Journal of Statistical Research *ISSN 0256 - 422 X*
2007, Vol. 41, No. 2, pp. 83-92
Bangladesh

ESTIMATION OF MEAN OF A SENSITIVE QUANTITATIVE VARIABLE

ZAWAR HUSSAIN
Department of Statistics, Quaid–i–Azam University 45320, Islamabd 44000, Pakistan
Email: zhlangah@yahoo.com

JAVID SHABBIR
Department of Statistics, Quaid–i–Azam University 45320, Islamabd 44000, Pakistan
Email: jsqau@yahoo.com

SUMMARY

The present study introduces an unbiased estimator of population mean of a sensitive quantitative variable based on multiple selections of numbers from a scrambling distribution to confound the actual response on sensitive variable with some unrelated variable. The proposed method may be viewed as more protective against the reprisal or stigma. A simulation study is done to observe the performance of the proposed method relative to Ryu et al. (2005) Randomized Response (RR) method. The relative efficiency of proposed estimator with respect to Ryu et al. (2005) RR method is calculated and found appreciable.

Keywords and phrases: randomized response technique, sensitive quantitative variable, estimation of mean, evasive answer bias.

AMS Classification: 94A20

1 Introduction

One may doubt the truth of the responses gathered by direct questioning while conducting a survey to estimate the intensity of a sensitive attribute or mean of a sensitive quantitative variable (e.g. induced abortion, drug usage, tax evasion, shop lifting, cheating in exams, sexual abuse, etc.) in a population. The reason of falsified answers might be the fear of reprisals, getting punishment from the authorities, or simply the embarrassment that's why, in social surveys, lack of a reliable measure of incidence or prevalence of both qualitative and quantitative variable becomes serious issue. Generally, individuals in the population avoid to be stigmatized and fear of reprisals by revealing the truth to the strangers. This usually results in lying by the respondents when approached with the conventional or direct–response

survey method. As a consequence of falsification of the responses an avoidable estimation bias creeps into the estimates. In an effort to reduce the evasive answer bias, Warner (1965) proposed a randomized response method to estimate proportion of prevalence of the sensitive characteristic 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 et al. (1998), Singh (1999), Singh et al. (2001), Chang and Haung (2001),Gupta et al. (2002), Bar–Lev et al. (2004), Ryu et al. (2005) and many others.

In this paper we present an unbiased estimator of the mean and compare it with the estimator proposed by the Ryu et al. (2005). The organization of the paper is as follows: The Ryu et al. (2005) estimation procedure is outlined in Section 2. In Section 3 we present our proposed model. Section 4 contains the efficiency comparison of the proposed procedure. A short discussion is given in Section 5.

2 Ryu et al. Proposed Model

Ryu et al. (2005) proposed a randomized response (RR) model to estimate the mean of the sensitive quantitative variable, based on Mangat and Singh (1990) two–stage randomized response model. 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 A_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 A_i of sensitive question" and (ii) "Report the scrambled response A_iS_i of sensitive question" represented with probabilities T and 1–T respectively. Let Y_i be the response of the i^{th} respondent, then it can be written as

$$Y_i = \alpha A_i + (1-\alpha)[\beta A_i + (1-\beta)A_iS_i], \tag{2.1}$$

where $\alpha = 1$, if a respondent is randomly assigned to statement (i) in R_1, and $\alpha = 0$, if a respondent is randomly assigned to statement (ii) in R_1. Also $\beta = 1$, if a respondent is randomly assigned to statement (i) in R_2, and $\beta = 0$, if a respondent is randomly assigned to statement (ii) in R_2. Also α and β are variables with means P, T and variances $P(1–P)$, $T(1–T)$, respectively. Using the assumption of known distribution of scrambling variable such that $\mu_S = 1$ and $\sigma_S^2 = \theta^2$, the expected value of the observed response is given by

$$\begin{aligned} E(Y_i) &= E(\alpha)E(A_i) + E(1-\alpha)[E(\beta)E(A_i) + E(1-\beta)E(A_i)E(S_i)] \\ &= P\mu_A + (1-P)[T\mu_A + (1-T)\mu_A] \\ &= \mu_A. \end{aligned} \tag{2.2}$$

Based on the responses $Y_i, i = 1, 2, \ldots, n$, Ryu et al. (2005) suggested an unbiased estimator of the mean of the sensitive variable as

$$\hat{\mu}_{AR} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{2.3}$$

with variance

$$V(\hat{\mu}_{AR}) = \frac{\sigma_A^2}{n} + \frac{1}{n}(\mu_A^2 + \sigma_A^2)(1-P)(1-T)\theta^2. \tag{2.4}$$

3 Proposed Procedure

In the proposed procedure each respondent selected in the sample of size n is requested to select k random numbers $S_1, S_2, \ldots, S_k$ from a probability distribution with any known mean $-\infty < \mu_S < \infty$ and known variance $\sigma_S^2 = \theta^2$, and requested to report mean of selected numbers plus k^k times of his/her actual response A_i. (We included the term k^k to be multiplied with A_i because it plays a role of decreasing the variance of the estimator to be proposed). Let us suppose that d_i be the response of the i^{th} respondent, it can then be written as

$$d_i = \frac{1}{k}\sum_{j=1}^{k} S_j + k^k A_i. \tag{3.1}$$

For the i^{th} respondent we have

$$E(d_i) = \mu_S + k^k \mu_A. \tag{3.2}$$

This suggests defining an unbiased estimator of the population mean μ_A as

$$\hat{\mu}_{AP} = \frac{[\bar{d} - \mu_S]}{k^k}. \tag{3.3}$$

To derive an expression for variance of the proposed estimator, we proceed as follows. By definition

$$\begin{aligned} V(d_i) &= E(d_i^2) - (E(d_i))^2 \\ &= E[\frac{1}{k}\sum_{j=1}^{k} S_j + k^k A_i]^2 - [E[\frac{1}{k}\sum_{j=1}^{k} S_j + k^k A_i]]^2 \\ &= \frac{\psi^2}{k} + k^{2k}\sigma_A^2. \end{aligned} \tag{3.4}$$

where $\psi^2 = \theta^2 = \sigma_S^2$. Hence the variance of the proposed estimator $\hat{\mu}_{AP}$ is given by

$$\begin{aligned} V(\hat{\mu}_{AP}) &= V(\frac{\bar{d}}{k^{2k}}) \\ &= V(\frac{\bar{d}}{nk^{2k}}) \\ &= \frac{\sigma_A^2}{n} + \frac{\psi^2}{nk^{2k+1}}. \end{aligned} \tag{3.5}$$

From (3.5) we can observe that the variance of the proposed estimator is a bounded decreasing function of the constant k , therefore, no optimum k can be derived to obtain the

minimal estimator in the proposed class of estimators. By design, minimum value of k is 2. For practical purposes one can use k=3 or 4 also, but k=2 works well as we will show in the Section 4. Instead of fixing $\mu_S = 1$, any value of the mean of the scrambling variable, μ_S can be set depending on the size of the values of the study variable. So the proposed model provides more flexibility in choosing the scrambling variable. If the values of the study variable in the population are suspected to be larger, one can set μ_S larger, otherwise smaller. It may help decreasing the suspicion amongst the respondents to confound the actual response with scrambling variable. Also, the variance of the proposed estimator does not depend on the unknown mean of the study variable, as is the case with the Ryu et al. RR model. Unlike the Ryu et el. estimator, the variance of the proposed estimator is not affected by the unknown mean of the study variable.

4 Efficiency Comparison

The proposed estimator is more efficient than Ryu et al. (2005) estimator if

$$V(\hat{\mu}_{AP}) - V(\hat{\mu}_{AP}) \geq 0,$$

that is

$$\frac{\sigma_A^2}{n} + \frac{1}{n}(\mu_A^2 + \sigma_A^2)(1-P)(1-T)\psi^2 - \frac{\sigma_A^2}{n} - \frac{\psi^2}{nk^{2k+1}} \geq 0,$$

or if

$$(\mu_A^2 + \sigma_A^2)(1-P)(1-T) \geq \frac{1}{k^{2k+1}} \tag{4.1}$$

The inequality (4.1) can easily be made true by suitably choosing the constant k to achieve the desired efficiency.

4.1 Numerical Example

In the following Tables 1–5 we give the relative efficiency of proposed estimator relative to the Ryu et al. (2005) estimator for different practicable values of P, T and μ_A . We fix $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

4.2 Simulation Study

To study the behavior of the proposed RR model compared to Ryu et al (2005) RR model we performed a simulation study for some parametric values and sample size $n = 100$. We assumed that the sensitive variable A_i follows a Gamma(1,2) distribution , so that the mean of the study variable is 2. The scrambling variable S_i is assumed to have a standard normal distribution. For $k = 3$ and $P, T = 0.1, 0.3, 0.5, 0.7, 0.9$. we performed 5000 simulations and the simulated means and standard deviations of the both the proposed estimator and the Ryu et al estimator are given in the following Tables 6–10.

Table 1: Relative efficiency of the $\hat{\mu}_{AP}$ relative to $\hat{\mu}_{AR}$ when $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

	μ_A						μ_A				
P	T	2	4	6	8	P	T	2	4	6	8
	0.1	4.30	13.73	29.44	51.43		0.1	3.93	12.31	26.27	45.82
	0.2	3.93	12.31	26.27	45.82		0.2	3.60	11.05	23.46	40.84
	0.3	3.56	10.59	23.11	40.22		0.3	3.27	9.79	20.65	35.85
	0.4	3.19	9.47	19.95	34.61		0.4	2.94	8.53	17.84	30.87
0.1	0.5	2.82	8.06	16.78	29.00	0.2	0.5	2.61	7.27	15.03	25.89
	0.6	2.45	6.64	13.62	23.39		0.6	2.28	6.01	12.21	20.90
	0.7	2.08	5.22	10.46	17.79		0.7	1.95	4.75	9.40	15.92
	0.8	1.71	3.80	7.29	12.18		0.8	1.62	3.49	6.59	10.93
	0.9	1.34	2.38	4.13	6.57		0.9	1.29	2.23	3.78	5.95

Table 2: Relative efficiency of the $\hat{\mu}_{AP}$ relative to $\hat{\mu}_{AR}$ when $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

	μ_A						μ_A				
P	T	2	4	6	8	P	T	2	4	6	8
	0.1	3.56	10.89	23.11	40.22		0.1	3.19	9.47	19.95	34.61
	0.2	3.29	9.79	20.65	35.85		0.2	2.94	8.53	17.84	30.87
	0.3	2.98	8.69	18.19	31.49		0.3	2.70	7.58	15.73	27.13
	0.4	2.70	7.58	15.73	27.13		0.4	2.45	6.64	13.62	23.39
0.3	0.5	2.41	6.48	13.27	22.77	0.4	0.5	2.20	5.69	11.51	16.66
	0.6	2.12	5.38	10.81	18.41		0.6	1.95	4.75	9.40	15.92
	0.7	1.83	4.27	8.35	14.05		0.7	1.71	3.80	7.29	12.18
	0.8	1.54	3.17	5.89	9.69		0.8	1.46	2.86	5.18	8.44
	0.9	1.25	2.07	3.43	5.33		0.9	1.21	1.91	3.07	4.70

Table 3: Relative efficiency of the $\hat{\mu}_{AP}$ relative to $\hat{\mu}_{AR}$ when $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

	μ_A						μ_A				
P	T	2	4	6	8	P	T	2	4	6	8
	0.1	2.82	8.06	16.78	29.00		0.1	2.45	6.64	13.62	23.39
	0.2	2.61	7.27	15.03	25.89		0.2	2.28	6.01	12.21	20.90
	0.3	2.41	6.48	13.27	22.77		0.3	2.12	5.38	10.80	18.41
	0.4	2.20	5.69	11.51	19.66		0.4	1.95	4.75	9.40	15.92
0.5	0.5	2.00	4.90	9.75	16.54	0.6	0.5	1.79	4.12	8.00	13.43
	0.6	1.79	4.12	8.00	13.43		0.6	1.62	3.49	6.59	10.93
	0.7	1.58	3.33	6.24	10.31		0.7	1.46	2.86	5.18	8.44
	0.8	1.38	2.54	4.48	7.20		0.8	1.29	2.23	3.78	5.95
	0.9	1.17	1.75	2.72	4.08		0.9	1.13	1.60	2.37	3.46

Table 4: Relative efficiency of the $\hat{\mu}_{AP}$ relative to $\hat{\mu}_{AR}$ when $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

	μ_A						μ_A				
P	T	2	4	6	8	P	T	2	4	6	8
	0.1	2.08	5.22	10.46	17.79		0.1	1.71	3.80	7.29	12.18
	0.2	1.95	4.75	9.40	15.92		0.2	1.62	3.49	6.59	10.93
	0.3	1.83	4.27	8.35	14.05		0.3	1.54	3.17	5.89	9.69
	0.4	1.71	3.80	7.29	12.18		0.4	1.46	2.86	5.18	8.44
0.7	0.5	1.58	3.33	6.24	10.31	0.8	0.5	1.38	2.54	4.48	7.20
	0.6	1.46	2.86	5.18	8.44		0.6	1.29	2.23	3.78	5.95
	0.7	1.34	2.38	4.13	6.57		0.7	1.21	1.91	3.07	4.70
	0.8	1.21	1.91	3.07	4.70		0.8	1.13	1.60	2.37	3.46
	0.9	1.09	1.44	2.02	2.83		0.9	1.05	1.28	1.67	2.21

Table 5: Relative efficiency of the $\hat{\mu}_{AP}$ relative to $\hat{\mu}_{AR}$ when $n = 100$, $\theta^2 = \psi^2 = 0.5, k = 2$ and $\sigma_A^2 = 0.5$.

			μ_A		
P	T	2	4	6	8
	0.1	1.34	2.78	4.13	6.57
	0.2	1.29	2.23	3.78	5.95
	0.3	1.25	2.07	3.43	5.33
	0.4	1.21	1.91	3.07	4.70
0.9	0.5	1.17	1.75	2.72	4.08
	0.6	1.13	1.60	2.37	3.46
	0.7	1.09	1.44	2.02	2.83
	0.8	1.05	1.28	1.67	2.21
	0.9	1.01	1.12	1.32	1.59

Table 6: Simulated Means and Variances of the $\hat{\mu}_{AR}$ and $\hat{\mu}_{AP}$ for $K = 3$ and $T = 0.1$

			$T = 0.1$	
P	$Mean(\hat{\mu}_{AR})$	$Stdev(\hat{\mu}_{AR})$	$Mean(\hat{\mu}_{AP})$	$Stdev(\hat{\mu}_{AP})$
0.1	1.9985	0.2298	1.9992	0.1992
0.3	2.0005	0.2244	2.0016	0.1995
0.5	2.0036	0.2204	2.0020	0.2033
0.7	1.9967	0.2158	1.9966	0.2022
0.9	1.9988	0.2143	1.9978	0.2023

Table 7: Simulated Means and Variances of the $\hat{\mu}_{AR}$ and $\hat{\mu}_{AP}$ for $K = 3$ and $T = 0.3$

			$T = 0.3$	
P	$Mean(\hat{\mu}_{AR})$	$Stdev(\hat{\mu}_{AR})$	$Mean(\hat{\mu}_{AP})$	$Stdev(\hat{\mu}_{AP})$
0.1	2.0015	0.2193	2.0003	0.2018
0.3	1.9992	0.2176	1.9979	0.2011
0.5	1.9964	0.2088	1.9980	0.1982
0.7	1.9968	0.2101	1.9957	0.2018
0.9	2.0022	0.2072	2.0013	0.2013

Table 8: Simulated Means and Variances of the $\hat{\mu}_{AR}$ and $\hat{\mu}_{AP}$ for $K = 3$ and $T = 0.5$

			$T = 0.5$	
P	$Mean(\hat{\mu}_{AR})$	$Stdev(\hat{\mu}_{AR})$	$Mean(\hat{\mu}_{AP})$	$Stdev(\hat{\mu}_{AP})$
0.1	1.9954	0.2068	1.9946	0.1977
0.3	2.0012	0.2017	2.0013	0.2027
0.5	2.0030	0.2062	2.0019	0.2003
0.7	1.9974	0.2011	1.9973	0.1968
0.9	2.0003	0.2017	2.0006	0.1994

Table 9: Simulated Means and Variances of the $\hat{\mu}_{AR}$ and $\hat{\mu}_{AP}$ for $K = 3$ and $T = 0.7$

			$T = 0.7$	
P	$Mean(\hat{\mu}_{AR})$	$Stdev(\hat{\mu}_{AR})$	$Mean(\hat{\mu}_{AP})$	$Stdev(\hat{\mu}_{AP})$
0.1	2.0009	0.2055	2.0010	0.204
0.3	1.9963	0.2045	1.9962	0.2021
0.5	1.9979	0.2018	1.9978	0.1998
0.7	2.0068	0.2033	2.0077	0.2019
0.9	2.0024	0.2018	2.0021	0.2011

Table 10: Simulated Means and Variances of the $\hat{\mu}_{AR}$ and $\hat{\mu}_{AP}$ for $K = 3$ and $T = 0.9$

			$T = 0.9$	
P	$Mean(\hat{\mu}_{AR})$	$Stdev(\hat{\mu}_{AR})$	$Mean(\hat{\mu}_{AP})$	$Stdev(\hat{\mu}_{AP})$
0.1	1.9983	0.2008	1.9984	0.2009
0.3	1.9996	0.2017	1.9995	0.2017
0.5	1.9953	0.1985	1.9954	0.1986
0.7	2.0037	0.2030	2.0035	0.2033
0.9	1.9988	0.1989	1.9985	0.1989

5 Discussion

A new randomized response procedure for estimating the mean of a quantitative sensitive quantitative variable is presented. It has been observed that the relative efficiency of proposed estimator increases as k increases for the same parametric values we used in computing the relative efficiency for $k = 2$ in Section 4. We also did a simulation study and observed that both the estimators are unbiased but the proposed estimator is more efficient. We did the same for increasing k and observed that relative efficiency of the proposed estimator increases. e..g. for $k = 3$ the relative efficiency of the proposed estimator is 1.33 and for $k = 4$ it is 1.36 keeping the other parameters fixed. We also did the simulation study for increased sample size and observed the similar result therefore we did not represent them in Section 4 to save the space. The efficiency condition given by (4.1) depends on the unknown mean and variance of the study variable. If it is suspected that mean as well as variance of the study variable is small, we suggest to use larger k, otherwise a smaller value of k should preferably be used . As S_i could be chosen any real valued random variable, any of the responses $\frac{1}{k}\sum_{i=1}^{n} S_j + k^k A_i$ can not be traced back to the actual response A_i. Also a greater flexibility is provided in choosing k. It is shown by Ryu et al (2005) that their estimator is superior to the estimators given by Greenberg et al (1971), Eichhorn and Hayre (1983), and Gupta et al (2002) in term of relative efficiency. Therefore, the proposed procedure is not less efficient than Greenberg et al (1971), Eichhorn and Hayre (1983), and Gupta et al (2002) estimators. Although the comparison of proposed estimator with Ryu et al (2005) estimator looks arbitrary but we made comparison on the premise that both estimators estimate the same parameter μ_A . As our k is fixed, Ryu et al can claim that they can choose P and T in such a way that their estimator has smaller variance but it is for the non–practicable values of $P$and T(e.g.$P > 0.9$ and $T > 0.9$). Ryu et al (2005) have shown that the estimator given in (2.3) is more efficient than Greenberg et al (1971), Eichhorn and Hayre (1983) and Gupta et al (2002) estimators. Therefore, we can conclude that our proposed estimator is not less efficient than Greenberg et al (1971), Eichhorn and Hayre (1983) and Gupta et al (2002) estimators. It is important to note that the proposed model provides more privacy as compared to Ryu et al. (2005) to the respondents by giving a choice of selecting $k \geq 2$ random numbers and multiply his/her actual response to k^k .

Acknowldgements

First auothor is thankful to the anonymous referee for his valuable comments and suggestions.

References

[1] Bar–Lev, S. K., Bobovitch, E., and Boukai, B (2004). A note on randomized response models. *Metrika,* **60***, 255–260.*

[2] Chang, H–J. and Haung, K–C (2001). A note on an addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn. *Biometrical Journal,***43(4)**, *497–500.*

[3] Eichorn, 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.*

[4] Greenberg, B. G., Kuebler, R. R., Jr., Abernathy, J. R., and Hovertiz, D. G (1971). Application of the randomized response techniques in obtaining quantitative data. *Journal of the American Statistical Association,***66**, *243–250.*

[5] 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.*

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

[7] Ryu, J.–B., Kim, J.–M. Heo, T.–Y., and Park, C. G (2005). On stratified randomized response sampling. *Model Assisted Statistics and Application,***1(1)** *1–6.*

[8] Singh, S., Horn, S., and Chowdhury, S (1998). Estimation of stigmatized characteristics of a hidden gang in a finite population. *Australian and New Zealand Journal of Statistics,* **40 (3)**, *291–297.*

[9] Singh, S(1999). An addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn. *Biometrical Journal,* **41(8)**, *955–966.*

[10] 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.*

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