Imputing and Jackknifing Scrambled Responses

Sarjinder Singh[1*], Jong-Min Kim[2] and Inderjit Singh Grewal[3]

[1]Department of Mathematics, The University of Texas at Brownsville and Texas Southmost College, Brownsville, Texas, TX 78520, U.S.A (*-Current address: Department of Mathematics, Texas A&M University-Kingsville, Kingsville, Texas 78363, U.S.A.) (**E-mail:** sarjinder@yahoo.com)
[2]Statistics, Division of Science and Mathematics University of Minnesota--Morris, Morris, MN 56267, U.S.A. (**E-mail:**jongmink@umn.edu)
[3]Department of Mathematics, Statistics and Physics, Punjab Agricultural University, Ludhiana 141004, India (**E-mail**:isg1969@yahoo.com)

"Scrambled data are as good as scrambled eggs." In the present investigation, it has been shown that scrambled responses on sensitive variables such as income, drugs used, induced abortions etc. can also be imputed by following Singh, Joarder and King (1996) and can be Jackknifed to estimate the variance of the resultant ratio type estimator by following Rao and Sitter (1995). Results have been simulated under different levels of untruthful reporting by following Singh, Joarder and King (1996) and are compared with those from Rao and Sitter (1995) study.

Keywords: Estimation of mean, estimation of variance, Jackknifing, sensitive variables.

1. INTRODUCTION

You may have traveled by Greyhound Bus, which makes morning stops. You may have realized that on these morning stops you can find scrambled eggs sold at several bus stands. As these scrambled eggs are sufficient to prevent starvation during the journey, scrambled data are also equally good in protecting the privacy of respondents during a survey. The collection of data through personal interview surveys on sensitive issues such as induced abortions, drug abuse, and family income is a serious issue. Warner (1965) considered the case where the respondents in a population can be divided into two mutually exclusive groups: one group with stigmatizing/sensitive characteristic A and the other group without it. For estimating π , the proportion of respondents in the population belonging to the sensitive group A, a simple random sample of n respondents is selected with replacement from the population. For collecting information on the sensitive characteristic, Warner (1965) made use of a randomization device. One such device could be a deck of cards with each card having one of the following two statements: (i) "I belong to group A" , (ii) "I do not belong to group A." The statements occur with relative frequencies p_0 and $(1-p_0)$ respectively in the deck of cards. Each respondent in the sample is asked to select a card at random from the well-shuffled deck. Without showing the card to the interviewer, the interviewee answers the question, "Is the statement true for you?" 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, by means of a randomization device, one of two questions. However, there are several difficulties, which arise when using this unrelated question method. The main one is choosing the unrelated question. As Greenberg et al. (1971) note, 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. However, the mean and variance of the responses to the sensitive question are unknown, making it difficult to choose good unrelated questions. A second possibility is that in some cases the answers to the unrelated question may be more rounded or regular, making it possible to recognize which question was answered. For example, Greenberg *et al.* (1971) considered the sensitive question: "how much money did the head of this household earn last year?" This was paired with the question: "how much money do you think the average head of a household of your size earns in a year?" An answer such as \$26,350 is more likely to be in response to the unrelated question, while as answer such as \$18,618 is almost certainly in response to the sensitive question. A third difficulty is that some people are hesitant to disclose their answer to the sensitive question even though they know that the interviewer cannot be sure that the sensitive question was selected. For example, some respondents may not want to reveal their income even though they know that the interviewer can only be 75% certain, say, that the figure given is the respondent's income. These difficulties are no longer present in the scrambled randomized response method introduced by Eichhorn and Hayre (1983). This method can be summarized as follows: each respondent scrambles in response Y by multiplying it by a random scrambling variable S and only then reveals the scrambled result $Z = YS$ to the interviewer. The mean of the response, $E(Y)$ can be estimated from a sample of Z and the knowledge of the distribution of the scrambling variable S. This method may also be used to estimate the median or other parameters of the distribution function of Y as reported by Ahsanullah and Eichhorn (1988). If some auxiliary information is available then such scrambled responses can also be used in regression analysis by following Singh, Joarder and King (1996), Strachan, King and Singh (1998), and Singh and King (1999). We consider a practicable randomization device proposed by Chaudhuri and Adhikary (1990). According to this device, the i^th respondent in the sample is required to choose independently at random two tickets numbered S_{1j} and S_{2k} out of boxes proposed by the investigator containing the tickets numbered (i) $A_1, A_2, ..., A_m$ with known mean A and known variance σ_A^2 and (ii) $B_1, B_2, ..., B_t$ with known mean B and variance σ_B^2. The respondent is required to report the response as

$Z_i = S_{1j}Y_i + S_{2j}$. The problems of underreporting and non-response on the sensitive issues are very common diseases in most of the surveys due to our social setup. For examples, students do not want to disclose their true GPA, politicians do not want to disclose the true number of votes, girls do not want to disclose their number of boyfriends, drug users do not want to disclose the amount of drugs they use, businessmen do not want to disclose their true income or tax dodging, doctors do not want to disclose the true number of AIDS patients in their areas because of confidentiality and politicians do not want to disclose the true number of murders committed by them. In fact, there is no end of such issues where underreporting and non-response is not expected. It is not a fault of anybody that he/she underreports or refuses to report on some personal questions related to him/her, because in some cases lying is necessary to maintain one's social status. Thus there is a strong need to develop techniques that would ensure respondents' privacy if they respond truthfully concerning personal questions. It is fact that medical doctors have not paid much attention to develop any medicine or tablets that we could give birth to honest kids, we feel that some survey techniques could be developed which should ensure our kids that even if they will tell truth through the randomization devices then their privacy could be maintained. Recent applications of randomized response sampling techniques can be found in Gjestvang and Singh (2006, 2007), Elffers *et al* (2003) and Clark and Desharnais (1998).

In the present paper, we consider the situation if some of the respondents refuse to give scrambled responses related to their sensitive questions. We show that if no scrambling is applied then the proposed imputing method leads to the Rao and Sitter (1995) method of imputation.

2. IMPUTING THE SCRAMBLED RESPONSES

Consider that we selected a simple random without replacement (SRSWOR) sample of n respondents from a population consisting of N units. Let y_i be the true response, for example income, of the ith respondent in the sample. The ith respondent selected in the sample is requested to draw two numbers S_1 and S_2 from the two independent randomization devises say R_1 and R_2 respectively, and report the scrambled response as:

$$Z_i = (S_1 y_i + S_2 - B)/A \tag{2.1}$$

where $E_R(S_1) = A$ and $E_R(S_2) = B$, and A and B are known. Also let $V_R(S_1) = \sigma_A^2$ and $V_R(S_2) = \sigma_B^2$ be known. Let r be the number of respondents in the sample s_1 who responded to the sensitive question with the help of the above randomization device and the remainder of the $(n-r)$ selected units be in the sample s_2 who refused to respond using the above randomization device, such that $s = s_1 \cup s_2$. Thus we have the following situation:

$$\hat{Z}_i = \begin{cases} Z_i, & i \in s_1 \\ b_s x_i, & i \in s_2 \end{cases} \tag{2.2}$$

Let x_i be an auxiliary variable related to the study variable y_i . Following Singh, Joarder and King (1996), under the models:

$$Z_i = bx_i + \eta_i \text{ and } y_i = bx_i + e_i \tag{2.3}$$

where $e_i \sim N(0, \sigma^2)$ and the distribution of η_i is unknown, a linear unbiased estimator of the regression coefficient b is:

$$\hat{b}_s = \bar{z}/\bar{x} \tag{2.4}$$

The goodness of fit of the regression model (2.3) with scrambled responses could easily be assessed by following Singh and King (1999). Under the above mechanism of imputing the scrambled responses, the point estimator $\bar{y}_s$ of the population mean $\bar{Y}$ defined as:

$$\bar{y}_s = \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_i \tag{2.5}$$

becomes:

$$\bar{y}_{sr} = \bar{z}_r\left(\frac{\bar{x}_n}{\bar{x}_r}\right) \tag{2.6}$$

where $\bar{z}_r = \frac{1}{r}\sum_{i=1}^{r} Z_i$, $\bar{x}_r = \frac{1}{r}\sum_{i=1}^{r} x_i$, and $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. The estimator $\bar{y}_{sr}$ in (2.6) is clearly a ratio estimator, with scrambled responses (sr), of the population mean. In the next section, we develop theoretical derivations of the bias and variance of the proposed estimator to the first order of approximation.

3. BIAS AND VARIANCE OF THE RATIO ESTIMATOR

The following theorems have been devoted to study the bias and variance of the ratio estimator under scrambled responses.

Theorem 3.1. The bias in the scrambled ratio estimator $\bar{y}_{sr}$, to the first order of approximation, is

$$B(\bar{y}_{sr}) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left[C_x^2 - \rho_{xy} C_x C_y\right] \tag{3.1}$$

Proof. See Singh *et al* (2008).

Theorem 3.2. The mean squared error of the scrambled ratio estimator $\bar{y}_{sr}$, to the first order of approximation, is given by:

$$\text{MSE}(\bar{y}_{sr}) = \text{MSE}(\bar{y}_{rr}) + \frac{\bar{Y}^2}{n}\left\{C_A^2(1 + C_y^2) + (B/A)^2 C_B^2\right\} \tag{3.2}$$

where

$$\text{MSE}(\bar{y}_{rr}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right] \tag{3.3}$$

and $R = \bar{Y}/\bar{X}$.

Proof. See Singh *et al* (2008).

Thus the mean squared error (MSE) of the ratio estimator $\bar{y}_{sr}$ under the scrambled responses has one more term than the MSE of the unscrambled ratio estimator $\bar{y}_{rr} = \bar{y}_r \bar{x}_n / \bar{x}_r$ (see Cochran (1997)). The estimation of variance of the ratio estimator under scrambled responses is more tedious, and there are more chances that the jackknife estimator of variance may lead to serious underestimation.

4. JACKKNIFE ESTIMATOR OF VARIANCE

Following Rao and Sitter (1995), we apply the Jackknife technique to the scrambled ratio estimator. Let $s = s_1 \cup s_2$, and define:

$$\bar{z}_r(j) = \begin{cases} \dfrac{r\,\bar{z}_r - z_j}{r-1}, & \text{if} \quad j \in s_1 \\ \bar{z}_r, & \text{if} \quad j \in s_2 \end{cases}, \quad \bar{x}_r(j) = \begin{cases} \dfrac{r\,\bar{x}_r - x_j}{r-1}, & \text{if} \quad j \in s_1 \\ \bar{x}_r, & \text{if} \quad j \in s_2 \end{cases} \text{ and } \bar{x}_n(j) = \frac{n\,\bar{x}_n - x_j}{n-1}, \quad \text{if} \quad j \in s \tag{4.1}$$

Following Rao and Sitter (1995), we define a Jackknife estimator of the variance of the ratio estimator $\bar{y}_{sr}$ under the scrambled responses (sr) given by:

$$\hat{v}_{\text{jack}}(\text{sr}) = \frac{(n-1)}{n}\sum_{j=1}^{n}\left[\bar{y}_{\text{sr(j)}} - \bar{y}_{sr}\right]^2 \tag{4.2}$$

where

$$\bar{y}_{sr(j)} = \bar{z}_{r(j)}\left[\frac{\bar{x}_{n(j)}}{\bar{x}_{r(j)}}\right], \quad j = 1,2,\ldots,n \tag{4.3}$$

It can be easily observed that:

$$E_R\left[\bar{y}_{\text{sr}(j)} - \bar{y}_{\text{sr}}\right]^2 = \begin{cases} \dfrac{1}{(n-1)^2}\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^2 (y_j - \hat{R}x_j)^2 + \hat{R}^2\dfrac{(x_j - \bar{x}_n)^2}{(n-1)^2} + 2\hat{R}\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)\dfrac{(x_j - \bar{x}_n)^2}{(n-1)} \\ + C_A^2\left[\dfrac{1}{(n-1)^2}\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^2 (y_j - \hat{R}x_j)^2 + \hat{R}^2\dfrac{(x_j - \bar{x}_n)^2}{(n-1)^2}\right] \\ + \dfrac{\sigma_B^2}{A^2}\left[\dfrac{1}{(n-1)^2}\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^2\left(\dfrac{x_j}{\bar{x}_r} - 1\right)^2 + \dfrac{1}{(n-1)^2}\left(\dfrac{x_j}{\bar{x}_n} - 1\right)^2\right], & \text{if} \quad j \in s_1 \\ \\ \hat{R}^2\dfrac{(x_j - \bar{x}_n)^2}{(n-1)^2} + \dfrac{(x_j - \bar{x}_n)^2}{(n-1)^2}\left\{C_A^2\hat{R}^2 + \sigma_B^2/(\bar{x}_r^2 A^2)\right\}, & \text{if } j \in s_2 \end{cases} \tag{4.4}$$

If $C_A = 0$ and $\sigma_B = 0$, then (4.4) reduces to the same expression due to Rao and Sitter (1995). In other words, if no scrambling is used then the jackknife estimator of variance in (4.2) reduces to the estimator due to Rao and Sitter (1995). In order to study the performance of the proposed imputation method of scrambled responses we have performed an extensive simulation study in the next section.

5. SIMULATION STUDY AND DISCUSSION OF RESULTS

For the purpose of simulation study, two different populations having different amounts of correlation values between the study and auxiliary variable, given on the CD with the book by Lohr (1999), were investigated thoroughly under different three mechanism: (a) direct questions on the study variable are feasible; (b) direct questions may cause different types of underreporting; (c) scrambled responses are obtained. The data set given in file **agpop.dat** has been used in this empirical study after dropping the data values marked as –99 by Lohr (1999), after cleaning the data (see Singh *et al* (2008)). In the first population, we considered the study variable: $Y_i =$ Number of acres devoted to farms during 1992 (ACRES92), and the independent variable: $X_i =$ Number of acres devoted to farms during 1982 (ACRES82), which has a very high correlation coefficient with the study variable. Note that the number of acres devoted to farms by the ith family in a particular year could be sensitive or non-sensitive depending upon the situation. For example, if a farmer is expected to get benefits from the Government by reporting the number of acres destroyed by any storm such as hurricane, tornado etc, then the farmer will try to report more acres. In contrast, if a farmer knows that the Government is expected to put more taxes based on the number of acres devoted to farming then the farmer is expected to report fewer acres than the actual. At the same time, the number of acres devoted by the ith farmer about 10 years ago may be known from some secondary sources of data.

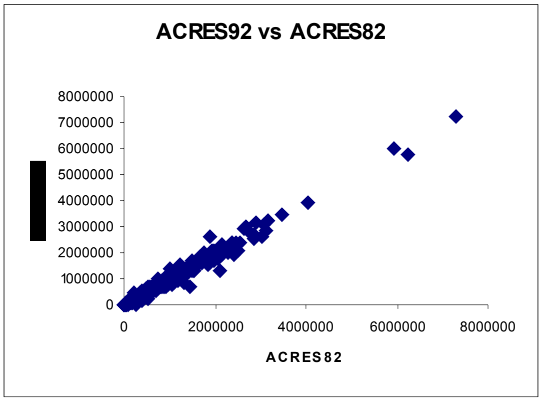

Fig. 5.1. ACRES92 vs ACRES82 from agpop.dat

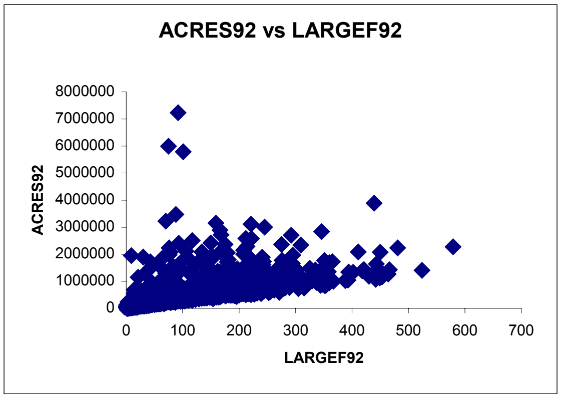

Fig. 5.2. ACRES92 vs LARGEF92 from agpop.dat

In the second population, we keep the same study variable (ACRES92) but we consider a different auxiliary variable, X_i =Number of large farmers during 1992 (LARGEF92), which has a low correlation with the study variable. Now let us explain the simulation procedure with the help of one population as follows. We used a subroutine CALL RNUN (N, R) from the IMSL subroutines to generate $N = 3050$ uniform random numbers between 0 and 1 and assigned to the population units from 1 through $N = 3050$. We set a condition that if the random number R is less than or equal to $f_1 = n/N$ then we retained the unit in the sample, otherwise we rejected it. In other words, we selected the first phase large simple random sample with a finite population correction factor (f.p.c) f_1 which means that $n = f_1 N$ units have been selected in the first phase sample. Again we assigned another set of uniform random numbers by calling RNUN (N, R) subroutine to all the units in the sample and later we retained only $f_2 = r/N$, such that $f_2 < f_1$, of the sample in the second phase sampling. The rest of the $(1 - f_2)\%$ we treated as a random nonresponse in the sample, which implies $r = f_2 N$ units are selected in the second phase sample of responding. In the first phase of the large sample of n size, we measured the values of the auxiliary variable x_i for the units included in the sample and computed the first phase sample mean as $\bar{x}_n$. In the second phase sample of r size, we generated two scrambling variables $S_1 \sim \text{Beta}(\alpha, \beta)$ and $S_2 \sim N(B, \sigma_B)$. We generated the scrambling variable S_1 by using the subroutine CALL RNBET(r, PIN, QIN, S1) with $\text{PIN} = \alpha = 0.6$ and $\text{QIN} = \beta = 0.05$. Thus, the mean $E(S_1) = A = \alpha/(\alpha + \beta)$ and the variance $V(S_1) = \sigma_A^2 = (\alpha\beta)/\left[(\alpha + \beta)^2(\alpha + \beta + 1)\right]$ of the scrambling variable are known. We used the subroutine CALL RNNOR (r, S) to generate another independent scrambling variable S from the standard normal distribution and later used a transformation $S_2 = B + \sigma_B S$ with $B = 1{,}000$ and $\sigma_B = 0.025$ to generate a scrambling variable $S_2 \sim N(B, \sigma_B)$. Then we noted three measures from all the units in the second phase sample: the study variable y_i, the auxiliary variable x_i, and the scrambled response $z_i = (S_1 y_i + S_2 - B)/A$, for $i = 1,2,...,r$. The fourth measure we modeled that the respondents may provide underreporting when asked direct questions on the study variable as:

$$y_i^{(u)} = y_i u \exp(-c(1-u) y_i) \tag{5.1}$$

where $u = 0.7,\ 0.8,\ 0.9$, 1 and $c = 0.000001,\ 0.000003,\ 0.000007,\ 0.000009$. The choice of c depends upon the value of the study variable that a respondent with a higher value of the study variable may provide a lower response. The value of $u = 1$

means that there is no underreporting, and we are comparing a true response model with a scrambled response model. We generated K = 100,000 samples. Then from each one of the k^{th} sample, for $k = 1,2,...,K$, we computed three different ratio estimates of the true population mean $\overline{Y}$ as:

(i) Full true response in the second phase sample as:

$$\overline{y}_{\text{rr}} |_k = \overline{y}_r \left(\frac{\overline{x}_n}{\overline{x}_r} \right) \tag{5.2}$$

(ii) Full but underreporting in the second phase sample as:

$$\overline{y}_{\text{ur}} |_k = \overline{y}_r^{(u)} \left(\frac{\overline{x}_n}{\overline{x}_r} \right) \tag{5.3}$$

where $\overline{y}_r^{(u)} = \frac{1}{r} \sum_{i=1}^{r} y_i^{(u)}$. (5.4)

(iii) Full but scrambled reporting in the second phase sample as:

$$\overline{y}_{\text{sr}} |_k = \overline{z}_r \left(\frac{\overline{x}_n}{\overline{x}_r} \right) \tag{5.5}$$

We estimated the variance of these three estimators with the concept of Jackknifing due to Rao and Sitter (1995) as explained in Section 4 as follows:

$$\hat{v}_{\text{jack}}(\text{hh}) |_{\text{k}} = \frac{(n-1)}{n} \sum_{j=1}^{n} \left[\overline{y}_{\text{hh}}(j) - \overline{y}_{hh} \right]^2 \tag{5.6}$$

for $hh = rr, ur,$ and sr respectively.

Then, we computed the proportion of times the true population mean, $\overline{Y}$, falls in the following three 95% confidence interval estimates:

$$\overline{y}_{\text{hh}} |_k \quad \pm \quad 1.96 \sqrt{\hat{v}_{\text{jack}}(\text{hh}) |_k} \tag{5.7}$$

which we denote as CCI(rr), CCI(ur), and CCI(sr), $hh = rr, ur,$ and sr, respectively. We also computed the true mean squared errors of the usual ratio estimator and the estimator under scrambled responses as:

$$\text{MSE}(\overline{y}_{rr}) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n} \right) \left[S_y^2 + R^2 S_x^2 - 2RS_{xy} \right] \tag{5.8}$$

and

$$\text{MSE}(\overline{y}_{\text{sr}}) = \text{MSE}(\overline{y}_{\text{rr}}) + \frac{1}{n} \left\{ C_A^2 (R^2 \overline{X}^2 + S_y^2) + \overline{Y}^2 (B/A)^2 C_B^2 \right\} \tag{5.9}$$

Then we computed the empirical MSE of the three jackknifed estimators of the variance of the ratio estimator as:

$$\text{MSE}\left[\hat{v}_{\text{jack}}(\text{rr}) \right] = \frac{1}{K} \sum_{k=1}^{K} \left[\hat{v}_{\text{jack}}(\text{rr}) |_{\text{k}} - \text{MSE}(\overline{y}_{rr}) \right]^2 \tag{5.10}$$

$$\text{MSE}\left[\hat{v}_{\text{jack}}(\text{ur}) \right] = \frac{1}{K} \sum_{k=1}^{K} \left[\hat{v}_{\text{jack}}(\text{ur}) |_{\text{k}} - \text{MSE}(\overline{y}_{rr}) \right]^2 \tag{5.11}$$

and

$$\text{MSE}\left[\hat{v}_{\text{jack}}(\text{sr}) \right] = \frac{1}{K} \sum_{k=1}^{K} \left[\hat{v}_{\text{jack}}(\text{sr}) |_{\text{k}} - \text{MSE}(\overline{y}_{sr}) \right]^2 \tag{5.12}$$

The relative efficiency of the jackknifed estimator of variance of the ratio estimator with respect to that based on underreporting and scrambled responses has been computed as:

$$\text{RE}\left[\hat{v}_{ur}, \hat{v}_{rr} \right] = \frac{\text{MSE}\left[\hat{v}_{\text{jack}}(\text{ur}) \right]}{\text{MSE}\left[\hat{v}_{\text{jack}}(\text{rr}) \right]} \times 100\% \tag{5.13}$$

and

$$\text{RE}\left[\hat{v}_{sr}, \hat{v}_{rr} \right] = \frac{\text{MSE}\left[\hat{v}_{\text{jack}}(\text{sr}) \right]}{\text{MSE}\left[\hat{v}_{\text{jack}}(\text{rr}) \right]} \times 100\% \tag{5.14}$$

We also computed the empirical variance and empirical mean squared errors of the four estimators as:

$$V(\overline{y}_r)_e = \frac{1}{K} \sum_{k=1}^{K} \left[\overline{y}_r |_k - \overline{Y} \right]^2 \tag{5.15}$$

and $$\text{MSE}\left(\bar{y}_{hh}\right)_e = \frac{1}{K}\sum_{k=1}^{K}\left[\bar{y}_{hh}\,|_k - \bar{Y}\right]^2 \tag{5.16}$$

for $hh = rr, ur,$ and sr respectively. Then the empirical relative efficiencies of the ratio ($\bar{y}_{rr}$), underreporting ($\bar{y}_{ur}$) and scrambled responses ($\bar{y}_{sr}$) estimators with respect to the sample mean estimator ($\bar{y}_r$) were computed as:

$$\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right] = \frac{\text{V}\left(\bar{y}_r\right)_e}{\text{MSE}\left[\bar{y}_{rr}\right]_e} \times 100\% \tag{5.17}$$

$$\text{RE}\left[\bar{y}_{ur}, \bar{y}_r\right] = \frac{\text{V}\left(\bar{y}_r\right)_e}{\text{MSE}\left[\bar{y}_{ur}\right]_e} \times 100\% \tag{5.18}$$

and $$\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right] = \frac{\text{V}\left(\bar{y}_r\right)_e}{\text{MSE}\left[\bar{y}_{sr}\right]_e} \times 100\% \tag{5.19}$$

The results so obtained are presented in the Appendix with the Fig I and Fig II for the first population, and Fig III and Fig IV for the second population. In Fig I we compared the ratio estimator when there is no underreporting on the sensitive variable with the ratio estimator under a scrambled variable when the correlation between the study and auxiliary variable is positive and high. Thus the ratio estimator under scrambled responses is expected to be less efficient than the ratio estimator under truthful responses. Here $f_1 = 0.03$ means that we selected 3% of the population as the first phase sample consisting of $n = 92$ units, and $f_2 = 0.01$ showing that we selected 1% of the population as the second phase sample of $r = 30$ units which results in non-response of $n - r = 62$ units. This process was repeated 100,000 times. The missing units were imputed with the ratio method of imputation for both the scrambled responses and the true responses. It is interesting to note from Fig I(a) that the 95% coverage by the scrambled responses is 91.024%, while the ratio estimator under the true responses is 90.868%. This means the coverage by the scrambled response is better than the ratio estimator under true responses, however this it is an unexpected result. We changed the values of f_1 from 0.03 to 0.21, with several increments as shown in Fig I(a), which implies that we studied sample sizes 3% to 21%. In the second phase sample size the values changed from 1% to 5%, with an increment of 2%, which would be the most practicable situation in real surveys. For the first population having a high value of correlation coefficient, if $f_1 = 0.15$ and $f_2 = 0.03$ then the value of $\text{CCI}(\text{rr})$ becomes 95.043% which is very close to the expected coverage of 95%. At this times the coverage $\text{CCI}(\text{sr})$ due to the ratio estimator under the scrambled responses becomes 95.102%. Thus the ratio estimator under the scrambled responses shows a bit higher coverage than the expected coverage of 95%. Further note that for this choice of sampling fractions, the relative efficiency as shown in Fig I(b) of the jackknife estimator of the variance under true responses with respect to the ratio estimator under scrambled responses defined as $\text{RE}\left[\hat{v}_{sr}, \hat{v}_{rr}\right]$ is given by 110.96%. The relative efficiency of the ratio estimator under true responses with respect to the sample mean estimator defined as $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ is given by 536.37% and that of the ratio estimator under scrambled responses defined as $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ is given by 513.68%. In Fig II(a) it is interesting to note that the value of the 95% coverage by the ratio estimator under different models of underreporting defined as $\text{CCI}(\text{ur})$ for different sampling fractions remains much lower than expected. From this study one can conclude that underreporting is more dangerous in real surveys, but the scrambled responses are safe. In Fig II(b) and Fig II(c) if we critically analyze the last two values of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ and $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$, we can see that they indicate that the use of scrambled responses remains slightly less efficient than the ratio estimator under true responses. The loss in efficiency is compromised with the privacy of the respondents. For $f_2 = 0.01$, as the value of f_1 changes from 0.03 to 0.21, the value of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ increases from 307.01% to 2001.59% and the value of $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ changes from 300.94% to 1757.00%. For $f_2 = 0.03$, as the value of f_1 changes from 0.05 to 0.21, the value of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ increases from 169.08% to 785.37% and the value of $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ changes from 166.78% to 738.94%. For $f_2 = 0.05$, as the value of f_1 changes from 0.07 to 0.21, the value of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ increases from 142.16% to 480.50% and the value of $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ changes from 140.49% to 462.43%. Fig II(c) shows that if $u = 0.7$ (or $u = 0.8$), and $0.000001 \le c \le 0.000009$, the values of $\text{RE}\left[\bar{y}_{ur}, \bar{y}_r\right]$ always remain less than 100% indicating that too much underreporting is dangerous and that the ratio estimator remains less efficient than the sample mean estimator. Note that if $u = 0.9$ then at certain places the value of $\text{RE}\left[\bar{y}_{ur}, \bar{y}_r\right]$ is more than 100% indicates that if there is not too much underreporting then it may be possible that the ratio estimator may adjust the underreporting sample mean estimator $\bar{y}_r^{(u)}$ to perform better than the sample mean estimator $\bar{y}_r$ under true responses. For example, in Fig II(c) if $u = 0.9$, $f_1 = 0.03$, and $f_2 = 0.01$ then the value of $\text{RE}\left[\bar{y}_{ur}, \bar{y}_r\right]$ is 158.43%. In this case the value of $\text{RE}\left[\hat{v}_{ur}, \hat{v}_{rr}\right]$ remains only 59.59% indicating that the jackknife estimator of variance for underreporting remains less efficient than the jackknife estimator of variance for true responses for the case of the ratio estimator. The rest of the plotted values in Fig I(a) and Fig I(b) can also be compared with each other and one could determine to use either the scrambled response or direct response question depending upon the sampling fractions available. The results reported in Fig. III and Fig. IV could be used to study the effect of the value of correlation coefficient between the study and the auxiliary variable as it changes from $\rho_{xy} = 0.993$ to $\rho_{xy} = 0.677$. Fig. I(b) and Fig. III(b) indicate that for $f_1 = 0.03$ and $f_2 = 0.01$ the value of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ changes from 307.01% to 127.75%, and the value of $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ changes from 300.94% to 126.42% as the value of ρ_{xy} changes from 0.993 to 0.677. For a low value of correlation of 0.677, the value of $\text{RE}\left[\bar{y}_{rr}, \bar{y}_r\right]$ changes from 127.75% to 150.09% as the value of f_1 changes from 0.03 to 0.21 and f_2 remains as 0.01. The value of $\text{RE}\left[\bar{y}_{sr}, \bar{y}_r\right]$ changes from 126.42% to 147.92%. Fig II(c) and Fig IV(c) indicates that the value of $\text{RE}\left[\bar{y}_{ur}, \bar{y}_r\right]$ changes from 10.36% to 10.11%. Thus, the results in Fig II(c) to Fig IV(c) could be used to study the sensitivity of the results affecting the value of the correlation coefficient between the study and auxiliary variable. After critically examining the results, we conclude that the imputation of scrambled responses may perform very well if any auxiliary variable highly correlated to the study variable is available and could be used as a imputing variable. A modification for the above results is always feasible by following Arnab and Singh (2005), but we are leaving this as an exercise to the readers. We used IMSL subroutines in FORTRAN for doing the entire simulation.

6. CONCLUSION

We conclude that the use of randomized response sampling remains better than underreporting by respondents in case of sensitive variables. It is also possible to impute the missing scrambled data and Jackknifing the scrambled data at the estimation stage. This theory developed here has been validated based on datasets available in Lohr (1999), thus an application of the proposed method with a real dataset remains acknowledged.

REFERENCES

AHSANULLAH, M. and EICHHORN, B.H. (1988). On estimation of response from scrambled quantitative data. *Pak. J. Statist.*, 4(2), A, 83-91.

ARNAB, R. and SINGH, S. (2005). A new method for estimating variance from data imputed with ratio method of imputation. *Statistics and Probability Letters,* 513-519**.**

CHAUDHURI, A. and ADHIKARY, A.K. (1990). Variance estimation with randomized response. *Commun. Statist. – Theory Meth.*, 19(3), 1119-1126.

CLARK, S.J. and DESHARNAIS, R.A. (1998). Honest answers to embarrassing questions: detecting cheating in the randomized response model**.** *Psychology Methods,* 3, 160-168.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd Ed, Wiley.

EICHHORN, B.H. and HAYRE, L.S. (1983). Scrambled randomized response method for obtaining sensitive quantitative data. *J. Statist. Planning and Infer.,* 7, 307-316.

ELFFERS, E., VAN DER HEIJDEN, P.G.M. and HEZEMANS, M. (2003). Explaining regulatory non-compliance: a survey study of rule transgression for two *Deutch* instrumental laws, applying the randomized response method. *J. Quantitative Criminology*, 19, 409-439.

GJESTVANG, C.R. and SINGH, S. (2006). A new randomized response model. *J. Roy. Stat. Soc., B,* 68**,** 523-530**.**

GJESTVANG, C.R. and SINGH, S**.** (2007). Forced quantitative randomized response model: A new device. *Metrika,* 66(2)*,* 243-256.

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R. and HORVITZ D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Jour. Amer. Statist. Assoc.*, 66, 243-250.

HORVITZ, D.G., SHAH, B.V., and SIMMONS, W.R. (1967). The unrelated question randomized response model. *Proc. Social Statist. Sec., Amer. Statist. Assoc.,* 65-72.

LOHR, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

RAO, J.N.K. and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika***,** 82, 453-460.

SINGH, S., JOARDER, A.H. and KING, M.L. (1996). Regression analysis using scrambled responses**.** *Australian J. Statist.* 38(2), 201-211.

SINGH, S., KIM, J.-M. and Grewal, I.S. (2008). Imputing and Jackknifing scrambled responses. *Metron,* LXVI(2), 183-204.

SINGH, S. and KING, M.L. (1999). Estimation of coefficient of determination using scrambled responses. *J. Indian Soc. Agric. Statist*. 52(3), 338-343.

STRACHAN, R., KING, M.L. and SINGH, S. (1998). Likelihood based estimation of the regression model with scrambled responses. *Australian & New Zealand J. Statist.*, 40(3), 279-290.

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

APPENDIX

Fig I. Comparison of the ratio estimator with the scrambled responses estimator for the first population with high value of correlation coefficient.

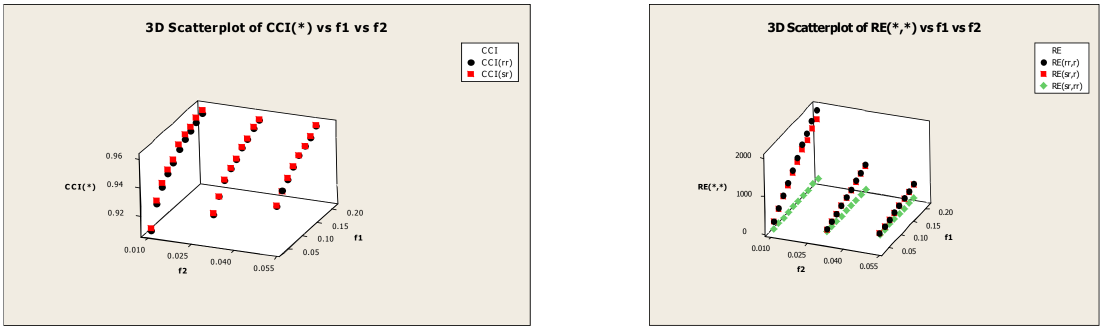

Fig. I(a) The values of CCI(rr) and CCI(sr)　　**Fig. I (b)** The values of RE(rr,r), RE(sr, r) and RE(sr, rr)

Fig II. Comparison of the ratio estimator with underreporting with respect to the sample mean estimator for the first population with high value of correlation coefficient.

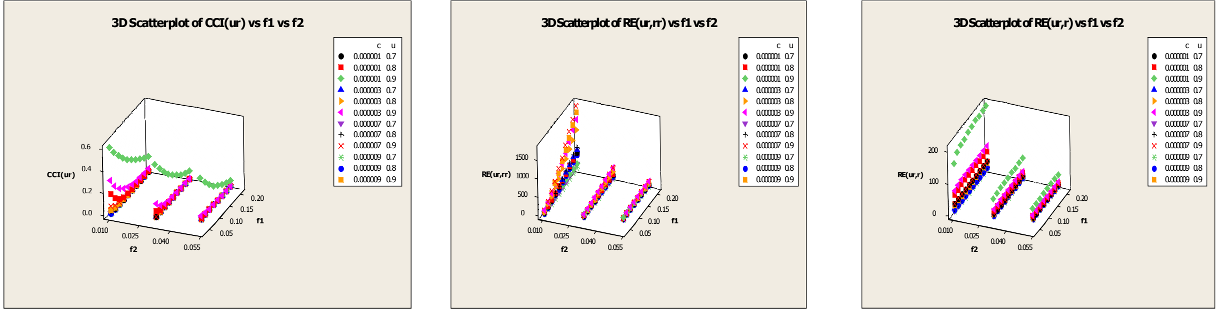

Fig. II(a) The values of CCI(ur)　　**Fig. II(b)** The values of RE(ur, rr)　　**Fig. II (c)** The values of RE(ur, r)

Fig III. Comparison of the ratio estimator with the scrambled responses estimator for the second population with low value of correlation coefficient.

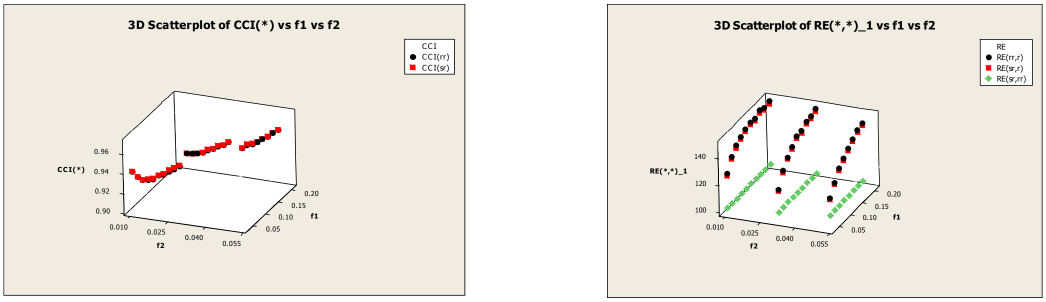

Fig. III(a) The values of CCI(rr) and CCI(sr)　　**Fig. III (b)** he values of RE(rr,r), RE(sr, r) and RE(sr, rr)

Fig IV. Comparison of the ratio estimator with underreporting with respect to the sample mean estimator for the second population with low value of correlation coefficient.

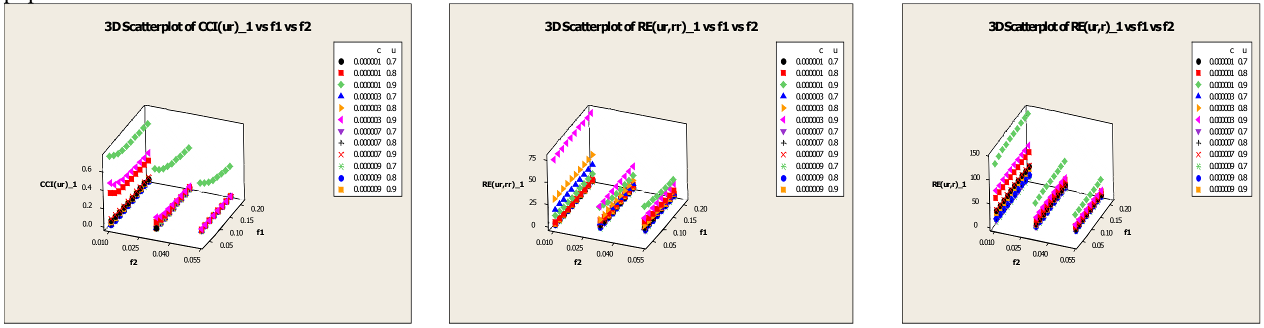

Fig. IV(a) The values of CCI(ur)　　**Fig. IV(b)** The values of RE(ur, rr)　　**Fig. IV(c)** The values of RE(ur, r)