Journal of Informatics and Mathematical Sciences
Volume 4 (2012), Number 3, pp. 311–317

http://www.rgnpublications.com

Estimation of Sensitive Characteristic using Two-Phase Sampling for Regression Type Estimator

M. Javed and M.L. Bansal

Abstract. The problem of underreporting and non response on the sensitive issues are very common in most of the surveys due to our social setup. The randomized response (RR) models reduce rates of non-response and biased response that would ensure respondents' privacy if they respond truthfully concerning personal questions. Here RR device is proposed for regression type estimator in which two independent samples are drawn from the population. The estimator of population mean of sensitive variable has been developed. Its bias and variance and optimum variance have also been derived.

Introduction

Warner (1965) in his pioneer model suggested the technique of randomized response where every person in the population belongs either to sensitive group A, or its compliment. Enquirer often feels embarrassed in asking direct questions about sensitive issues. Such questions might be about drug usage, tax evasion, history of induced abortion, illegal income etc. Horvitz *et al.* (1967) and Greenberg *et al.* (1971) have extended the Warner (1965) model to the case where the responses to the sensitive question are quantitative rather than a simple 'yes' or 'no'. The respondent selects one of the two questions by means of a randomization device. According to Greenberg *et al.* (1971), it is essential that the mean and variance of the responses to the unrelated question be close to those for the sensitive questions, otherwise, it will often be possible to recognize from the response which question was selected. Two sample single alternate RR model proposed by Greenberg and co-workers for obtaining the data on continuous type sensitive random variable is more practicable and easy in handling. If some auxiliary information is available then responses can be used in regression analysis by following Singh *et al.* (1996), Strachan *et al.* (1998), and Singh and King (1999). Recent publications on RR techniques among others are, Singh *et al.* (2000), Chaudhuri (2001, 2004), Singh (2002), Elffers *et al.* (2003), Huang

Key words and phrases. RR device; Ratio type estimator; Two-phase sampling; Bias and variance.

(2004), Javed and Grewal (2006), Grewal *et al.* (2006) and Sidhu *et al.* (2007) and Gjestvang and Singh (2006, 2007), Samuel (2008) and Carel *et al.* (2010).

Proposed Randomized Response Device

Here an RR device is proposed for ratio type estimator in which two independent samples, sample-I and sample-II of sizes 'n' and 'k' with replacement are drawn from the population. The respondents in the first sample are given the RR device which contained statements regarding 'A' based on sensitive characteristic X and 'Q_2' based on unrelated characteristic 'Y_2' with respective probabilities 'p' and '$(1-p)$'. Then the respondents are asked to answer the selected statement in terms of numerical figures without revealing to the interviewer which one of the two statements is answered. The respondents in the second sample are asked to answer in numerical figures (without using the RR device) for two direct statements 'Q_1' and 'Q_2' based on unrelated characteristics 'Y_1' and 'Y_2' respectively. The statements 'Q_1' and 'Q_2' are correlated. The characteristics 'Y_1'and 'Y_2'are unrelated with the sensitive character X.

Let

p	$=$	probability that sensitive question is selected by the first respondent in the first sample.
$1-p$	$=$	probability that non-sensitive question 'Q_2'is selected by the first respondent in the first sample.
	$=$	q
Z_i	$=$	observed response from individual 'i' in first sample.
X_i	$=$	response of individual 'i' in case he/she selects sensitive question through RR device in first sample.
Y_{2i}	$=$	response of individual 'i' in case he/ she selects alternate question 'Q_2' through RR device in first sample.
Y_{1j}	$=$	response of first question 'Q_1' directly asked from j-th respondent in second sample.
Y_{2j}	$=$	response of second question 'Q_2' directly asked from j-th respondent in second sample.

Estimator based on proposed randomized device

Here a regression type estimator for two-phase sampling is proposed using the suggested RR device.

A regression type estimator $\widehat{\mu}_{xlr}$ of mean $\widehat{\mu}_x$ is proposed in two-phase sampling is as below

$$\widehat{\mu}_{xlr} = \frac{1}{p}\left[p\bar{X} + (1-p)\bar{y}_2' - (1-p)\{\bar{y}_2 + \widehat{\beta}(\bar{y}_1' - \bar{y}_1\}\right] \tag{1}$$

where

$$\bar{y}_1 = \frac{1}{m}\sum_{j=1}^{m} y_{1j}, \quad \bar{y}_2 = \frac{1}{m}\sum_{j=1}^{m} y_{2j}, \quad \bar{y}_1' = \frac{1}{k}\sum_{j=1}^{k} y_{1j} \text{ and } \bar{y}_2' = \frac{1}{n}\sum_{i=1}^{n} y_{2i}, \quad (2)$$

$$\widehat{\beta} = \frac{s_{y_1y_2}}{s_{y_1}{}^2}, \quad (3)$$

$$s_{y_1y_2} = \frac{1}{m-1}\sum_{j=1}^{m}(y_{1j}-\bar{y}_1)(y_{2j}-\bar{y}_2), \quad s_{y_1}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{1j}-\bar{y}_1)^2 \text{ and}$$

$$s_{y_2}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{2j}-\bar{y}_2)^2. \quad (4)$$

Let

$$\epsilon_1 = \frac{\bar{y}_1}{\bar{Y}_1}-1, \quad \epsilon_2 = \frac{\bar{y}_1'}{\bar{Y}_1}-1, \quad \epsilon_3 = \frac{s_{y_1y_2}}{S_{y_1y_2}}-1 \text{ and } \epsilon_4 = \frac{s_{y_1}^2}{S_{y_1}^2}-1 \quad (5)$$

such that $E(\epsilon_3) = E(\epsilon_4) = 0$

$$E(\epsilon_1\epsilon_3) = \frac{\mu_{21}}{\bar{y}_1\mu_{11}} \text{ and } E(\epsilon_1\epsilon_4) = \frac{\mu_{30}}{\bar{y}_1\mu_{20}} \quad (6)$$

where

$$\mu_{ij} = E(y_1-\bar{y}_1)^i(y_2-\bar{y}_2)^j. \quad (7)$$

Bias of the Estimator

The proposed estimator is biased. The bias of the estimator $\widehat{\mu}_{xlr}$ is given in the following theorem:

Theorem 1. *The bias of the estimator $\widehat{\mu}_{xlr}$ to the order $O(n^{-2})$ is given by*

$$B(\widehat{\mu}_{xlr}) = -\left[\frac{1-p}{p}\right]\left(\frac{1}{m}-\frac{1}{k}\right)\left[\frac{\mu_{21}}{\mu_{20}}-\frac{\mu_{11}\mu_{30}}{\mu_{20}^2}\right]. \quad (8)$$

Proof. We have

$$E(\widehat{\mu}_{xlr}) = \frac{1}{p}E[p\bar{X}+(1-p)\bar{y}_2'-(1-p)\{\bar{y}_2+\widehat{\beta}(\bar{y}_1'-\bar{y}_1\}]$$
$$= \frac{1}{p}[p\bar{X}+(1-p)\bar{Y}_2-(1-p)\{\bar{Y}_2+\beta\bar{Y}_1E((1+\epsilon_3)(1+\epsilon_4)^{-1}(\epsilon_2-\epsilon_1))\}].$$

Considering the binomial expansion of $(1+\epsilon_4)^{-1}$ up to the order of $O(n^{-2})$ and substituting the values of $E(\epsilon_1\epsilon_3)$ and $E(\epsilon_1\epsilon_4)$ in the above expression, we have

$$E(\widehat{\mu}_{xlr}) = \bar{X} - \left[\frac{1-p}{p}\right]\left(\frac{1}{m}-\frac{1}{k}\right)\left[\frac{\mu_{21}}{\mu_{20}}-\frac{\mu_{11}\mu_{30}}{\mu_{20}^2}\right]. \quad (9)$$

As

$$B(\widehat{\mu}_{xlr}) = E(\widehat{\mu}_{xlr}) - \bar{X}. \quad (10)$$

Using (9) in (10) we get (8). □

Variance of the Estimator

The variance of the estimator $\widehat{\mu}_{xlr}$ is given in the following theorem:

Theorem 2. *The variance of the estimator $\widehat{\mu}_{xlr}$ to the order of $O(n^{-2})$ is given by*

$$V(\widehat{\mu}_{xlr}) = \frac{1}{p^2}\left[\frac{\sigma_z^2}{n} + (1-p)^2\left\{\left(\frac{1}{k} - \frac{1}{N}\right)\sigma_{y_2}^2 + \left(\frac{1}{m} - \frac{1}{k}\right)\sigma_{y_2}^2(1-\rho^2)\right\}\right] \quad (11)$$

where $\sigma_z^2 = p\sigma_x^2 + q\sigma_{y_2}^2 + pq(\bar{X} - \bar{Y}_2)^2$.

Proof. We have

$$V(\widehat{\mu}_{xlr}) = \frac{1}{p^2}[V(\bar{z}) + (1-p)^2 V(\bar{y}_{lrd})] \quad (12)$$

where $\bar{z} = p\bar{X} + (1-p)\bar{y}_2'$ and $\bar{y}_{lrd} = \bar{y}_2 + \widehat{\beta}(\bar{y}_1' - \bar{y}_1)$

$$V(\widehat{\mu}_{xlr}) = \frac{1}{p^2}\left[\frac{\sigma_z^2}{n} + (1-p)^2 V(\bar{y}_{lrd})\right]. \quad (13)$$

Now

$$V(\bar{y}_{lrd}) = E_1 V_2(\bar{y}_{lrd}/ps) + V_1 E_2(\bar{y}_{lrd}/ps)$$
$$= E_1 V_2(\bar{y}_{lrd}/ps) + V_1(y_2)$$

where '*ps*' means preliminary sample and

$$V_2(\bar{y}_{lrd}/ps) = E(\bar{y}_{lrd}^2/ps) - [E(\bar{y}_{lrd}/ps)]^2$$
$$= E(\{\bar{y}_2 + \widehat{\beta}(\bar{y}_1' - \bar{y}_1)\}^2) - (E\{\bar{y}_2 + \widehat{\beta}(\bar{y}_1' - \bar{y}_1)\})^2.$$

For two-phase sampling one can refer Sukhatme *et al.* (1984) to obtain the variance expression

$$V_2(\bar{y}_{lrd}/ps) = \left(\frac{1}{m} - \frac{1}{k}\right)s_{y_2}^2(1-\rho^2)$$

where ρ is the coefficient of correlation between y_1 and y_2 and $s_{y_2}^2$ is defined at (4).

$$V(\bar{y}_{lrd}) = E_1\left[\left(\frac{1}{m} - \frac{1}{k}\right)s_{y_2}^2(1-\rho^2)\right] + V_1(y_2)$$
$$= \left(\frac{1}{m} - \frac{1}{k}\right)s_{y_2}^2(1-\rho^2) + \left(\frac{1}{k} - \frac{1}{N}\right)\sigma_{y_2}^2. \quad (14)$$

Using the result obtained at (14) in (13), we get (11). □

Optimum Variance of the Estimator

We should choose that procedure which for a fixed cost C_0 minimizes the variance of the estimate. If C_1, C_2 and C_3 are the costs per unit of collecting information on the variable Z, auxiliary variable y_1 and y_2 under study respectively. Then, the total cost of the survey can be expressed as

$$C_0 = nC_1 + kC_2 + mC_3\,. \quad (15)$$

We shall now determine 'n', 'k' and 'm' so that for a fixed cost C_0, the variance of the estimator $\widehat{\mu}_{xlr}$ is least.
Consider therefore, the expression

$$\phi = V(\widehat{\mu}_{xlr}) - \lambda(C_0 - nC_1 - kC_2 - mC_3)\,. \tag{16}$$

On differentiating ϕ partially with respect to 'n', 'k' and 'm' and equating them equal to zero, we have

$$\frac{\partial \phi}{\partial n} = \frac{1}{p^2}\left(\frac{-\sigma_z^2}{n^2}\right) + \lambda C_1\,, \tag{17}$$

$$\frac{\partial \phi}{\partial k} = \left(\frac{1-p}{p}\right)^2 \left[\frac{-\sigma_{y2}^2(1-\rho^2)}{k^2}\right] + \lambda C_2\,, \tag{18}$$

$$\frac{\partial \phi}{\partial m} = \left(\frac{1-p}{p}\right)^2 \left[\frac{-\sigma_{y2}^2(1-\rho^2)}{m^2}\right] + \lambda C_3\,. \tag{19}$$

Solving (17), (18) and (19) for 'n', 'k' and 'm' respectively, we have

$$n = \frac{\sigma_z}{p\sqrt{\lambda C_1}}\,, \tag{20}$$

$$k = \left(\frac{1-p}{p}\right)\sigma_{y2}\sqrt{\frac{1-\rho^2}{\lambda C_2}}\,, \tag{21}$$

$$m = k\sqrt{\frac{C_2}{C_3}}\,. \tag{22}$$

Again differentiating ϕ partially with respect to λ and equating the result equal to zero, we have

$$\frac{\partial \phi}{\partial \lambda} = C_0 - nC_1 - kC_2 - mC_3\,. \tag{23}$$

Substituting the values of 'n', 'k' and 'm' in (23), we have

$$\sqrt{\lambda} = \frac{\sigma_z\sqrt{C_1} + \sqrt{C_2}[qA' + k\sqrt{C_3}]}{pC_0}$$

where $A' = \sigma_{y2}\sqrt{1-\rho^2}$.
Now substituting the values of $\sqrt{\lambda}$ in (20), (21) and (22), we get the optimum values of 'n', 'k' and 'm' respectively as

$$n_{opt} = \frac{C_0\sigma_z}{B'\sqrt{C_1}}\,,$$

$$k_{opt} = \frac{C_0 qA'}{B'\sqrt{C_2}}\,,$$

$$m_{opt} = k_{opt}\sqrt{\frac{C_2}{C_3}}$$

and

$$B' = \sigma_z \sqrt{C_1} + \sqrt{C_2}[qA' + k\sqrt{C_3}].$$

where A' is defined above and $q = 1 - p$.

Theorem 3. *The optimum variance of the estimator $\widehat{\mu}_{xlr}$ to the order of $O(n^{-2})$ is given by*

$$V(\widehat{\mu}_{xlr}) = \frac{1}{p^2}\left[\frac{\sigma_z B'\sqrt{C_1}}{C_0} + q^2\left\{\left(\frac{B'\sqrt{C_2}}{C_0 qA'} - \frac{1}{N}\right)\sigma^2_{y2} + \frac{B'}{C_0 qA'}(\sqrt{C_3} - \sqrt{C_2})A'\right\}\right]. \tag{24}$$

Proof. We have

$$V(\widehat{\mu}_{xlr}) = \frac{1}{p^2}\left[\frac{\sigma_z^2}{n} + (1-p)^2\left\{\left(\frac{1}{k} - \frac{1}{N}\right)\sigma^2_{y_2} + \left(\frac{1}{m} - \frac{1}{k}\right)\sigma^2_{y_2}(1-\rho^2)\right\}\right].$$

Substituting the optimum values of 'n', 'k' and 'm' obtained earlier in the above expression, we get (24). □

References

[1] F.W. Carel, Gerty Peeters, J.L.M. Lensvelt-Mulders and Karin Lasthuizen (2010), A note on a simple and practical randomized response framework for eliciting dichotomous and quantitative information, *Sociological Meth. Res.*, **39**, 283–296.

[2] A. Chaudhuri (2001), Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population, *J. Statist. Planning Inference* **94**, 37–42.

[3] A. Chaudhuri (2004), Christofides' randomized response technique in complex sample survey, *Metrika* **60**, 223–228.

[4] E. Elffers, P.G.M. Van Der Heijden and M. Hezemans (2003), Explaining regulatory non-compliance: a survey study of rule transgression for two Deutch instrumentl laws, applying the randomized response method, *J. Quantitative Criminology* **19**, 409–439.

[5] C.R. Gjestvang and S. Singh (2006), A new randomized response model, *J. Roy. Stat. Soc., B*, **68**, 523–530.

[6] C.R. Gjestvang and S. Singh (2007), Forced quantitative randomized response model: a new device, *Metrika* **66(2)**, 243–256.

[7] D.G. Horvitz, B.V. Shah and W.R. Simmons (1967), The unrelated question randomized response model, *Proc. Social Statist. Sec., Am. Statist. Ass.*, 65–72.

[8] C.H. Huang (2004), A survey technique for estimating the proportion and sensitivity in a dichotomous finite population, *Statistica Neerlandica* **58**, 75–82.

[9] B.G. Greenberg, R.R. Kuebler, J.R. Abernathy and D.G. Horvitz (1971), Application of the randomized response technique in obtaining quantitative data, *J. Am. Statist. Ass.* **66**, 243–250.

[10] I.S. Grewal, M. Bansal and S.S. Sidhu (2006), Population mean estimator corresponding to Horvitz and Thompson's estimator for multi-characteristics using randomized response technique, *Model Assisted Statistics and Applications* **1**(4), 215–220.

[11] M. Javed and I.S. Grewal (2006), On the relative efficiencies of randomized response devices with Greenberg unrelated question model, *Model Assisted Statistics and Applications* **1**(4), 291–297.

[12] H. Samuel (2008), The multi-item randomized response technique, *Sociological Meth. Res.* **36**(4), 495–514.

[13] S.S. Sidhu, M.L. Bansal and S. Singh (2007), Estimation of sensitive multi-characters using unknown value of unrelated question, *Applied Mathematical Sciences* **1**(37), 1803–1820.

[14] S. Singh (2002), A new stochastic randomized response model, *Metrika* **56**, 131–142.

[15] S. Singh, A.H. Joarder and M.L. King (1996), Regression analysis using scrambled responses, *Australian J. Statist.* **38**(2), 201–211.

[16] S. Singh and M.L. King (1999), Estimation of coefficient of determination using scrambled responses, *J. Indian Soc. Agric. Statist.* **52**(3), 338–343.

[17] S. Singh, R. Singh and N.S. Mangat (2000), Some alternative strategies to Moor's model in randomized response sampling, *J. Statist. Planning Inference* **83**, 243–255.

[18] R. Strachan, M.L. King and S. Singh (1998), Likelihood based estimation of the regression model with scrambled response, *Australian Newzealand J. Statist.* **40**(3), 279–290.

[19] P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme and C. Asok (1984), *Sampling Theory of Surveys with Applications*, Iowa State University Press, USA and Indian Society of Agricultural Statistics, New Delhi, India.

[20] S.L. Warner (1965), Randomized response: a survey technique for eliminating evasive answer bias, *J. Am. Statist. Ass.* **60**, 63–69.

M. Javed, *Department of Mathematics, Statistics & Physics, Punjab Agricultural University, Ludhiana 141004, India*.
E-mail: mjaved@pau.edu

M.L. Bansal, *Department of Mathematics, Statistics & Physics, Punjab Agricultural University, Ludhiana 141004, India*.

Received May 2, 2011
Accepted May 29, 2012