A REVIEW OF RANDOMIZED RESPONSES PROCEDURES: THE QUALITATIVE VARIABLE CASE

Carlos N. Bouza[1]* Carmelo Herrera** and Pasha G. Mitra***
*Facultad de Matemática y Computación. Universidad de La Habana
**Facultad de Ingenieria, Universidad de Las Palas de Gran Canaria
*** Department of Computer Sciences, College of Management Sciences and Business Administration

ABSTRACT
In this paper a review on the results on randomized responses for qualitative variables is presented. It incorporates recently published papers. It tries to fulfill the lack of a review on the theme and aims completing the oeuvre of Chaudhuri, A . & R. Mukerjee. (1988), (Randomized response: theory and techniques, Marcel Dekker, New York).

KEY WORDS: randomized responses, truthful response, response bias

MSC 62D05

RESUMEN
Este trabajo presenta una revisión crítica de resultados sobre respuestas aleatorizadas para variables cualitativas. Este incorpora trabajos recientemente publicados en el tema. Tratamos de rellenar la falta de revisiones en el tema y se aspira a completar la presentada en la obra de Chaudhuri, A . & R. Mukerjee. (1988), (Randomized response: theory and techniques, Marcel Dekker, New York).

1. INTRODUCTION

Randomized Response (RR) techniques were developed for the purpose of protecting surveyee's privacy and avoiding answer bias mainly. They were introduced by Warner (1965) as a technique to estimate the percentage of people in a population *U* that has a stigmatizing attribute *A.* In such cases respondents may decide not to reply at all or to incorrectly answer. The usual problem faced by researchers is to encourage participants to respond, and then to provide truthful response in surveys. The RR technique was designed to reduce both response bias and non-response bias, in surveys which ask sensitive questions. It uses probability theory to protect the privacy of an individual's response, and has been used successfully in several sensitive research areas, such as abortion, drugs and assault. The basic idea of RR is to scramble the data in such a way that the real status of the respondent can not be identified.

Different modifications of Warner's (1965) RR were developed by various authors, including Greenberg et al. (1969), Horvitz et al. (1967), Moors (1971), Raghavarao (1978), Mangat and Singh (1990), Kuk (1990), Mangat (1994), Mangat, Singh and Singh (2000), Haung (2004), Kim and Warde (2004), Chang et al. (2004) and Gjestvang and Singh (2006) among many others.

[1] Corresponding autor
E-mail: *bouza@matcom.uh.cu*

The paper is divided into two section 1 is devoted to presenting the main procedures developed during the XXth century. The second one is concerned with the procedures developed since 2001. The nature of the procedures makes impossible to develop a comparison analytically, and any numerical experiment would depend on a series of parameters that must be fixed by the surveyer. Hence, to implement a global comparison using Monte Carlo experiments would not bring any concluding result.

We present this review for interested readers and looks to complete other existing ones, as Chaudhury and Mukharjee (1988).

2. PROPOSALS DURING THE XXth CENTURY

Warner(1965) introduced the method of randomized response (RR) for surveying human populations for obtaining information on variables of sensitive protecting the anonymity of the respondents. Warner's model is as follows.

Consider a dichotomous finite population U of size $/U/= N$ stratified into two strata with respect to the possible responses. $U(A)$ is the stratum that contains the sensitive group A or the non-sensitive group A *belongs to* $U(B)=U\backslash A$. The purpose is to estimate population proportion θ of persons in $U(A)$. A simple random sampling with replacement (SRSWR) or without replacement (SRSWOR) is the sampling design. A sample s of size $/s/= n$. Each individual performs a Random trial consisting of two questions:

P(W).RR procedure of Warner (1965)

- (a) I belong to $U(A)$, (I am a member of the stigmatizing group A)
- (b) I do not belong to $U(A)$.

Each interviewed answers the question

(a) with probability P and the question
(b) with probability $1 - P=Q$

Using a randomized device, such as a spinner, balls of different colors etc. Therefore, the individual does not reveal whether he/she belong to the stigmatizing group or not. The response reported is

$$Z_i = \begin{cases} 1 \text{ if the response is 'yes'} \\ 0 \text{ if the response is 'not'} \end{cases}$$

Computing

$$n_Y = \sum_{i=1}^{n} Z_i$$

Warner (1965) proposed that the *ith i*nterviewed in a SRSWR sampling is provided of a randomization device which selects between two questions

The respondent does not declare the outcome of the randomization device. He/she only reports "yes or no". A truthful response is expected because the privacy is granted.
Take $\theta=P(A)$ as the probability of belonging to A. The proposal of Warner (1965) as

$$P_Y = P\theta + Q(1-\theta)$$

we have the following results

Proposition (Warner, 1965). Use procedure *P(W)* and take *P≠1/2*, then $\hat{\theta}(W) = \frac{\frac{n_Y}{n} - Q}{2P-1}$, where n_Y is the number of "Yes" in a srswr simple of size *n* , is unbiased with variance $V(\hat{\theta}(W)) = \frac{\theta(1-\theta)}{n} + \frac{PQ}{n(2P-1)^2}$

Greenberg et al. (1969) proposed an unbiased estimator of *θ(A)* . Two independent samples are selected using srswr, say *s(1)* and *s(2)*, with sizes n_1 and n_2, $n=n_1+n_2$. The individuals in sample *s(i)* are provided with a randomization device *R(j)*. It selects between two questions. The improved RR procedure is described as follows

P(G) . RR procedure of Greenberg (1971)

- Select two independent samples s(1) and s(2) and select one of them with probability p_j or question or s(2) with probability $q_j=1-p_j$.
- Take P(Y|j) as the probability of obtaining a "yes" ´in the sample s(j).

The Bernoulli random variable

$$Z(i|j) = \begin{cases} 1 \text{ if individual i}\in\text{s(j) responds "yes"} \\ 0 \text{ if individual i}\in\text{s(j) responds "not"} \end{cases}$$

Allows calculating the estimator of the probability of "Yes" in each sample

$$p(Y|j) = \sum_{i=1}^{n_j} \frac{Z(i|j)}{n_j} = \frac{n_j(Y)}{n_j}$$

Proposition (Greenberg et al. (1969)) The estimator

$$\theta\ (G) = \frac{q_2 p(Y|1) + q_2 p(Y|2)}{p_1 - p_2}, \qquad p_1 \neq p_2$$

Is unbiased and

$$V(\theta(G)) = \frac{\frac{q2_2 P(Y|1)Q(Y|1) +}{n_1} + \frac{q2_2 P(Y|2)Q(Y|2)}{n_2}}{(p_1 - p_2)^2}, \qquad p_1 \neq p_2$$

When srswr is used and for srswor

$$V(\theta(G)|srswor) = V(\theta(G)) - \frac{\frac{q2_2(n_1-1)[p_1^2 PQ + p_1^2 P(Y)Q(Y)} + \frac{q2_2 P(Y|2)Q(Y|2)}{n_1} \quad n_2}{(p_1-p_2)^2}$$

Greenberg et al. (1969) recommended that one of the values of the design probabilities should be close to zero and the other close to one.

Chaudhuri-Mukerjee's (1988) modified Warner's (1965) method. The procedure is

P(CM) RR procedure of Chaudhuri-Mukerjee's (1988)

- Perform a random experiment and report *A* with probability T_1 or its complement *A** with probability T_2.
- Report $r = (I - T_2)/(T_1 - T_2)$,

The variable generated is

$$I = \begin{cases} 1 & \text{if the characteristic A or } A* \text{ matches} \\ 0 & \text{if the characteristic A or } A* \text{ does not matches} \end{cases}$$

Under the procedure its expectation is

$E\ (I\backslash P(CM)\ ) = T_1 y + T_2(1 - y\ )$

Hence the expectation of the report is

$E(\ r_i) = (T_1 y_i + T_2(1 - y_i\ ) - T_2)/(T_1 - T_2) = y_i,$

and

$$V_i = V\ (r_i) = E\ (I_i)(1 - E\ (I_i))/(T_1 - T_2)^2 = \begin{cases} \frac{T_1(1-T_1)}{(T_1-T_2)^2} & \text{if } y_i = 1 \\ \frac{T_2(1-T_2)}{(T_1-T_2)^2} & \text{if } y_i = 0 \end{cases}$$

Proposition (Chaudhuri-Mukerjee's (1988)) Use P(CM) each respondent reports r_i, i=1,…,n. Then

$$\hat{\theta}_{CM}(A) = \frac{\sum_{i=1}^{n} r_i}{n}$$

It is unbiased and its variance is

$$V(\hat{\theta}_{CM}(A)) = \begin{cases} \frac{T_1(1-T_1)}{n(T_1-T_2)^2} & \text{if } y_i = 1 \\ \frac{T_2(1-T_2)}{n(T_1-T_2)^2} & \text{if } y_i = 0 \end{cases}$$

Mangat (1990) proposed to use SRWR and two random devices $R`$ and $R``$. Using the random device $R`$ the interviewed selects between the statements

P(M1) RR procedure of Mangat (1990)

- I belong to stigmatizing group A"
- Use the device R ".

With probabilities P_1 and $Q_1=1-P_1$
The random second device $R``$ poses the alternative questions

- I am a member of the stigmatizing group A
- I am not a member of A.

with probabilities P_1 and $Q_2=1-P_2$. The respondent is asked to use first the random device and redirected with probability Q_2.

In this case

$$P_Y = P_1\theta + Q_1(P_2\theta + Q_2(1 - \theta))$$

Mangat (1994) proposed the simpler procedure

P(M2) RR procedure of Mangat (1990)

- Report a "yes" if you are a member of the stigmatizing group A
- If not go to P(W).

The results associated with the work of Mangat are resumed in the following proposition.

Proposition (Mangat , 1990, 1994). Take n_Y as the number of "Yes" in a srswr simple of size n

- If P(M1) is used and $P_2 \neq 1/2$ then an unbiased estimotro of the P(A) is $\hat{\theta}(M1) = \frac{\frac{n_Y}{n} - Q_1 Q_2}{2P_2 - 1 + 2P_1 Q_2}$, with variance $V(\hat{\theta}(Y)) = \frac{\theta(1-\theta)}{n} + \frac{P_1 Q_2 (1 - Q_1 Q_2)}{n(2P_2 - 1 + 2P_1 Q_2)^2}$.
- If PM2) is used, then an unbiased estimotro of P (A) is $\hat{\theta}(M2) = \frac{\frac{n_Y}{n} - Q}{P}$, with variance $(\hat{\theta}(Y)) = \frac{\theta(1-\theta)}{n} + \frac{Q\ (1-\theta)}{nP}$.□

RR procedures have been considered as a method for reducing also no responses rates, see Bouza (1981) and (1985), Van der Heijden et al. (1998) for examples.

3. PROPOSALS MADE DURING THIS CENTURY

A variation of Warner`s method was proposed by Christofides (2003). A set of of numbers {1,…,L} is fixed by the surveyer. A random mechanism selects a number form that set with an assigned probability. Then we know the mean and variance are $\mu(e), \sigma^2(e)$. He proposed the use of the following procedure

P(Ch) RR procedure Christofides (2003).

Generate $e \in \{1,...,L\}$ with probability $P(e=j)=p_j, j=1,...L$
Report if you carry the stigma $d=L+1-e$
Report $d=0$ otherwise.

The distributional problem is described by noting that the response is

$$d_i = \begin{cases} L_i + 1 - e_i \text{ if the respondent belongs to A} \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}$$

As $P(X_i=L+1)=\theta$, $P(X_i=0)=1-\theta$ we have that the probability that a random report be equal to k is $P(d_i=k)=\theta p_k + (1-\theta)p_{L+1-k}$, $k=1,.,,L$

Then $E(d_i)=\mu_e + \theta(L+1-2\mu_e)=\mu_d, E(e_i)=\sum_{k=1}^{L} kp_k = \mu_e$ and $V(e_i)=\sum_{k=1}^{L}(k-\mu_e)^2 p_k = \sigma_e^2$

The behavior of this procedure for quantitive variables is fixed as follows

Proposition (Christofides (2003)) The estimator

$$\hat{\theta}_{Ch}(A) = \frac{d_s - \mu(e)}{L + 1 - 2\mu(e)}$$

where, $d_s=\sum_{i=1}^{n} d_i/n$, is unbiased an its variance is determined as

$$V(\hat{\theta}_{Ch}(A)) = \frac{V(d_s)}{(L+1-2\mu(e))^2} = \frac{\theta(A)(1-\theta(A))}{n} + \frac{\sigma^2(e)}{n(L+1-2\mu(e))^2}$$□

Another proposal was made by Haung (2004) when srswr is used for selecting a sample of size n. It can be described as follows

P(H1) RR procedure of Huang (2004)

- Question 1: Report if you belong to the sensitive group or not with probability T.
- Question 2: If the answer is "no", use a randomization device and select one of the following statements

(i) "I am a member of group A",
(ii) "I am not a member of group A",
with probabilities S and 1- S respectively.

Of course the surveyer assumes that the respondent belonging to group *A* respond honestly
The strategy due Haung's (2004) procedure is characterized as follows

Proposition (Huang 2004). Take n_{Yt} as the number of "Yes" in a srswr simple of size *n* to the question *t=1* if and use P(H1). The probability of obtaining a "Yes" to the first question is
$\theta_i=T\ \theta(A)$
and to the randomized question
$\theta_{ii}=S(1-T)+(1-S)\ \theta(A)$
An unbiased estimator of the $\theta(A)$ is
$\hat{\theta}(H)=\frac{\frac{Sn_{Y1}+n_{Y2}}{n}-(1-S)}{2S-1}$, with variance $\left(\hat{\theta}(H)\right)=\frac{\theta(1-\theta)}{n}+\frac{S(1-S)(1-\theta T)}{n(2S-1)^2}$.□

The proposed estimator is more efficient Warner for any *S* and *T*.
In the same year Kim and Warde (2004) suggested to the RR procedure

P(KW) RR procedure of Kim-Warde (2004)

- Use a Warner´s randomization device *R* that selects with probabilities *T* and *1-T* between the direct sensitive question "I a member of the stigmatizing group" and its counter part "I am not a member of the stigmatizing group" .
- If the answers is 'no' use again *P(W)* with probabilities *S* and *1-S* .

Each set of respondent provides an estimator of *P(A)*. Fort he sample directed by the randomization device *t* the probability of yes is $P^t{}_Y=P_t\ \theta+Q_t$. $Q_t=1-P_t$, Hence
$\hat{\theta}(KW)=\frac{\frac{n_{Y(1)}}{n(1)}-Q_1}{P_1}$ and $V(\hat{\theta}(HW(1)))=\frac{(1-\theta)(P_1\theta+Q_1)}{n_1P_1}$

Proposition (Kim-Warde, 2004) Use P(KW) and assume that $n(1)\in]0, n[$. Take $\hat{\theta}(KW(1))=\frac{\frac{n_{Y(1)}}{n(1)}-Q_1}{P_1}$ and
$\hat{\theta}(KW(2))=\frac{\frac{n_{Y(2)}}{n(2)}-Q_2}{2Q_2-1}$ as estimators usign the sets. The estimator

$$\hat{\theta}(KW)=\sum_{t=1}^{2}\frac{n(t)}{n}\hat{\theta}(HW(t))$$

is unbiased and its variance is

$$V(\hat{\theta}(HW))=\frac{\theta(1-\theta)}{n}+\frac{Q_1(n(1)P_1(1-)\theta+1-\theta}{n^2P_1^2}$$□

This result follows by taking the variances of the individual estimators

$V(\hat{\theta}(HW(1)))=\frac{(1-\theta)(P_1\theta+Q_1)}{n(1)P_1}$ and $V(\hat{\theta}(HW(2)))=\frac{\theta(1-\theta)}{n(2)}+\frac{P_2Q_2}{n(2)(2Q_2-1)^2}$

As the proposed estimator is a linear function of the estimators each one is itself an estimator of P(A) .

Gjestvang-Singh (2006) proposed that if the interviewed selected belongs to the sensitive group a randomization device must be used. The surveyer fixes Let α_h and β_h, real numbers and a non-directional scrambling variable S_h, h=1,2, with $V(S_h)=\delta_h^2$ known

P(GS) RR procedure of Gjestvang-Singh (2006)

- If the interviewed carries the stigma use the randomization procedure R(1), that chooses with probability $t=\alpha_1/(\alpha_1+\beta_1)$, for reporting a scrambled response $y=1+\beta_1 S_1$, and with $q=1-p$ report $y=1-\alpha_1 S_1$,
- If the interviewed does not carry the stigma use the randomization procedure R(2)) that chooses with probability $p^*=\alpha_2/(\alpha_2+\beta_2)$ reporting a scrambled response $y=1+\beta_2 S_2$, and with $y=1-p$ report $y=-\alpha_2 S_2$,

Proposition (Gjestvang-Singh (2006)). Use P(GS) in a srswr simple of size n

$$\hat{\theta}_{GS}(A)=\frac{\sum_{i=1}^{n} Y_i}{n}$$

is unbiased for the true probability $\theta(A)$ and if

$$\frac{\alpha_2\beta_2}{\alpha_1\beta_1}=\frac{\delta_1^2+\tau_1^2}{\delta_2^2+\tau_2^2},\ \delta_2^2+\tau_2^2=1$$

Then

$$V(\hat{\theta}_{GS}(A))=\frac{\theta(A)(1-\theta(A))}{n}+\frac{\alpha_2\beta_2}{n}$$

This variance can be made smaller than the variances of the estimators proposed by Warner (1965) and the Mangat and Singh (1990) fixing adequately the values of α_2 and β_2 .

RR has been used in complex designs. See for example the development of it in a two-stage model, Saha (2006) and Bouza (2008) in ranked set sampling.

Acknowledgments. This paper was developed under the auspices of collaboration between the universities of La Habana, Cuba, and Las Palmas de Gran Canarias, Spain.

RECEIVED FEBRUARY 2010
REVISED JUNE 2010

REFERENCES

[1] BOUZA,C. (1981): Bias and non response rate reduction using a randomized response model among the nr. **Survey Statistician**, 5, 3-5.

[2] BOUZA,C. (1985): Evaluation of the subsample from the nr by a random response design. **Test,** 27, 799-805.

[3] BOUZA, C. (2008): Ranked set sampling and randomized response procedures for estimating the mean of a sensitive quantitative character . **Metrika**, doi 10.1007/s00184-008-0191-6.

[4] CHANG, H., WANG, C. and HAUNG, K. (2004): On estimating the proportion of a qualitative sensitive character using Randomized response sampling. **Quality and Quantity** 38,675-680.

[5] CHAUDHURI, A . and R. MUKERJEE (1988): **Randomized response: theory and techniques**, Marcel Dekker, New York.

[6] CHRISTOFIDES, T.C. (2003): A generalized response technique. **Metrik**a. 57, 195-200.

[7] GJESTVANG, C. R. and SINGH, S. (2006): A new randomized response model. **J. Royal Statist. Soc Ser. B**, 68, 523-530.

[8] GREENBERG, B. G., KUEBLER, R. R., JR., ABERNATHY, J. R., and HOVERTIZ, D. G. (1969): The unrelated question randomized response model: theoretical framework. **J. Amer. Statist.Assoc**., 64, 520-539.

[9] HAUNG, K. (2004): A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. **Statistica Neerlandica** 58, 75-82.

[10] HORVITZ, D. G., SHAH, B.V., and SIMMMONS, W. R.(1967): The unrelated question randomized response model. **Proceedings of Social Stati. Sec. Amer. Statist. Assoc**., 65-72.

[11] KIM J. and WARDE W.D., (2004): A mixed randomized response model. **J., Statist. Plann.Inference**, 110, 1-11.

[12] KUK, A. Y. C. (1990): Asking sensitive questions directly. **Biometrika** 77, 436-438.

[13] MANGAT, N. S. and SINGH, R. (1990): An alternative randomized response procedure. **Biometrika** 77, 439-442.

[14] MANGAT, N. S. (1994): An improved randomized response strategy. **J. Royal Statist. Soc Ser. B**,56, 93-95.

[15] MOORS, J. J. A. (1971): Optimization of the unrelated question randomized response model. J**. Amer. Statist**. **Assoc**., 66, 627-629.

[16] RAGHAVARAO, D. (1978): On an estimation problem in Warner's randomized response technique. **Biometrics** 34, 87-90.

[17] RYU J.-B, KIM J.-M, HEO T-Y and PARK C G (2005): On stratified Randomized response sampling, **Model Assisted Statistics and Application** 1, 31-36.

[18] SAHA, AMITA (2006): A generalized two-stage randomized response procedure in complex sample surveys**. Aust. N. Z. J. Stat**. 48**,** 429–443.

[19] SINGH, S. (2002): Randomized response model. **Metrika**, 56, 131-142.

[20] SINGH S, R SINGH, and N. S. MANGAT (2000) Some alternative strategies to Moor's model in randomized response sampling. **Journal of Statistical Planning and inference** 83, 243-255.

[21] VAN DER HEIJDEN, P.G.M., VAN GILS, G., BOUTS, J. and HOX, J. (1998): A comparison of randomised response, CASAQ, and direct questioning; eliciting sensitive information in the context of social security fraud. **Kwantitative Methoden** 19, 15-34.

[22] WARNER, S. L. (1965):Randomized response: a survey technique for eliminating evasive answer bias. **J. Amer. Statist. Assoc**. 60, 63-69