Estimation of Sensitive Characteristic using Two-Phase Sampling for Regression Type Estimator

M. Javed and M.L. Bansal

Abstract. The problem of underreporting and non response on the sensitive issues are very common in most of the surveys due to our social setup. The randomized response (RR) models reduce rates of non-response and biased response that would ensure respondents’ privacy if they respond truthfully concerning personal questions. Here RR device is proposed for regression type estimator in which two independent samples are drawn from the population. The estimator of population mean of sensitive variable has been developed. Its bias and variance and optimum variance have also been derived.

Introduction

Warner (1965) in his pioneer model suggested the technique of randomized response where every person in the population belongs either to sensitive group A, or its compliment. Enquirer often feels embarrassed in asking direct questions about sensitive issues. Such questions might be about drug usage, tax evasion, history of induced abortion, illegal income etc. Horvitz et al. (1967) and Greenberg et al. (1971) have extended the Warner (1965) model to the case where the responses to the sensitive question are quantitative rather than a simple 'yes' or 'no'. The respondent selects one of the two questions by means of a randomization device. According to Greenberg et al. (1971), it is essential that the mean and variance of the responses to the unrelated question be close to those for the sensitive questions, otherwise, it will often be possible to recognize from the response which question was selected. Two sample single alternate RR model proposed by Greenberg and co-workers for obtaining the data on continuous type sensitive random variable is more practicable and easy in handling. If some auxiliary information is available then responses can be used in regression analysis by following Singh et al. (1996), Strachan et al. (1998), and Singh and King (1999). Recent publications on RR techniques among others are, Singh et al. (2000), Chaudhuri (2001, 2004), Singh (2002), Elffers et al. (2003), Huang
Proposed Randomized Response Device

Here an RR device is proposed for ratio type estimator in which two independent samples, sample-I and sample-II of sizes ‘n’ and ‘k’ with replacement are drawn from the population. The respondents in the first sample are given the RR device which contained statements regarding ‘A’ based on sensitive characteristic X and ‘Q2’ based on unrelated characteristic ‘Y2’ with respective probabilities ‘p’ and ‘(1 − p)’. Then the respondents are asked to answer the selected statement in terms of numerical figures without revealing to the interviewer which one of the two statements is answered. The respondents in the second sample are asked to answer in numerical figures (without using the RR device) for two direct statements ‘Q1’ and ‘Q2’ based on unrelated characteristics ‘Y1’ and ‘Y2’ respectively. The statements ‘Q1’ and ‘Q2’ are correlated. The characteristics ‘Y1’ and ‘Y2’ are unrelated with the sensitive character X.

Let

\[ p = \text{probability that sensitive question is selected by the first respondent in the first sample.} \]
\[ 1 - p = \text{probability that non-sensitive question ‘Q2’ is selected by the first respondent in the first sample.} \]
\[ q = \text{observed response from individual ‘i’ in first sample.} \]
\[ X_i = \text{response of individual ‘i’ in case he/she selects sensitive question through RR device in first sample.} \]
\[ Y_{2i} = \text{response of individual ‘i’ in case he/she selects alternate question ‘Q2’ through RR device in first sample.} \]
\[ Y_{1j} = \text{response of first question ‘Q1’ directly asked from j-th respondent in second sample.} \]
\[ Y_{2j} = \text{response of second question ‘Q2’ directly asked from j-th respondent in second sample.} \]

Estimator based on proposed randomized device

Here a regression type estimator for two-phase sampling is proposed using the suggested RR device.

A regression type estimator \( \hat{\mu}_{str} \) of mean \( \mu \) is proposed in two-phase sampling is as below

\[
\hat{\mu}_{str} = \frac{1}{p} \left[ p \bar{X} + (1 - p) \bar{y}_2' - (1 - p) \{ \bar{y}_2 + \beta (\bar{y}_1' - \bar{y}_1) \} \right]
\] (1)
where
\[\hat{y}_1 = \frac{1}{m} \sum_{j=1}^{m} y_{1j}, \quad \hat{y}_2 = \frac{1}{m} \sum_{j=1}^{m} y_{2j}, \quad \hat{y}'_1 = \frac{1}{k} \sum_{j=1}^{k} y_{1j} \quad \text{and} \quad \hat{y}'_2 = \frac{1}{n} \sum_{i=1}^{n} y_{2i}, \quad (2)\]
\[\hat{\beta} = \frac{s_{y_1 y_2}}{s_{y_1}^2}, \quad (3)\]
\[s_{y_1 y_2} = \frac{1}{m-1} \sum_{j=1}^{m} (y_{1j} - \hat{y}_1)(y_{2j} - \hat{y}_2), \quad s^{2}_{y_1} = \frac{1}{m-1} \sum_{j=1}^{m} (y_{1j} - \hat{y}_1)^2 \quad \text{and} \quad (4)\]
\[s^{2}_{y_2} = \frac{1}{m-1} \sum_{j=1}^{m} (y_{2j} - \hat{y}_2)^2.\]

Let
\[e_1 = \frac{\hat{y}_1}{\bar{y}_1} - 1, \quad e_2 = \frac{\hat{y}'_1}{\bar{y}_1} - 1, \quad e_3 = \frac{s_{y_1 y_2}}{S_{y_1 y_2}} - 1 \quad \text{and} \quad e_4 = \frac{s^2_{y_1}}{S^2_{y_1}} - 1 \quad (5)\]
such that
\[E(e_3) = E(e_4) = 0 \quad E(e_1 e_3) = \frac{\mu_{21}}{\hat{y}_1 \mu_{11}} \quad \text{and} \quad E(e_1 e_4) = \frac{\mu_{30}}{\hat{y}_1 \mu_{20}} \quad (6)\]
where
\[\mu_{ij} = E((y_1 - \hat{y}_1)^i (y_2 - \hat{y}_2)^j). \quad (7)\]

### Bias of the Estimator

The proposed estimator is biased. The bias of the estimator \(\hat{\mu}_{xlr}\) is given in the following theorem:

**Theorem 1.** The bias of the estimator \(\hat{\mu}_{xlr}\), to the order \(O(n^{-2})\) is given by

\[B(\hat{\mu}_{xlr}) = -\left[\frac{1-p}{p}\right] \left(\frac{1}{m} - \frac{1}{k}\right) \left[\frac{\mu_{21}}{\mu_{20}} - \frac{\mu_{11} \mu_{30}}{\mu_{20}}\right]. \quad (8)\]

**Proof.** We have

\[E(\hat{\mu}_{xlr}) = \frac{1}{p} \left[ E[p \bar{X} + (1-p)\bar{y}'_2 - (1-p)\{\bar{y}_2 + \hat{\beta}(\bar{y}'_1 - \bar{y}_1)\}] \right.\]
\[= \frac{1}{p} \left[ E[p \bar{X} + (1-p)\bar{y}_2 - (1-p)\{\bar{y}_2 + \hat{\beta} \bar{y}_1 (1 + e_3)(1 + e_4)^{-1}(e_2 - e_1)\}] \right].\]

Considering the binomial expansion of \((1 + e_4)^{-1}\) up to the order of \(O(n^{-2})\) and substituting the values of \(E(e_1 e_3)\) and \(E(e_1 e_4)\) in the above expression, we have

\[E(\hat{\mu}_{xlr}) = \bar{X} - \left[\frac{1-p}{p}\right] \left(\frac{1}{m} - \frac{1}{k}\right) \left[\frac{\mu_{21}}{\mu_{20}} - \frac{\mu_{11} \mu_{30}}{\mu_{20}}\right]. \quad (9)\]

As

\[B(\hat{\mu}_{xlr}) = E(\hat{\mu}_{xlr}) - \bar{X}. \quad (10)\]

Using (9) in (10) we get (8).
Variance of the Estimator

The variance of the estimator $\hat{\mu}_{xlr}$ is given in the following theorem:

**Theorem 2.** The variance of the estimator $\hat{\mu}_{xlr}$ to the order of $O(n^{-2})$ is given by

$$V(\hat{\mu}_{xlr}) = \frac{1}{p^2} \left[ \frac{\sigma^2_s}{n} + (1-p)^2 \left\{ \left( \frac{1}{k} - \frac{1}{N} \right) \sigma^2_{y_2} + \left( \frac{1}{m} - \frac{1}{k} \right) \sigma^2_{y_2} (1-\rho^2) \right\} \right]$$

(11)

where $\sigma^2_z = p \sigma^2_x + q \sigma^2_y + pq (\bar{X} - \bar{Y})^2$.

**Proof.** We have

$$V(\hat{\mu}_{xlr}) = \frac{1}{p^2} [V(\bar{z}) + (1-p)^2 V(\bar{y}_{lrd})]$$

(12)

where $\bar{z} = p \bar{X} + (1-p) \bar{y}_2'$ and $\bar{y}_{lrd} = \bar{y}_2 + \hat{\beta}(\bar{y}_1' - \bar{y}_1)$

$$V(\hat{\mu}_{xlr}) = \frac{1}{p^2} \left[ \frac{\sigma^2_s}{n} + (1-p)^2 V(\bar{y}_{lrd}) \right].$$

(13)

Now

$$V(\bar{y}_{lrd}) = E_1 V_2(\bar{y}_{lrd}/ps) + V_1 E_2(\bar{y}_{lrd}/ps)$$

$$= E_1 V_2(\bar{y}_{lrd}/ps) + V_1(\bar{y}_2)$$

where ‘ps’ means preliminary sample and

$$V_2(\bar{y}_{lrd}/ps) = E(\bar{y}_{lrd}/ps) - [E(\bar{y}_{lrd}/ps)]^2$$

$$= E(\{\bar{y}_2 + \hat{\beta}(\bar{y}_1' - \bar{y}_1)\}^2) - (E(\bar{y}_2 + \hat{\beta}(\bar{y}_1' - \bar{y}_1)))^2.$$

For two-phase sampling one can refer Sukhatme et al. (1984) to obtain the variance expression

$$V_2(\bar{y}_{lrd}/ps) = \left( \frac{1}{m} - \frac{1}{k} \right) s^2_{y_2} (1-\rho^2)$$

where $\rho$ is the coefficient of correlation between $y_1$ and $y_2$ and $s^2_{y_2}$ is defined at (4).

$$V(\bar{y}_{lrd}) = E_1 \left[ \left( \frac{1}{m} - \frac{1}{k} \right) s^2_{y_2} (1-\rho^2) \right] + V_1(\bar{y}_2)$$

$$= \left( \frac{1}{m} - \frac{1}{k} \right) s^2_{y_2} (1-\rho^2) + \left( \frac{1}{k} - \frac{1}{N} \right) \sigma^2_{y_2}.$$ 

(14)

Using the result obtained at (14) in (13), we get (11). □

**Optimum Variance of the Estimator**

We should choose that procedure which for a fixed cost $C_0$ minimizes the variance of the estimate. If $C_1$, $C_2$ and $C_3$ are the costs per unit of collecting information on the variable $Z$, auxiliary variable $y_1$ and $y_2$ under study respectively. Then, the total cost of the survey can be expressed as

$$C_0 = nC_1 + kC_2 + mC_3.$$ 

(15)
We shall now determine ‘\(n\)’, ‘\(k\)’ and ‘\(m\)’ so that for a fixed cost \(C_0\), the variance of the estimator \(\hat{\mu}_{xlr}\) is least.

Consider therefore, the expression

\[
\phi = V(\hat{\mu}_{xlr}) - \lambda(C_0 - nC_1 - kC_2 - mC_3).
\]

On differentiating \(\phi\) partially with respect to ‘\(n\)’, ‘\(k\)’ and ‘\(m\)’ and equating them equal to zero, we have

\[
\frac{\partial \phi}{\partial n} = \frac{1}{p^2} \left( \frac{-\sigma_z^2}{n^2} \right) + \lambda C_1, \tag{17}
\]

\[
\frac{\partial \phi}{\partial k} = \left( \frac{1 - p}{p} \right)^2 \left[ -\frac{\sigma_y^2 (1 - \rho^2)}{k^2} \right] + \lambda C_2, \tag{18}
\]

\[
\frac{\partial \phi}{\partial m} = \left( \frac{1 - p}{p} \right)^2 \left[ -\frac{\sigma_y^2 (1 - \rho^2)}{m^2} \right] + \lambda C_3. \tag{19}
\]

Solving (17), (18) and (19) for ‘\(n\)’, ‘\(k\)’ and ‘\(m\)’ respectively, we have

\[
n = \frac{\sigma_z}{p \sqrt{\lambda C_1}}, \tag{20}
\]

\[
k = \left( \frac{1 - p}{p} \right) \sigma_y \sqrt{\frac{1 - \rho^2}{\lambda C_2}}, \tag{21}
\]

\[
m = k \sqrt{\frac{C_2}{C_3}}. \tag{22}
\]

Again differentiating \(\phi\) partially with respect to \(\lambda\) and equating the result equal to zero, we have

\[
\frac{\partial \phi}{\partial \lambda} = C_0 - nC_1 - kC_2 - mC_3. \tag{23}
\]

Substituting the values of ‘\(n\)’, ‘\(k\)’ and ‘\(m\)’ in (23), we have

\[
\sqrt{\lambda} = \frac{\sigma_z \sqrt{C_1} + \sqrt{C_2} [qA' + k \sqrt{C_3}]}{p C_0}
\]

where \(A' = \sigma_y \sqrt{1 - \rho^2}\).

Now substituting the values of \(\sqrt{\lambda}\) in (20), (21) and (22), we get the optimum values of ‘\(n\)’, ‘\(k\)’ and ‘\(m\)’ respectively as

\[
n_{opt} = \frac{C_0 \sigma_z}{B' \sqrt{C_1}},
\]

\[
k_{opt} = \frac{C_0 q A'}{B' \sqrt{C_2}},
\]

\[
m_{opt} = k_{opt} \sqrt{\frac{C_2}{C_3}}.
\]
and

\[ B' = \sigma_z \sqrt{C_1} + \sqrt{C_2} \left[ qA' + k \sqrt{C_3} \right]. \]

where \( A' \) is defined above and \( q = 1 - p \).

**Theorem 3.** The optimum variance of the estimator \( \bar{\mu}_{xlr} \) to the order of \( O(n^{-2}) \) is given by

\[
V(\bar{\mu}_{xlr}) = \frac{1}{p^2} \left[ \frac{\sigma_z B' \sqrt{C_1}}{C_0} + q^2 \left\{ \left( \frac{B' \sqrt{C_2}}{C_0 qA'} - \frac{1}{N} \right) \sigma_{\gamma z}^2 + \frac{B'}{C_0 qA'} (\sqrt{C_1} - \sqrt{C_2}) A' \right\} \right].
\]

(24)

**Proof.** We have

\[
V(\bar{\mu}_{xlr}) = \frac{1}{p^2} \left[ \frac{\sigma_z^2}{n} + (1 - p)^2 \left\{ \left( \frac{1}{k} - \frac{1}{N} \right) \sigma_{\gamma z}^2 + \left( \frac{1}{m} - \frac{1}{k} \right) \sigma_{\gamma z}^2 (1 - \rho^2) \right\} \right].
\]

Substituting the optimum values of ‘\( n \)’, ‘\( k \)’ and ‘\( m \)’ obtained earlier in the above expression, we get (24).

\[ \square \]

**References**


M. Javed, Department of Mathematics, Statistics & Physics, Punjab Agricultural University, Ludhiana 141004, India. E-mail: mjaved@pau.edu

M.L. Bansal, Department of Mathematics, Statistics & Physics, Punjab Agricultural University, Ludhiana 141004, India.

Received May 2, 2011
Accepted May 29, 2012