A NEW METHOD OF ESTIMAION IN OPTIONAL RANDOMIZED RESPONSE TECHNIQUES FOR QUANTITATIVE CHARACTERISTICS

R. Arnab

University of Botswana, Botswana

Abstract

Gupta et al. (2002) and Huang (2010) considered optional randomized response techniques where the probability of choosing the randomized (or direct) response is fixed for all the respondents. In this paper the assumption of the constant probability of choosing the option has been relaxed by dividing respondents into two groups: one group provides direct response and the second a randomized response.

The method of estimation of the population mean and variances under the modified assumptions are obtained. Relative efficiencies of the proposed techniques are compared theoretically and empirically.

1. Introduction

In surveys related to sensitive issues such as domestic violence, induced abortions and use of the illegal drugs, direct questioning methods of interview, the respondents deliberately provide socially desirable answers or refuse to respond entirely due to social stigma and/or fear that their personal information may be disclosed to the third parties.

To improve cooperation from respondents and get more truthful answers from them, Warner (1965) proposed the randomized response (RR) technique where respondents provide indirect responses.

Thus RR techniques provide reliable data, protect respondents' confidentiality and avoid high rate of nonresponses.

1.1. The pioneering method (Warner, 1965)

The proposed RR technique was used to estimate π, the proportion of units in a population possessing a certain stigmatized character A such as HIV infection status.

A sample of size n is selected from a population of size N by simple random sampling with replacement (SRSWR) procedure.

Randomized device:

The respondent has to select a card at random from a pack of cards consists of two types of cards

Card type1:“I belong to the group A” with proportion $P(\neq 1/2)$

Card type 2 “I belong to group $\bar{A}$” with proportion $1 - P$

The respondent will supply a truthful answer

“Yes” if the statement matches his/her status and

“No” if the statement does not matches his/her status.

Here probability of obtaining the true response is

= P(Respondent ∈ A) × (Yes answer|

respondent ∈ A)

+ P (Respondent ∈ $\bar{A}$) P(No answer|

respondent ∈ $\bar{A}$)

= $\pi P + (1 - \pi)P = P$

2. Optional Randomized Response Technique (ORT)

In an ORT, most of the respondents feel that the subject of enquiry is sensitive, but a minority may feel that it is not sensitive and are therefore willing to provide direct response (DR). For example, HIV/AIDS infection status is a sensitive issue for most people but some respondents are nevertheless willing to reveal their status to the interviewer.

A randomized response technique which provides the opportunity to give DR instead of making RR compulsory to all the respondents is known as an ORT.

Accordingly in an ORT, respondents provide RR if they feel the subject of enquiry is sensitive, but provide DR if they feel that the enquiry is not sufficiently sensitive to require anonymity. ORT was introduced by Chaudhuri and Mukherjee (1988).

ORT can be classified into two categories

(see Arnab and Rueda, 2016)

Full optional RR technique (FORT) and

Partial optional RR technique (PORT)

FORT: Population is divided into two groups:

G: Respondents always provide RR

$\bar{G}$: Respondents always provide DR

PORT: Respondents provide RR (or DR) with certain probability W (say) depending on their state of mind.

FORT was considered by Chaudhuri and Mukherjee (1988), Arnab (2004), Chaudhuri and Saha (2005) amongst others.

PORT was considered by Gupta (2002), Gupta et al. (2002), Pal (2008), among others.

2.1. Gupta et al.’s PORT(multiplicative model)

In Gupta et al.’s (2002) PORT, a sample of n respondents is selected from a population by the SRSWR method. Each of the selected respondents of the sample was asked to choose one of the following options:

(a) Report the true response y

(b) Provide a randomised response

$$z = y(x/\mu_x)$$

where x is a random sample from a pre-assigned distribution such as chi-square, Exponential, Poisson, etc.

The mean μ_x and variance σ_x^2 of x are known.

Gupta et al. (2002) assumed that each of the respondents of the population provides randomized response (i.e. choose option (a)) and direct response (i.e. choose option (b)) with probability W and $1 - W$ respectively.

Let z_i be the response obtained from the ith respondent. Then

$$z_i = \begin{cases} y_i & with\,probability\;\; W \\ y_i Q_i & with\,probability\;\; 1 - W \end{cases}$$

where $Q_i = x_i/\mu_x$ and x_i = scrambled response

Theorem 2.1.

(i) $\hat{\mu}_{1y}$ is an unbiased estimator of μ_y

(ii) The variance of $\hat{\mu}_{1y}$ is

$$V\left(\hat{\mu}_{1y}\right) = \frac{\sigma_{z1}^2}{n}$$

(iii) An unbiased estimator of $V\left(\hat{\mu}_{1y}\right)$ is

$$\hat{V}\left(\hat{\mu}_{1y}\right) = \frac{1}{n}\left[\hat{\sigma}_y^2 + \widehat{W}C_x^2\left(\hat{\sigma}_y^2 + \hat{\mu}_{1y}^2\right)\right]$$

where $\sigma_{z1}^2 = \sigma_y^2 + WC_x^2\left(\sigma_y^2 + \mu_y^2\right)$,

$$\hat{\sigma}_y^2 = \frac{s_z^2 - \widehat{W}C_x^2\mu_y^2}{1+\widehat{W}C_x^2}, s_z^2 = \frac{1}{n-1}\sum_{i\in s}\left(z_i - \bar{z}^2\right)^2$$

$$\widehat{W} = \frac{\frac{1}{n}\sum_{i\in s} log(z_i) - log\left[\frac{1}{n}\sum_{i\in s} z_i\right]}{\delta} \text{ and } \delta = E(logx)$$

2.2. Huang's PORT (2010): Additive model

Two independent samples s_1, s_2 of sizes n_1, n_2 are selected by the SRSWR method.

The jth respondent selected in the sample $s_i (i = 1,2)$ provides:

The true value: y_j with probability W

Randomized response: $z_j(i) = (x(i)/\mu_x(i))y_i + t(i)$

with probability $1 - W$

where $x(i)$ and $t(i)$ are independent random samples from pre-assigned distributions $X(i)$ and $T(i)$. The means $\mu_x(i)$, $\mu_t(i)$ and variances $\sigma_x^2(i), \sigma_t^2(i)$ of $X(i)$ and $T(i)$ are assumed to be known.

Theorem 2.2

(i) $\hat{\mu}_{2y} = \frac{\mu_t(2)\bar{z}(1) - \mu_t(1)\bar{z}(2)}{\mu_t(2) - \mu_t(1)}$ with $\mu_t(2) \neq \mu_t(1)$ is an

unbiased estimator of μ_y

(ii) The variance of $\hat{\mu}_{2y}$ is

$$V(\hat{\mu}_{2y}) = \frac{1}{\left(\mu_t(2) - \mu_t(1)\right)^2}\left[\frac{\mu_t^2(2)\sigma_{z(1)}^2}{n_1} + \frac{\mu_t^2(1)\sigma_{z(2)}^2}{n_2}\right]$$

where

$$\sigma^2_{z(i)} = \sigma^2_y + WC^2_{x(i)}\left(\sigma^2_y + \mu^2_y\right) + \ W\sigma^2_t(i)$$

$$+W(1-W)\mu^2_t(i),\ i = 1,2$$

and

$$C^2_{x(i)} = \sigma^2_x(i)/\mu^2_x(i)$$

(iii) An unbiased estimator of $V\left(\hat{\mu}_{2y}\right)$ is

$$\hat{V}\left(\hat{\mu}_{2y}\right) = \frac{1}{\left(\mu_t(2) - \mu_t(1)\right)^2}\left[\frac{\mu_t^2(2)s_{z(1)}^2}{n_1} + \frac{\mu_t^2(1)s_{z(2)}^2}{n_2}\right]$$

where $s_{z(i)}^2 = \sum_{j\in s_i}\left(z_j(i) - \bar{z}(i)\right)^2/(n_i - 1)\ ;\ i = 1,2$

3. Proposed FORT

Now consider full optional randomized response techniques (FORT) based on Gupta et al.’s (2002) multiplicative and Huang’s (2010) additive RR techniques respectively.

Under the FORT, it is assumed that the respondents are classified into two mutually exclusive and exhaustive categories G and $\bar{G}$.

Respondents belonging to the sensitive group G always provide randomized responses while respondents belonging to the non-sensitive group $\bar{G}$ provide exclusively direct responses.

3.1. Randomized Response: R1 (multiplicative model)

Under the proposed RR Technique R1, a sample s of size n is selected by Gupta et al. (2002).

i.e. $i \in \bar{G}$ provides the true value y_i

$i \in G$ provides a RR $y_i(x_i/\mu_x)$

Let z_i be the RR obtained from the ith respondent. Then

$$z_i = \begin{cases} y_i & for\, i \in \bar{G} \\ y_i Q_i & for\, i \in G \end{cases}$$

where $Q_i = x_i/\mu_x$

Using Arnab’s (2004) notation:

- $$z_i = \delta_i y_i Q_i + (1 - \delta_i) y_i$$
- where
- $$\delta_i = \begin{cases} 0 \; for \; i \in \bar{G} \\ 1 \; for \; i \in G \end{cases}$$

Theorem 3.1.

(i) $\bar{z} = \frac{1}{n}\sum_{i\in s} z_i$ is an unbiased estimator of μ_y

(ii) Variance of $\bar{z}$ is

$$Var(\bar{z}) = \frac{\bar{\sigma}_z^2}{n}$$

(iii) An unbiased estimator of $Var(\bar{z})$ is

$$\widehat{Var}(\bar{z}) = \frac{1}{n(n-1)}\sum_{i\in s}(z_i - \bar{z})^2$$

where

$$\bar{\sigma}_z^2 = \sigma_y^2 + C_x^2 W_G \mu_{yG}^2 \left(1 + C_{yG}^2\right)$$

C_x = CV of x for the entire population,

C_{yG} = CV of y for the group G , μ_{yG} = mean of y for the group G, and W_G is the proportion of persons belonging to the group G.

3.2. Comparison with Gupta et al. (2002)

The estimators for the population mean μ_y for the FORT and PORT for Gupta et al.'s (2002) RR techniques are identical

i.e. $\bar{z} = \hat{\mu}_{1y}$

However, their variances under the assumptions of FORT and PORT are not equal.

The variance of the proposed estimator $\bar{z}$ will be higher than $\hat{\mu}_{1y}$ if

$$Var(\bar{z}) - Var(\hat{\mu}_{1y}) \geq 0$$

i.e. $\sum_{i \in G} y_i^2 - W \sum_{i \in U} y_i^2 \geq 0$

i.e $W_G \mu_{yG}^2 (1 + C_{yG}^2) \geq W \mu_y^2 (1 + C_y^2)$

Particular cases A:

(i) $W = W_G$ and (ii) $C = C_{yG}$

Then $$Var(\bar{z}) \geq Var\left(\hat{\mu}_{1y}\right)$$

i.e. if $$\mu_{yG} \geq \mu_y$$

i.e. the mean of the sensitive characteristic of y for the group G is higher than the entire population mean μ_y.

The condition holds for personal incomes, incidence of involvement in domestic violence, or number of sexual partners after being diagnosed as HIV positive.

On the other hand if $\mu_{yG} \leq \mu_y$ the variance of $\bar{z}$ will be smaller than that of $\hat{\mu}_{1y}$.

Particular case B

$$W = W_G = 1$$

In this situation all respondents provide a randomized response and we have

$$\mu_{yG} = \mu_y, \;\; C_{yG} = C_y$$

and $$Var(\bar{z}) = \mathrm{Var}\left(\hat{\mu}_{1y}\right)$$

$$= \frac{1}{n}\left[\sigma_y^2 + C_x^2\mu_y^2\left(1 + C_y^2\right)\right]$$

Particular case C

$W = W_G = 0$: everybody provides a direct response and we get

$$\mu_{yG} = \mu_{y}, \quad C_{yG} = C_{y}, \quad \text{and}$$

$$Var(\bar{z}) = Var(\hat{\mu}_{1y}) = C_y^2/n$$

3.3. Estimation under general sampling scheme

The Horvitz-Thompson estimator (HTE) of the population mean μ_y under the FORT is given by

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$$

$$Var(\hat{\mu}_{HT}) = \frac{1}{N^2} \left[\sum_{i \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \right]$$

$$+ C_x^2 \sum_{i \in G} \frac{y_i^2}{\pi_i}$$

$$\widehat{Var}(\hat{\mu}_{HT}) = \frac{1}{N^2}\left[\sum_{i\in s}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right)^2\right]$$

$$+\frac{C_x^2}{1+C_x^2}\sum_{i\in s_G}\frac{z_i^2}{\pi_i}$$

where $s_G = s \cap G$

FOR SRSWOR: $\hat{\mu}_s = \frac{1}{n}\sum_{i\in s} z_i = \bar{z}_s$

$$Var(\bar{z}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \frac{1}{Nn} C_x^2 \sum_{i\in G} y_i^2$$

$$\widehat{Var}(\bar{z}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) s_z^2 + \frac{1}{Nn} \frac{C_x^2}{1 + C_x^2} \sum_{i\in s_G} z_i^2$$

where $S_y^2 = \frac{1}{N-1}\sum_{i\in U}\left(y_i - \mu_y\right)^2$ and

$$S_z^2 = \frac{1}{n-1}\sum_{i\in s}(z_i - \bar{z}_s)^2$$

4. Randomized Response, Huang (2010): Additive model (additive model)

Two independent samples s_1, s_2 of sizes n_1, n_2 are selected by the SRSWR method.

The jth respondent selected in the sample $s_i (i = 1,2)$ provides the true value y_j, if the respondent belongs to the non-sensitive group $\bar{G}$, and if the respondent belongs to the sensitive group G, report randomized response

-

$$z_j(i) = (x(i)/\mu_x(i))y_i + t(i)$$

where $x(i)$ and $t(i)$ are independent random samples from pre-assigned distributions $X(i)$ and $T(i)$. The means $\mu_x(i)$, $\mu_t(i)$ and variances $\sigma_x^2(i), \sigma_t^2(i)$ of $X(i)$ and $T(i)$ are assumed to be known.

Theorem 4.1

(i) $\hat{\mu}_{2y} = \frac{\mu_t(2)\bar{z}(1)-\mu_t(1)\bar{z}(2)}{\mu_t(2)-\mu_t(1)}$ with $\mu_t(2) \neq \mu_t(1)$ is an unbiased estimator of μ_y

(ii) The variance of $\hat{\mu}_{2y}$ is

$$V(\hat{\mu}_{2y}) = \frac{1}{(\mu_t(2)-\mu_t(1))^2}\left[\frac{\mu_t^2(2)\sigma_{z(1)}^2}{n_1} + \frac{\mu_t^2(1)\sigma_{z(2)}^2}{n_2}\right]$$

Where

$$\sigma_{z(i)}^2 = \sigma_y^2 + WC_{x(i)}^2\left(\sigma_y^2 + \mu_y^2\right) + W\sigma_t^2(i)$$

$+W(1-W)\mu_t^2(i)$, and $C_{x(i)}^2 = \sigma_x^2(i)/\mu_x^2(i)$

(iii) An unbiased estimator of $V\left(\hat{\mu}_{2y}\right)$ is

$$\hat{V}\left(\hat{\mu}_{2y}\right) = \frac{1}{\left(\mu_t(2)-\mu_t(1)\right)^2}\left[\frac{\mu_t^2(2)s_{z(1)}^2}{n_1} + \frac{\mu_t^2(1)s_{z(2)}^2}{n_2}\right]$$

where $s^2_{z(i)} = \sum_{j\in s_i}\left(z_j(i) - \bar{z}(i)\right)^2 / (n_i - 1)$

$$i = 1,2$$

4.3. Comparison with Huang (2010) estimator

(i) The proposed estimator $\bar{\tilde{z}}$ for FORT is identical to the Huang et al. (2010) estimator $\hat{\mu}_{2y}$ but their variances are not equal.

(ii) If $C_y = C_{yG}$ and $\mu_y = \mu_{yG}$ both the estimators $\bar{\tilde{z}}$ and $\hat{\mu}_{2y}$ are equally efficient.

(iii) $W = W_G = 1$: In this situation all respondents provides randomized response and $V(\bar{\tilde{z}}) = V(\hat{\mu}_{2y})$.

(iv) $W = W_G = 0$: In this situation all respondents provides randomized response and $V(\bar{\tilde{z}}) = V(\hat{\mu}_{2y})$.

$$\mu_{yG} = \sum\nolimits_{j\in G} y_j \,/N_G,\ \sigma_{yG}^2 = \frac{\sum_{j\in G} y_j^2}{N_G} - \mu_{yG}^2$$

and

$$s_{\tilde{z}(i)}^2 = \sum\nolimits_{j\in s_i}\left(\tilde{z}_j(i) - \bar{\tilde{z}}(i)\right)^2 /(n_i - 1)\ ;$$

$$i = 1{,}2$$

5. Concluding Remarks

In surveys relating to sensitive subjects, respondents often provide socially desirable answers due to social stigma or fear.

Randomized response (RR) techniques may be used to collect a better quality of data and reduce instances of nonresponse, as this method protects respondents' privacy.

In an optional randomized response technique (ORT), respondents are asked to choose one of the two options: (a) provide direct response or (b) provide a randomized response.

In the partial optional response technique (PORT), respondents choose option (a) with the constant probability W. In the full optional randomized response technique (FORT), each of the respondents belonging to group G provide a RR while the respondents belonging to the complementary group $\bar{G}$ provide direct responses.

Gupta et al. (2002) and Huang (2010) proposed multiplicative and additive RR models for the PORT for estimating the population mean μ_y of the sensitive characteristic y under SRSWR sampling only.

It was pointed out by Huang (2010) that both models can produce a scrambled response outside of the range of the sensitive variable y and he provided a remedy for the removal of such limitations.

Here the multiplicative and additive models are used under the assumptions of FORT. The proposed estimators of μ_y and W of the FORT are identical to the corresponding estimators of PORT but their variances differ significantly.

It is found for both the multiplicative and additive models that the variances of the estimator of μ_y under the assumption of FORT are larger than the variances computed under the assumption of FORT if $\mu_{yG} \geq \mu_y$.

Simulation studies reveal that variance of the proposed estimator based on the model R2 performs better than R1 if the multiplicative part of the model R2 is kept constant. On the other hand the proposed model R1 performs better than R2 if the multiplicative term varies significantly. The proposed ORT techniques R1 and R2 are also extended to complex survey designs.

- **References:**

- Arnab, R. (2018). Optiona randomized response techniques for quantitative characteristics. Communications in Statistics-Theory and methods (in Press)
- Arnab, R. (2004). Optional randomized response techniques for complex survey designs. *Biometrical Journal, 46,* 1, 114-124.
- Arnab, R. and Rueda, M. (2016). Optional Randomized Response: A Critical Review, *Hand book of Statistics*, 34, 253-271, edited by Chaudhuri, A., Christofides, T. C. and Rao, C.R., Elsevier, U.K.
- Chaudhuri, A. and Mukherjee, R. (1988). Randomized response: Theory and Techniques. *Marcel Dekker,* New York.
- Chaudhuri, A. and Saha, A. (2005). Optional versus compulsory randomized response techniques in complex surveys. *Journal of Statistical Planning and Inference,* 135, 516-527.
- Gupta, S. (2002). Qualifying the sensitivity level of binary response personal interview survey questions. *Journal of Combinatorics, Information & System Sciences*, 26(1–4), 101–109.

- Gupta, S., Gupta, B., Singh,S. (2002). Estimation of sensitivity level of personal interview survey question. *Journal of Statistical Planning and Inference,* 100, 239-247
- Huang, K.C., (2010). Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling. *Metrika,* 71, 341-352.
- Pal, S. (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. *Statistical Papers*. 49, 157-164.
- Warner, S.L. (1965). Randomize response: a survey technique for eliminating evasive answer bias. *American Statistical Association,* 60, 63-69

Thank you