On stratified randomized response sampling

Jea-Bok Ryu\textsuperscript{a,∗}, Jong-Min Kim\textsuperscript{b}, Tae-Young Heo\textsuperscript{c} and Chun Gun Park\textsuperscript{d}
\textsuperscript{a}Statistics, Division of Life Science and Genetic Engineering and Statistics, Cheongju University, Cheongju, Chungbuk, 360-764, Republic of Korea
\textsuperscript{b}Statistics, Division of Science and Mathematics, University of Minnesota, Morris, MN, 56267, USA
\textsuperscript{c}Department of Statistics, North Carolina State University, Raleigh, NC, 27695, USA
\textsuperscript{d}National Cancer Center, Ilsan, Goyang-si, Gyeonggi-do, 411-769, Republic of Korea

Abstract. In this paper, we propose a new quantitative randomized response model based on Mangat and Singh [7] two-stage randomized response model. We derive the estimator of the sensitive variable mean, and show that our method is more efficient than other randomized response models suggested by Greenberg et al. [3] and Gupta et al. [4] estimators.

Keywords: Quantitative randomized response technique, sensitive characteristics, stratified sampling

1. Introduction

The randomized response technique is a procedure for collecting the information on sensitive characteristics without exposing the identity of the respondent. It was first introduced by Warner [8] as an alternative survey technique for socially undesirable or incriminating behavior questions. Greenberg et al. [3] have proposed and developed the unrelated question randomized response design for estimating the mean and the variance of the distribution of a quantitative variable. Gupta et al. [4], and Arnab [1] have showed that optional randomized response model is more accurate while being less intrusive. Hong et al. [5] suggest a stratified randomized response model using a proportional allocation. However their model may have a high costs due to the difficulty in obtaining a proportional sample from each stratum. To rectify this problem, Kim and Warde [6] suggest a stratified randomized response model using an optimal allocation which is more efficient than that of using the proportional allocation.

2. A review of quantitative randomized response methods

Unrelated question randomized response method proposed by Greenberg et al. [3] is a survey procedure that a respondent could be asked one of two questions depending on the outcome of a randomization device. For example, an interviewee performs a randomization device with two outcomes each with pre-assigned probabilities \( P \) and \( 1 - P \) which will answer one of the following questions:

\[ S : \text{How many abortions have you had during your lifetime?} \]
\[ N : \text{How many magazines do you subscribe to?} \]

where we denotes \( S \) as the sensitive question and \( N \) as the non-sensitive question. Two independent, non-overlapping samples of sizes \( n_1 \) and \( n_2 \) are used (size \( n_1 \) need not be equal to size \( n_2 \)). Let the population mean of both the sensitive and non-sensitive distributions be \( \mu_A, \mu_Y \), respectively. Let the population variance of both the sensitive and non-sensitive distributions be \( \sigma^2_A, \sigma^2_Y \), respectively. Unbiased estimators for the means of the sensitive and non-sensitive random variables, \( \mu_A \) and \( \mu_Y \), are

\[ \hat{\mu}_1 = \frac{(1 - P_2)\hat{T}_1 - (1 - P_1)\hat{T}_2}{P_1 - P_2} \]  

(1)
\[ \hat{\mu}_2 = \frac{P_2 \bar{T}_1 - P_1 \bar{T}_2}{P_2 - P_1}, \]  

(2)

where \( \bar{T}_i \) is total sample mean computed from the responses in the \( i_{th} \) samples and \( P_i \) is the selection probability for the sensitive question in the \( i_{th} \) sample, for \( i = 1, 2 \) (\( P_1 \neq P_2 \)).

The variance of \( \hat{\mu}_1 \) is given by

\[ \Var(\hat{\mu}_1) = \frac{(1 - P_2)^2 \Var(\bar{T}_1) + (1 - P_1)^2 \Var(\bar{T}_2)}{(P_2 - P_1)^2}, \]

(3)

where \( \Var(\bar{T}_j) = \frac{1}{n_j} (\sigma_Y^2 + P_j (\sigma_A^2 - \sigma_Y^2) + P_j (1 - P_j)(\mu_A - \mu_Y)^2). \)

In this method, if \( \mu_Y \) and \( \sigma_Y^2 \) are known in advance, only one sample is needed. So we define \( P_1 = P \) and \( T_1 = T \), then Eqs (1) and (3) are simplified as

\[ \hat{\mu}_1 = \bar{T} - (1 - P)\mu_Y \]

(4)

and

\[ \Var(\hat{\mu}_1) = \frac{\Var(\bar{T})}{P^2} \]

\[ = \frac{1}{nP^2} (\sigma_Y^2 + P(\sigma_A^2 - \sigma_Y^2)) + P(1 - P)(\mu_A - \mu_Y)^2. \]

(5)

Eichhorn and Hayre [2] introduce a scrambled randomized response method for estimating the mean \( \mu_A \) and the variance \( \sigma_A^2 \) of the sensitive question \( A \). According to them, each respondent selected in the sample is instructed to use a randomization device and generate a random number, say \( B \), from some pre-assigned distribution. The distribution of the random variable \( B \), also called a scrambling variable, is assumed to be known. The mean \( \mu_B \) and the variance \( \sigma_B^2 \) of the scrambling variable are also assumed to be known. The \( i_{th} \) respondent selected in the sample of size \( n \), drawn by using simple random sampling with replacement (SRSWR), is requested to report the value \( Z_i \), \( B_i \mu_A / \mu_B \), as a scrambled response on the sensitive variable, \( A \). They show that an unbiased estimator of the population mean, \( \mu_A \), is given by

\[ \hat{\mu}_G = \frac{1}{n} \sum_{i=1}^{n} Z_i, \]

(6)

with variance

\[ \Var(\hat{\mu}_G) = \frac{1}{n} \left( \sigma_A^2 + \sigma_B^2 (\sigma_A^2 + \mu_A^2) \right), \]

(7)

where \( \sigma_B^2 \) is the standard deviation of the scrambling variable \( B \), and \( C_B = \sigma_B / \mu_B \) denotes the known coefficient of variation of the scrambling variable \( B \).

Gupta et al. [4] propose an optional randomized response technique, which is more efficient than the scrambled randomized response technique suggested by Eichhorn and Hayre [2]. In the optional randomized response technique, where each respondent selected by SRSWR, can choose one of the following two options: (a) The respondent can report the correct response \( A \), or (b) The respondent can report the scrambled response \( BA \), where \( B \) denotes the independent scrambling variable. In optional procedure, they assumed that both \( B \) and \( A \) are positive random variables and \( \mu_B = 1 \). The optional randomized response model can be written as

\[ Z = B^T A, \]

(8)

where \( I \) is an indicator random variable defined as

\[ I = \begin{cases} 1 & \text{if the response is scrambled} \\ 0 & \text{otherwise} \end{cases} \]

If \( W \) denotes the probability that a person will report the scrambled response, then \( I \) is a Bernoulli random variable with \( E(I) = W \), where \( W \) can be called the sensitivity of the question. They showed an unbiased estimator of population mean, \( \mu_A \), is given by

\[ \hat{\mu}_G = \frac{1}{n} \sum_{i=1}^{n} Z_i, \]

(9)

with variance

\[ \Var(\hat{\mu}_G) = \frac{1}{n} \left( \sigma_A^2 + WC_B^2 (\sigma_A^2 + \mu_A^2) \right), \]

(10)

where \( C_B = \sigma_B / \mu_B \) denotes the known coefficient of variation of the scrambling variable \( B \).

3. Two-stage quantitative randomized response model

In this section, we propose a two-stage quantitative randomized response model. We assume that a sample of size \( n \) is selected by SRSWR. The method is described as follows.

Stage 1 An individual respondent in the sample is instructed to use the randomization device \( R_1 \) which consists of two statements:

(i) “Report the correct response \( A \) of a sensitive question” and

(ii) “Go to the randomization device \( R_2 \) in the second stage”
represented with probabilities $P$ and $1 - P$.

**Stage 2** The randomization device $R_2$ consists of two statements:

(i) “Report the correct response $A$ of a sensitive question”

(ii) “Report the scrambled response $AB$ of a sensitive question”

represented with probabilities $T$ and $1 - T$.

The respondent should not report to an interviewer which steps are taken to protect the respondent’s privacy. We assumed that both $B$ and $A$ are positive random variables, $\mu_B = 1$, and $\sigma_B^2 = \psi^2$. Similar to Eichhorn and Hayre [2] approach, the distribution of random variable $B$, the mean $\mu_B$ and the variance $\sigma_B^2$ of the scrambling variable are all assumed to be known. Based on two-stage procedures, the $i^{th}$ respondent selected in the sample of size $n$, drawn by using SRSWR, is requested to report the value,

$$U = \alpha A + (1 - \alpha) (\beta A + (1 - \beta) AB), \quad (11)$$

where

$$\alpha = \begin{cases} 
1 \text{ if a respondent chooses a statement 1 in } R_1 \\
0 \text{ if a respondent chooses a statement 2 in } R_1 
\end{cases}$$

and

$$\beta = \begin{cases} 
1 \text{ if a respondent chooses a statement 1 in } R_2 \\
0 \text{ if a respondent chooses a statement 2 in } R_2 
\end{cases}$$

The expected value of the observed response is,

$$E(U) = E(\alpha A + (1 - \alpha)(\beta A + (1 - \beta) AB))$$

$$= P\mu_A + (1 - P) (T\mu_A + (1 - T)\mu_A\mu_B)$$

$$= \mu_A, \quad (12)$$

where $\alpha$ is a Bernoulli random variable with $E(\alpha) = P$, $Var(\alpha) = P(1 - P)$ and $\beta$ is Bernoulli with $E(\beta) = T$, $Var(\beta) = T(1 - T)$.

**Theorem 3.1.** An unbiased estimator, $\hat{\mu}_A$, of the population mean $\mu_A$ is given by,

$$\hat{\mu}_A = \frac{1}{n} \sum_{i=1}^{n} U_i \quad (13)$$

**Theorem 3.2.** The variance of the proposed estimator $\hat{\mu}_A$ is given by

$$Var(\hat{\mu}_A) = \frac{1}{n} \left[ \sigma_A^2 + (1 - P)(1 - T)\psi^2(\mu_A^2 + \sigma_A^2) \right]. \quad (14)$$

If $0 < P < 1$ in Eq. (14), then, obtain the relative efficiency of $\hat{\mu}_A$ with respect to $\hat{\mu}_G$, we compare $Var(\hat{\mu}_G)$ and $Var(\hat{\mu}_A)$ as follows:

$$Var(\hat{\mu}_G) - Var(\hat{\mu}_A)$$

$$= \frac{1}{n} \left( \sigma_A^2 + W\psi^2(\mu_A^2 + \sigma_A^2) \right)$$

$$- \frac{1}{n} \left[ \sigma_A^2 + (1 - P)(1 - T)\psi^2(\mu_A^2 + \sigma_A^2) \right]$$

$$= \frac{1}{n} \left( (\mu_A^2 + \sigma_A^2)(1 - T)P\psi^2 \right) \geq 0.$$ 

We have shown that the proposed estimator $\hat{\mu}_A$ is more efficient than the estimator $\hat{\mu}_G$ suggested by Gupta et al. [4].

4. **Stratified two-stage quantitative randomized response model**

In this section, we newly propose a two-stage quantitative randomized response technique in stratified sampling. The main advantage of the stratified approach is that the technique overcome the limitation of the loss of individual characteristics of the respondents. We assume that the population is partitioned into strata, and a sample $n_h$ is selected by the SRSWR from each stratum. We assume that the number of units in each stratum is known. Let $n_h$ denote the number of units in the sample from stratum $h$ and $n$ denote the total number of units in the samples from all strata so that $n = \sum_{h=1}^{k} n_h$.

**Stage 1** An individual respondent in the sample is instructed to use the randomization device $R_{1h}$ which consists of two statements:

(i) “Report the correct response $A$ of a sensitive question” and

(ii) “Go to the randomization device $R_{2h}$ in the second stage”

represented with probabilities $P_h$ and $1 - P_h$.

**Stage 2** The randomization device $R_{2h}$ consists of two statements:

(i) “Report the correct response $A$ of a sensitive question” and

(ii) “Report the scrambled response $AB$ of a sensitive question”
\[
\frac{n_h}{n} = \frac{w_h \sqrt{(\mu^2_{A_h} + \sigma^2_{A_h}) (P_h + (1 - P_h)T_h + (1 - P_h)(1 - T_h)(1 + \psi_h^2)) - \mu^2_{A_h}}}{\sum_{h=1}^{k} w_h \sqrt{(\mu^2_{A_h} + \sigma^2_{A_h}) (P_h + (1 - P_h)T_h + (1 - P_h)(1 - T_h)(1 + \psi_h^2)) - \mu^2_{A_h}}}
\]

represented with probabilities \(T_h\) and \(1 - T_h\).

Under the assumption that respondent reports truthfully and \(P_h\) and \(T_h\) are set by the researcher, the distribution of random variable \(B_h\), the mean \(\mu_{B,h}\) and the variance \(\sigma^2_{B,h}\) of the scrambling variable are all assumed to be known. We assume that \(\mu_{B,h} = 1\) and \(\sigma^2_{B,h} = \psi_h^2\) for all \(h = 1, 2, \ldots, k\). The \(i^{th}\) respondent selected in the sample of size \(n_h\) in stratum \(h\), drawn by using SRSWR, is requested to report the value,

\[
U_h = \alpha_h A_h + (1 - \alpha_h) \beta_h A_h + (1 - \beta_h)A_h B_h,
\]

where,

\[
\alpha_h = \begin{cases} 
1 & \text{if a respondent chooses a statement 1 in } R_{1h} \\
0 & \text{if a respondent chooses a statement 2 in } R_{1h}
\end{cases}
\]

\[
\beta_h = \begin{cases} 
1 & \text{if a respondent chooses a statement 1 in } R_{2h} \\
0 & \text{if a respondent chooses a statement 2 in } R_{2h}
\end{cases}
\]

Similar to Eq. (12), the expected value of the observed response is given by,

\[
E(U_h) = E(\alpha_h A_h + (1 - \alpha_h) (\beta_h A_h + (1 - \beta_h)A_h B_h))
= P_h \mu_{A,h} + (1 - P_h)(T_h \mu_{A,h} + (1 - T_h)\mu_{A,h} \mu_{B,h}) = \mu_{A,h}
\]

where \(\alpha_h\) is a Bernoulli random variable with \(E(\alpha_h) = P_h\), \(\text{Var}(\alpha_h) = P_h(1 - P_h)\) and \(\beta_h\) is a Bernoulli random variable with \(E(\beta_h) = T_h\), \(\text{Var}(\beta_h) = T_h(1 - T_h)\). By Theorem 3.1, an unbiased estimator of the population mean \(\mu_{A,h}\) in stratum \(h\) is,

\[
\hat{\mu}_{A,h} = \frac{1}{n_h} \sum_{i=1}^{n} U_{hi}
\]

and its variance is

\[
\text{Var}(\hat{\mu}_{A,h}) = \frac{1}{n_h} \sum_{h=1}^{k} \frac{w^2_h}{n_h} (\mu^2_{A,h} + \sigma^2_{A,h})(P_h + (1 - P_h)T_h + (1 - P_h)(1 - T_h)(1 + \psi_h^2)) - \mu^2_{A,h}
\]

\[
+ (1 - P_h)(1 - T_h)(1 + \psi_h^2)) - \mu^2_{A,h}\]

Since the selections in different strata are made independently, the mean estimators for individual strata can be added together to obtain a mean estimator for the whole population. The mean estimator of \(\mu_A\) for stratified sampling scheme is:

\[
\hat{\mu}^* = \sum_{h=1}^{k} \frac{w_h}{n_h} \mu_{A,h} = \sum_{h=1}^{k} \frac{w_h}{n_h} n_h \sum_{i=1}^{n} U_{hi}
\]

where \(w_h = (N_h/N)\) for \(h = 1, 2, \ldots, k\), so that \(w = \sum_{h=1}^{k} w_h = 1\), \(N\) is the number of units in the whole population and \(N_h\) is the total number of units in stratum \(h\). It is easily shown that the proposed mean estimator \(\hat{\mu}^*\) is an unbiased estimate for the population mean \(\mu_A\). The variance of the mean estimator \(\hat{\mu}^*\) is:

\[
\text{Var}(\hat{\mu}^*) = \sum_{h=1}^{k} \frac{w^2_h}{n_h} (\mu^2_{A,h} + \sigma^2_{A,h})(P_h + (1 - P_h)T_h + (1 - P_h)(1 - T_h)(1 + \psi_h^2)) - \mu^2_{A,h}
\]

Information on \(\mu_{A,h}\) and \(\sigma^2_{A,h}\) is usually unavailable, however if prior information on \(\mu_{A,h}\) and \(\sigma^2_{A,h}\) is available from past experience, then we may derive the following optimal allocation formula.

Using the optimal-allocation approach based on Kim and Warde [6], one can show that the variance in Eq. (19) is minimized when \(n_1, n_2, \ldots, n_k\) are chosen such that (the first equation on the top of the page).

Under this optimal-allocation assumption, the variance in Eq. (19) becomes in Eq. (20).

**Theorem 4.1.** Assuming optimal allocation, when \(w_1 = w_2 = 1/2\) and \(D = (D_1 + D_2)/2\), the stratified estimator \(\hat{\mu}^*\) is more efficient than the proposed model.
estimator $\hat{\mu}_A$, where,
\[
D = \left(\mu_A^2 + \sigma_A^2\right)(P + (1 - P)T) + (1 - P)\left((1 + P)\left(1 + \psi^2\right)\right) - \mu_A^2,
\]
\[
D_h = \left(\mu_{A_h}^2 + \sigma_{A_h}^2\right)(P_h + (1 - P_h)T_h) + (1 - P_h)(1 - T_h)\left(1 + \psi_h^2\right) - \mu_{A_h}^2,
\]
for $h = 1$ and 2.

For other cases, we can easily check the relative efficiency by a variance comparison with various settings of $\mu_{A_h}$, $\sigma_{A_h}$, $P_h$, and $T_h$. Theorem 4.1 guarantees that when optimal- allocation is used, the stratified estimator, $\hat{\mu}_A$, is more efficient than the estimator, $\hat{\mu}_A$, which ignores the stratification.

5. Comparison and Discussion

In this section, we present a numerical study of the two-stage quantitative randomized response model. The purpose of the simulation is to confirm that the proposed technique is more efficient. We compare the original quantitative randomized response model proposed by Greenberg et al. [3] ($\hat{\mu}_1$) with the proposed model ($\hat{\mu}_A$) in terms of variance. From Eqs (4) and (5), the mean and the variance of Greenberg et al. [3]'s sensitive mean estimator (when $\mu_Y$ is known) are

$$\hat{\mu}_1 = \frac{T - (1 - P)\mu_Y}{P},$$

$$\text{Var}(\hat{\mu}_1) = \frac{\text{Var}(T)}{P^2} = \frac{1}{nP^2}(\psi^2 + P(\sigma_A^2 - \sigma_Y^2)) + P(1 - P)(\mu_A - \mu_Y)^2).$$

Under the assumptions with $\mu_Y = \mu_B = 1$ and $\sigma_Y^2 = \sigma_B^2 = \psi^2$,

$$\hat{\mu}_1 = \frac{T - (1 - P)}{P},$$

$$\text{Var}(\hat{\mu}_1) = \frac{\text{Var}(T)}{P^2} = \frac{1}{nP^2}(\psi^2 + P(\sigma_A^2 - \sigma_Y^2)) + P(1 - P)(\mu_A - 1)^2).$$

The relative efficiency of $\hat{\mu}_A$ with respect to $\hat{\mu}_1$ is as follows:

$$RE_2 = \frac{\text{Var}(\hat{\mu}_1)}{\text{Var}(\hat{\mu}_A)} = [\psi^2 + P(\sigma_A^2 - \psi^2)].$$

Fig. 1. The relative efficiency of $\hat{\mu}_A$ with respect to $\hat{\mu}_1$ as a function of $P = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and $T = 0.2, 0.4, 0.6, 0.8$ with $\sigma_A^2 = 1, \sigma_Y^2 = \sigma_B^2 = 0.5$ and $\mu_A = 2$ (top left); $\mu_A = 4$ (top right); $\mu_A = 6$ (bottom left); $\mu_A = 8$ (bottom right).
\[ + P(1 - P)(\mu_A - 1)^2]/\]
\[ [P^2(\mu_A^2 + \sigma_A^2)(P + (1 - P)T)
+ (1 - P)(1 - T)(1 + \psi^2) - \mu_A^2] \].

Figure 1 shows that the proposed estimator, \( \hat{\mu}_A \), is more efficient than the Greenberg et al. [3] estimator, \( \hat{\mu}_1 \), with \( \sigma_A^2 = 1 \). We can show that the proposed method is more efficient than the Greenberg et al. [3] method if the coefficient of variation, \( C_B = \sigma_B/\mu_B = \sigma_B \leq 1.0 \).

Our newly proposed two-stage quantitative randomized response model improves the performance by taking advantage of randomized response information provided by second stage. We have shown that our model is much more efficient than other models (Greenberg et al. [3] and Gupta et al. [4]). Additionally, we have provided a comprehensive description of the two-stage quantitative stratified randomized response model and its statistical properties. The use of stratified quantitative randomized response model can overcome the limitations of randomized response model which can lose the individual characteristics of the respondents.

References