二分類變量的敏感性水準評估之研究

On Measuring the Sensitivity Level of Dichotomous Characters

黃國忠[1]

Abstract

Randomized response techniques are useful for promoting respondent cooperation and reducing the inflated response bias experienced in direct response surveys with respect to potentially sensitive issues. But the primary problem is that nobody knows what an issue of survey inquiry should be regarded as sensitive. This study suggests a new technique, which is more useful due to its advance in measuring the sensitivity of survey issues. And it may be utilized to check whether or not a randomized response technique is better in practice. Circumstances in which the proposed technique can be applied is studied and illustrated using a numerical example.

***Keywords*:** estimation of proportion, randomized response, sampling surveys

1. Introduction

In some socioeconomic investigations, conducting direct response (DR) surveys on sensitive topics, such as illegal and/or unethical practices, and some issues that are customarily disapproved of by society, it is likely to yield refusals or untruthful answers. Consider a dichotomous population in which every person belongs either to a sensitive group A, or to its complement $\overline{\mathrm{A}}$. Let T_A and $T_{\bar{A}}$ be the corresponding probability that the respondents belonging to A and $\overline{\mathrm{A}}$ report the truth. The problem of interest is to estimate the population proportion π of individuals who are members of A. For a DR survey of size n, the interviewee is asked if he/she is a member of A. As an estimator of π, one may take $\hat{\lambda}_D$, the sample proportion of 'yes' responses. The bias and mean square error of the estimator $\hat{\lambda}_D$ are respectively given by

$$Bias(\hat{\lambda}_D) = \pi(T_A + T_{\bar{A}} - 2) + (1 - T_{\bar{A}}),$$

$$MSE(\hat{\lambda}_D) = \frac{\lambda_D(1-\lambda_D)}{n} + (\lambda_D - \pi)^2,$$

[1]Chungyu Institute of Technology Department of Business Administration Associate Professor

where $\lambda_D = \pi T_A + (1-\pi)(1-T_{\bar{A}})$. Note that the expressions for the bias and mean square error of $\hat{\lambda}_D$ involve the unknown parameters T_A and $T_{\bar{A}}$ for which there are no sample analogues, and $Bias(\hat{\lambda}_D)$ along with $MSE(\hat{\lambda}_D)$ cannot be estimated.

To improve respondent cooperation and to procure reliable data, Warner (1965) proposed the following randomized response (RR) technique. A randomization device used to collect sample information consists of two statements: (a) I am a member of group A, and (b) I am not a member of group A, represented with probabilities p_w and $(1-p_w)$ respectively. Following this device, the interviewee chooses a statement and then simply replies 'yes' or 'no' to the statement chosen. The process of selecting one of the statements is unobserved by the interviewer. Thus although the interviewer gets a 'yes' or 'no' response, because of the randomization procedure, he/she cannot identify a particular respondent with group A or $\overline{\text{A}}$ on the basis of such a reply. This maintains an interviewee's privacy, and he/she may then be expected to cooperate and respond truthfully. Let $\hat{\lambda}_w$ be the proportion of 'yes' answers in a random sample of n respondents. Assuming the true reporting, Warner obtained an unbiased estimator of π as $(p_w \neq 0.5)$

$$\hat{\pi}_w = \frac{\hat{\lambda}_w - (1-p_w)}{(2p_w - 1)},$$

and the expression for the variance of $\hat{\pi}_w$ is given by

$$Var(\hat{\pi}_w) = \frac{\lambda_w(1-\lambda_w)}{(2p_w-1)^2 n},$$

where $\lambda_w = \pi p_w + (1-\pi)(1-p_w)$. A good exposition of modifications on Warner's (1965) pioneering technique and other related work could be referred to Chaudhuri and Mukerjee (1988).

It is obvious that RR techniques are designed for sensitive characters to achieve honest sample information. But intuition and experience suggest that certain issues apprehended by an investigator to be sensitive in general may not be so felt by certain sampled subjects. If a character under study is sensitive is therefore not easy to determine. Certainly, theoretical comparison with a DR survey is not complex to implement in measuring a RR technique's efficacy. Nevertheless, it is not easily to gauge how well it may fare in practice. For a DR survey, Chaudhuri and Mukherjee (1988, p.8) remarked that 'However, it is unlikely that one will have any way to guess correctly about the magnitudes of T_A and $T_{\bar{A}}$ so as to be able to

judge the extent of bias involved and the effect on the accuracy in estimation'. To measure the sensitivity for certain items of inquiry in practice, we suggest a scheme to get admissible estimators for the truthful reporting probabilities T_A and $T_{\bar{A}}$, which may be viewed as a measurement of sensitivity. The proposed procedure also enables us to unbiasedly estimate the mean square errors for DR and RR surveys simultaneous. We show how the estimators and the estimated measure of their sampling errors may be developed. Also, a numerical illustration is carried out to demonstrate the practicability of the proposed technique.

2. The Proposed Procedure

In the proposed procedure, three independent sub-samples of size n_j, $j = 1, 2, 3$, are drawn from the population using simple random sampling with replacement such that $\sum_{j=1}^{3} n_j = n$, the total sample size required. The person in the first sub-sample is instructed to directly respond whether he/she is a member of A. In the second sub-sample, each respondent is given a suitable randomization device R_1, which comprises of two statements (a) and (b), represented with probabilities p_1 and $(1-p_1)$ respectively. The respondent randomly chooses one statement and report 'yes' or 'no' with respect to his/her actual status, without revealing to the interviewer which of the two statements he/she has chosen. The person in the third sub-sample is instructed to reply whether he/she belongs to A. If the answer is 'no', the respondent is required to use a randomization device R_2 consisting of (a) and (b) with probabilities p_2 and $(1-p_2)$ respectively. Then he/she simply gives a 'yes' or 'no' answer depending on the outcome of randomized device R_2 without revealing the statement selected.

The probability of getting a 'yes' answer in the j th sub-sample, $j = 1, 2, 3$, is therefore given by

$$\lambda_1 = \pi T_A + (1-\pi)(1-T_{\bar{A}}),$$

$$\lambda_2 = p_1\pi + (1-p_1)(1-\pi),$$

$$\lambda_3 = \pi T_A + (1-\pi)(1-T_{\bar{A}}) + p_2\pi(1-T_A) + (1-p_2)(1-\pi)T_{\bar{A}}.$$

By the method of moments, the estimator of π can be easily obtained as

$$\hat{\pi} = \frac{\hat{\lambda}_2 - (1-p_1)}{2p_1 - 1},$$

and the estimators of T_A and $T_{\bar{A}}$ are respectively given by

$$\hat{T}_A = \frac{(2p_1-1)p_2\hat{\lambda}_1 + (2p_2-1)\hat{\lambda}_2 - (2p_1-1)\hat{\lambda}_3 + (p_1-p_2)}{(2p_2-1)[\hat{\lambda}_2 - (1-p_1)]},$$

$$\hat{T}_{\bar{A}} = \frac{(2p_1-1)[(1-p_2)\hat{\lambda}_1 - \hat{\lambda}_3 + p_2]}{(2p_2-1)(p_1-\hat{\lambda}_2)},$$

where $\hat{\lambda}_j$ is the observed proportion of 'yes' answers obtained from the jth sub-sample, and is distributed as the binomial distribution $B(n_j, \lambda_j)$, $j = 1, 2, 3$. Clearly, the estimator $\hat{\pi}$ is unbiased with variance given by

$$Var(\hat{\pi}) = \frac{\lambda_2(1-\lambda_2)}{(2p_1-1)^2 n_2},$$

and the corresponding unbiased variance estimator can be obtained as

$$\hat{Var}(\hat{\pi}) = \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)}{(2p_1-1)^2(n_2-1)}. \qquad (2.1)$$

To derive the mean square errors of the estimators $\hat{T}_A$ and $\hat{T}_{\bar{A}}$, let us define

$$d_1 = (2p_1-1)p_2\hat{\lambda}_1 + (2p_2-1)\hat{\lambda}_2 - (2p_1-1)\hat{\lambda}_3 + (p_1-p_2),$$

$$d_2 = (2p_2-1)[\hat{\lambda}_2 - (1-p_1)],$$

$$\delta_1 = (2p_1-1)[(1-p_2)\hat{\lambda}_1 - \hat{\lambda}_3 + p_2],$$

$$\delta_2 = (2p_2-1)(p_1-\hat{\lambda}_2),$$

then we have $\hat{T}_A = d_1/d_2$ and $\hat{T}_{\bar{A}} = \delta_1/\delta_2$. In addition, it can be verified that

$$E(d_1) = (2p_1-1)(2p_2-1)\pi T_A,$$
$$E(d_2) = (2p_1-1)(2p_2-1)\pi,$$
$$E(\delta_1) = (2p_1-1)(2p_2-1)(1-\pi)T_{\bar{A}},$$
$$E(\delta_2) = (2p_1-1)(2p_2-1)(1-\pi),$$

it follows that $T_A = E(d_1)/E(d_2)$ and $T_{\bar{A}} = E(\delta_1)/E(\delta_2)$. Further, we define the following

quantities:

$$e_1 = \frac{d_1 - E(d_1)}{E(d_1)},\ e_2 = \frac{d_2 - E(d_2)}{E(d_2)},\ \varepsilon_1 = \frac{\delta_1 - E(\delta_1)}{E(\delta_1)} \text{ and } \varepsilon_2 = \frac{\delta_2 - E(\delta_2)}{E(\delta_2)},$$

where $|e_2|$ and $|\varepsilon_2|$ are assumed to be less than unity such that the functions $(1+e_2)^{-1}$ and $(1+\varepsilon_2)^{-1}$ can be expressed as power series. One can easily show that

$$E(e_1^2) = \frac{(2p_1-1)^2 p_2^2 \lambda_1(1-\lambda_1)n_1^{-1} + (2p_2-1)^2\lambda_2(1-\lambda_2)n_2^{-1} + (2p_1-1)^2\lambda_3(1-\lambda_3)n_3^{-1}}{(2p_1-1)^2(2p_2-1)^2\pi^2 T_A^2},$$

$$E(e_2^2) = \frac{\lambda_2(1-\lambda_2)n_2^{-1}}{(2p_1-1)^2\pi^2},\ E(e_1e_2) = \frac{\lambda_2(1-\lambda_2)n_2^{-1}}{(2p_1-1)^2\pi^2 T_A},$$

$$E(\varepsilon_1^2) = \frac{(2p_1-1)^2[(1-p_2)^2\lambda_1(1-\lambda_1)n_1^{-1} + \lambda_3(1-\lambda_3)n_3^{-1}]}{(2p_1-1)^2(2p_2-1)^2(1-\pi)^2 T_{\bar{A}}^2},$$

$$E(\varepsilon_2^2) = \frac{\lambda_2(1-\lambda_2)n_2^{-1}}{(2p_1-1)^2(1-\pi)^2},\ E(\varepsilon_1\varepsilon_2) = 0.$$

The expressions for mean square error of the estimators $\hat{T}_A$ and $\hat{T}_{\bar{A}}$ are as follows.

Theorem 2.1. To the first degree of approximation, the mean square errors of the estimators $\hat{T}_A$ and $\hat{T}_{\bar{A}}$ are respectively given by

$$MSE(\hat{T}_A) = \frac{1}{(2p_1-1)^2(2p_2-1)^2\pi^2}\left[\frac{(2p_1-1)^2 p_2^2\lambda_1(1-\lambda_1)}{n_1} + \frac{(2p_2-1)^2(1-T_A)^2\lambda_2(1-\lambda_2)}{n_2} + \frac{(2p_1-1)^2\lambda_3(1-\lambda_3)}{n_3}\right], \tag{2.2}$$

$$MSE(\hat{T}_{\bar{A}}) = \frac{1}{(2p_1-1)^2(2p_2-1)^2(1-\pi)^2}\left[\frac{(2p_1-1)^2(1-p_2)^2\lambda_1(1-\lambda_1)}{n_1} + \frac{(2p_2-1)^2 T_{\bar{A}}^2\lambda_2(1-\lambda_2)}{n_2} + \frac{(2p_1-1)^2\lambda_3(1-\lambda_3)}{n_3}\right]. \tag{2.3}$$

Proof. Since the estimator $\hat{T}_A$ can be written as $\hat{T}_A = T_A(1+e_1)(1+e_2)^{-1}$, to the first order approximation, we have $MSE(\hat{T}_A) = T_A^2 E(e_1^2 - 2e_1e_2 + e_2^2)$. Substituting the corresponding expected values and then after some simple algebra yields the result (2.2). Expression (2.3) follows similarly. Hence the proof of the theorem.

Having obtained the sample analogues for those mentioned population parameters, one may estimate $MSE(\hat{T}_A)$ and $MSE(\hat{T}_{\bar{A}})$ respectively by

$$\hat{MSE}(\hat{T}_A) = \frac{1}{(2p_1-1)^2(2p_2-1)^2\hat{\pi}^2}\left[\frac{(2p_1-1)^2 p_2^2\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1}\right.$$

$$\left.+\frac{(2p_2-1)^2(1-\hat{T}_A)^2\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2}+\frac{(2p_1-1)^2\hat{\lambda}_3(1-\hat{\lambda}_3)}{n_3}\right],$$

$$\hat{MSE}(\hat{T}_{\bar{A}}) = \frac{1}{(2p_1-1)^2(2p_2-1)^2(1-\hat{\pi})^2}\left[\frac{(2p_1-1)^2(1-p_2)^2\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1}\right.$$

$$\left.+\frac{(2p_2-1)^2\hat{T}_{\bar{A}}^2\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2}+\frac{(2p_1-1)^2\hat{\lambda}_3(1-\hat{\lambda}_3)}{n_3}\right].$$

Furthermore, since in the proposed procedure, the survey technique used in the first sub-sample is actual direct response, one can straightforwardly establish the unbiased estimators of the bias and mean square error of the estimator $\hat{\lambda}_1$, which is outlined in the following theorem.

Theorem 2.2. The unbiased estimators of $Bias(\hat{\lambda}_1)$ and $MSE(\hat{\lambda}_1)$ are given by

$$\hat{Bias}(\hat{\lambda}_1) = \hat{\lambda}_1 - \hat{\pi}, \tag{2.4}$$

$$\hat{MSE}(\hat{\lambda}_1) = (\hat{\lambda}_1 - \hat{\pi})^2 - \hat{Var}(\hat{\pi}). \tag{2.5}$$

Proof. Since the sub-samples are drawn independently, and we have $E(\hat{\lambda}_1) = \lambda_1$ as well as $E(\hat{\pi}) = \pi$, the expression (2.4) is unbiased. The unbiasedness of the estimator (2.5) follows from that

$$E(\hat{\lambda}_1 - \hat{\pi})^2 = \frac{\lambda_1(1-\lambda_1)}{n_1} + (\lambda_1 - \pi)^2 + Var(\hat{\pi}),$$

and $\hat{Var}(\hat{\pi})$, given in (2.1), is an unbiased estimator of $Var(\hat{\pi})$. Hence the theorem.

3. Numerical Illustration

Consider a hypothetical example to support the proposed procedure. It is supposed that a small population consists of $N = 20$ persons and their true status is given in Table 3.1.

Table 3.1. Data for a Small Population of 20 persons

Person Number	1	2	3	4	5	6	7	8	9	10
Member of A or $\overline{A}$	A	A	$\overline{A}$	$\overline{A}$	A	$\overline{A}$	$\overline{A}$	A	A	$\overline{A}$
Truthful (1) or not (0)	0	0	1	1	1	0	1	0	0	0

Person Number	11	12	13	14	15	16	17	18	19	20
Member of A or $\overline{A}$	$\overline{A}$	A	$\overline{A}$	$\overline{A}$	A	$\overline{A}$	A	A	$\overline{A}$	A
Truthful (1) or not (0)	1	0	0	0	0	1	0	0	0	0

We are interested in estimating π, the proportion of persons being a member of A, T_A and $T_{\overline{A}}$, the corresponding probability that the persons bearing A and $\overline{A}$ report the truth. In this example π is actually 0.5, T_A is 0.1 and $T_{\overline{A}}$ is 0.5.

Suppose we draw three independent and non-overlapping sub-samples of sizes $n_1 = n_2 = n_3 = 5$ and consisting of first five, second five and next five persons in the list shown above. Further let $p_1 = 0.8$ and $p_2 = 0.9$ for the randomization devices used during the survey. Then the basic data are as follows.

$\lambda_1 = 0.3$, $\lambda_2 = 0.5$, $\lambda_3 = 0.73$,

$MSE(\hat{T}_A) = 0.909$, $MSE(\hat{T}_{\overline{A}}) = 0.388$,

$Var(\hat{\pi}) = 0.139$, $Bias(\hat{\lambda}_1) = 0.2$, $MSE(\hat{\lambda}_1) = 0.082$.

It is obvious that $Var(\hat{\pi}) > MSE(\hat{\lambda}_1)$, implying the superiority of a DR over a RR survey. The responses obtained from the three sub-samples are given in Table 3.2.

Table 3.2. Sample Data for the Three Sub-samples

Sub-sample 1		Sub-sample 2		Sub-sample 3	
Person No.	Response	Person No.	Response	Person No.	Response
1	No	6	No	11	(No, No)
2	No	7	Yes	12	(No, Yes)
3	No	8	Yes	13	(Yes, —)
4	No	9	Yes	14	(Yes, —)
5	Yes	10	No	15	(No, Yes)

From Table 3.2, we have $\hat{\lambda}_1 = 0.2$, $\hat{\lambda}_2 = 0.6$ and $\hat{\lambda}_3 = 0.8$. Using the proposed estimators and after some simple algebra, we get $\hat{\pi} = 0.667$, $\hat{T}_A = 0.025$ and $\hat{T}_{\overline{A}} = 0.45$. In addition, one can also obtain

$\hat{MSE}(\hat{T}_A) = 0.489$, $\hat{MSE}(\hat{T}_{\overline{A}}) = 0.698$,

$\hat{Var}(\hat{\pi}) = 0.167$, $\hat{Bias}(\hat{\lambda}_1) = 0.2$, $\hat{MSE}(\hat{\lambda}_1) = 0.051$.

From this example, we conclude that it is possible to find admissible estimators of those

mentioned population parameters with the help of the proposed procedure. And the proposed method makes it possible to guess about the relative efficiencies of estimators based on RR and DR survey techniques.

4. Concluding Remarks

Clearly, in the proposed procedure, the randomization device used is identical to the Warner's (1965) RR device. In fact, one can easily extend to conduct a better RR technique to achieve higher gains in efficiency. As an instance, instead of Warner's (1965) technique, one may utilize Kuk's (1990) procedure, which allows repeated trials. As the number of repetitions increases, the efficiency of the resulting estimator of population proportion will be improved. The problem mentioned above could also be overcome by proceeding on the lines of the present paper. Similar modifications may be brought on the other RR techniques as well. We omit details about them here.

References

1. Chaudhuri, A. & R. Mukerjee (1988.), *Randomized Response: Theory and Techniques*, New York: Marcel Dekker.
2. Kuk, A. Y. C. (1990), "Asking Sensitive Questions Indirectly," *Biometrika*, 77, pp.436-438.
3. Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60, pp.63-69.