Estimation of sensitive quantitative characteristics in randomized response sampling

Kuo-Chung Huang∗

Department of Business Administration
Chungyu Institute of Technology
Keelung
Taiwan 201
R.O.C.

Chun-Hsiung Lan

Mei-Pei Kuo

Department of Business Administration
Nanhua University
Dalin, Chiayi
Taiwan 622
R.O.C.

Abstract

This paper considers the problem of procuring honest responses for sensitive quantitative characteristics. An alternative survey technique is proposed, which enables us to estimate the population mean unbiasedly and to gauge how sensitive a survey topic is. An asymptotically unbiased estimator of sensitivity level is proposed, and conditions for which unbiased estimation for population variance being available is also studied. In addition, an efficiency comparison is worked out to examine the performance of the proposed procedure. It is found that higher estimation efficiency results from higher variation of randomization device.

Keywords: Percent relative efficiency, privacy protection, sample size allocation, simple random sampling with replacement.

∗E-mail: kchuang.tw@gmail.com
1. Introduction

Accurate information on the survey topics is extremely relevant for parameter estimations. However, it is difficult to obtain valid and reliable information in the area of sensitive topics. If a direct survey research is employed to assess a sensitive characteristic, respondents often refuse to take part, or reply untruthfully, especially when they have committed sensitive behavior. To improve respondent cooperation and to procure reliable data, Warner [15] first introduced the randomized response (RR) technique for qualitative characteristics. Subsequently, Greenberg et al. [8] suggested an extension of Warner’s RR technique to quantitative characteristics. An excellent exposition of modifications on RR techniques and other related works could be referred to Chaudhuri and Mukerjee [6]. Some resent developments are Bhargava and Singh [1], Chua and Tsui [7], Padmawar and Vijayan [12], Singh et al. [14], Chang and Huang [2], Chaudhuri [5], Singh et al. [13], Huang [10], and Chang et al. [3, 4], etc. In particular, to quantify the sensitivity level for certain items of inquiry in practice, Gupta et al. [9] first shown how an estimator may be developed on the quantitative characteristics.

Consider a finite population in which every person has a positive value for the sensitive characteristic $X$. The problem of interest is to estimate the mean $\mu_x$, the variance $\sigma^2_x$, and the sensitivity level $W$ of $X$ from a with-replacement simple random sample of size $n$. In Gupta et al. [9] procedure, each sampled respondent is instructed to utilize a randomization device and generate a positive-valued random number $S$ from a pre-assigned distribution with known mean $\mu_S = 1$ and known variance $\gamma^2_S$. Then he or she chooses one of the following options: (a) The respondent can report the correct response $X$, or (b) The respondent can report the scrambled response $SX$. The optional randomized response model considered is $Z = SYX$, where $Y = 1$ or 0 according as the response is scrambled or not. Here $Y$ is a Bernoulli variate with $E(Y) = W$, where $W$ is the probability that a respondent will report the scrambled response rather than the actual response $X$. It is shown that the usual sample mean $\hat{\mu}_x = \bar{Z}$ is unbiased with variance given by

$$\text{Var}(\hat{\mu}_x) = n^{-1} \sigma^2_Z = n^{-1}[\sigma^2_x + W\gamma^2_S(\sigma^2_x + \mu^2_x)].$$

(1.1)

Denote by $s^2_Z = (n - 1)^{-1} \sum_{j=1}^n (Z_j - \bar{Z})^2$ the sample variance, an unbiased estimator of $\text{Var}(\hat{\mu}_x)$ is given by $\text{Var}(\hat{\mu}_x) = n^{-1}s^2_Z$. And, they
suggested

\[ \hat{W}_G = \frac{n^{-1} \sum_{j=1}^{n} \log(Z_j) - \log Z}{E[\log(S)]} \quad \text{and} \quad \hat{\sigma}^2_{xG} = \frac{s^2_Z - \hat{W}_G \gamma^2_{S \hat{G}}}{1 + \hat{W}_G \gamma^2_{S}} , \]

as an estimator of \( W \) and \( \sigma^2_x \), respectively.

Although it creates an appropriate environment to estimate some unknown population parameters, the Gupta et al. [9] development in finding an estimator for \( W \) remains biased because of the presence of log term. They did address estimation efficiency and biasedness in the \( W \) estimate using empirical estimators, but analytical variance expression of \( W \) remains unidentified. In addition, the estimator \( \hat{\sigma}^2_{xG} \) is merely a biased estimator of \( \sigma^2_x \). In these regards, we intend to suggest an survey procedure that provides an unbiased estimator of \( \sigma^2_x \) and an asymptotically unbiased estimator of \( W \) together with its variance. The proposed estimators with principal properties are given in the following section. In section 3, an efficiency comparison is carried out to study the performance of the proposed procedure.

2. The proposed procedure

In the proposed procedure, two independent samples of size \( n_i \), \( i = 1, 2 \), are drawn from the population using simple random sampling with replacement such that \( n_1 + n_2 = n \), the total sample size required. In the \( i \)th sample, each respondent is instructed to use a randomization device and generate a random number, say \( S_i \), from some pre-assigned distributions such as Chi-square, uniform, or Weibull etc. The RR procedures are set up in such a way that the interviewer does not know what the respondent has selected. Then the respondent is requested to report one of the following two options: (a) The correct response \( X \), or (b) The scrambled response \( S_i X \). It is supposed that \( S_i \) is a positive-valued random variate with known mean \( \mu_{S_i} = \theta_i \neq 1 \) and known variance \( \sigma^2_{S_i} = \gamma^2_{S} \). The optional randomized response model for the \( i \)th sample, \( i = 1, 2 \), is given by

\[ Z_i = (1 - Y)X + YS_i X , \]

where \( Y \) is a random variate, with expectation \( E(Y) = W \), defined as

\[ Y = \begin{cases} 1, & \text{if the response is scrambled,} \\ 0, & \text{otherwise.} \end{cases} \]
Under the proposed procedure, the sample response $Z_i$ for the $i$th sample, $i = 1, 2$, has expectation
\[ E(Z_i) = (1 - W)\mu_x + W\theta_i\mu_x = \mu_x + W(\theta_i - 1)\mu_x. \] (2.1)

If $\bar{Z}_1$ and $\bar{Z}_2$ are the observed means for the two samples, the proposed estimators of $\mu_x$ and $W$ are respectively given by
\[ \hat{\mu}_x = \frac{(1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2}{\theta_1 - \theta_2} \quad \text{and} \quad \hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{(1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2}, \]
whence provided that $\theta_1 \neq \theta_2$.

**Theorem 1.** The estimator $\hat{\mu}_x$ is unbiased with variance given by
\[ \text{Var}(\hat{\mu}_x) = \frac{1}{\theta_1 - \theta_2} \left[ (1 - \theta_2)^2 \frac{\sigma_x^2}{n_1} + (1 - \theta_1)^2 \frac{\sigma_x^2}{n_2} \right], \] (2.2)
where $\sigma_{Z_i}^2 = \sigma_x^2 + W(\gamma_i^2 + \theta_i^2 - 1)\sigma_x^2 + W(\gamma_i^2 + (1 - W)(1 - \theta_i)^2)\mu_x^2$, $i = 1, 2$. (2.3)

**Proof.** Taking expectation for $\hat{\mu}_x$, the unbiasedness follows from using (2.1). From the distribution of $Z_i^2$, we have
\[ E(Z_i^2) = \sigma_x^2 + \mu_x^2 + W(\gamma_i^2 + \theta_i^2 - 1)(\sigma_x^2 + \mu_x^2). \] (2.4)

Using (2.1), (2.4) and the fact that $\sigma_{Z_i}^2 = E(Z_i^2) - [E(Z_i)]^2$, we get (2.3). Expression (2.2) then follows from the independence of the two samples. Hence the proof. \(\square\)

**Theorem 2.** An unbiased estimator of the variance of $\hat{\mu}_x$ is given by
\[ \text{Var}(\hat{\mu}_x) = \frac{1}{(\theta_1 - \theta_2)^2} \left[ (1 - \theta_2)^2 \frac{s_{Z_i}^2}{n_1} + (1 - \theta_1)^2 \frac{s_{Z_i}^2}{n_2} \right], \]
where $s_{Z_i}^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2$, $i = 1, 2$.

**Proof.** It follows from that $E(s_{Z_i}^2) = \sigma_x^2$, $i = 1, 2$. \(\square\)

For deriving the properties of the estimator $\hat{W}$, let us define $d_1 = \bar{Z}_1 - \bar{Z}_2$ and $d_2 = (1 - \theta_2)\bar{Z}_1 - (1 - \theta_1)\bar{Z}_2$. Since $E(d_1) = (\theta_1 - \theta_2)W\mu_x$ and $E(d_2) = (\theta_1 - \theta_2)\mu_x$, we then have $\hat{W} = d_1/d_2$ and $W = E(d_1)/E(d_2)$. Additionally, if we further denote by $e_1 = |d_1 - E(d_1)|/E(d_1)$ and $e_2 = |d_2 - E(d_2)|/E(d_2)$, assuming that $|e_2| < 1$ so that $(1 + e_2)^{-1}$ can be validly expanded as power series, it follows that
\[ E(e_1) = E(e_2) = \frac{(1 - \theta_2)\sigma_{Z_i}^2 + (1 - \theta_1)^2\sigma_{Z_i}^2}{(\theta_1 - \theta_2)^2W\mu_x^2}, \]
some algebraic simplification, we get (2.7). Replacing $E_e$ the second and then taking expectation, we have

$$E(\hat{\theta}_1) = \frac{\sigma_1^2 + \sigma_2^2}{(\theta_1 - \theta_2)^2 \mu_2^2}.$$  

And the estimation error can be written in terms of $e_1$ and $e_2$ as

$$V W - W = W(e_1 - e_2) + o(n^{-1}).$$  

Then we have the following theorem.

**Theorem 4.** The estimator $\hat{\theta}_1$ is asymptotically unbiased with variance given by

$$\text{Var}(\hat{\theta}_1) = \frac{1}{(\theta_1 - \theta_2)^2 \mu_2^2} \left\{ [1 + W(\theta_2 - 1)]^2 \frac{\sigma_2^2}{n_1} + [1 + W(\theta_1 - 1)]^2 \frac{\sigma_2^2}{n_2} \right\},$$  

which can be estimated by

$$\text{Var}(\hat{\theta}_1) = \frac{1}{(\theta_1 - \theta_2)^2 \mu_2^2} \left\{ [1 + \hat{W}(\theta_2 - 1)]^2 \frac{s_2^2}{n_1} + [1 + \hat{W}(\theta_1 - 1)]^2 \frac{s_2^2}{n_2} \right\}.  \tag{2.8}$$

**Proof.** The asymptotically unbiasedness follows from $E(e_1) = E(e_1) = 0$. Squaring expression (2.6), omitting terms with power in $e_i$’s higher than the second and then taking expectation, we have $E(\hat{W} - W)^2 = W^2 E(\hat{e}_2^2 - 2e_1e_2 + e_2^2)$. On substituting the expected values given in (2.5) and after some algebraic simplification, we get (2.7). Replacing $\mu_2$, $W$ and $\sigma_2^2$ in (2.7) by the corresponding sample analogue, (2.8) then follows. Hence the theorem.

Next, let us consider the problem of unbiased estimation for the variance $\sigma_2^2$ of the sensitive characteristic $X$. For the sake of notational convenience, let us denote $a = 2(\theta_2 - 1)\gamma_1^2 + (2 - \theta_1 - \theta_2)\gamma_2^2 + (1 - 2\theta_1 - \theta_2) \times (\theta_1 - \theta_2)$, $b = (\theta_1 - \theta_2)(1 - \gamma_1^2 - \theta_1^2)$ and $c = -2[(\theta_2 - 1)\gamma_1^2 - (\theta_1 - 1)\gamma_2^2 + (\theta_1 - 1)(\theta_2 - 1)(\theta_1 - \theta_2)]$. An unbiased estimator of $\sigma_2^2$ is given in the following theorem.

**Theorem 4.** If $(\theta_2 - 1)(1 - 2\theta_1 + \theta_2)\gamma_1^2 + (\theta_1 - 1)^2 \gamma_2^2 = 2\theta_1(\theta_1 - 1)(\theta_2 - 1) \times (\theta_1 - \theta_2)$, an unbiased estimator of $\sigma_2^2$ is given by

$$\sigma_2^2 = \frac{as_2^2 + bs_2^2 + c \left[ \left( n^{-1} \sum_{i=1}^n Z_{1i}^2 \right) - Z_1 Z_2 \right]}{(\gamma_2^2 + \theta_1^2 - \gamma_1^2 - \theta_1^2)(\theta_1 - \theta_2)},$$

whence provided that $\gamma_1^2 + \theta_1^2 \neq \gamma_2^2 + \theta_2^2$ and $\theta_1 \neq \theta_2$.
unbiasedness of \( \hat{\mu} \) and efficiency of the proposed estimator compared with Gupta et al. [9] procedure. The percent relative efficiency of the proposed estimator \( \hat{\mu} \) with respect to Gupta et al. estimator is remarkable that the actual values of \( \alpha \) and the minimum value of \( \text{Var}(\hat{\mu}) \) given in (2.2) and (2.7) respectively, given by

\[
\text{Var}(\hat{\mu}, \hat{W}) = \alpha_1 \text{Var}(\hat{\mu}_x) + \alpha_2 \text{Var}(\hat{W})
\]

\[
= \frac{1}{(\theta_1 + \theta_2)^2 \mu_x^2} \left[ \{\alpha_1(1 - \theta_2)^2 \mu_x^2 + \alpha_2[1 + W(\theta_2 - 1)]^2\} \sigma_{Z_1}^2/n_1 
+ \{\alpha_1(1 - \theta_1)^2 \mu_x^2 + \alpha_2[1 + W(\theta_1 - 1)]^2\} \sigma_{Z_2}^2/n_2 \right]
\]

where \( \alpha_1 \) and \( \alpha_2 \) are non-negative coefficients chosen by the investigator.

From using (2.9), (2.10) and the fact that \( E(s_i^2) = \sigma_i^2 \), \( i = 1, 2 \), the unbiasedness of \( \hat{\mu} \) then follows. Hence the theorem.

We now move on to study the appropriate sample size allocations for various objectives in sampling surveys. Consider a linear combination of \( \hat{\mu}_x, \hat{W} \), given in (2.2) and (2.7) respectively, for which

\[
\text{Var}(\hat{\mu}_x, \hat{W}) = \alpha_1 \text{Var}(\hat{\mu}_x) + \alpha_2 \text{Var}(\hat{W})
\]

\[
= \frac{1}{(\theta_1 + \theta_2)^2 \mu_x^2} \left[ \{\alpha_1(1 - \theta_2)^2 \mu_x^2 + \alpha_2[1 + W(\theta_2 - 1)]^2\} \sigma_{Z_1}^2/n_1 
+ \{\alpha_1(1 - \theta_1)^2 \mu_x^2 + \alpha_2[1 + W(\theta_1 - 1)]^2\} \sigma_{Z_2}^2/n_2 \right]
\]

where \( \alpha_1 \) and \( \alpha_2 \) are non-negative coefficients chosen by the investigator. If the total sample size \( n \) (= \( n_1 + n_2 \)) is fixed, then through a simple application of the Cauchy-Schwarz inequality, the sample allocation for which \( \text{Var}(\hat{\mu}_x, \hat{W}) \) attains its minimum is given by

\[
n_1/n_2 = \left[ \{\alpha_1(1 - \theta_2)^2 \mu_x^2 + \alpha_2[1 + W(\theta_2 - 1)]^2\} \sigma_{Z_1}^2 
+ \{\alpha_1(1 - \theta_1)^2 \mu_x^2 + \alpha_2[1 + W(\theta_1 - 1)]^2\} \sigma_{Z_2}^2 \right]^{-1/2}
\]

and the minimum value of \( \text{Var}(\hat{\mu}_x, \hat{W}) \) will be

\[
\text{Min} \text{ Var}(\hat{\mu}_x, \hat{W}) = \left( \left[ \{\alpha_1(1 - \theta_2)^2 \mu_x^2 + \alpha_2[1 + W(\theta_2 - 1)]^2\} \sigma_{Z_1}^2 
+ \{\alpha_1(1 - \theta_1)^2 \mu_x^2 + \alpha_2[1 + W(\theta_1 - 1)]^2\} \sigma_{Z_2}^2 \right]^{1/2} \right)^2/n(\theta_1 - \theta_2)^2 \mu_x^2
\]

It is remarkable that the actual values of \( \sigma_{Z_1}, \sigma_{Z_2}, \) and \( W \) are always unknown, information regarding them can be obtained from past experience or a pilot survey, which is helpful for practical applications (Murthy [11], pp. 96–99).

3. Efficiency comparison

In what follows, we study the performance of the proposed procedure compared with Gupta et al. [9] procedure. The percent relative efficiency of the proposed estimator \( \hat{\mu} \) with respect to Gupta et al. estimator

\[ E(Z_1 Z_2) = [1 + (\theta_1 + \theta_2 - 2) W + (\theta_1 - 1)(\theta_2 - 1) W^2] \mu_x^2. \]
\( \hat{\mu}_G \) is defined as
\[
PRE = \frac{(\theta_1 - \theta_2)^2 \sigma_Z^2}{(|1 - \theta_2| \sigma_{Z_1} + |1 - \theta_1| \sigma_{Z_2})^2} \times 100,
\]
where \( \sigma_Z^2 \) and \( \sigma_{Z_i}^2 \) are respectively given in (1.1) and (2.3). Since the above \( PRE \) expression is infested with too many parameters, here the comparison is performed on an empirical investigation using the same probability distribution in the competing randomizing devices. To examine how the variation of randomization devices will affect the efficiency, it is assumed that scrambled variables follow the distributions with mean \( \mu \) and variance \( \gamma^2 \) satisfying \( \gamma^2 = k \mu \). Without loss of generality, here we simply choose \((\mu_S, \gamma_S^2) = (1, k), (\theta_1, \gamma_1^2) = (k, k^2) \) and \((\theta_2, \gamma_2^2) = (0.5, 0.5 k)\), where \( k = 2, 3, \ldots, 10 \). It is remarkable that the \( PRE \) value remains unchanged if the coefficient of variation \( CV_x = \sigma_x \mu^{-1} \) is fixed. The values of \( CV_x \) are then chosen to be 0.2, 0.4, 0.6, 0.8 and 1. The percent relative efficiencies thus obtained are outlined in Table 1 for different values of \( W \).

**Table 1**

Percent relative efficiency of the two competing procedures

<table>
<thead>
<tr>
<th>CV_x</th>
<th>W</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>92.2</td>
<td>96.1</td>
<td>100.8</td>
<td>105.2</td>
<td>109.1</td>
<td>112.6</td>
<td>115.7</td>
<td>118.4</td>
<td>120.9</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>95.9</td>
<td>101.0</td>
<td>106.1</td>
<td>110.6</td>
<td>114.5</td>
<td>117.9</td>
<td>120.9</td>
<td>123.5</td>
<td>125.9</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>100.0</td>
<td>105.7</td>
<td>111.0</td>
<td>115.5</td>
<td>119.3</td>
<td>122.5</td>
<td>125.4</td>
<td>127.9</td>
<td>130.1</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>104.5</td>
<td>110.9</td>
<td>116.2</td>
<td>120.6</td>
<td>124.3</td>
<td>127.4</td>
<td>130.1</td>
<td>132.5</td>
<td>134.6</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>109.4</td>
<td>116.5</td>
<td>121.9</td>
<td>126.2</td>
<td>129.8</td>
<td>132.7</td>
<td>135.3</td>
<td>137.6</td>
<td>139.6</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>93.4</td>
<td>95.4</td>
<td>99.0</td>
<td>102.7</td>
<td>106.2</td>
<td>109.5</td>
<td>112.4</td>
<td>115.1</td>
<td>117.5</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>96.3</td>
<td>100.7</td>
<td>105.5</td>
<td>109.8</td>
<td>113.6</td>
<td>116.9</td>
<td>119.8</td>
<td>122.4</td>
<td>124.8</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>100.0</td>
<td>105.4</td>
<td>110.4</td>
<td>114.8</td>
<td>118.5</td>
<td>121.7</td>
<td>124.6</td>
<td>127.1</td>
<td>129.3</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>104.1</td>
<td>110.2</td>
<td>115.3</td>
<td>119.7</td>
<td>123.3</td>
<td>126.4</td>
<td>129.1</td>
<td>131.5</td>
<td>133.6</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>108.4</td>
<td>115.3</td>
<td>120.5</td>
<td>124.8</td>
<td>128.3</td>
<td>131.2</td>
<td>133.8</td>
<td>136.1</td>
<td>138.1</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>94.6</td>
<td>95.4</td>
<td>98.0</td>
<td>101.0</td>
<td>104.0</td>
<td>106.8</td>
<td>109.5</td>
<td>112.1</td>
<td>114.4</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>96.6</td>
<td>100.4</td>
<td>104.8</td>
<td>108.9</td>
<td>112.5</td>
<td>115.7</td>
<td>118.6</td>
<td>121.2</td>
<td>123.6</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>100.0</td>
<td>105.0</td>
<td>109.8</td>
<td>114.0</td>
<td>117.6</td>
<td>120.8</td>
<td>123.6</td>
<td>126.1</td>
<td>128.3</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>103.6</td>
<td>109.4</td>
<td>114.4</td>
<td>118.6</td>
<td>122.1</td>
<td>125.2</td>
<td>127.9</td>
<td>130.2</td>
<td>132.4</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>107.5</td>
<td>113.9</td>
<td>119.0</td>
<td>123.1</td>
<td>126.6</td>
<td>129.5</td>
<td>132.1</td>
<td>134.3</td>
<td>136.3</td>
<td></td>
</tr>
</tbody>
</table>

(Contd. Table 1)
From Table 1, it is seen that the proposed procedure performs better than Gupta et al. [9] procedure for most of the practical situations. Even though in some cases the proposed procedure is less efficient, it can be viewed as a tradeoff for being able to get an unbiased estimator of $\sigma^2$ and an asymptotically unbiased estimator of $W$. Note that the PRE value increases with $k$ and/or $W$, whereas decreases with $CV_x$, if other parameters are unchanged.

References


Received April, 2005