Fjrs.futa.edu.ng
FUTA Journal of Research in Sciences
ISSN: 2315 – 8239 (Print); E-ISSN: 2489 - 0413

FUTA Journal of Research in Sciences, Vol. 12 (2) 2016: 334-340

A COMPARATIVE STUDY OF SOME METHODS OF IMPROVING THE EFFICIENCY OF RANDOMIZED RESPONSE MODELS

F. B. ADEBOLA [1] and O.ALO DAMILOLA [2*]

Department of Statistics, Federal University of Technology, Akure. Ondo State, Nigeria

* E-mail of the corresponding author: alooluwadamilola@gmail.com

ABSTRACT

This paper compares different randomized response designs by their efficiency condition and also using the concept of Jeopardy Function. A population is divided into two sensitive groups, $A\ and\ A^C$ with unknown proportion $\pi\ and\ (1-\pi)$, respectively. Considering a dichotomous response model where a typical response *R* is *yes* (say, *y*) or *no* (say, *n*). The conditional probabilities that a person *R* comes from individual of groups, $A\ and\ A^C$, are $P(R|A)\ and\ P(R|A^C)$, respectively. These probabilities are at the researcher's disposal and are called *Design probabilities*. Taking into consideration these design probabilities, a natural measure was proposed and it was carried by $R\ about\ A\ and\ A^C$, respectively.

Considering the Percentage Relative Efficiency of Mangat *et al*. and Bhargava and Singh, it is found that the Bhargava and Singh model is better than Mangat *et al*.'s model when $\pi > \frac{1}{2}$, and when $\pi < \frac{1}{2}$, Mangat *et al*.'s model is better than Bhargava and Singh's model.

Keywords: Response Error, Privacy Protection, Jeopardy Function, Relative Efficiency (PRE),

INTRODUCTION

RR technique was introduced by Warner (1965) for estimating π_y, the proportion of population possessing a certain stigmatized character y (say, yes). Warner (1965) technique has been modified by Horvitz *et al.* (1967), Greenberg *et al.* (1969), Mangat and Singh (1990), Kim and Warde (2004, 2005a,b), and Singh (2002), among other researchers, for yielding better response and efficiency.

Results from surveys are affected mainly by two sources of error. The first is sampling error that results from taking a sample instead of enumerating the whole population. The second type of error is non-sampling error that cannot be attributed to sample-to-sample variability. Non-sampling error has two different errors, which are random error and non-random error. Random error, which results from a reduction in the reliability of measurements, can be minimized over repeated measurements. However, non-random error, which is bias in the survey data, is difficult to cancel out over repeated measurements.

The main sources of non-sampling error in any survey are non-response bias and response bias. Non-response bias arises from subjects' refusal to respond and response bias arises from giving incorrect responses.

Randomized response procedure is used in sample surveys of human populations for estimating the proportion of a population possessing a given attribute. They are mostly appropriately used when the attribute under study is such that people who has the attribute, show reluctance to accept having it when being asked in a direct-question designed by the interviewer. In a typical situation, the attribute may concern the respondent's involvement or non-involvement in illegal or socially stigmatizing behavior. Since the degree of privacy is an essential part of the randomized response procedure and greater privacy will be attained in general, in terms of variances, only when the required degree of protection is held constant. To achieve this, we will take into consideration a measure of privacy protection which was given by Leysieffer and Warner (1976). This measure of privacy protection has also been used by Bhargava and Singh (2002) and Tung-Hai Lin (2005).

SOME RANDOMIZED RESPONSE TECHNIQUES

MANGAT AND SINGH RRT

Mangat and Singh (1990) proposed a two stage Randomized Response Technique (RRT). In this technique, each sample respondents, selected by Simple Random Sampling With Replacement, is provided with two random devices R_1 *and* R_2. The random device R_1 consists of two statements: (i) "I belong to sensitive group *A*", and (ii) "Go to and Singh (1990) two-stage Randomized Response model. The sample of size *n* is selected by Simple Randomized Sampling With Replacement. Each sample respondent is asked to report a *yes* if he/she possesses the sensitive characteristics. If the respondent does not possess the sensitive characteristics, he/she is directed to use Warner's randomization device.

The probability of a *yes* answer for this procedure is given by

R_2", represented with probabilities U_1 and $(1-U_1)$ respectively. The random device R_2 is exactly the same as used by Warner (1965) with probabilities of the two statements U_2 and $(1-U_2)$. The respondent is asked to use first the random device R_1 and then R_2 if directed by the outcome of R_1. The probability of *yes* answer, then, is

$$\alpha = U_1\pi + (1-U_1)\{U_2\pi + (1-U_2)(1-\pi)\} \quad (1)$$

An unbiased estimator of π (the proportion of population possessing a certain stigmatized character) is given by Mangat and Singh as:

$$\hat{\pi}_{M_S} = \frac{\hat{\alpha}-(1-U_1)(1-U_2)}{2U_2-1+2U_1(1-U_2)}, \quad (2)$$

where $\hat{\alpha}$ is the sample proportion of *yes* responses. Its variance is given by

$$Var(\hat{\pi}_{M_S}) = \frac{\pi(1-\pi)}{n} + \frac{(1-U_1)(1-U_2)\{1-(1-U_1)(1-U_2)\}}{n(2U_2-1+2U_1(1-U_2))^2} \quad (3)$$

Mangat RRT

Mangat (1994) proposed a procedure, which has benefit of simplicity over that of Mangat

$$\delta = \pi + (1-\pi)(1-p) \quad (4)$$

An unbiased estimator of π is given by

$$\hat{\pi}_{M_2} = \frac{\delta-1+p}{p}. \quad (5)$$

The variance of $\hat{\pi}_{M_2}$ is given by

$$Var(\hat{\pi}_{M_2}) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p)}{np}. \quad (6)$$

Review of work done by previous Authors on RRT

Overcoming the difficulty in getting valid response from an interviewee, Warner (1965) started a randomized response technique (RRT) for estimating the proportion of persons belonging to a group with sensitive attribute or having a specific sensitive characteristic, on the basis of a SRSWR (simple random sampling with replacement) scheme of respondents (Chaudhuri and Mukerjee, 1988).

Horvitz *et al.* (1967) and Greenberg *et al.* (1969) came up with the unrelated question model, which is based on asking question that is unrelated to the sensitive question. Abernathy *et al.* (1970) used the randomized response technique to estimate the induced abortion in urban North Carolina; Shimizu and Bonham (1978) estimated the incidence of abortion within a 12-month period in the national survey of family growth where the tactic of two unrelated questions randomized response model in separate half samples was used.

Anderson (1976) took a measured risk of the true value revealed which is the probability that a respondent is a member of the sensitive group and to the non-sensitive group, when he or she reports a randomized response and the relation of that risk with the variance of the proportion estimator. Dowling and Shachtman (1975) made a comparison of the variance of the maximum likelihood estimator in each of the one-sample and two-sample alternate question model with Warner (1965) model. Greenberg *et al.* (1977) measured the gain due to randomization in protecting privacy with truthful response from a person who belongs to a sensitive group as well as the corresponding loss for one in the complementary group. A procedure was recommended of how to limit the risks of revealing true values.

Himmelfarb and Edgell (1980) advise to associate the answer to a sensitive question with a scrambling variable of known mean and variance. The estimators depend upon the parameters of the scrambling variable. Mangat and Singh (1990) modified Warner (1965) model by giving a chance to the respondents to reveal their true values with a known probability and studied conditions under which this modification led to improved efficiency in estimation.

There is a reduction in the efficiency of Randomized response methods when compared to direct question designs due to larger variance generated by randomness. To overcome this problem of large variance, a researcher has to use larger samples, which leads to increase in completion time and costs.

Singh et al. (2012) considered a study, in which a comparison was made of three different estimators for estimating the proportion of a sensitive attribute in survey sampling at equal protection of the respondents. The three estimators considered are due to work done by Odumade and Singh (2009), Singh and Sedory (2011) and a new estimator was obtained by minimizing a chi-squared distance. A SAS Macro was developed to compare these three estimators using a simulation study at equal protection of the respondents. A set of data from a real face-to-face interview was collected using two decks of cards and has been analyzed.

A Measure of Privacy Protection

Leysieffer and Warner (1976) assumed that it is the member of the sensitive group A who may hesitate to reveal the group he actually belongs to. On the other hand, a person who belongs to A^C, is expected to be quite willing to acknowledge the fact. It means that the membership in A may be embarrassing while the membership in A^C might not be considered so. Thus, the larger the conditional probability of belonging to A given a certain answer, the greater is the embarrassment caused by giving that response. The conditional probability that the interviewee belongs to group A when he has given

the response "yes" is denoted by g(ylA) and defined by:

$$g(y|A) = \frac{\pi P(y|A)}{\pi P(y|A)+(1-\pi)P(y|A^C)} \quad (7)$$

Similarly, when the response is “no”, then

$$g(n|A) = \frac{\pi P(n|A)}{\pi P(n|A)+(1-\pi)P(n|A^C)} \quad (8)$$

These posterior probabilities are also called *revealing probabilities.* The probabilities $P(y|A)$, $P(y|A^C)$, $P(n|A)$ and $P(n|A^C)$ are at the investigator's disposal and are called design probabilities.

Hence, for randomized response procedures, one strategy may be considered to be more protective than the other strategy if the conditional probability of the maximum response that the interviewee belongs to group A when he has given the response "yes" and when the response is “no” given as:

$$P = max\ [g(y|A), g(n|A)] \quad (9)$$

is smaller for the previous strategy than for the latter. Therefore, according to Leysieffer and Warner (1976), a measure of protection of privacy is given by (9). Let there be two randomized response strategies (say, 1 and 2), each involving two possible responses "yes" and "no", where $g(y|A) > g(n|A)$ holds for both the strategies. We shall consider these two strategies equally protective for a certain value of π, if

$$P_1(y|A) = P_2(y|A) \quad (10)$$

(where P_1 is the probability of strategy 1 and P_2 is the probability of strategy 2) for that value of π. If this equality holds for all values of π, then the strategies 1 and 2 will be said to be equally protective.

Efficiency Condition for some RRTs

In this section, based on the concept of jeopardy function, we derived efficiency conditions for Mangat *et al.* (1994) and Bhargava and Singh (2000) model then compared both at equal level of privacy protection.

Efficiency Conditions for Mangat *et al.* RR Model at Equal Level of Privacy Protection

Mangat *et al* designed three types of cards, which has these words imprinted on it “I do not belong to group A”, “I belong to group A”, and the third one is left blank, and these are represented with probabilities P_1, P_2, P_3, respectively. Where $P_1 + P_2 + P_3 = 1$. When a blank card is pulled by a respondent, the respondent must say “No” irrespective of the group he/she belongs to. In regards to this process, an unbiased estimator of π is given as:

$$\hat{\pi}_M = \frac{\hat{\lambda}_1 - P_2}{P_1 - P_2}, \quad P_1 \neq P_2, \quad (11)$$

Where $\hat{\lambda}_1$ is the observed proportion of “Yes” answers in a sample of size n. the variance of this estimator $\hat{\pi}_m$ is given as;$V(\hat{\pi}_M) = \frac{\pi(1-\pi)}{n} + \frac{\pi P_3}{n(P_1 - P_2)} + \frac{P_2(1-P_2)}{n(P_1-P_2)^2}$.

Then, taking $P(y|A) = P_1$ and $P(y|A^C) = P_2$, the jeopardy function for Mangat *et al.*’s model is

$$g_m(y|A) = \frac{P(y|A)}{P(y|A^C)} = \frac{P_1}{P_2} \quad (12)$$

$$g_m(n|A^C) = \frac{P(n|A^C)}{P(n|A)} = \frac{P_1+P_3}{P_2+P_3} = \frac{1-P_2}{1-P_1} \quad (13)$$

This clearly shows that, if $P_1 > P_2$, it assures that “yes” and “no” are jeopardy for A and A^C respectively.

Hence, the measure of protection of privacy in Mangat *et al.*'s model is given by

$$P_M = P_M(y|A) \quad (14)$$

where suffix M, is used for Mangat *et al.*'s model.

Efficiency Conditions for Bhargava and Singh RR Model at Equal Level of Privacy Protection

There was a proposition made on a randomized response procedure similar to that of Mangat *et al.* (1995) model with only a little alteration which is, when a blank card is pulled, the respondent is expected to say "Yes" only. Assuming these three cards has probabilities $P_1^{'}, P_2^{'}, P_3^{'}$, respectively. Where $P_1^{'} + P_2^{'} + P_3^{'} = 1$, and $\hat{\lambda}_2$ is the proportion of "Yes" answers in a sample of size n. An unbiased estimator of π is obtained according to Lin (2005) by;

$$\hat{\pi}_{BS} = \frac{\hat{\lambda}_2 - (P_2^{'} + P_3^{'})}{P_1^{'} - P_2^{'}}, \qquad P_1^{'} \neq P_2^{'}, \qquad (15)$$

With variance, $V(\hat{\pi}_{BS}) = \frac{1-\pi}{n} - \frac{\pi P_3^{'}}{n(P_1^{'} - P_2^{'})} + \frac{P_1^{'}(1-P_1^{'})}{n(P_1^{'} - P_2^{'})^2}$

If $P(y|A) = P_1^{'} + P_3^{'}$, and $P(y|A^C) = P_2^{'} + P_3^{'}$, the jeopardy function for Bhargava and Singh's model is

$$g_{BS}(y|A) = \frac{P(y|A)}{P(y|A^C)} = \frac{P_1^{'} + P_3^{'}}{P_2^{'} + P_3^{'}} = \frac{1 - P_2^{'}}{1 - P_1^{'}} \qquad (16)$$

$$g_{BS}(n|A^C) = \frac{P(n|A^C)}{P(n|A)} = \frac{P_1^{'}}{P_2^{'}} \qquad (17)$$

Definitely, $P_1^{'} > P_2^{'}$ assures that "yes" and "no" are jeopardizing for A and A^C.

Hence, the measure of protection of privacy in Bhargava and Singh's model is given by

$$P_{BS} = P_{BS}(y|A) \qquad (18)$$

where suffix BS, is used for Bhargava and Singh's model.

Consider the table 1 below, the values of the variances of Mangat et al. (1995) model and Bhargava and Singh's (2002) model are tabulated and the Percentage Relative Efficiency (PRE) of the estimator $\hat{\pi}_M$ with respect to $\hat{\pi}_{BS}$ for the whole range of π, when $K_1 = 1.1, 1.3, 1.5, 1.7, 1.9$ and $n = 50, 100$ at $\pi = 0.1, 0.3, 0.5, 0.7, 0.9$ was deduced. From the result given, the closer the values are to 100 the more accurate the compared models are. While the values above 100 show that there is a slight or major variation when the models were compared at a particular value of π.

Table 1: Percentage Relative Efficiency of estimator $\hat{\pi}_M$ with respect to $\hat{\pi}_{BS}$

			K_1		
π	1.1	1.3	1.5	1.7	1.9
0.1	100.7358	105.5730	113.3005	122.7567	133.2556
0.3	100.4446	103.3437	107.8905	113.3174	119.1730
0.5	100.2268	101.7013	104	106.7215	109.6314
0.7	100.0817	100.6141	101.4493	102.4461	103.5216
0.9	100.0091	100.0688	100.1642	100.2809	100.4106

CONCLUSION

It appears that the results of a randomized response design become more valid when the topic under investigation becomes more sensitive. So an advantage of using RRTs to obtain information on sensitive topics is that the responses are less distorted than when direct question–answer designs are used, making the randomized response method more effective. A second advantage of using randomized response method when conducting sensitive research is that, although the individual 'yes'-answer becomes biased, researchers are still able to link population estimates to explaining variables using a logistic regression approach.The optimal choice of the design parameters for Mangat *et al.* (1995) model compared to Bhargava and Singh (2000) model has been drawn and the variances derived by these choices has been compared using the different values of π (the proportion of population possessing a certain stigmatized character) and the design that yields the smallest value is considered as the most efficient (see Table 1). It may be concluded that the Bhargava and Singh model is better than Mangat *et al.*'s model when $\pi > \frac{1}{2}$, and when $\pi < \frac{1}{2}$, Mangat *et al.*'s model is better than Bhargava and Singh's model, when the level of privacy protection is at various sample sizes.

REFERENCES

Abernathy, J.R., Greenberg, B.G., and Horvitz, D.G. (1970). Estimates of induced abortion in Urban North Carolina. *Demography*, 7: 19–29.

Anderson, H. (1976). Estimation of a proportion through randomized response, *International Statistical Review / Revue Internationale de Statistique* 44(2): 213–217.

Bhargava and R. Singh (2002). On the efficiency of the certain randomized response strategies, *Metrika*, 55: 191-197.

Chaudhuri, A. and Mukerjee, R. (1988). Randomized Response: Theory and Techniques. *New York: Marcel Dekker*.

Dowling, T.A. and Shachtman, R.H. (1975). On the relative efficiency of randomized response models. *Journal of the American Statistical Association*, 70(349): 84–87.

Greenberg, B. V., Abdul-Ela, A. A., Simmons, W. R. & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association* 64: 529 – 539, 66: 243–250.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., and Horvitz, D.G. (1977). Respondent hazards in the unrelated question randomized response model. *Journal of Statistical Planning and Inference*, 1: 53–60.

Himmelfarb, S. and Edgell, S.E. (1980). Additive constant model: A randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, 87(3): 525–530.

Horvitz, D.G., Saha, B.V., and Simmons, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section of American Statistical Association*, 65–72.

Kim, J-M. &Warde, W.D. (2004). "A Stratified Warner's Randomized Response Model", *Journal of Statistical Planning and Inference* 120 (1-2): 155-165.

Kim, J-M. &Elam, M.E. (2005). "A Two-Stage Stratified Warner's Randomized Response Model Using Optimal Allocation", *Metrika* 61: 1-7.

Leysieffer F.W., and Warner S.L. (1976). Respondent Jeopardy and Optimal Designs in Randomized Response Models. *Journal of the American Statistical Association* 71: 649-656

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of Royal Statistics Society Ser. B*, 56, 93-95.

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

Mangat, N.S. and Singh, S. (1994). Optional randomized response model. *Journal of Indian Statistical Association*, 32(3): 71–75.

Odumade and Singh, S. (2009). Theory Methods. *Journal of Communication Statistics.*

Shimizu, I.M. and Bonham, G.S. (1978). Randomized response technique in a national survey. *Journal of the American Statistical Association*, 73(361): 35–39.

Singh, S. (2002). Randomized Response Model. *Metrika*, 56, 131-142.

Singh, S. and Sedory, S. A. (2011). Cramer-Rao lower bound of variance in randomized response sampling. *Journal of Sociological Methods and Research*, 40(3) 536–546.

Singh P.H., Chandra P., Grewal I.S., Singh S., Chen C.C., Sedory S.A., and Kim J. (2012). Estimation of population ratio, product and mean using multiauxiliary information with random nonresponse. *Statistica*, 72 (4): 449 - 450.

Tung-Hai. Lin (2005). The efficiency comparison on Warner's and two modified models, *Journal of Information and Management Sciences*, 16, 73-78.

Warner S.L. (1965). "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association* 60: 63 – 69.