensitivity level.

4. SIMULATION RESULTS

We now report some simulation results based on 10,000 iterations using samples of size 100 and 500. Statistical software package SAS is used for running the simulations. We have used $X \sim \text{Poisson}(\lambda = 5)$ and $S \sim \chi^2(1)$.

By looking at the estimated variances $\hat{V}(\hat{\mu})$ and $\hat{V}(\hat{W})$, one can note that the estimation of the mean in a two-stage model is better as compared to a one-stage model but the estimation of the sensitivity in the two-stage model is less precise as compared to the one-stage model.

Table 1. One stage mean and sensitivity estimation model

n	W	$\hat{\mu}$	$\hat{V}(\hat{\mu})$	$\hat{W}$	$\hat{V}(\hat{W})$
100	0.1	5.0004	0.09814	0.0981	0.00305
	0.3	5.0029	0.21622	0.2887	0.00830
	0.5	5.0028	0.32962	0.4856	0.01257
	0.7	5.0060	0.44784	0.6832	0.01635
500	0.1	5.0025	0.01929	0.0918	0.00062
	0.3	5.0033	0.04182	0.2910	0.00170
	0.5	5.0040	0.06641	0.4914	0.00258
	0.7	5.0041	0.09156	0..6915	0.00330

Table 2. Two stage mean and sensitivity estimation model with T = 0.1

n	W	$\hat{\mu}$	$\hat{V}(\hat{\mu})$	$\hat{W}$	$\hat{V}(\hat{W})$
100	0.1	5.0015	0.09300	0.0901	0.00349
	0.3	5.0010	0.19914	0.2878	0.00935
	0.5	5.0031	0.30510	0.4853	0.01473
	0.7	5.0045	0.40486	0.6823	0.01839
500	0.1	5.0025	0.01800	0.0909	0.00069
	0.3	5.0024	0.03803	0.2902	0.00190
	0.5	5.0031	0.05900	0.4902	0.00297
	0.7	5.0037	0.08278	0.6907	0.00376

Table 3. Two stage mean and sensitivity estimation model with T = 0.3

n	W	$\hat{\mu}$	$\hat{V}(\hat{\mu})$	$\hat{W}$	$\hat{V}(\hat{W})$
100	0.1	5.0009	0.08086	0.0872	0.00449
	0.3	5.0021	0.16285	0.2851	0.01226
	0.5	5.0050	0.24499	0.4834	0.01947
	0.7	5.0053	0.33025	0.6809	0.02600
500	0.1	5.0021	0.01586	0.0881	0.00090
	0.3	5.0032	0.03133	0.2876	0.00253
	0.5	5.0035	0.04766	0.4872	0.00402
	0.7	5.0034	0.06349	0.6875	0.00525

5. CONCLUSION

If the estimation objective is to only estimate the mean of the sensitive variable, then a two-stage model does produce a smaller variance as compared to a one-stage model. However, one should note that in such a situation, the two-stage model of Ryu *et al.* (2006) is same as the "partial RRT model" discussed by Gupta and Thornton (2002). If the objective is to also estimate the sensitivity level of the sensitive question, an optional RRT model, such as the one discussed in (3.2) is needed. But one should be awared of that the gain in estimation of the mean may be somewhat nullified in estimating the sensitivity level.

REFERENCES

Eichhorn, B.H. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Plann. Inf.*, **7**, 307-316.

Greenberg, B.G., Keubler, R.T., Jr., Abernathy, J.R., and Horvitz, D.G. (1971). Application of randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.*, **66**, 243-250.

Gupta, S.N., Gupta, B.C. and Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *J. Statist. Plann. Inf.*, **100**, 239-247.

Gupta, S.N., and Thornton, B. (2002). Circumventing social desirability response bias in personal interview surveys. *Amer. Jour. Math. Manag. Sci.*, **22**, 369-383.

Hossiene, J.C. and Armacost R.L. (1993). Gathering sensitive information in organization. *American Behavioral Scientist*, **36**, 443-471.

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

Ryu, J.B., Kim, J.M., Heo, T.Y. and Park, C.G. (2006). On stratified randomized response sampling. *Model Assisted Stat. Appl.*, **1**, 31-36.

Scheers, N. (1992). A review of randomized response technique. *Measurement Evaluation Counselling Dev.*, **25**, 27-41.

Umesh, U.N. and Peterson, R.A. (1991). A critical evaluation of the randomized response method: Application, validation and research agenda. *Socio. Methods Res.*, **20**, 104-138.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60**, 63-69.