Estimation of Finite Population Variance Using Scrambled Responses in the Presence of Auxiliary Information

SARJINDER SINGH,1 STEPHEN A. SEDORY,1 AND RAGHUNATH ARNAB2

1Texas A & M University-Kingsville, Kingsville, TX, USA
2Department of Statistics, University of Botswana, Gaborone, Botswana

In this article, a new estimator for estimating the finite population variance of a sensitive variable based on scrambled responses collected using a randomization device is introduced. The estimator is then improved by using known auxiliary information. The estimators due to Das and Tripathi (1978: Sankhya) and Isaki (1983: JASA) are shown to be special cases of the proposed estimator. Numerical simulations are performed to study the magnitude of the gain in efficiency when using the estimator with auxiliary information with respect to the estimator based only on the scrambled responses. An idea to extend the present work from SRSWOR design to more complex design is also given.

Keywords Auxiliary information; Efficiency; Finite population; Randomized response sampling; Sensitive variables; Variance estimation.

Mathematics Subject Classification 62D05.

1. Introduction

The collection of data through personal interview surveys on sensitive issues such as induced abortions, drug abuse, and family income is a serious issue. For example, some questions are sensitive: (a) By how much did you underreport your income on your 2009 tax return? (b) How many abortions have you had? (c) How many children have you molested? (e) Do you use illegal drugs? Randomized response techniques are one way to get people to answer truthfully. Horvitz et al. (1967) and Greenberg et al. (1971) have extended Warner’s (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 the two questions: one being the sensitive question, the other being unrelated. As pointed out in Eichhorn and Hayre (1983) there are several difficulties that arise when using this unrelated question method. The main one is that of 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 of the sensitive question: 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 question. A second difficulty 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: about how much money did the head of this household earn last year. This was paired with the question: about 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 an 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 0.75 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 we summarize as follows:

Each respondent scrambles their response $Y$ by multiplying it by a random variable $S$ and then reveals only the scrambled result $Z = YS$ to the interviewer. Thus, the scrambled randomized response model maintains the privacy of the respondents. Assumptions of the model: (a) The variable $S$ is called a scrambling variable and its distribution is assumed to be known. (b) The study variable $Y$ and the scrambling variable $S$ are independent. (c) The selection of the sample units and the randomization procedure are independently carried out, (d) The randomization procedure is independently performed on each individual, and (e) In particular, the quantities $E(S) = \theta$ and $\gamma_a = E(S - \theta)^a$ for $a = 2, 3, 4$ are assumed to be known.

Diana and Perri (2009, 2010, 2011, 2012) and Perri (2008) have rightly pointed out that in direct question survey techniques, when dealing with nonsensitive questions, it is very common to use auxiliary information to improve estimation strategies. A very limited effort has been made to make use of auxiliary information to improve the estimators of sensitive variables, as one might see in referring to Singh et al. (1996), Strachan et al. (1998), Tracy and Singh (1999), Ryu et al. (2005/2006), Mahajan (2005/2006, 2007), Son et al. (2008), Sidhu and Bansal (2008), Zaizai et al. (2008), and Singh and Kim (2011). An extensive review of the literature on randomized response sampling can be found in a recent monograph by Chaudhuri (2011).

To our knowledge, no one has made any attempt to study an estimator of the finite population variance of the study variable using scrambled responses and making use of an auxiliary variable to improve the estimator.

1.1. Notation

Assume that a simple random sample done without replacement (SRSWOR) of size $n$ is drawn from the given population of $N$ units. Let the value of the sensitive study variable, $Y$ and the auxiliary variable, $X$, for the $i$th unit ($i = 1, 2, ..., N$) of the population be denoted by $Y_i$ and $X_i$. We define a few parameters of these variables as follows. Let $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$
denote the population mean of the sensitive study variable \( Y \) and let \( \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \) be the population mean of the auxiliary variable, \( X \). In this article, we consider the problem of estimating the finite population variance (or population mean square error) of the sensitive study variable \( Y \) defined by, 
\[
S^2_y = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2,
\]
in the presence of the known population variance, \( S^2_x = (N - 1)^{-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 \), of the auxiliary variable, \( X \). Let the higher ordered central moments of the study variable \( Y \) and the auxiliary variable \( X \) be given by:
\[
\mu_{ab} = \frac{1}{(N - 1)} \sum_{i=1}^{N} (Y_i - \bar{Y})^a(X_i - \bar{X})^b
\]
for \( a, b = 0, 1, 2, 3, 4, \) etc.

Let \( Z_i = Y_i S_i \) denote the scrambled response from the \( i \)th sampled unit for \( i = 1, 2, ..., n \). Here, we differ from the Das and Tripathi (1978) and Isaki (1983) estimators in that, instead of directly observing responses \( Y_i \) on the sensitive study variable, we observe the scrambled responses, \( Z_i = Y_i S_i \).

### 2. Naive Estimator of the Finite Population Variance

Following Eichhorn and Hayre (1983) the mean of the response, \( \bar{Y} \), can be estimated from a sample of scrambled \( Z \) values by using the knowledge of the distribution of the scrambling variable \( S \). The sample variance of the scrambled responses is given by:
\[
s^2_{z} = \frac{1}{n - 1} \left[ \sum_{i=1}^{n} Z_i^2 - \frac{(\sum_{i=1}^{n} Z_i)^2}{n} \right]
= \frac{1}{n - 1} \left[ \sum_{i=1}^{n} Y_i^2 S_i^2 - \frac{\sum_{i=1}^{n} Y_i^2 S_i^2 + \sum_{i \neq j} \sum_{j=1}^{n} Y_i S_i Y_j S_j}{n} \right].
\]

Let \( E_R \) denote the expected value over the randomization device. Taking expected value \( E_R \) on both sides above, we get:
\[
E_R \left( s^2_{z} \right) = \frac{Y^2}{n} \sum_{i=1}^{n} Y_i^2 + \theta^2 s^2_{y}.
\]

An unbiased estimator of the finite population variance is given by
\[
s^2_{y} = \frac{1}{\theta^2} \left[ s^2_{z} - \frac{\gamma^2}{n(\gamma^2 + \theta^2)} \sum_{i=1}^{n} Z_i^2 \right].
\]

**Theorem 2.1.** The variance of the estimator \( s^2_{y} \) is given by
\[
V(s^2_{y}) = \left( \frac{1}{n} - \frac{1}{N} \right) (\mu_{40} - \mu_{20}^2) + \frac{(\gamma_4 + 4\gamma_3\theta + 12\theta^2 \gamma_2 - 2\gamma_2^2) 1}{N} \sum_{i=1}^{N} Y_i^4
+ \frac{2(\gamma_2^2 + 2\gamma_2\theta^2)}{n(n - 1)\theta^4} \frac{1}{N(N - 1)} \sum_{i \neq j}^{N} Y_i^2 Y_j^2.
\]
where estimator of the finite population variance, as:

Following Das and Tripathi (1978) and Isaki (1983), we define a new, difference-type,

\[ 3. \text{ Difference Type Estimator of the Finite Population Variance} \]

\[ \hat{\sigma}^2 = \frac{4\gamma_2(n-2)}{n(n-1)\theta^2N(N-1)(N-2)} \sum_{i \neq j} \sum_{k=1}^{N} Y_i^2 Y_j Y_k \]

\[ - \frac{4(\gamma_2\theta + 2\gamma_2\theta^2)}{n(\gamma_2 + \theta^2)\theta^2N(N-1)} \sum_{i \neq j} \sum_{k=1}^{N} Y_i^3 Y_j, \]

(3)

**Proof. (See Appendix-A).**

**Theorem 2.2.** An estimator of variance of the estimator \( s_y^2 \) is given by:

\[ \hat{\nu}(s_y^2) = \left(1 - \frac{1}{N}\right) (\hat{\mu}_{40} - s_y^4) + \frac{(\gamma_4 + 4\gamma_3\theta + 4\theta^2\gamma_2 - \gamma_2^2)}{n(\gamma_2 + \theta^2)^2} \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i^4 \]

\[ + \frac{2(\gamma_2^2 + 2\gamma_2\theta^2)}{n^2(n-1)^2\theta^4} \sum_{i \neq j} \sum_{k=1}^{n} \hat{Y}_i^2 \hat{Y}_j^2 \]

\[ + \frac{4\gamma_2}{n^2(n-1)^2\theta^2} \sum_{i \neq j} \sum_{k=1}^{n} \hat{Y}_i^2 \hat{Y}_j \hat{Y}_k \]

\[ - \frac{4(\gamma_3\theta + 2\gamma_2\theta^2)}{n^2(\gamma_2 + \theta^2)^2(n-1)} \sum_{i \neq j} \sum_{k=1}^{n} \hat{Y}_i^3 \hat{Y}_j, \]

(4)

where using the method of moments \( \hat{\mu}_{40} \) is a rescaled estimator of \( \mu_{40} \) based on scrambled responses, and \( \hat{Y}_i, \) for \( t = 1, 2, 3, 4 \) is a rescaled estimator of \( Y_i \) based on scrambled responses. For example, we suggest using \( \hat{Y}_i = \frac{Z_i}{\sqrt{E(S^{(5)})}}, \hat{Y}_i^2 = \frac{Z_i}{\sqrt{E(S^{(5)})}}, \hat{Y}_i^3 = \frac{Z_i}{\sqrt{E(S^{(5)})}}, \hat{Y}_i^4 = \frac{Z_i}{\sqrt{E(S^{(5)})}} \) and \( \hat{\mu}_{40} = \frac{(n-1)^3}{E(S^{(5)})} \sum_{i=1}^{n} (Z_i - \bar{Z})^4 \), etc. The expected values \( E(S^k) \) for \( k = 1, 2, 3, 4 \) can be had from Singh and Chen (2009) by using the concept of higher order moments of scrambling variables.

### 3. Difference Type Estimator of the Finite Population Variance

Following Das and Tripathi (1978) and Isaki (1983), we define a new, difference-type, estimator of the finite population variance, as:

\[ \hat{\sigma}_v^2 = s_y^2 + B(S_v^2 - s_v^2), \]

(5)

where \( B \) is a constant to be determined such that the variance of the estimator \( \hat{\sigma}_v^2 \) is minimum. The proposed estimator \( \hat{\sigma}_v^2 \) is an unbiased estimator of the finite population variance \( S_v^2 \), for any fixed value of \( B \), and so in, particular, is unbiased, and has minimum variance, for the optimal value of \( B \) given by:

\[ B = \frac{\mu_{22} - \mu_{20}\mu_{02}}{\mu_{04} - \mu_{02}^2}. \]

(6)

The minimum variance of the proposed estimator \( \hat{\sigma}_v^2 \) then becomes:

\[ \text{Min.} \text{Var} \left( \hat{\sigma}_v^2 \right) = \left(1 - \frac{1}{N}\right) (\mu_{40} - \mu_{20}^2) \left[ 1 - \frac{(\mu_{22} - \mu_{20}\mu_{02})^2}{(\mu_{40} - \mu_{20}^2)(\mu_{04} - \mu_{02}^2)} \right] \]

\[ + \frac{(\gamma_4 + 4\gamma_3\theta + 4\theta^2\gamma_2 - \gamma_2^2)}{n(\gamma_2 + \theta^2)^2} \frac{1}{N} \sum_{i=1}^{n} Y_i^4 \]

\[ + \frac{2(\gamma_2^2 + 2\gamma_2\theta^2)}{n^2(n-1)^2\theta^4} \frac{1}{N(N-1)} \]
\[ \sum_{i \neq j=1}^{N} Y_i^2 Y_j^2 + \frac{4 \gamma_2 (n-2)}{n(n-1) \theta^2 N(N-1)(N-2)} \sum_{i \neq j \neq k=1}^{N} Y_i Y_j Y_k \]

\[ - \frac{4(\gamma_3 \theta + 2 \gamma_2 \theta^2)}{n(\gamma_2 + \theta^2) \theta^2 N(N-1)} \sum_{i \neq j=1}^{N} Y_i Y_j. \] 

(7)

4. Regression Type Estimator

In practice the value of \( B \) is unknown so the proposed estimator \( \hat{\sigma}^2_y \) becomes difficult to implement in practice. Thus, we suggest a linear regression type estimator \( \hat{\sigma}^2_y \) given by:

\[ \hat{\sigma}^2_y = s^2_y + \hat{B}(S^2_x - s^2_x) \] 

where

\[ \hat{B} = \frac{\hat{\mu}_{22}^* - s^2_y s^2_x}{\hat{\mu}_{04} - s^4_x} \] 

(9)

with

\[ (n-1)s^2_x = \sum_{i=1}^{n} (X_i - \hat{X})^2, \quad (n-1)\hat{\mu}_{04} = \sum_{i=1}^{n} (X_i - \hat{X})^4, \quad s^2_y \]

\[ = \frac{1}{\theta^2} \left[ s^2_z - \frac{\gamma^2}{n(\gamma_2 + \theta^2)} \sum_{i=1}^{n} Z_i^2 \right] \]

\[ \hat{\mu}_{22}^* = \frac{1}{(n-1)(\gamma_2 + \theta^2)} \sum_{i=1}^{n} (Z_i - \hat{Z})^2 (X_i - \hat{X})^2 \] and \( \hat{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \).

The value of \( \hat{B} \) depends upon scrambled responses, thus, it will increase the variance of the linear regression estimator \( \hat{\sigma}^2_y \) compared to \( s^2_y \). It will also make it difficult to find the value of \( V_R(\hat{\sigma}^2_y) \), which would be equal to:

\[ V_R(\hat{\sigma}^2_y) = V_R(s^2_y) + \left( \frac{S^2_x - s^2_x}{\hat{\mu}_{04} - s^4_x} \right)^2 V_R(\hat{\mu}_{22} - s^2_y s^2_x) \]

\[ + 2 \left( \frac{S^2_x - s^2_x}{\hat{\mu}_{04} - s^4_x} \right) \text{Cov}_R(s^2_y, \hat{\mu}_{22} - s^2_y s^2_x) \]

and in fact it makes it difficult to find the exact variance \( V(\hat{\sigma}^2_y) \) of the linear regression type estimator. Thus, in Section 6, we consider comparing the linear regression type estimator \( \hat{\sigma}^2_y \) to the naive unbiased estimator \( s^2_y \) through simulation study.

Remarks. If \( \hat{B} = s^2_y / s^2_x \), then the estimator \( \hat{\sigma}^2_y \) becomes the ratio type estimator given by

\[ \hat{\sigma}_{rat}^2 = s^2_y \left( \frac{S^2_x}{s^2_x} \right) \] 

(10)
If \( \hat{B} = \frac{\alpha s^2}{\alpha s^2 + (1 - \alpha)S^2} \), where \( \alpha \) is a constant, then the estimator \( \hat{\sigma}^2 \) becomes the Das and Tripathi (1978) type estimator given by:

\[
\hat{\sigma}^2_{\text{DT}} = \frac{s^2_x S^2}{\alpha s^2 + (1 - \alpha)S^2}
\]

5. Relative Efficiency of the Difference Estimator

The proposed difference type estimator \( \hat{\sigma}^2 \) will be more efficient than the naive estimator \( s^2_y \) if:

\[
\text{Min.} V \left( \hat{\sigma}^2 \right) < V \left( s^2_y \right)
\]

that is, if

\[
\left( \frac{1}{n} - \frac{1}{N} \right) \left( \mu_{40} - \mu_{20}^2 \right) \left[ 1 - \frac{(\mu_{22} - \mu_{20} \mu_{02})^2}{(\mu_{40} - \mu_{20}^2)(\mu_{04} - \mu_{02}^2)} \right] < \left( \frac{1}{n} - \frac{1}{N} \right) \left( \mu_{40} - \mu_{20}^2 \right)
\]

or equivalently if

\[
\frac{(\mu_{22} - \mu_{20} \mu_{02})^2}{(\mu_{40} - \mu_{20}^2)(\mu_{04} - \mu_{02}^2)} > 0
\]

which is always true. Thus, the proposed estimator \( \hat{\sigma}^2 \) is always more efficient than the naive estimator \( s^2_y \).

6. Simulation Study for the Regression Estimator

In this section, we consider the case where the sensitive variable \( Y \) and the auxiliary variable \( X \) are related to each other by the linear model defined as:

\[
Y_i = RX_i + e_i X_i^g
\]

where \( e_i \sim N(0, 1) \). The auxiliary variable \( X_i \sim G(a, b) \) is generated from the gamma distribution with parameters \( a = 2.2 \) and \( b = 3.5 \). We generate a population of size \( N = 3000 \) units from the model for a given value of \( g \) and \( R \). Then from the given population of size \( N = 3,000 \), we select an SRSWOR sample of size \( n \) and both the study variable \( Y_i \) and the auxiliary variable \( X_i \) are observed for \( i = 1, 2, 3, ..., n \). Next we generate \( n \) values of a scrambling variable \( S_i, i = 1, 2, ..., n \) from the beta distribution with a given choice of \( \alpha \) and \( \beta \). In other words, we assumed the scrambling variable \( S \sim B(\alpha, \beta) \). Then, we obtained the scrambled responses on the study variable as \( Z_i = Y_i S_i, i = 1, 2, ..., n \) from a given sample. We repeated this process \( T = 5,000 \) times. By using information from the ordered pairs \( (X_i, Z_i), i = 1, 2, ..., n, \) of the \( t \)th sample, \( t = 1, 2, ..., T \), we computed the three estimators \( \hat{\theta}_{0t} = s^2_{y^2} \), \( \hat{\theta}_{1t} = \hat{\sigma}^2_{y^2} \) and \( \hat{\theta}_{2t} = \hat{\sigma}^2_{lr} \). Note that we used the first four moments of the scrambling variable \( S \) given by:

\[
E(S) = \theta = \frac{\alpha}{\alpha + \beta}
\]
\[ V(S) = \gamma_2 = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \]  
\[ \gamma_3 = E(S - \theta)^3 = \gamma_2^{3/2} \left[ \frac{2(\beta - \alpha)\sqrt{\alpha + \beta + 1}}{\sqrt{\alpha\beta(\alpha + \beta + 2)}} \right] \]  

and
\[ \gamma_4 = E(S - \theta)^4 = \gamma_2^{2} \left[ \frac{6\{\alpha^3 - \alpha^2(2\beta - 1) + \beta^2(\beta + 1) - 2\alpha\beta(\beta + 2)\}}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)} + 3 \right] \]

No doubt theoretically the estimators \( \hat{\theta}_{0|t} \) and \( \hat{\theta}_{1|t} \) are unbiased estimators of the parameter of interest \( S^2_y \), while the estimator \( \hat{\theta}_{2|t} \) is a biased estimator. In order to see the performance of the proposed estimators, we computed the simulated relative bias in each of the three estimators as follows:

\[ RB(\hat{\theta}_j) = \frac{1}{T} \sum_{t=1}^{T} \left( \hat{\theta}_{j|t} - S^2_y \right) \times 100\% \]  

We also computed the relative efficiencies of the linear regression and the difference type estimator with respect to the usual estimator as:

\[ RE(0, j) = \frac{\sum_{t=1}^{T} \left( \hat{\theta}_{0|t} - S^2_y \right)^2}{\sum_{t=1}^{T} \left( \hat{\theta}_{j|t} - S^2_y \right)^2} \times 100\%, \text{ for } j = 1, 2. \]  

The results obtained are presented in Table 1. The FORTRAN codes used in this simulation are given in Appendix B.

In the simulation study, we fixed \( \alpha = 6.5, \beta = 0.5, \) and \( g = 0.5 \); two values of \( R = 0.5 \) and 1.5 were considered. We also used different values of the sample size \( n \) ranging between 10 and 100 with a step of 5 units. An absolute value of the percent relative bias less than 10% is regarded as acceptable by following Cochran (1977). The values of \( RB(\hat{\theta}_0), RB(\hat{\theta}_1), \) and \( RB(\hat{\theta}_2) \) fall within this range in all the cases considered in the simulation study and so relative bias is considered negligible. The value of \( RE(0, 2) \) is lower than the value of \( RE(0, 1) \) in all situations. For \( R = 0.5 \), the average \( RE(0, 1) \) value is 175.51% with a standard deviation of 1.57%, and with a minimum value of 171.70%, maximum of 178.10%, and median value 175.70%. For \( R = 1.5 \), the average \( RE(0, 1) \) value is 476.76% with a standard deviation of 4.40%, and with a minimum value of 465.50%, maximum of 485.10%, and median value 476.10%. For \( R = 0.5 \), the average \( RE(0, 2) \) value is 157.16% with a standard deviation of 16.35%, and with a minimum value of 105.70%, maximum of 170.50%, and median value 163.90%. For \( R = 1.5 \), the average \( RE(0, 2) \) value is 414.15% with a standard deviation of 28.68%, and with a minimum value of 328.40%, maximum of 436.70%, and median value 425.90%. Fig. 1 shows that the value of the simulated percent relative efficiency of the difference and regression estimators with respect to the usual unbiased estimator is almost free from the sample size.
Randomized Response Sampling

Table 1
Comparison of three estimators of the finite population variance

| $R$ | $n$ | $RB(\hat{\theta}_0)$ | $RB(\hat{\theta}_1)$ | $RB(\hat{\theta}_2)$ | $RE(0, 1)$ | $RE(0, 2)$ |
|-----|-----|----------------------|----------------------|----------------------|-------------|-------------|
| 0.5 | 10  | 0.732                | 0.475                | -2.266               | 171.8       | 105.7       |
|     | 15  | -0.202               | -0.222               | -1.203               | 174.7       | 131.7       |
|     | 20  | -0.005               | 0.042                | -0.696               | 177.9       | 141.6       |
|     | 25  | -0.148               | -0.197               | -0.842               | 176.6       | 146.9       |
|     | 30  | 0.067                | 0.075                | -0.652               | 174.1       | 153.1       |
|     | 35  | -0.055               | 0.009                | -0.683               | 174.1       | 152.4       |
|     | 40  | 0.105                | 0.152                | -0.835               | 175.7       | 161.0       |
|     | 45  | 0.262                | 0.115                | -0.653               | 175.1       | 160.8       |
|     | 50  | 0.245                | 0.237                | -0.554               | 174.3       | 160.3       |
|     | 55  | -0.146               | -0.139               | -1.049               | 174.2       | 163.9       |
|     | 60  | -0.087               | -0.143               | -0.914               | 176.0       | 164.4       |
|     | 65  | 0.030                | 0.038                | -0.722               | 174.2       | 164.3       |
|     | 70  | 0.173                | 0.133                | -0.755               | 176.9       | 170.5       |
|     | 75  | -0.073               | -0.058               | -0.845               | 177.6       | 166.4       |
|     | 80  | -0.080               | -0.072               | -0.827               | 176.0       | 165.1       |
|     | 85  | -0.062               | -0.085               | -0.861               | 178.1       | 170.5       |
|     | 90  | 0.127                | 0.113                | -0.647               | 176.0       | 169.0       |
|     | 95  | 0.093                | 0.113                | -0.641               | 175.7       | 168.1       |
|     | 100 | -0.063               | -0.029               | -0.751               | 175.7       | 170.4       |
| 1.5 | 10  | 0.125                | 0.131                | 1.565                | 480.9       | 328.4       |
|     | 15  | 0.378                | 0.037                | 1.730                | 465.5       | 361.2       |
|     | 20  | 0.146                | 0.012                | 1.623                | 476.3       | 385.8       |
|     | 25  | 0.147                | 0.121                | 1.489                | 485.1       | 401.1       |
|     | 30  | 0.141                | 0.132                | 1.511                | 474.6       | 398.9       |
|     | 35  | 0.117                | 0.067                | 1.113                | 476.1       | 415.5       |
|     | 40  | 0.124                | -0.049               | 1.021                | 474.2       | 414.7       |
|     | 45  | 0.016                | -0.034               | 0.861                | 474.5       | 419.3       |
|     | 50  | -0.128               | 0.026                | 0.923                | 480.6       | 423.8       |
|     | 55  | -0.150               | 0.086                | 0.890                | 482.3       | 425.9       |
|     | 60  | -0.052               | 0.030                | 0.736                | 480.3       | 434.5       |
|     | 65  | -0.093               | 0.012                | 0.663                | 478.1       | 429.1       |
|     | 70  | -0.033               | 0.015                | 0.661                | 479.0       | 432.6       |
|     | 75  | 0.138                | 0.047                | 0.640                | 479.5       | 433.5       |
|     | 80  | -0.097               | 0.043                | 0.579                | 474.5       | 428.0       |
|     | 85  | 0.171                | 0.057                | 0.581                | 476.0       | 436.1       |
|     | 90  | -0.185               | 0.063                | 0.571                | 471.1       | 428.8       |
|     | 95  | -0.243               | -0.010               | 0.449                | 475.3       | 435.0       |
|     | 100 | 0.280                | 0.035                | 0.424                | 474.5       | 436.7       |

The Fig. 2 shows that the percent relative bias in the proposed estimators are between −2% to +2% for two different values of $R = 0.5$ and 1.5 and sample size between 10 to 100.

In the next section, we consider a simulation study based on a real dataset given in the book by Rosner (2006).
7. Application Based on a Real Dataset

We used the dataset, FEV.DAT, available on the CD that accompanies the text by Rosner (2006) and that contains data on 654 children from the Childhood Respiratory Disease Study done in Boston. Among the variables are age and height. While age is not usually a sensitive characteristic, for illustration purpose, we will consider it sensitive and take height as the nonsensitive characteristic. It should be noted that any variable could be a sensitive variable depending on a situation being considered as reported in Singh et al. (2008). Here, we consider the problem of estimating the variance of age of the population utilizing height

![Figure 1. Percent relative efficiency versus sample size.](image)

![Figure 2. Percent relative biases versus sample sizes.](image)
at the estimation stage. Although exact date of birth a child is available in hospitals, but still sometimes parents do not want to disclose it in public domain because when a child will grew up then all his/her credit history will be related to the exact date of birth. It is also very common among girls that most the girls like to hide their age because girls feel it is sensitive information for them telling others they are not young. The rest of the procedure of the simulation study has been kept the same except that the synthetic data considered in Section 6 is replaced by the real dataset, and the sample size was changed from 10 to 70. The results obtained are presented in Table 2.

For $n = 10$ and 15, the linear regression estimator shows percent relative efficiency less than 100%, but for sample size more than 20 both the linear regression estimator and the difference estimators are performing very well in comparison to the proposed naive estimator. The percent relative bias values remain negligible. For $n > 20$, the average percent RE(0, 1) value is 121.50% with a standard deviation of 0.86%; the minimum value is 119.50% and maximum value is 122.90% with median value of 121.70%. In the same way, for $n > 20$, the average percent RE(0, 2) value is 114.70% with a standard deviation of 3.70%; the minimum value is 106.40% and maximum value is 118.30% with median value of 115.20%.

8. Complex Survey Design

Let a sample $s$ of size $n$ be selected from a population by using an arbitrary sampling design $P(s)$ with inclusion probabilities $\pi_i$ for the $i$th unit and $\pi_{ij}$ for the $i$th and $j$th unit $i \neq j = 1, 2, ..., N$. Assuming $\pi_{ij}$’s are positive for all $i \neq j$, an unbiased estimator of the population variance is given by:

$$S_z^2 = \frac{1}{N(N - 1)\theta^2} \sum_{i \neq j} \frac{Z_i Z_j}{\pi_{ij}} - \frac{1}{N(N - 1)\theta^2} \sum_{i \neq j} \sum_{j \in s} \frac{Z_i Z_j}{\pi_{ij}}$$

(20)
A Das and Tripathi (1978) and Isaki (1983) type variance estimator for the complex surveys design can be obtained as:

$$S^2_{\xi} = S^2_{\hat{\xi}} + B^* (S^2_{\xi} - \hat{S}^2_{\xi})$$

where $\hat{S}^2_{\xi} = \frac{1}{N} \sum_{i \in s} \frac{k_i}{\pi_i} - \frac{1}{N(N-1)} \sum_{i \neq j \in s} \frac{k_i k_j}{\pi_i \pi_j}$ and $B^*$ is a suitably chosen constant.

9. Conclusion

In this article, three estimators of finite population variance based on scrambled responses and in the presence of an auxiliary variable are investigated through extensive simulation study. We conclude that both the difference and the linear regression type estimators of the finite population variance, based on scrambled responses, lead to quite satisfactory results. We acknowledge that a simulation study is a must when carrying out a real survey, both for making a choice of a scrambling variable and for determining a minimum sample size.

Acknowledgments

The authors are thankful to the Editor in Chief Professor Dr. N. Balakrishnan, ADM: Debbie Iscoe and three learned referees to bring the original manuscript in the present form.

Appendix A: Derivation of Variance of the Proposed Estimator

Proof of Theorem 2.1: Let $E_P$ and $V_P$ denote the expected value and variance respectively over all possible samples. Similarly let $E_R$ and $V_R$ denote the expected value and variance respectively over the randomization device. Then the variance of $s^2_{\gamma}$ over the randomization device is given by:

$$V_R(s^2_{\gamma}) = \frac{(\gamma_4 + 4\gamma_3\theta + 4\theta^2\gamma_2 - \gamma_2^2)}{n^2(\gamma_2 + \theta^2)^2} \sum_{i=1}^{n} Y_i^4 + \frac{2(\gamma_2 + 2\gamma_2\theta^2)}{n^2(n-1)^2\theta^4} \sum_{i \neq j \neq k \neq l} Y_i^2 Y_j Y_k Y_l
\hspace{1cm} + \frac{4\gamma_2}{n^2(n-1)^2\theta^2} \sum_{i \neq j \neq k} Y_i^2 Y_j Y_k - \frac{4(\gamma_3\theta + 2\gamma_2\theta^2)}{n^2(n-1)(\gamma_2 + \theta^2)^2} \sum_{i \neq j} Y_i^3 Y_j$$

(A.1)

The expected value of $V_R(s^2_{\gamma})$ over the sampling design $P(s)$ is given by:

$$E_P[V_R(s^2_{\gamma})] = \frac{(\gamma_4 + 4\gamma_3\theta + 4\theta^2\gamma_2 - \gamma_2^2)}{n(\gamma_2 + \theta^2)^2} \frac{1}{N} \sum_{i=1}^{N} Y_i^4 + \frac{2(\gamma_2 + 2\gamma_2\theta^2)}{n(n-1)^2\theta^4} \frac{1}{N(N-1)}
\hspace{1cm} \times \sum_{i \neq j} \sum_{j=1}^{N} Y_i^2 Y_j^2 + \frac{4\gamma_2(n - 2)}{n(n-1)^2 \theta^2 N(N-1)(N-2)} \sum_{i \neq j} \sum_{j=1}^{N} Y_i^2 Y_j Y_k
\hspace{1cm} - \frac{4(\gamma_3\theta + 2\gamma_2\theta^2)}{n(\gamma_2 + \theta^2)^2 \theta^2 N(N-1)} \sum_{i \neq j} \sum_{j=1}^{N} Y_i^3 Y_j$$

(A.2)
Now, we have the theorem with the fact that:

\[ V (s_y^2) = V_P E_R (s_y^2) + E_P V_R (s_y^2) = V_P (s_y^2) + E_P V_R (s_y^2). \]

**Appendix B: FORTRAN Codes Used in the Simulation Study**

```
!PROGRAM: ARNAB5.F95 (USED FOR SIMULATION)
USE NUMERICAL_LIBRARIES
IMPLICIT NONE
INTEGER I, J, NP, NS, IR(5000), IIII, NITR, KOUT1
REAL E(5000), XP(5000), YP(5000)
REAL TH1, TH2, G, R, ANP, SUMYP, SUMXP, YMP, XMP
REAL VARYP, VARXP, AM22P, AM04P, YS(5000), XS(5000)
REAL A, B, TH, G2, S(5000), ANS, SMEAN, VARS, SUMS
REAL Z(5000), SZ, SZ2, ZM, VARZ, SYSTAR2
REAL BV0, RBV0, BTRUE, SUMXS, XSM, VARXS, BV1
REAL RBV1, V1, AM04S, AM22S, V2, RBV2, BV2, BEST
REAL AMSEV0, AMSEV1, AMSEV2
REAL RE01, RE02
CHARACTER*20 OUT_FILE
WRITE(*, '(A)') 'NAME OF THE OUTPUT FILE'
READ(*, '(A20)') OUT_FILE
OPEN(42, FILE = OUT_FILE, STATUS = 'UNKNOWN')
NP = 3000
G = 0.5
R = 1.5
TH1 = 2.2
TH2 = 3.5
CALL RNGAM(NP, TH1, XP)
CALL SSCAL(NP, TH2, XP, 1)
CALL RNNOR(NP, E)
DO 11 I = 1, NP
  11 YP(I) = R*XP(I) + E(I)*XP(I)**G
! 12 I = 1, NP
!12 WRITE(42, 101)I, YP(I), XP(I)
!101 FORMAT(2X, I4, 2X, F9.2, 2X, F9.2)
ANP = NP
SUMYP = 0.0
SUMXP = 0.0
DO 121 I = 1, NP
  121 SUMYP = SUMYP + YP(I)
  SUMXP = SUMXP + XP(I)
YMP = SUMYP/ANP
XMP = SUMXP/ANP
! WRITE(42, 102)YMP, XMP
! 102 FORMAT(2X, F9.3, 2X, F9.3)
VARYP = 0.0
```
VARXP = 0.0
AM22P = 0.0
AM04P = 0.0
DO 1 I = 1, NP
AM22P = AM22P + (YP(I)-YMP)**2*(XP(I)-XMP)**2
AM04P = AM04P + (XP(I)-XMP)**4
VARYP = VARYP + (YP(I)-YMP)**2
1 VARXP = VARXP + (XP(I)-XMP)**2
AM22P = AM22P/(ANP-1)
AM04P = AM04P/(ANP-1)
VARYP = VARYP/(ANP-1)
VARXP = VARXP/(ANP-1)
BTRUE = (AM22P-VARYP**VARXP)/(AM04P-VARXP**2)
! WRITE(42, 103)AM22P, AM04P, VARYP, VARXP, BTRUE
! 103 FORMAT(2X, F12.3, 2X, F9.3, 2X, F9.3, 2X, F9.3)
DO 7777 NS = 10, 170, 5
ANS = NS
BV1 = 0.0
BV2 = 0.0
BV0 = 0.0
KOUT1 = 0
AMSEV0 = 0.0
AMSEV1 = 0.0
AMSEV2 = 0.0
NITR = 50000
DO 9999 IIII = 1, NITR
CALL RNSRI(NS, NP, IR)
DO 13 I = 1, NS
YS(I) = YP(IR(I))
13 XS(I) = XP(IR(I))
SUMXS = 0.0
DO 41 I = 1, NS
41 SUMXS = SUMXS + XS(I)
XSM = SUMXS/ANS
VARXS = 0.0
DO 42 I = 1, NS
42 VARXS = VARXS + (XS(I)-XSM)**2
VARXS = VARXS/(ANS-1)
! DO 14 I = 1, NS
! 14 WRITE(42, 104)I, YS(I), XS(I)
! 104 FORMAT(5X, I4, 2X, F9.3, 2X, F9.3)
A = 6.5
B = 0.5
TH = A/(A+B)
G2 = (A*B)/((A+B)**2*(A+B+1))
CALL RNBEET(NS, A, B, S)
SUMS = 0.0
DO 15 I = 1, NS
15 SUMS = SUMS+S(I)
SMEAN = SUMS/ANS
VARS = 0.0
DO 16 I = 1, NS
16 VARS = VARS+(S(I)-SMEAN)**2
       VARS = VARS/(ANS-1)
! WRITE(42, 106) TH, G2, SMEAN, VARS
! 106 FORMAT(2X, F8.2, 1X, F8.4, 1X, F8.2, 1X, F8.4)
       DO 17 I = 1, NS
17 Z(I) = YS(I)*S(I)
       SZ = 0.0
       SZ2 = 0.0
       DO 18 I = 1, NS
18 SZ = SZ + Z(I)
       SZ2 = SZ2 + Z(I)**2
       ZM = SZ/ANS
       VARZ = 0.0
       DO 19 I = 1, NS
19 VARZ = VARZ + (Z(I)-ZM)**2
       VARZ = VARZ/(ANS-1)
       SYSTAR2 = (VARZ-G2*SZ2/(ANS*(G2+TH**2)))/(TH**2)
       V1 = SYSTAR2 + BTRUE*(VARXP-VARXS)
       BV0 = BV0 + SYSTAR2
       BV1 = BV1 + V1
! WRITE(42, *) SYSTAR2, V1, BTRUE, VARXP, VARXS
! REGRESSION TYPE ESTIMATOR
       AM04S = 0.0
       AM22S = 0.0
       DO 51 I = 1, NS
51 AM04S = AM04S + (XS(I)-XSM)**4
       AM04S = AM04S/(ANS-1)
       AM22S = AM22S/((ANS-1))
       AM22S = AM22S/(G2+TH**2)
       BEST = (AM22S-SYSTAR2*VARXS)/(AM04S-VARXS**2)
       V2 = SYSTAR2 + BEST*(VARXP-VARXS)
       BV2 = BV2 + V2
       AMSEV0 = AMSEV0 + (SYSTAR2-VARYP)**2
       AMSEV1 = AMSEV1 + (V1-VARYP)**2
       AMSEV2 = AMSEV2 + (V2-VARYP)**2
9999 CONTINUE
       RBV0 = (BV0/DBLE(NITR)-VARYP)/VARYP*100
       RBV1 = (BV1/DBLE(NITR)-VARYP)/VARYP*100
       RBV2 = (BV2/DBLE(NITR)-VARYP)/VARYP*100
       RE01 = AMSEV0*100/AMSEV1
       RE02 = AMSEV0*100/AMSEV2
       WRITE (*, 109) A, B, R, G, NS, RBV0, RBV1, RBV2, RE01, RE02
       WRITE (42, 109) A, B, R, G, NS, RBV0, RBV1, RBV2, RE01, RE02
109 FORMAT(1X, F6.3, 1X, F6.3, 1X, F6.3, 1X, F6.3, 1X, F6.3, 1X, F6.3, 1X, F6.3, 1X, I5, 1X, F7.4, 11X, F7.4, 1X, F7.4, 1X, F7.2, 1X, F7.2)
References

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton, FL: A Chapman & Hall, CRC Press.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York, NY: John Wiley and Sons.

Das, A. K., Tripathi, T. P. (1978). Use of auxiliary information in estimating finite population variance. *Sankhyā, C*, 40:139–148.

Diana, G., Perri, P. F. (2009). Estimating a sensitive proportion through randomized response procedures based on auxiliary information. *Statistical Papers* 50:661–672.

Diana, G., Perri, P. F. (2010). New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics* 37:1875–1890.

Diana, G., Perri, P. F. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers* 52:633–650.

Diana, G., Perri, P. F. (2012). A calibration-based approach to sensitive data: A simulation study. *Journal of Applied Statistics* 39:53–65.

Eichhorn, B. H., Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference* 7:307–316.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association* 66:243–250.

Horvitz, D. G., Shah, B. V., Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section ASA*:65–72.

Isaki, C. T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association* 78:117–123.

Mahajan, P. K. (2005–2006). Optimum stratification for scrambled response with ratio and regression methods of estimation. *Model Assisted Statistics and Applications* 1(1):17–22.

Mahajan, P. K., Sharma, P., Gupta, R. K. (2007). Optimum stratification for allocation proportional to strata totals for scrambled response. *Model Assisted Statistics and Applications* 2(2):81–88.

Perri, P. F. (2008). Modified randomized devices for Simmons’ model. *Model Assisted Statistics and Applications* 3(3):233–239.

Rosner, B. (2006). *Fundamentals of Biostatistics*. 6th ed. Belmont, CA: Thompson: Brooks/Cole.

Ryu, J. B., Kim, J. M., Hoe, T. Y., Park, C. G. (2006/2006). On stratified randomized response sampling. *Model Assisted Statistics and Applications* 1(1):31–36.

Sidhu, S. S., Bansal, M. L. (2008). Estimator of population total using Rao, Hartley and Cochran’s scheme using optimal randomized response technique in multi-character surveys. *Model Assisted Statistics and Applications* 3(3):259–267.

Singh, S., Chen, C. (2009). Utilization of higher order moments of scrambling variables in randomized response sampling. *Journal of Statistical Planning and Inference* 139:3377–3380.

Singh, S., Joarder, A. H., King, M. L. (1996). Regression analysis using scrambled responses. *Australian Journal of Statistics* 38:201–211.

Singh, S., Kim, J.-M. (2011). Pseudo empirical log-likelihood estimator using scrambled responses. *Statistics and Probability Letters* 81:345–351.

Singh, S., Kim, J.-M., Grewal, I. S. (2008). Imputing and jackknifing scrambled responses. *Metron* LXVI(2):183–204.

Son, C.-K., Hong, K.-H., Lee, G.-S., Kim, J.-M. (2008). The calibration of stratified randomized response estimators. *Communications of the Korean Statistical Society* 14:597–603.

Strachan, R., King, M. L., Singh, S. (1998). Likelihood-based estimation of the general model with scrambled responses. *Australian and New Zealand Journal of Statistics* 40:279–290.
Tracy, D. S., Singh, S. (1999). Calibration estimators in randomized response survey. *Metron* LVII:47–68.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60:63–69.

Zaizai, Y., Jingyu, W., Junfeng, L., Hua, W. (2008). Ratio imputation method for handling item-nonresponse in Eichhorn model. *Model Assisted Statistics and Applications* 3(2):89–98.