Jour. Ind. Soc. Ag. Statistics
54(1), 2001 : 19–28

Efficiency Comparison of Certain Randomized Response Schemes with U-Model

Manoj Bhargava and Ravindra Singh[1]
HP Krishi Vishvavidyalaya, Palampur, H.P.
(Received: February, 2000)

SUMMARY

In the recent years, randomized response procedures have been introduced in an attempt to improve the accuracy and honesty in surveys involving sensitive characteristics, while protecting respondent's privacy. Several research workers have made comparisons based on variances of the estimators but they have not taken into account the degree of respondent's privacy. In the present paper, an attempt has been made in this direction and two randomized response schemes have been compared with the U-model, taking into account the protection afforded the respondent.

Key words : Randomized response technique, U-model, Design probabilities, Jeopardy functions, Efficiency comparison.

1. Introduction

To procure reliable data for estimating the proportion π of population possessing a sensitive attribute, Warner [9] proposed an ingenious procedure, called randomized response (RR) technique. This procedure enables the respondents to provide truthful information while protecting their privacy. To enhance the confidence of the respondents, Horvitz *et al.* [4] developed an unrelated question model (U-model) by providing the respondent the opportunity of replying to one of the two questions in which one question was completely innocuous and unrelated to the sensitive attribute. The theoretical framework for this model was given by Greenberg *et al.* [3]. Their model was further improved by Moors [8] and Folsom *et al.* [2]. Mangat [6] and Mangat *et al.* [7] have also suggested certain modifications to the U-model.

Comparisons of efficiency for different schemes have already been performed by several workers e.g., Greenberg *et al.* [3], Moors [8] and, Dowling and Shachtman [1]. These comparisons of schemes based on the variances of

1 B-80, Moti Kunj Extension, Mathura-281 001.

the estimators may result in misleading conclusions about the relative performances of the schemes.

As the degree of privacy is an essential component of the RR procedure, hence, an obvious basis for comparing randomized response models is to compare variances, only when the required degree of respondent privacy is held constant.

In the present article, an attempt has been made in this direction. To achieve this, a measure of privacy protection has been considered which was given by Leysieffer and Warner [5]. According to them, a population is divided into complementary sensitive groups, A and A^c with unknown proportions, π and $(1-\pi)$, respectively. Let us consider a dichotomous response model where a typical response R is "yes" (say, y) or "no" (say, n). The conditional probabilities that a response R comes from an individual of groups, A and A^c, are $P(R|A)$ and $P(R|A^c)$, respectively. These probabilities are at the investigator's disposal and are called *design probabilities.*

Using these design probabilities, Leysieffer and Warner [5] proposed the natural measures of jeopardy carried by R about A and A^c, respectively. These measures are as follows:

$$g(R \mid A) = \frac{P(R \mid A)}{P(R \mid A^c)}; \quad g(R \mid A^c) = \frac{1}{g(R \mid A)} \tag{1.1}$$

In the sequel, we will call these measures as *jeopardy functions.*

They have also shown that an unbiased estimator of π is defined if and only if

$$P(y \mid A) - P(y \mid A^c) \neq 0$$

and the existence of an unbiased estimator for π necessarily makes a response jeopardic with respect to either A or A^c.

Assuming, without loss of generality, that

$$P(y \mid A) > P(y \mid A^c) \tag{1.2}$$

so that a "yes" answer increases the odds of A and is jeopardizing with respect to A, i.e.

$$g(y \mid A) = \frac{P(y \mid A)}{P(y \mid A^c)} > 1$$

while a "no" answer increases the odds of A^c and is jeopardizing with respect to A^c, i.e.

$$g(n \mid A^c) = \frac{P(n \mid A^c)}{P(n \mid A)} > 1$$

Therefore, for the sake of efficiency, one needs as large magnitudes as possible for $g(y \mid A)$ and $g(n \mid A^c)$ and both above unity. Hence, from the practical point of view, regarding protection of privacy, one can fix some maximal allowable levels of $g(y \mid A)$ and $g(n \mid A^c)$ (say, k_1 and k_2), respectively. After fixing $g(y \mid A)$ and $g(n \mid A^c)$ at k_1 and k_2, the optimal choice of the design parameters for the particular randomized response model can be worked out. These design parameters will now be in terms of k_1 and k_2. In this way, we can derive the variance expressions for each randomized response model by substituting the values of design parameters and then these variances can be compared at the same level of protection of privacy.

In the present investigation, the schemes given by Mangat [6] and Mangat *et al.* [7] have been compared with the Greenberg's [3] U-model, using privacy protection measure as discussed above. Before coming to the question of efficiency comparison of these randomized response schemes with respect to protection of privacy, we shall discuss, in brief, these schemes in the subsequent sections.

2. *U-model*

While developing the theory of this model, Greenberg *et al.* [3] considered two cases when π_y, the proportion of population belonging to non-sensitive (innocuous) group Y, is known and when it is unknown. We shall restrict ourselves to the first case, i.e. when π_y is known due to its simplicity.

When π_y is known, each sampled-respondent is provided with a random device. This device consists of two statements, (i) I belong to sensitive group A, and (ii) I belong to non-sensitive group Y, represented with probabilities p_1 and $(1 - p_1)$, respectively. The respondent selects randomly one of these two statements, unobserved by the interviewer and reports "yes" or "no" with respect to his/her actual status. The probability of "yes" answer is

$$\theta_1 = p_1 \pi + (1 - p_1) \pi_y$$

An unbiased estimator of π is, therefore, given by

$$\hat{\pi}_U = \frac{\hat{\theta}_1 - (1 - p_1)\,\pi_y}{p_1}$$

where $\hat{\theta}_1$ is the observed proportion of "yes" answers, obtained from the n sampled-respondents.

The variance of the estimator $\hat{\pi}_U$ is given by

$$V(\hat{\pi}_U) = \frac{\pi(1-\pi)}{n} + \frac{(1-p_1)(1-2\pi_y)\pi}{n p_1} + \frac{(1-p_1)[1-(1-p_1)\pi_y]\pi_y}{n p_1^2} \qquad (2.1)$$

3. *Scheme 1*

Mangat [6] proposed a two-stage randomized response unrelated question scheme. In this method, each interviewee is provided with two randomization devices R_1 and R_2. The randomization device R_1 consists of two statements, namely:

(i) I belong to sensitive group A, and

(ii) Go to randomization device R_2

represented with probabilities T and $(1 - T)$, respectively. The randomization device R_2 is the same as used in the U-model represented with probabilities p_2 and $(1 - p_2)$, respectively. The rest of the procedure remains unchanged.

Then θ_2, the probability of "yes" answer, is given by

$$\theta_2 = T\pi + (1 - T)[p_2\pi + (1 - p_2)\pi_y]$$

The unbiased estimator of π is then given by

$$\hat{\pi}_1 = \frac{\hat{\theta}_2 - (1 - T)(1 - p_2)\pi_y}{[T + p_2(1 - T)]}$$

where $\hat{\theta}_2$ is the observed proportion of "yes" answers in the sample and π_y is assumed to be known.

The variance of the estimator $\hat{\pi}_1$ is given by

$$V(\hat{\pi}_1) = \frac{\pi(1-\pi)}{n} + \frac{\pi(1-T)(1-p_2)(1-2\pi_y)}{n[T+p_2(1-T)]}$$

$$+ \frac{(1-T)(1-p_2)\pi_y[1-(1-T)(1-p_2)\pi_y]}{n[T+p_2(1-T)]^2} \tag{3.1}$$

4. Scheme 2

Mangat *et al.* [7] proposed an improved U-model. In this model, each respondent in a sample of n individuals, is instructed to say "yes" if he/she belongs to sensitive group A, otherwise he is to report "yes" or "no" according to the outcome of the using of the randomization device as suggested by Greenberg *et al.* [3]. The two statements of this device are represented with the probabilities p_3 and $(1-p_3)$, respectively. The rest of the procedure remains unchanged.

Then, the probability of "yes" answer is given by

$$\theta_3 = \pi + (1-\pi)(1-p_3)\pi_y$$

The unbiased estimator of π is, therefore, obtained as

$$\hat{\pi}_2 = \frac{\hat{\theta}_3 - (1-p_3)\pi_y}{[1-(1-p_3)\pi_y]}$$

where $\hat{\theta}_3$ is the observed proportion of "yes" answers from n sampled-respondents.

The variance of $\hat{\pi}_2$ is, thus given by

$$V(\hat{\pi}_2) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p_3)\pi_y}{n[1-(1-p_3)\pi_y]} \tag{4.1}$$

Mangat showed that both the schemes 1 and 2 were always more efficient than that of usual U-model with known π_y.

Now, we come to the efficiency aspect with respect to the protection of privacy.

5. *Efficiency Comparisons*

In this section, the scheme 1 and scheme 2 have been compared with the U-model, for their efficiencies, using the measure of privacy protection discussed in Section 1.

Let us consider the U-model, which is defined in Section 2. Here, π_y is known and assuming that the attribute A and Y are independent, we have the design probabilities as

$$\begin{aligned} P(y \mid A) &= p_1 + (1 - p_1)\,\pi_y \\ P(n \mid A) &= (1 - p_1)(1 - \pi_y) \\ P(y \mid A^c) &= (1 - p_1)\,\pi_y \\ P(n \mid A^c) &= 1 - (1 - p_1)\,\pi_y \end{aligned} \tag{5.1}$$

Clearly, $P(y \mid A) > P(y \mid A^c)$, if $p_1 > 0$ which is always true.

Now, from (1.1) and (5.1), we have the jeopardy functions as

$$g_u(y \mid A) = \frac{p_1 + (1 - p_1)\,\pi_y}{(1 - p_1)\,\pi_y}$$

$$g_u(n \mid A^c) = \frac{1 - (1 - p_1)\,\pi_y}{(1 - p_1)(1 - \pi_y)}$$

If k_1 and k_2 be the maximum allowable values for $g_u(y \mid A)$ and $g_u(n \mid A^c)$ respectively, then the optimal choice of the design parameters, π_y and p_1 is seen to be

$$\pi_y = \frac{k_2 - 1}{k_1 + k_2 - 2} \qquad \text{and} \qquad p_1 = \frac{(k_1 - 1)(k_2 - 1)}{k_1 k_2 - 1}$$

It sometimes happens that A^c is innocuous and only A is stigmatizing. Then we can allow $k_2 \to \infty$ (i.e. k_2 is as large as infinity). In the sequel, we shall make all comparisons assuming the non-sensitivity of A^c.

Therefore, the optimal choice of design parameters for the U-model turns out to be

$$\pi_y = 1 \quad \text{and} \qquad p_1 = \frac{k_1 - 1}{k_1}$$

With this optimal choice of the design parameters, the variance of the unbiased estimator $\hat{\pi}_U$ (2.1) becomes

$$V^*(\hat{\pi}_U) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(k_1-1)^{-1}}{n} \tag{5.2}$$

5.1 Comparison of Scheme 1 with the U-model

Considering scheme 1, discussed in Section 3, we have the design probabilities as

$$\begin{aligned} P(y \mid A) &= T + (1-T)[p_2 + (1-p_2)\pi_y] \\ P(n \mid A) &= (1-T)(1-p_2)(1-\pi_y) \\ P(y \mid A^c) &= (1-T)(1-p_2)\pi_y \\ P(n \mid A^c) &= 1-(1-T)(1-p_2)\pi_y \end{aligned} \tag{5.3}$$

To check the condition (1.2), we have

$$T + (1-T)[p_2 + (1-p_2)\pi_y] > (1-T)(1-p_2)\pi_y$$

or, $$T + (1-T)p_2 > 0$$

which is always true.

Now, from (1.1) and (5.3), we have the following jeopardy functions

$$g_1(y \mid A) = \frac{T + (1-T)[p_2 + (1-p_2)\pi_y]}{(1-T)(1-p_2)\pi_y}$$

and

$$g_1(n \mid A^c) = \frac{1-(1-T)(1-p_2)\pi_y}{(1-T)(1-p_2)(1-\pi_y)}$$

If we take maximal allowable limits for $g_1(y \mid A)$ and $g_1(n \mid A^c)$ as k_1 and k_2, respectively, then the optimal choice of the design parameters, π_y and p_2, after some simplification, is obtained as

$$\pi_y = \frac{k_2 - 1}{k_1 + k_2 - 2} \qquad \text{and} \qquad p_2 = 1 - \frac{k_1 + k_2 - 2}{(k_1 k_2 - 1)(1-T)}$$

Now, if we assume that A^c is innocuous and we allow k_2 as large as infinity, then the optimal choice of design parameters for scheme 1, turns out to be

$$\pi_y = 1 \quad \text{and} \quad p_2 = 1 - \frac{1}{k_1 (1 - T)}$$

With this optimal choice of the design parameters, the variance of the unbiased estimator $\hat{\pi}_1$ (3.1) comes out to be

$$V^* (\hat{\pi}_1) = \frac{\pi (1 - \pi)}{n} + \frac{(1 - \pi) (k_1 - 1)^{-1}}{n} \tag{5.4}$$

Since variance expressions in (5.2) and (5.4) are equal, hence it can be concluded that both the estimators i.e. $\hat{\pi}_U$ and $\hat{\pi}_1$, given by Greenberg *et al.* [3] and Mangat [6], respectively, are equally efficient, when compared at the same level of protection of privacy, provided that A^c is innocuous and A & Y are independent. It is also suggested that an easy way to achieve $\pi_y = 1$ is to so design that with a probability p_1 or p_2 a respondent is to divulge his/her truth about A and to be instructed to report "yes" with the complementary probability $(1 - p_1)$ or $(1 - p_2)$, respectively. This conclusion is given in the following theorem.

Theorem 5.1. Scheme 1 and U-model are equally efficient, when compared at the same level of protection of privacy.

5.2 Comparison of Scheme 2 with the U-model

Let us consider the scheme 2, in which the design probabilities are given by

$$\begin{aligned} P (y \mid A) &= 1 \\ P (n \mid A) &= 0 \\ P (y \mid A^c) &= (1 - p_3) \pi_y \\ P (n \mid A^c) &= 1 - (1 - p_3) \pi_y \end{aligned} \tag{5.5}$$

Clearly, the condition, $P (y \mid A) > P (y \mid A^c)$ is satisfied for every value of p_3 and π_y.

Therefore, using (1.1) and (5.5), the jeopardy functions are obtained as

$$g_2 (y \mid A) = \frac{1}{(1 - p_3) \pi_y}$$

$$g_2 (n \mid A^c) = \infty$$

Since, $g_2(n \mid A^c)$ is infinite, then taking maximal allowable limit for $g_2(y \mid A)$ as k_1, we have

$$\frac{1}{(1-p_3)\,\pi_y} = k_1$$

or,
$$p_3 = 1 - \frac{1}{k_1 \pi_y} \tag{5.6}$$

For $p_3 > 0$, k_1 and π_y should be such that $k_1\pi_y > 1$.

Hence, (5.6) is the optimal choice of design parameter for scheme 2 subject to the condition, $k_1\pi_y > 1$.

With this optimal choice of the design parameter, the variance of the unbiased estimator $\hat{\pi}_2$ (4.1), in case of scheme 2, becomes

$$V^*(\hat{\pi}_2) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(k_1-1)^{-1}}{n} \tag{5.7}$$

Now, on comparing $V^*(\hat{\pi}_2)$ with $V^*(\hat{\pi}_U)$, we conclude that the estimators $\hat{\pi}_U$ and $\hat{\pi}_2$ are equally efficient when compared at the same level of protection of privacy, provided that A^c is innocuous and A & Y are independent. This result can be stated in the form of the following theorem.

Theorem 5.2. The scheme 2 and the U-model are equally efficient, when compared at the same level of privacy protection.

6. Conclusions

In this paper, we have considered two different schemes; one has been proposed by Mangat [6] and the other is due to Mangat *et al.* [7]. They have compared their schemes with the U-model based on the variances of the estimators and have shown that the schemes proposed by them are always more efficient than the U-model. But we have compared these two schemes with the usual U-model, when the level of privacy protection is held constant, using the measure of privacy protection proposed by Leysieffer and Warner [5]. It is observed that at the same level of privacy protection, both the schemes are as efficient as the Greenberg's [3] U-model.

REFERENCES

[1] Dowling, T.A. and Shachtman, R.H., (1975). On the relative efficiency of randomized response models. *J. Am. Statist. Assoc.*, **70**, 84-87.

[2] Folsom, R.E., Greenberg, B.G., Horvitz, D.G. and Abernathy, J.R., (1973). The two alternate questions randomized response model for human surveys. *J. Am. Statist. Assoc.*, **68**, 525-530.

[3] Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz, D.G., (1969). The unrelated question randomized response model: Theoretical framework. *J. Am. Statist. Assoc.*, **64**, 520-539.

[4] Horvitz, D.G., Shah, B.V. and Simmons, W.R., (1967). The unrelated question randomized response model. *Proc. Am. Statist. Assoc. Social Statist. Sect.*, 65-72.

[5] Leysieffer, F.W. and Warner, S.L., (1976). Respondent jeopardy and optimal designs in randomized response models. *J. Am. Statist. Assoc.*, **71**, 649-656.

[6] Mangat, N.S., (1992). Two stage randomized response sampling procedure using unrelated question. *J. Ind. Soc. Agric. Stat.*, **44**, 82-87.

[7] Mangat, N.S., Singh, R., and Singh, S., (1992). An improved unrelated question randomized response strategy. *Calcutta Statist. Assoc. Bull.*, **42**, 277-281.

[8] Moors, J.J.A., (1971). Optimization of the unrelated question randomized response model. *J. Am. Statist. Assoc.*, **66**, 627-629.

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