

Estimation of sensitive quantitative characteristics in randomized response sampling

Kuo-Chung Huang*

*Department of Business Administration
Chungyu Institute of Technology
Keelung
Taiwan 201
R.O.C.*

Chun-Hsiung Lan

Mei-Pei Kuo

*Department of Business Administration
Nanhua University
Dalin, Chiayi
Taiwan 622
R.O.C.*

Abstract

This paper considers the problem of procuring honest responses for sensitive quantitative characteristics. An alternative survey technique is proposed, which enables us to estimate the population mean unbiasedly and to gauge how sensitive a survey topic is. An asymptotically unbiased estimator of sensitivity level is proposed, and conditions for which unbiased estimation for population variance being available is also studied. In addition, an efficiency comparison is worked out to examine the performance of the proposed procedure. It is found that higher estimation efficiency results from higher variation of randomization device.

Keywords : Percent relative efficiency, privacy protection, sample size allocation, simple random sampling with replacement.

*E-mail: kchuang.tw@gmail.com

1. Introduction

Accurate information on the survey topics is extremely relevant for parameter estimations. However, it is difficult to obtain valid and reliable information in the area of sensitive topics. If a direct survey research is employed to assess a sensitive characteristic, respondents often refuse to take part, or reply untruthfully, especially when they have committed sensitive behavior. To improve respondent cooperation and to procure reliable data, Warner [15] first introduced the randomized response (RR) technique for qualitative characteristics. Subsequently, Greenberg et al. [8] suggested an extension of Warner's RR technique to quantitative characteristics. An excellent exposition of modifications on RR techniques and other related works could be referred to Chaudhuri and Mukerjee [6]. Some recent developments are Bhargava and Singh [1], Chua and Tsui [7], Padmawar and Vijayan [12], Singh et al. [14], Chang and Huang [2], Chaudhuri [5], Singh et al. [13], Huang [10], and Chang et al. [3, 4], etc. In particular, to quantify the sensitivity level for certain items of inquiry in practice, Gupta et al. [9] first shown how an estimator may be developed on the quantitative characteristics.

Consider a finite population in which every person has a positive value for the sensitive characteristic X . The problem of interest is to estimate the mean μ_x , the variance σ_x^2 , and the sensitivity level W of X from a with-replacement simple random sample of size n . In Gupta et al. [9] procedure, each sampled respondent is instructed to utilize a randomization device and generate a positive-valued random number S from a pre-assigned distribution with known mean $\mu_S = 1$ and known variance γ_S^2 . Then he or she chooses one of the following options: (a) The respondent can report the correct response X , or (b) The respondent can report the scrambled response SX . The optional randomized response model considered is $Z = S^Y X$, where $Y = 1$ or 0 according as the response is scrambled or not. Here Y is a Bernoulli variate with $E(Y) = W$, where W is the probability that a respondent will report the scrambled response rather than the actual response X . It is shown that the usual sample mean $\hat{\mu}_{xG} = \bar{Z}$ is unbiased with variance given by

$$\text{Var}(\hat{\mu}_{xG}) = n^{-1}\sigma_Z^2 = n^{-1}[\sigma_x^2 + W\gamma_S^2(\sigma_x^2 + \mu_x^2)]. \quad (1.1)$$

Denote by $s_Z^2 = (n-1)^{-1} \sum_{j=1}^n (Z_j - \bar{Z})^2$ the sample variance, an unbiased estimator of $\text{Var}(\hat{\mu}_{xG})$ is given by $\hat{\text{Var}}(\hat{\mu}_{xG}) = n^{-1}s_Z^2$. And, they

suggested

$$\hat{W}_G = \frac{n^{-1} \sum_{j=1}^n \log(Z_j) - \log \bar{Z}}{E[\log(S)]} \quad \text{and} \quad \hat{\sigma}_{xG}^2 = \frac{s_Z^2 - \hat{W}_G \gamma_S^2 \hat{\mu}_{xG}^2}{1 + \hat{W}_G \gamma_S^2},$$

as an estimator of W and σ_x^2 , respectively.

Although it creates an appropriate environment to estimate some unknown population parameters, the Gupta et al. [9] development in finding an estimator for W remains biased because of the presence of log term. They did address estimation efficiency and biasedness in the W estimate using empirical estimators, but analytical variance expression of W remains unidentified. In addition, the estimator $\hat{\sigma}_{xG}^2$ is merely a biased estimator of σ_x^2 . In these regards, we intend to suggest a survey procedure that provides an unbiased estimator of σ_x^2 and an asymptotically unbiased estimator of W together with its variance. The proposed estimators with principal properties are given in the following section. In section 3, an efficiency comparison is carried out to study the performance of the proposed procedure.

2. The proposed procedure

In the proposed procedure, two independent samples of size n_i , $i = 1, 2$, are drawn from the population using simple random sampling with replacement such that $n_1 + n_2 = n$, the total sample size required. In the i th sample, each respondent is instructed to use a randomization device and generate a random number, say S_i , from some pre-assigned distributions such as Chi-square, uniform, or Weibull etc. The RR procedures are set up in such a way that the interviewer does not know what the respondent has selected. Then the respondent is requested to report one of the following two options: (a) The correct response X , or (b) The scrambled response $S_i X$. It is supposed that S_i is a positive-valued random variate with known mean $\mu_{S_i} = \theta_i \neq 1$ and known variance $\sigma_{S_i}^2 = \gamma_i^2$. The optional randomized response model for the i th sample, $i = 1, 2$, is given by

$$Z_i = (1 - Y)X + YS_i X,$$

where Y is a random variate, with expectation $E(Y) = W$, defined as

$$Y = \begin{cases} 1, & \text{if the response is scrambled,} \\ 0, & \text{otherwise.} \end{cases}$$

Under the proposed procedure, the sample response Z_i for the i th sample, $i = 1, 2$, has expectation

$$E(Z_i) = (1 - W)\mu_x + W\theta_i\mu_x = \mu_x + W(\theta_i - 1)\mu_x. \quad (2.1)$$

If \bar{Z}_1 and \bar{Z}_2 are the observed means for the two samples, the proposed estimators of μ_x and W are respectively given by

$$\hat{\mu}_x = \frac{(1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2}{\theta_1 - \theta_2} \quad \text{and} \quad \hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{(1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2},$$

whence provided that $\theta_1 \neq \theta_2$.

Theorem 1. *The estimator $\hat{\mu}_x$ is unbiased with variance given by*

$$\text{Var}(\hat{\mu}_x) = \frac{1}{\theta_1 - \theta_2} \left[(1 - \theta_2)^2 \frac{\sigma_{Z_1}^2}{n_1} + (1 - \theta_1)^2 \frac{\sigma_{Z_2}^2}{n_2} \right], \quad (2.2)$$

where $\sigma_{Z_i}^2 = \sigma_x^2 + W(\gamma_i^2 + \theta_i^2 - 1)\sigma_x^2 + W[\gamma_i^2 + (1 - W)(1 - \theta_i)^2]\mu_x^2$, $i = 1, 2$. (2.3)

Proof. Taking expectation for $\hat{\mu}_x$, the unbiasedness follows from using (2.1). From the distribution of Z_i^2 , we have

$$E(Z_i^2) = \sigma_x^2 + \mu_x^2 + W(\gamma_i^2 + \theta_i^2 - 1)(\sigma_x^2 + \mu_x^2). \quad (2.4)$$

Using (2.1), (2.4) and the fact that $\sigma_{Z_i}^2 = E(Z_i^2) - [E(Z_i)]^2$, we get (2.3). Expression (2.2) then follows from the independence of the two samples. Hence the proof. \square

Theorem 2. *An unbiased estimator of the variance of $\hat{\mu}_x$ is given by*

$$\hat{\text{Var}}(\hat{\mu}_x) = \frac{1}{(\theta_1 - \theta_2)^2} \left[(1 - \theta_2)^2 \frac{s_{Z_1}^2}{n_1} + (1 - \theta_1)^2 \frac{s_{Z_2}^2}{n_2} \right],$$

where $s_{Z_i}^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2$, $i = 1, 2$.

Proof. It follows from that $E(s_{Z_i}^2) = \sigma_{Z_i}^2$, $i = 1, 2$. \square

For deriving the properties of the estimator \hat{W} , let us define $d_1 = \bar{Z}_1 - \bar{Z}_2$ and $d_2 = (1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2$. Since $E(d_1) = (\theta_1 - \theta_2)W\mu_x$ and $E(d_2) = (\theta_1 - \theta_2)\mu_x$, we then have $\hat{W} = d_1/d_2$ and $W = E(d_1)/E(d_2)$. Additionally, if we further denote by $e_1 = [d_1 - E(d_1)]/E(d_1)$ and $e_2 = [d_2 - E(d_2)]/E(d_2)$, assuming that $|e_2| < 1$ so that $(1 + e_2)^{-1}$ can be validly expanded as power series, it follows that

$$E(e_1) = E(e_2) = 0, \quad E(e_1 e_2) = \frac{(1 - \theta_2)\sigma_{Z_1}^2 + (1 - \theta_1)^2\sigma_{Z_2}^2}{(\theta_1 - \theta_2)^2 W \mu_x^2},$$

$$E(e_1^2) = \frac{\sigma_{Z_1}^2 + \sigma_{Z_2}^2}{(\theta_1 - \theta_2)^2 W^2 \mu_x^2}, \quad E(e_2^2) = \frac{(1 - \theta_2)^2 \sigma_{Z_1}^2 + (1 - \theta_1)^2 \sigma_{Z_2}^2}{(\theta_1 - \theta_2)^2 \mu_x^2}. \quad (2.5)$$

And the estimation error can be written in terms of e_1 and e_2 as

$$\hat{W} - W = W(e_1 - e_2) + o(n^{-1}). \quad (2.6)$$

Then we have the following theorem.

Theorem 3. *The estimator \hat{W} is asymptotically unbiased with variance given by $\text{Var}(\hat{W})$*

$$= \frac{1}{(\theta_1 - \theta_2)^2 \mu_x^2} \left\{ [1 + W(\theta_2 - 1)]^2 \frac{\sigma_{Z_1}^2}{n_1} + [1 + W(\theta_1 - 1)]^2 \frac{\sigma_{Z_2}^2}{n_2} \right\}, \quad (2.7)$$

which can be estimated by

$$\hat{\text{Var}}(\hat{W}) = \frac{1}{(\theta_1 - \theta_2)^2 \hat{\mu}_x^2} \left\{ [1 + \hat{W}(\theta_2 - 1)]^2 \frac{s_{Z_1}^2}{n_1} + [1 + \hat{W}(\theta_1 - 1)]^2 \frac{s_{Z_2}^2}{n_2} \right\}. \quad (2.8)$$

Proof. The asymptotically unbiasedness follows from $E(e_1) = E(e_2) = 0$. Squaring expression (2.6), omitting terms with power in e_i 's higher than the second and then taking expectation, we have $E(\hat{W} - W)^2 = W^2 E(e_1^2 - 2e_1e_2 + e_2^2)$. On substituting the expected values given in (2.5) and after some algebraic simplification, we get (2.7). Replacing μ_x , W and $\sigma_{Z_i}^2$ in (2.7) by the corresponding sample analogue, (2.8) then follows. Hence the theorem. \square

Next, let us consider the problem of unbiased estimation for the variance σ_x^2 of the sensitive characteristic X . For the sake of notational convenience, let us denote $a = 2(\theta_2 - 1)\gamma_1^2 + (2 - \theta_1 - \theta_2)\gamma_2^2 + (1 - 2\theta_1 - \theta_2) \times (\theta_1 - \theta_2)$, $b = (\theta_1 - \theta_2)(1 - \gamma_1^2 - \theta_1^2)$ and $c = -2[(\theta_2 - 1)\gamma_1^2 - (\theta_1 - 1)\gamma_2^2 + (\theta_1 - 1)(\theta_2 - 1)(\theta_1 - \theta_2)]$. An unbiased estimator of σ_x^2 is given in the following theorem.

Theorem 4. *If $(\theta_2 - 1)(1 - 2\theta_1 + \theta_2)\gamma_1^2 + (\theta_1 - 1)^2\gamma_2^2 = 2\theta_1(\theta_1 - 1)(\theta_2 - 1) \times (\theta_1 - \theta_2)$, an unbiased estimator of σ_x^2 is given by*

$$\hat{\sigma}_x^2 = \frac{as_1^2 + bs_2^2 + c \left[\left(n_1^{-1} \sum_{j=1}^{n_i} Z_{1j}^2 \right) - \bar{Z}_1 \bar{Z}_2 \right]}{(\gamma_2^2 + \theta_2^2 - \gamma_1^2 - \theta_1^2)(\theta_1 - \theta_2)},$$

whence provided that $\gamma_1^2 + \theta_1^2 \neq \gamma_2^2 + \theta_2^2$ and $\theta_1 \neq \theta_2$.

Proof. Since the two samples are independent, expression (2.1) leads

$$E(\bar{Z}_1\bar{Z}_2) = [1 + (\theta_1 + \theta_2 - 2)W + (\theta_1 - 1)(\theta_2 - 1)W^2]\mu_x^2. \quad (2.9)$$

From (2.4), we have

$$E\left(n_1^{-1} \sum_{j=1}^{n_1} Z_{1j}^2\right) = [1 + W(\gamma_1^2 + \theta_1^2 - 1)](\sigma_x^2 + \mu_x^2). \quad (2.10)$$

From using (2.9), (2.10) and the fact that $E(s_i^2) = \sigma_i^2$, $i = 1, 2$, the unbiasedness of $\hat{\sigma}_x^2$ then follows. Hence the theorem. \square

We now move on to study the appropriate sample size allocations for various objectives in sampling surveys. Consider a linear combination of the variances of $\hat{\mu}_x$ and \hat{W} , given in (2.2) and (2.7) respectively, given by

$$\begin{aligned} \text{Var}(\hat{\mu}_x, \hat{W}) &= \alpha_1 \text{Var}(\hat{\mu}_x) + \alpha_2 \text{Var}(\hat{W}) \\ &= \frac{1}{(\theta_1 + \theta_2)^2 \mu_x^2} \left[\left\{ \alpha_1 (1 - \theta_2)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_2 - 1)]^2 \right\} \frac{\sigma_{Z_1}^2}{n_1} \right. \\ &\quad \left. + \left\{ \alpha_1 (1 - \theta_1)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_1 - 1)]^2 \right\} \frac{\sigma_{Z_2}^2}{n_2} \right] \end{aligned}$$

where α_1 and α_2 are non-negative coefficients chosen by the investigator. If the total sample size n ($= n_1 + n_2$) is fixed, then through a simple application of the Cauchy-Schwarz inequality, the sample allocation for which $\text{Var}(\hat{\mu}_x, \hat{W})$ attains its minimum is given by

$$\frac{n_1}{n_2} = \left[\frac{\left\{ \alpha_1 (1 - \theta_2)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_2 - 1)]^2 \right\} \sigma_{Z_1}^2}{\left\{ \alpha_1 (1 - \theta_1)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_1 - 1)]^2 \right\} \sigma_{Z_2}^2} \right]^{1/2},$$

and the minimum value of $\text{Var}(\hat{\mu}_x, \hat{W})$ will be

$$\text{Min Var}(\hat{\mu}_x, \hat{W}) = \frac{\left(\left\{ \alpha_1 (1 - \theta_2)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_2 - 1)]^2 \right\}^{\frac{1}{2}} \sigma_{Z_1} \right) \left(\left\{ \alpha_1 (1 - \theta_1)^2 \mu_x^2 + \alpha_2 [1 + W(\theta_1 - 1)]^2 \right\}^{\frac{1}{2}} \sigma_{Z_2} \right)}{n(\theta_1 - \theta_2)^2 \mu_x^2}.$$

It is remarkable that the actual values of a σ_{Z_1} , σ_{Z_2} and W are always unknown, information regarding them can be obtained from past experience or a pilot survey, which is helpful for practical applications (Murthy [11], pp. 96–99).

3. Efficiency comparison

In what follows, we study the performance of the proposed procedure compared with Gupta et al. [9] procedure. The percent relative efficiency of the proposed estimator $\hat{\mu}_x$ with respect to Gupta et al. estimator

$\hat{\mu}_{xG}$ is defined as

$$PRE = \frac{(\theta_1 - \theta_2)^2 \sigma_Z^2}{(|1 - \theta_2| \sigma_{Z_1} + |1 - \theta_1| \sigma_{Z_2})^2} \times 100,$$

where σ_Z^2 and $\sigma_{Z_i}^2$ are respectively given in (1.1) and (2.3). Since the above *PRE* expression is infested with too many parameters, here the comparison is performed on an empirical investigation using the same probability distribution in the competing randomizing devices. To examine how the variation of randomization devices will affect the efficiency, it is assumed that scrambled variables follow the distributions with mean μ and variance γ^2 satisfying $\gamma^2 = k\mu$. Without loss of generality, here we simply choose $(\mu_S, \gamma_S^2) = (1, k)$, $(\theta_1, \gamma_1^2) = (k, k^2)$ and $(\theta_2, \gamma_2^2) = (0.5, 0.5k)$, where $k = 2, 3, \dots, 10$. It is remarkable that the *PRE* value remains unchanged if the coefficient of variation $CV_x = \sigma_x \mu_x^{-1}$ is fixed. The values of CV_x are then chosen to be 0.2, 0.4, 0.6, 0.8 and 1. The percent relative efficiencies thus obtained are outlined in Table 1 for different values of W .

Table 1
Percent relative efficiency of the two competing procedures

CV_x	W	k								
		2	3	4	5	6	7	8	9	10
0.2	0.1	92.2	96.1	100.8	105.2	109.1	112.6	115.7	118.4	120.9
	0.3	95.9	101.0	106.1	110.6	114.5	117.9	120.9	123.5	125.9
	0.5	100.0	105.7	111.0	115.5	119.3	122.5	125.4	127.9	130.1
	0.7	104.5	110.9	116.2	120.6	124.3	127.4	130.1	132.5	134.6
	0.9	109.4	116.5	121.9	126.2	129.8	132.7	135.3	137.6	139.6
0.4	0.1	93.4	95.4	99.0	102.7	106.2	109.5	112.4	115.1	117.5
	0.3	96.3	100.7	105.5	109.8	113.6	116.9	119.8	122.4	124.8
	0.5	100.0	105.4	110.4	114.8	118.5	121.7	124.6	127.1	129.3
	0.7	104.1	110.2	115.3	119.7	123.3	126.4	129.1	131.5	133.6
	0.9	108.4	115.3	120.5	124.8	128.3	131.2	133.8	136.1	138.1
0.6	0.1	94.6	95.4	98.0	101.0	104.0	106.8	109.5	112.1	114.4
	0.3	96.6	100.4	104.8	108.9	112.5	115.7	118.6	121.2	123.6
	0.5	100.0	105.0	109.8	114.0	117.6	120.8	123.6	126.1	128.3
	0.7	103.6	109.4	114.4	118.6	122.1	125.2	127.9	130.2	132.4
	0.9	107.5	113.9	119.0	123.1	126.6	129.5	132.1	134.3	136.3

(Contd. Table 1)

CV_x	W	k								
		2	3	4	5	6	7	8	9	10
0.8	0.1	95.5	95.7	97.5	100.0	102.6	105.1	107.6	109.9	112.1
	0.3	97.0	100.2	104.2	108.1	111.6	114.7	117.6	120.1	122.4
	0.5	100.0	104.6	109.2	113.3	116.8	120.0	122.7	125.2	127.4
	0.7	103.3	108.7	113.5	117.6	121.1	124.1	126.8	129.1	131.2
	0.9	106.7	112.8	117.6	121.7	125.1	128.0	130.5	132.7	134.7
1.0	0.1	96.0	95.9	97.4	99.5	101.7	104.1	106.3	108.5	110.5
	0.3	97.2	100.1	103.9	107.5	110.9	114.0	116.7	119.3	121.6
	0.5	100.0	104.4	108.8	112.7	116.2	119.3	122.0	124.5	126.7
	0.7	103.0	108.2	112.9	116.9	120.3	123.3	125.9	128.2	130.3
	0.9	106.1	111.9	116.6	120.5	123.9	126.7	129.2	131.5	133.5

From Table 1, it is seen that the proposed procedure performs better than Gupta et al. [9] procedure for most of the practical situations. Even though in some cases the proposed procedure is less efficient, it can be viewed as a tradeoff for being able to get an unbiased estimator of σ_x^2 and an asymptotically unbiased estimator of W . Note that the PRE value increases with k and/or W , whereas decreases with CV_x , if other parameters are unchanged.

References

- [1] M. Bhargava and R. Singh (2000), A modified randomization device for Warner's model, *Statistica*, Vol. 60, pp. 315–321.
- [2] H. J. Chang and K. C. Huang (2001), Estimation of proportion and sensitivity of a qualitative character, *Metrika*, Vol. 53, pp. 269–280.
- [3] H. J. Chang, C. L. Wang and K. C. Huang (2004), Using randomized response to estimate the proportion and truthful reporting probability in a dichotomous finite population, *Journal of Applied Statistics*, Vol. 31, pp. 565–573.
- [4] H. J. Chang, C. L. Wang and K. C. Huang (2004), On estimating the proportion of a qualitative sensitive character using randomized response sampling, *Quality & Quantity*, Vol. 38, pp. 675–680.
- [5] A. Chaudhuri (2001), Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population, *Journal of Statistical Planning and Inference*, Vol. 94, pp. 37–42.

- [6] A. Chaudhuri and R. Mukerjee (1988), *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.
- [7] T. C. Chua and A. K. Tsui (2000), Procuring honest responses indirectly, *Journal of Statistical Planning and Inference*, Vol. 90, pp. 107–116.
- [8] B. G. Greenberg, R. R. Kubler, J. R. Abernathy and D. G. Horvitz (1971), Applications of the RR technique in obtaining quantitative data, *Journal of the American Statistical Association*, Vol. 66, pp. 243–250.
- [9] S. Gupta, B. Gupta and S. Singh (2002), Estimation of sensitivity level of personal interview survey questions, *Journal of Statistical Planning and Inference*, Vol. 100, pp. 239–247.
- [10] K. C. Huang (2004), A survey technique for estimating the proportion and sensitivity in a dichotomous finite population, *Statistica Neerlandica*, Vol. 58, pp. 75–82.
- [11] M. N. Murthy (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.
- [12] V. R. Padmawar and K. Vijayan (2000), Randomized response revisited, *Journal of Statistical Planning and Inference*, Vol. 90, pp. 293–304.
- [13] S. Singh, M. Mahmood and D. S. Tracy (2001), Estimation of mean and variance of stigmatized quantitative variable using distinct units in randomized response sampling, *Statistical Papers*, Vol. 42, pp. 403–411.
- [14] S. Singh, R. Singh and N. S. Mangeao (2000), Some alternative strategies to Moor's model in randomized response sampling, *Journal of Statistical Planning and Inference*, Vol. 83, pp. 243–255.
- [15] S. L. Warner (1965), Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, Vol. 60, pp. 63–69.

Received April, 2005