CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report

Author (s): **T.J. Rao**

Title of the Report: **Some questions on Randomized Response Technique**

Research Report No.: **RR2017-02**

Date: **February 08, 2017**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Some questions on Randomized Response Technique

T.J.Rao
Retired Professor, Indian Statistical Institute, Kolkata
Adjunct Professor, C.R. Rao AIMSCS, Hyderabad
tjrao7@gmail.com

Abstract. In this paper, we shall revisit Warner's seminal paper on Randomized Response Technique introduced five decades ago, which over this period, led to a number of theoretical developments as well as practical applications.

1. Introduction

During the Fifties, initial Rounds of National Sample Survey of India consisted of topics of economic importance such as household consumer expenditure , employment and unemployment as well as topics of social policy implications, some of which are 'sensitive' relating to fertility, mortality and morbidity. Perhaps it was the training of investigators, field supervision and techniques adopted to control non-sampling errors initiated by Mahalanobis that allowed the respondents to wilfully participate in those surveys, though at times sensitive, thus contributing to a 'smooth flow of information'. It is said that the NSS population surveys during the fifties led to certain family planning measures in the country.

In the US and elsewhere, opinion polls were becoming popular and respondents were not reluctant to answer questions of a sensitive nature, such as their political affiliation or favourite candidate etc. However, by the early sixties, society faced the growing drug abuse in the United States and elsewhere, especially in schools and colleges, popularised by the hippie culture. It is felt that as a social scientist, Warner must have thought of the problem of measuring this phenomenon (Rao and Rao, 2016). In a social enquiry by a sample survey, it is difficult to elicit truthful information for some questions relating to a stigmatizing or a sensitive characteristic of an individual by a direct response. In order to circumvent this situation, having found that "*for reasons of modesty, fear of being thought bigoted, or merely a reluctance to confide secrets to*

strangers, many individuals attempt to evade certain questions put to them by interviewers", Warner devised the Randomized Response Technique (RRT) in 1965 (Warner, 1965). During the last five decades the technique has gained such an importance and application that several theoretical and practical contributions have appeared in the literature of sample surveys on this topic of RRT [see for example, the recent contributions in volume 34 of *Handbook of Statistics* published in 2016 and early reviews such as Horvitz *et al.*, 1975; Verdooren, 1976; Fox and Tracy, 1986; Chaudhuri and Mukerjee, 1987,1988; among others].

Being interested in applications, Warner did not seem to be bothered about the rigours of mathematics, but provided innovative answers to the questions he posed for collection of data on sensitive questions. His aim seems to be to "avoid unnecessary mathematical complications and details" (Christofides, 2016) in his publication of 1965.

We shall revisit Warner's technique and raise some simple questions following the theme of Christofides (2016).

2. Warner's technique

We shall first quickly revisit Warner's technique: Let π denote the population proportion of respondents who belong to the sensitive category. A random sample of n individuals is given a randomization device which lets them choose

Statement 1: I belong to the sensitive category (with probability $P \neq 1/2$, known to the surveyor based on the device); or

Statement 2: I do not belong to the sensitive category (with probability $1-P$).

Warner's random mechanism was a (pre-marked) spinner and the *"interviewee is asked to spin the spinner unobserved by the interviewer and report only whether are not the spinner points to the letter A with probability P and to the letter B with probability* $1-P$*."* However, in the literature that followed, his mechanism was reported as a pack of cards or marbles of two colours, etc.

Let φ be the number of population proportion of 'yes' responses under the above Randomized Response Technique (RRT) . We then have

$$\varphi = P\,\pi + (1-P)(1-\pi) = (2P-1)\,\pi + (1-P)$$

and

$$\hat{\pi} = [\hat{\varphi} - (1-P)] / (2P-1)$$

with

$$V(\hat{\pi}) = \varphi(1-\varphi) / n(2P-1)^2 .$$

2.1 *Independent sub samples*

Consider a simple situation in which two independent subsamples of sizes n each are taken and the same Warner's technique is applied on individuals from each sub sample. We then have

$$\widehat{\pi^i} = [\hat{\varphi}_i - (1-P)] / (2P-1)$$

with

$$V(\widehat{\pi^i}) = \varphi(1-\varphi) / n(2P-1)^2$$

based on sub sample i , i = 1,2. A simple pooled estimator is given by the average

$$t = (\widehat{\pi^1} + \widehat{\pi^2}) / 2$$

with

$$V(t) = \varphi(1-\varphi) / 2n\,(2P-1)^2 .$$

The pooled estimator keeps a check on the feasibility of the survey technique, with respect to non sampling errors, so long as $\widehat{\pi^1}$ and $\widehat{\pi^2}$ are not too different. It also provides a more efficient simple estimator. It may be noted that the independent samples considered here are different from those in Horvitz *et al.* (1967).

Concerned with practical questions such as '*how likely are people to cooperate and tell the truth'*, Warner perhaps did not wish to go through the rigours of parametric space for the maximum likelihood estimator and such finer details. He concentrated more in sending the message that his RRT can be profitably employed in the given situations. He exhibited several calculations and explanations in his short paper (cf. Rao and Rao, 2016). It may be noted that he *did* observe that '*possibility of* $\hat{\pi}$ *taking values outside the range cannot be ruled out'.* Various situations for the range of the estimate were discussed by several authors and rigorous maximum likelihood estimators are provided, which are described in standard text books (see for example, references in Mukhopadhyay, 1998).

In section 2 of the paper, Warner assumes that "*every person in a population belongs to either Group A or Group B and it is required to estimate by survey the proportion belonging to Group A"*. It may be noted that he did not use the notation A and its complement A^C . He might have wished to estimate population proportion of persons in 2 Groups A and B, both sensitive. But his assumption mentioned above makes B as A^C . If he had thought of two sensitive characteristics A and B, then under the current set up, he might have obtained

$\varphi = P\,\pi_A + (1-P)\,\pi_B$

which could not be solved and hence perhaps made his assumption that B is A^C using which the algebra is worked out. A couple of years later, Abu-Ela *et al.* (1967) extended Warner's technique to more than two categories 'each of which was in varying degrees potentially harmful or stigmatizing'. However, it was Greenberg *et al.* (1969) who instead of the complementary Statement of Warner, introduced Simmons' (see, Horvitz *et al.*, 1967) Statement relating to the 'membership in Group Y carrying no possible embarrassment or condemning quality' and nicely worked out the algebra in both situations when π', the population proportion belonging to Group Y is unknown as well as known. This case is considered in the next sub section.

2.2. *Unrelated model*

To increase the likelihood of cooperation by the respondent, having noticed that Warner's technique consists of statements both of which are sensitive in nature, Simpson [see Horvitz *et al*., 1967; Greenberg *et al*., 1969] altered Statement 2 to be innocuous and unrelated to Statement 1. Under this unrelated model, we now have

$$\hat{\pi} = [\hat{\varphi} - (1- P)\pi'] / P$$

with

$$V(\hat{\pi}) = \varphi(1-\varphi) / nP^2$$

where π' the population proportion of yes answers to the unrelated statement 2.

Greenberg *et al*. (1969) note that the respondent cooperates because his 'personal privacy with respect to the sensitive characteristic is maintained'. They further remark that 'the interviewer, not having the confidentiality privileges of the doctor or priest, is also protected because he cannot interpret with any certain assurance the meaning of respondent's reply'.

They also consider the case of estimating the population proportions of stigmatizing characteristic and the innocuous characteristic based on two independent samples of sizes n_1 and n_2 with randomization devices exhibiting different probabilities. As a special case they derive Warner's estimate based on only one sample of size *n*. This estimate is different from the pooled estimator of the above section.

2.3. *Independent subsamples* (π' known)

As in Warner's case, here also one can think of a simple situation of taking two independent subsamples of size n each. Greenberg *et al*.'s estimators of φ, the population proportion of 'yes' responses for the *i* th subsample are given by

$$\widehat{\pi^i} = [\widehat{\varphi_i} - (1- P)\pi'] / P$$

with

$V(\widehat{\pi^{\prime}}) = \varphi\,(1-\varphi) / nP^2$.

A simple pooled estimator is given by

$t = [\,\widehat{\varphi_1} + \widehat{\varphi_2} - 2\,(1-P)\,\pi'] / 2P$

with

$V(t) = \varphi\,(1-\varphi) / 2\,nP^2$.

If π' is unknown, for each subsample, two independent samples need be taken and the procedure of Greenberg *et al.* (1969) is followed.

In Warner's as well as Greenberg *et al.*'s techniques, it is possible that the respondent is unhappy about the *P* being chosen by the investigator. To counter this, Rao and Rao (2016) discussed the 'Reversed RRT'. In the next section we shall present an alternative method, where *P* is fixed by the respondent (subject to the usual specifications) and the interviewer has an addition task of estimating *P* as well.

2.4. *RRT with P chosen by the respondent*

First let us consider Warner's model:
In addition to the two statements of Warner, namely:

Statement 1: I belong to the sensitive category A
Statement 2: I do not belong to the sensitive category A ,

present to the respondent two more statements, *viz*.,

Statement 3: I belong to an unrelated category U (to be specified)
Statement 4: I do not belong to U.

The respondent is asked to make his own randomization device(for example, he is given a number of marbles of two colours , say Red (R) and White (W) and asked to make a device containing certain number of marbles of each colour adding up to, say, *N* . The respondent then generates a probability mechanism by choosing Statement 1 whenever R is randomly picked, *i.e.* with his own choice of *P*, the probability of Statement 1 being number of R's /*N*, (subject to the usual restrictions) .

Respondent chooses Statement 2 if W is picked (with probability 1– *P*, equal to number of W's/ *N*).He repeats this with Statements 3 and 4 (with same Probabilities *P* and 1–*P*) .Thus using his mechanism, he gives two answers (yes, yes), the first answer being for the first pair of two statements as in Warner's case and the second answer for the second pair, which are recorded by the investigator. The same device (and choice of *P*, not known to the investigator) is passed on to the next sampled respondent (possibly in a sealed envelope or by text/phone contact or by a prior resolve if the sample is from a group such as a college or a block of apartments etc.). Investigator now has an estimate of φ_i , the proportion of 'yes' answers for the i^{th} pair, *i*= 1,2. Thus we have

$$\varphi_1 = P\,\pi + (1-P)(1-\pi) \qquad (1)$$

and

$$\varphi_2 = P\,\pi' + (1-P)(1-\pi') \qquad (2)$$

with the same notation as before.

Here π' is assumed to be known or can be obtained by the usual techniques.

From (2) we get

$$\widehat{P} = (\widehat{\varphi_2} + \pi' - 1) / (2\pi' - 1)$$

$$2\widehat{P} - 1 = (2\widehat{\varphi_2} - 1) / (2\pi' - 1)$$

giving $\hat{\pi} = [\widehat{\varphi_1} - (1-P)] / (2P-1)$

$$= [\widehat{\varphi_1}(2\pi' - 1) + \widehat{\varphi_2} - \pi'] / (2\widehat{\varphi_2} - 1)$$

with conditions for feasible values of probabilities.

We observe that this is a ratio estimator and its mean sq uare error can be derived in the usual way. If the study is conducted by taking two independent subsamples (say by two investigators with the same *P*, or different P's, we can get two independent estimates of π and the usual simple estimate of the error of the pooled average.

This estimator has to be compared with and validated against Yu *et al*. (2008).

Remark. If *P* is chosen by the investigator, then based on the sample, he has

$\widehat{\pi'} = [\widehat{\varphi_2} - (1-P)] / (2P-1)$.

Furthermore, if we know the actual π', then we can make a ratio or regression estimate such as

$\hat{\pi}$ (ratio) = $(\hat{\pi} / \widehat{\pi'}) \pi'$

where $\hat{\pi}$ is as defined before.

Here the choice of known 'unrelated 'characteristic is important. The question is to find an 'unrelated' (in the sense of sensitivity) characteristic for which population parameters are known which is 'related' (in the sense of correlation) to the sensitive characteristic so that
ratio or regression adjustments work out well.

3. Conclusion

During the past 50 years several important contributions mostly theoretical and some practical have appeared in the literature on the subject of RRT. However, some simple classroom- notes type questions crop up from time to time. It is our aim in this paper to pose these questions and keep the spirit of Warner's paper alive.

At a time when society was changing and measurement of certain social phenomena by statistical sampling techniques of eliciting information through direct questioning was getting increasingly difficult, Warner came up with this elegant technique. In the past, questions relating to matters such as adoption, pay scales, frequent trips, health issues, etc. were not considered to be sensitive. Around seven decades ago, to avoid costly call-backs by collecting data in the first attempt itself, Hartley-Politz-Simmons (HPS) (Hartley, 1946; Politz and Simmons, 1949) suggested an ingenious technique. According to this technique, the respondent is asked a question about his/her presence at home at the same time of the interview during the preceding five week nights. However, in

the present day context of ever changing society, this question has become a 'sensitive' one [Rao, Sarkar and Sinha (RSS), 2016]. RSS (2016) show how RRT can be superimposed and present a Randomized HPS technique.

Increasing and sometimes indiscriminate use of 'Apps' on mobile phones for routine chores such as location based services has now become child's play and it is well known that there is a great risk to individual privacy. RR techniques and algorithms have found their way to guard against such attacks. During the recent years the subject is playing a major role in communications engineering (see Rao and Rao, 2016). For referrals in e- commerce, Collaborative Filtering (CF) algorithms are combined with RRT to protect privacy of users (Polat and Du, 2006). Recent research discusses further advances in the use of RR in privacy preserving data collection (for example, see Sun *et al.*, 2014; Wang *et al.*, 2016).

Finally, we may add that Internet of Things (IoT) is the buzz word of 2015 and years ahead. While there are tremendous advantages leading towards a changing and smart society, IoT and the resulting Big Data analytics may lead to several privacy concerns. Brown(2015) suggests that the best practice is to ensure 'security and privacy from outset of IoT system design process and development of co-regulation by all stakeholders to protect security and privacy' besides 'further development of privacy and consumer protection rules to ensure security testing of IoT systems that process sensitive personal data'. Guadagni *et al.* (2015) in a recent paper discuss the major challenges one faces with respect to security, data protection and privacy, especially in the context of e-health which is a very sensitive topic.

We hope to see the use of Randomized Response Technique in the applications of IoT in the near future.

Greenberg *et al*. (1969) speculated that "the legendary Pandora's box has been opened with Warner's technique"---It is not closed even after five decades!!

References

Abu-Ela, Abdel-Latif, A., Greenberg, B.G. and Horvitz, D.G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, **62**, 990-1008.

Brown, I. (2015). Regulation and the Internet of Things. *15th Global Symposium for Regulators (GSR15)-ppt presentation*.

Chaudhuri, A. and Mukerjee, R. (1987). Randomized response techniques: A review. *Statistica* Neerlandica, **41**, 27-44.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response*. Statistics Textbooks and Monographs, **85**, Mercel Dekker Inc., N.Y.

Christophides, T. (2016). The Classical Randomized Response Techniques: Reading Warner (1965) and Greenberg et al. (1969). 50 years later, In: *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Handbook of Statistics*, Volume 34, (Eds.) Chaudhuri, A., Christofides, T.C. and Rao, C.R., North Holland, Netherlands, 29-41

Fox, J.A. and Tracy, P.E. (1986). *Randomized Response: A method for sensitive surveys*. Beverly Hills, CA.

Greenberg, B.G., Abul-Ela, A.L.A., Simmons,W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, **64**, 520-539.

Guadagni, F., Scarpato, N., Ferroni, P. and Lisi, A. (2015). Personal and sensitive data in the e-health – IoT universe. *Second International Conference on Cognitive Internet of Things Technologies* (*COIOTE*2015), Rome.

Hartley, H.O.(1946). Discussion on paper by F. Yates. *Journal of the Royal Statistical Society*, **109**, 37–38.

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

Horvitz, D.G., Greenberg, B.G. and Abernathy, J.R. (1975). Recent developments in RR designs. In: J.N. Srivastava (Ed.), *A Survey of Statistical design and Linear Models,* North Holland, Amsterdam.

Mukhopadhyay, P. (1998). *Theory and Methods of Survey Sampling*. Prentice Hall of India.

Polat, H. and Du, W. (2006). Achieving private recommendations using randomized response techniques, In: *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference: Proceedings* (Eds.) Ng,W.K., Kitsuregawa, M. and Li, J., 637-646, Springer-Verlag, Berlin.

Politz, A., Simmons, W. (1949). An attempt to get "not at homes" into the sample without callbacks. *Journal of the American Statistical Association*, **44**, 9–16.

Rao, T.J. and Rao, C.R. (2016). Review of certain recent advances in Randomized Response Techniques. In: *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Handbook of Statistics*, Volume 34,(Eds.) Chaudhuri, A., Christofides,T.C. and Rao, C.R., North Holland, Netherlands,1-11.

Rao, T.J., Sarkar, J. and Sinha, B.K. (2016). Randomized response and new thoughts on Politz-Simmons technique. In: *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques*: *Qualitative and Quantitative Human Traits, Handbook of Statistics*, Volume 34,(Eds.) Chaudhuri, A., Christofides, T.C. and Rao, C.R., North Holland, Netherlands, 233-251.

Sun, C., Fu, Y. Zhou, J. and Gao, H. (2014). Personalized Privacy-Preserving Frequent Item set Mining Using Randomized Response. *Scientific World Journal*.

Verdooren, L.R. (1976). Loten bij delicate vragen een overzicht van 'randomized response'—technieken, *Statistica Neerlandica*, **30**, 7-24.

Wang, Y, Wu, X. and Hu, D. (2016). Using Randomized Response for differential privacy preserving data collection. Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference, Bordeaux, France.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

Yu, J-W., Tian, G-L. and Tang, M-L. (2008).Two new models for survey sampling with sensitive characteristic: design and analysis, *Metrika*, **67**, 251–263.