Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Improved exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information

Sat Gupta, Javid Shabbir, Rita Sousa & Pedro Corte-Real

IMPROVED EXPONENTIAL TYPE ESTIMATORS OF THE MEAN OF A SENSITIVE VARIABLE IN THE PRESENCE OF NON-SENSITIVE AUXILIARY INFORMATION

[1] Sat Gupta, [2] Javid Shabbir, [3] Rita Sousa and [4] Pedro Corte-Real

[1]University of North Carolina at Greensboro (sngupta@uncg.edu)

[2] Quaid-i-Azam University, Islamabad, Pakistan (jsqau@yahoo.com)

[3] New University of Lisbon & Statistics Portugal (rita.sousa@ine.pt)

[4] New University of Lisbon (parcr@fct.unl.pt)

Abstract

Recently Koyuncu et al. (2013) proposed an exponential type estimator to improve the efficiency of mean estimator based on RRT. In this paper, we propose an improved exponential type estimator which is more efficient than the Koyuncu et al. (2013) estimator, which in turn was shown to be more efficient than the usual mean estimator, ratio estimator, regression estimator, and the Gupta et al. (2012) estimator. Under simple random sampling without replacement (SRSWOR) scheme, bias and mean square error (MSE) expressions for the proposed estimator are obtained up to first order of approximation and comparisons are made with the Koyuncu et al. (2013) estimator. A simulation study is used to observe the performances of these two estimators. Theoretical findings are also supported by a numerical example with real data. We also show how to extend the proposed estimator to the case when more than one auxiliary variable is available.

Key-Words: Randomized response technique (RRT), Exponential estimator, Auxiliary variable, Bias, Mean square error (MSE).

 ACCEPTED MANUSCRIPT

Mathematics Subject Classification 62D05.

Communicating authors: Sat Gupta, University of North Carolina at Greensboro, sngupta@uncg.edu

1. Introduction

This study proposes an improved exponential type estimator for estimating the population mean of a sensitive variable when information about a non-sensitive auxiliary variable is available. A common problem in conducting a statistical sample survey is that of response bias in the face of sensitive questions. Warner (1965) introduced the Randomized Response Technique (*RRT*) in order to solve this problem. Our main purpose in this study is to improve the mean estimation of a sensitive variable based on a *RRT* model when one or more non-sensitive auxiliary variables are available.

Many authors such as Kadilar and Cingi (2004), Kadilar et al. (2007), Shabbir and Gupta (2007, 2010) and Nangsue (2009) have presented ratio and regression estimators when both the study variable and the auxiliary variable are directly observable.

In this study we propose an exponential type estimator for the mean of a sensitive variable using known information on a correlated but non-sensitive auxiliary variable. The proposed estimator performs better than the recently introduced estimator by Koyuncu et al. (2013) which was shown to outperform many existing estimators of this type. We also extend the proposed estimator to the case when more than one auxiliary variable is available.

2. Terminology

Consider a finite population with N units $U = (U_1, U_2, \ldots, U_N)$ from which a sample of size n is drawn using simple random sampling without replacement (SRSWOR). Let Y be the study variable, a sensitive variable which cannot be observed directly due to respondent bias. Let X be the non-sensitive auxiliary variable which is correlated with Y. Let S be a scrambling variable independent of Y and X. The respondent is asked to report a scrambled response for Y given by Z

= Y + S but is asked to provide a true response for X. Let $(\bar{y}, \bar{x})$ be the sample means corresponding to $(\bar{Y}, \bar{X})$, the population means of Y and X, respectively. Consider $\bar{Z}$ to be the population mean of the scrambled variable Z.

Let S_x^2 and s_x^2 respectively be the population variance and the sample variance of X. On the other hand, S_{zx}^2 and s_{zx}^2 are the population covariance and the sample covariance respectively between Z and X.

To obtain the *Bias* and *MSE* expressions for the proposed estimator, let us define

$$e_0 = (\bar{z} - \bar{Z})/\bar{Z}\,,\ e_1 = (\bar{x} - \bar{X})/\bar{X}\,,\ e_2 = (s_x^2 - S_x^2)/S_x^2, \text{ and } e_3 = (s_{zx}^2 - S_{zx}^2)/S_{zx}^2,$$

such that $E(e_i) = 0 \;\; i = 1,\ 2,\ 3$. To first degree of approximations, we have

$$E(e_0^2) = \lambda C_z^2 = v_{20},\ E(e_1^2) = \lambda C_x^2 = v_{02},\ E(e_0 e_1) = \lambda C_{zx} = \lambda \rho_{zx} C_z C_x = v_{11},$$

$$E(e_1 e_2) = \lambda \frac{\mu_{03}}{\bar{X}\mu_{02}}, \text{ and } E(e_1 e_3) = \lambda \frac{\mu_{12}}{\bar{X}\mu_{11}},$$

where $\lambda = \dfrac{1-f}{n}$, $f = n/N$, $C_{zx} = \rho_{zx} C_z C_x$ and $\mu_{rs} = \dfrac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{Z})^r (x_i - \bar{X})^s$.

3. Difference-cum-exponential Estimator (Koyuncu et al., 2013)

Recently Koyuncu et al. (2013) have suggested a combination of the difference estimator and the exponential estimator with some gain in the efficiency. This estimator is given by

$$\hat{\mu}_{DE} = [w_1 \bar{z} + w_2 (\bar{X} - \bar{x})] \exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right), \tag{3.1}$$

where w_1 and w_2 are constants.

The *Bias* and *MSE* of $\hat{\mu}_{DE}$, up to first degree of approximation, at optimum coefficient values

$$w_{1(opt)} = \frac{1-(\lambda C_x^2/8)}{1+\lambda C_z^2(1-\rho_{zx}^2)} \text{ and } w_{2(opt)} = \frac{\bar{Y}}{\bar{X}}\left\{\frac{1}{2} - w_{1(opt)}\left(1-\frac{\rho_{zx}C_z}{C_x}\right)\right\}$$

are given by

$$Bias(\hat{\mu}_{DE})_{opt} \cong (w_{1(opt)}-1)\bar{Y} + w_{1(opt)}\bar{Y}\lambda\left\{\frac{3}{8}C_x^2 - \frac{1}{2}\rho_{zx}C_zC_x\right\} + w_{2(opt)}\bar{X}\lambda C_x^2 \tag{3.2}$$

and

$$MSE(\hat{\mu}_{DE})_{opt} \cong \bar{Y}^2\left\{\left(1-\frac{1}{4}\lambda C_x^2\right) - \frac{\left(1-\frac{1}{8}\lambda C_x^2\right)^2}{1+\lambda C_z^2(1-\rho_{zx}^2)}\right\}, \text{ or}$$

$$MSE(\hat{\mu}_{DE})_{opt} \cong \bar{Y}^2\left\{\left(1-\frac{1}{4}v_{02}\right) - \frac{v_{02}\left(8-v_{02}\right)^2}{64\left(v_{02}+v_{20}v_{02}-v_{11}^2\right)}\right\}. \tag{3.3}$$

It is shown in Koyuncu et al. (2013) that this estimator is better than other similar estimators such as Sousa et al. (2010) and Gupta et al. (2012).

Koyuncu et al. (2013) also extended the estimator defined in (3.1) to the case when two non-sensitive auxiliary variables are available, as follows

$$\hat{\mu}_{DE2} = \left[w_1\bar{z} + w_2(\bar{X}_1 - \bar{x}_1) + w_3(\bar{X}_2 - \bar{x}_2)\right]\exp\left(\frac{\left(\bar{X}_1 - \bar{x}_1\right)+\left(\bar{X}_2 - \bar{x}_2\right)}{\left(\bar{X}_1 + \bar{x}_1\right)+\left(\bar{X}_2 + \bar{x}_2\right)}\right). \tag{3.4}$$

Koyuncu et al. (2013) provide optimal values of w_1, w_2 and w_3 which minimize the *MSE* of $\hat{\mu}_{DE2}$. The *Bias* and *MSE* for this extended estimator ($\hat{\mu}_{DE2}$), up to first degree of approximation, at optimum coefficient values

$$w_{1(opt)}=\frac{1}{2D}\frac{\begin{array}{l}A\left(D\lambda S_{x1x2}-FG\right)^2+\left(BDG+CDF+2AFG\right)\left(D\lambda S_{x1x2}-FG\right)\\-G\left(CD+AG\right)\left(D\lambda S_{x1}^2-F^2\right)-F\left(AF+BD\right)\left(D\lambda S_{x2}^2-G^2\right)\\-A\left(D\lambda S_{x1}^2-F^2\right)\left(D\lambda S_{x2}^2-G^2\right)\end{array}}{\left(D\lambda S_{x1}^2-F^2\right)\left(D\lambda S_{x2}^2-G^2\right)-\left(D\lambda S_{x1x2}-FG\right)^2},$$

$$w_{2(opt)}=\frac{1}{2}\frac{\left(AF+BD\right)\left(D\lambda S_{x2}^2-G^2\right)-\left(AG+CD\right)\left(D\lambda S_{x1x2}-FG\right)}{\left(D\lambda S_{x1}^2-F^2\right)\left(D\lambda S_{x2}^2-G^2\right)-\left(D\lambda S_{x1x2}-FG\right)^2},$$

and $w_{3(opt)}=\frac{1}{2}\frac{\left(AG+CD\right)\left(D\lambda S_{x1}^2-F^2\right)-\left(AF+BD\right)\left(D\lambda S_{x1x2}-FG\right)}{\left(D\lambda S_{x1}^2-F^2\right)\left(D\lambda S_{x2}^2-G^2\right)-\left(D\lambda S_{x1x2}-FG\right)^2}$

are given by

$$\begin{aligned}Bias\left(\hat{\mu}_{DE2}\right)_{opt}&=\Bigg\{\left(w_{1(opt)}-1\right)\bar{Z}\\&+\frac{w_{1(opt)}\lambda\bar{Z}}{2\left(\bar{X}_1+\bar{X}_2\right)}\left(-\bar{X}_1C_{zx1}-\bar{X}_2C_{zx2}+\frac{3\bar{X}_1^2}{4\left(\bar{X}_1+\bar{X}_2\right)}C_{x1}^2+\frac{3\bar{X}_2^2}{4\left(\bar{X}_1+\bar{X}_2\right)}C_{x2}^2+\frac{3\bar{X}_1\bar{X}_2}{2\left(\bar{X}_1+\bar{X}_2\right)}C_{x1x2}\right)\\&+\frac{w_{2(opt)}\lambda\bar{X}_1}{2\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_1C_{x1}^2+\bar{X}_2C_{x1x2}\right)+\frac{w_{3(opt)}\bar{X}_2}{2\left(\bar{X}_1+\bar{X}_2\right)}\lambda\left(\bar{X}_1C_{x1x2}+\bar{X}_2C_{x2}^2\right)\Bigg\}\end{aligned}\tag{3.5}$$

and

$$MSE\left(\hat{\mu}_{DE2}\right)_{opt}=\bar{Z}^2-\frac{A^2}{4D}-\frac{1}{4D}\frac{\begin{array}{l}\left(AG+CD\right)^2\left(D\lambda S_{x1}^2-F^2\right)+\left(AF+BD\right)^2\left(D\lambda S_{x2}^2-G^2\right)\\-2\left(AG+CD\right)\left(AF+BD\right)\left(D\lambda S_{x1x2}-FG\right)\end{array}}{\left(D\lambda S_{x1}^2-F^2\right)\left(D\lambda S_{x2}^2-G^2\right)-\left(D\lambda S_{x1x2}-GF\right)^2},\tag{3.6}$$

where

$$A=\bar{Z}^2\left(-2+\lambda\left\{\frac{\bar{X}_1C_{zx1}}{\left(\bar{X}_1+\bar{X}_2\right)}+\frac{\bar{X}_2C_{zx2}}{\left(\bar{X}_1+\bar{X}_2\right)}-\frac{3\bar{X}_1^2C_{x1}^2}{4\left(\bar{X}_1+\bar{X}_2\right)^2}-\frac{6\bar{X}_1\bar{X}_2C_{x1x2}}{4\left(\bar{X}_1+\bar{X}_2\right)^2}-\frac{3\bar{X}_2^2C_{x2}^2}{4\left(\bar{X}_1+\bar{X}_2\right)^2}\right\}\right),$$

$$B=\lambda\frac{\bar{Z}}{\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_1^2C_{x1}^2+\bar{X}_1\bar{X}_2C_{x1x2}\right),$$

$$C=\lambda\frac{\bar{Z}}{\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_2^2C_{x2}^2+\bar{X}_1\bar{X}_2C_{x1x2}\right),$$

$$D=\left(\bar{Z}^2+\lambda\left(\bar{Z}^2C_z^2+\frac{\bar{X}_1^2\bar{Z}^2C_{x1}^2}{\left(\bar{X}_1+\bar{X}_2\right)^2}+\frac{\bar{X}_2^2\bar{Z}^2C_{x2}^2}{\left(\bar{X}_1+\bar{X}_2\right)^2}-2\frac{\bar{X}_1\bar{Z}^2C_{zx1}}{\left(\bar{X}_1+\bar{X}_2\right)}-2\frac{\bar{X}_2\bar{Z}^2C_{zx2}}{\left(\bar{X}_1+\bar{X}_2\right)}+2\frac{\bar{X}_1\bar{X}_2\bar{Z}^2C_{x1x2}}{\left(\bar{X}_1+\bar{X}_2\right)^2}\right)\right),$$

$$F=\lambda\left(\frac{\bar{Z}\bar{X}_1^2}{\left(\bar{X}_1+\bar{X}_2\right)}C_{x1}^2+\frac{\bar{Z}\bar{X}_1\bar{X}_2}{\left(\bar{X}_1+\bar{X}_2\right)}C_{x1x2}-\bar{Z}\bar{X}_1C_{zx1}\right),\text{ and}$$

$$G=\lambda\left(\frac{\bar{Z}\bar{X}_2^2}{\left(\bar{X}_1+\bar{X}_2\right)}C_{x2}^2+\frac{\bar{X}_1\bar{X}_2\bar{Z}}{\left(\bar{X}_1+\bar{X}_2\right)}C_{x1x2}-\bar{Z}\bar{X}_2C_{zx2}\right).$$

3.5. Proposed estimator

The combined product estimators have shown advantages in efficiency in spite of being more biased than the traditional ratio or regression estimators (Koyuncu et al., 2013). Motivated by this fact, we propose a change in the difference-cum-exponential estimator in (3.1) so that when the sample mean ($\bar{x}$) of the auxiliary variable X is close to population mean ($\bar{X}$), the expected value of the proposed estimator is closer to the mean of the variable of interest Y. So, our proposed estimator is a modified version of the difference-cum-exponential estimator in (3.1) and is given by the following expression

$$\hat{\mu}_P=[d_1\bar{z}+d_2]\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right), \tag{3.7}$$

where d_1 and d_2 are constants.

Using Taylor's approximation and retaining terms of order up to 2, (3.7) can be rewritten as

$$\hat{\mu}_P-\bar{Z}=\left[(d_1-1)\bar{Z}+d_1\bar{Z}e_0+d_2\right]\left\{1-\frac{1}{2}e_1+\frac{3}{8}e_1^2\right\}. \tag{3.8}$$

Recognizing that $\bar{Z}=\bar{Y}$, the *Bias* and *MSE* of $\hat{\mu}_P$, to first degree of approximation, are given by

$$Bias(\hat{\mu}_P) \cong (d_1 - 1)\bar{Y} + d_1\bar{Y}\left(\frac{3}{8}v_{02} - \frac{1}{2}v_{11}\right) + d_2\left(1 + \frac{3}{8}v_{02}\right) \tag{3.9}$$

and

$$MSE(\hat{\mu}_P) \cong d_1^2\bar{Y}^2 A + d_2^2 B - 2d_1\bar{Y}^2 C - 2d_2\bar{Y}D + 2d_1 d_2\bar{Y}E + \bar{Y}^2, \tag{3.10}$$

where $A = 1 + v_{20} + v_{02} - 2v_{11}$, $B = 1 + v_{02}$, $C = 1 + \frac{3}{8}v_{02} - \frac{1}{2}v_{11}$, $D = 1 + \frac{3}{8}v_{02}$, $E = 1 + v_{02} - v_{11}$.

Using (3.10), the optimum values of d_1 and d_2 are

$d_{1(opt)} = \frac{BC - DE}{AB - E^2}$ and $d_{2(opt)} = \frac{\bar{Y}(AD - CE)}{AB - E^2}$.

Considering the *MSE* at optimum values we get

$$MSE(\hat{\mu}_P)_{opt} \cong \bar{Y}^2\left[1 - \frac{BC^2 + AD^2 - 2CDE}{AB - E^2}\right],$$

or

$$MSE(\hat{\mu}_P)_{opt} \cong \bar{Y}^2\left[1 - \frac{v_{20} + \frac{3}{4}v_{20}v_{02}(1-\rho_{zx}^2)\left(1 + \frac{3}{16}v_{02}\right) + \frac{1}{64}v_{02}v_{11}^2}{v_{20} + v_{02}v_{20}(1-\rho_{zx}^2)}\right], \tag{3.11}$$

where $\rho_{zx} = \frac{v_{11}}{\sqrt{v_{20}}\sqrt{v_{02}}}$.

Comparing the *MSE* of this estimator to the *MSE* of difference-cum-exponential estimator given in (3.3), we note that the proposed estimator will be more efficient if

$MSE(\hat{\mu}_P)_{opt} < MSE(\hat{\mu}_{DE})_{opt}$

This will be so if

$$\frac{64v_{20}-48v_{11}^2-8v_{02}v_{11}^2+48v_{20}v_{02}+9v_{20}v_{02}^2}{64(v_{20}+v_{02}v_{20}-v_{11}^2)}-\frac{v_{02}\left(8-v_{02}\right)^2}{64(v_{02}+v_{02}v_{20}-v_{11}^2)}-\frac{1}{4}v_{02}>0,$$

or if

$$[64v_{20}^2v_{02}+32v_{20}^2v_{02}^2+8v_{20}v_{02}^3-7v_{20}^2v_{02}^3-v_{20}v_{02}^4+15v_{02}^2v_{20}v_{11}^2-80v_{20}v_{02}v_{11}^2-16v_{20}v_{02}^2$$

$$-64v_{20}v_{11}^2-8v_{02}^2v_{11}^2-8v_{02}v_{11}^4+v_{02}^3v_{11}^2+48v_{11}^4+16v_{02}v_{11}^2]/M>0,$$

where $M=64\left[\left\{v_{20}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\left\{v_{02}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\right]$,

or if

$$\frac{v_{20}v_{02}\left(1-\rho^2\right)\left[8\left(8v_{20}+v_{02}v_{11}^2-6v_{11}^2\right)-v_{02}\left(4-v_{02}\right)^2+v_{20}v_{02}\left(32-7v_{02}\right)\right]}{64\left[\left\{v_{20}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\left\{v_{02}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\right]}>0. \qquad (3.12)$$

The above condition is true if numerator is positive. It may be noted that for the proposed estimator, any loss of privacy is due to the randomization of the study variable Y since we use here non-sensitive auxiliary information. And unlike the two-stage RRT models where a truth element is introduced, there is no extra loss of privacy here as compared to a simple additive RRT model.

The estimator proposed in (3.7) can be generalized to the case of multiple auxiliary variables. We consider below the case of two auxiliary non-sensitive variables. The generalized estimator is given by

$$\hat{\mu}_{P2}=[k_1\bar{z}+k_2]\exp\left(\frac{\bar{X}_1-\bar{x}_1}{\bar{X}_1+\bar{x}_1}+\frac{\bar{X}_2-\bar{x}_2}{\bar{X}_2+\bar{x}_2}\right), \qquad (3.13)$$

where k_1 and k_2 are constants.

To obtain the *Bias* and *MSE* expressions for the estimator in (3.13), let us still define

$$e_0 = (\bar{z} - \bar{Z}) / \bar{Z} \text{ and } e_1 = (\bar{x}_1 - \bar{X}_1) / \bar{X}_1,\ e_2 = (\bar{x}_2 - \bar{X}_2) / \bar{X}_2.$$

We note that $E(e_0^2) = \lambda C_z^2 = v_{200}$, $E(e_1^2) = \lambda C_{x_1}^2 = v_{020}$, $E(e_2^2) = \lambda C_{x_2}^2 = v_{002}$,

$E(e_0 e_1) = \lambda \rho_{zx_1} C_z C_{x_1} = v_{110}$,

$$E(e_0 e_2) = \lambda \rho_{zx_2} C_z C_{x_2} = v_{101},\ E(e_1 e_2) = \lambda \rho_{x_1 x_2} C_{x_1} C_{x_2} = v_{011},$$

where

$$\mu_{rst} = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{Z})^r (x_{1i} - \bar{X}_1)^s (x_{2i} - \bar{X}_2)^t \text{ and } v_{rst} = \frac{E(\bar{z} - \bar{Z})^r (\bar{x}_1 - \bar{X}_1)^s (\bar{x}_2 - \bar{X}_2)^t}{\bar{Z}^r \bar{X}_1^s \bar{X}_2^t}.$$

Expressing (3.13) in terms of e’s and retaining terms of order up to 2, (3.13) can be rewritten as

$$\hat{\mu}_{P2} - \bar{Z} = \left(k_1 - 1\right) \bar{Z} + k_1 \bar{Z} \left[e_0 - \frac{1}{2} e_1 - \frac{1}{2} e_2 - \frac{1}{2} e_0 e_1 - \frac{1}{2} e_0 e_2 + \frac{3}{8} e_1^2 + \frac{3}{8} e_2^2 + \frac{1}{4} e_1 e_2 \right]$$

$$+ k_2 \left[1 - \frac{1}{2} e_1 - \frac{1}{2} e_2 + \frac{3}{8} e_1^2 + \frac{3}{8} e_2^2 + \frac{1}{4} e_1 e_2 \right]. \tag{3.14}$$

For $\bar{Z} = \bar{Y}$ in (3.11), the *MSE* is given by

$$MSE(\hat{\mu}_{P2}) \cong \bar{Y}^2 + k_1^2 \bar{Y}^2 A^* + k_2^2 B^* - 2k_1 \bar{Y}^2 C^* - 2k_2 \bar{Y} D^* + 2k_1 k_2 \bar{Y} E^*, \tag{3.15}$$

where

$A^* = 1 + v_{200} + v_{020} + v_{002} + v_{011} - 2v_{110} - 2v_{101}$,

$B^* = 1 + v_{020} + v_{002} + v_{011}$,

$C^* = 1 - \frac{1}{2} v_{110} - \frac{1}{2} v_{101} + \frac{3}{8} v_{020} + \frac{3}{8} v_{002} + \frac{1}{4} v_{011}$,

$D^* = 1 + \frac{3}{8} v_{020} + \frac{3}{8} v_{002} + \frac{1}{4} v_{011}$,

$E^* = 1 + v_{020} + v_{002} - v_{110} - v_{101} + v_{011}$, The optimum values of k_1 and k_2 are

$$k_{1(opt)} = \frac{B^* C^* - D^* E^*}{A^* B^* - E^{*2}} \text{ and } k_{2(opt)} = \frac{\bar{Y}\left(A^* D^* - C^* E^*\right)}{A^* B^* - E^{*2}}.$$

Considering the *MSE* at optimum values we get

$$MSE(\hat{\mu}_{P2})_{opt} \cong \bar{Y}^2 \left[1 - \frac{B^* C^{*2} + A^* D^{*2} - 2C^* D^* E^*}{A^* B^* - E^{*2}}\right]. \quad (3.16)$$

This estimator can be generalized to the case of *m* auxiliary variables as follows:

$$\hat{\mu}_{Pm} = [k_1 \bar{z} + k_2] \exp\left(\frac{\bar{X}_1 - \bar{x}_1}{\bar{X}_1 + \bar{x}_1} + \frac{\bar{X}_2 - \bar{x}_2}{\bar{X}_2 + \bar{x}_2} + \ldots + \frac{\bar{X}_m - \bar{x}_m}{\bar{X}_m + \bar{x}_m}\right). \quad (3.17)$$

4. Simulation Results

In this section, we conduct a simulation study with particular focus on comparing the performance of the proposed combined estimator $\hat{\mu}_P$ to the estimator $\hat{\mu}_{DE}$ suggested by Koyuncu et al. (2013), using the *Bias* and *MSE* results, correct up to first order of approximation.

We consider 2 different bivariate normal distributions for (*Y*, *X*). The scrambling variable *S* is taken to be a normal distribution with mean equal to zero and standard deviation equal to 10% of the standard deviation of *X*. The reported response is given by *Z* = *Y*+*S*.

The summary statistics for the bivariate normal populations are given below.

Population Statistics:

I	$N = 1000$, $\mu_Y = 2$, $\sigma_Y = \sqrt{10}$, $\mu_Y = 2$, $\sigma_X = \sqrt{2}$, $\sigma_{XY} = 3$ and $\rho_{XY} = 0.6708$

II	$N = 1000$, $\mu_Y = 2$, $\sigma_Y = \sqrt{6}$, $\mu_Y = 2$, $\sigma_X = \sqrt{2}$, $\sigma_{XY} = 3$ and $\rho_{XY} = 0.8660$

We take samples of size *n* = 50, 100, 200 and 300 from each population to compare the results. We estimate the empirical *Bias* and *MSE* using 5000 samples of various sizes from the study populations. The absolute relative bias (*ARB*), used in the tables below, is given by

$$\left|\frac{Bias(\hat{\mu}_\alpha)}{\bar{Y}}\right|,$$

where $\alpha = DE$ and *P*.

The empirical and the theoretical results for the two estimators under study are presented in Table 1 and Table 2, respectively. From these tables we can observe that the proposed estimator shows reduced *Bias* as compared to other estimator.

As expected, the absolute relative bias generally decreases as the sample size increases, however this effect becomes less pronounced when the correlation between the auxiliary variable and the study variable is higher. Although the proposed estimator is not unbiased, the bias results show a very good performance for this estimator.

Table 3 below gives the empirical and theoretical *MSE's* for the two competing estimators.

The *MSE* values for the proposed estimators with one and two auxiliary variables are all less than the *MSE* values for the Koyuncu et al. (2013) estimator with one and two auxiliary variables, respectively. The estimators under study get more and more efficient as ρ_{XY} increases. Proposed exponential estimator with two auxiliary variables ($\hat{\mu}_{P2}$) performs better than the estimator with one auxiliary variable ($\hat{\mu}_P$), as expected.

5. Numerical Example

In this section, we use real data concerning enterprises for the Monthly Economic Survey (MES) in Portugal. The survey is conducted to provide an accurate picture of business trends of enterprises. It provides short-term indicators on a monthly basis compiled for 4 sectors: industry, retail trade, construction and service sector. The survey results are broken down by branches according to the Classification of Economic Activities in the European Community (NACE) Rev. 2 (Eurostat, 2008). In this survey the main questions refer to an assessment of recent trends in production, of the current levels of order books and stocks, as well as expectations about production, selling prices and employment. We consider as population the enterprises covered in the 2010 sample which provided results for the industry sector, taking the monthly salaries as study variable and number of employees as auxiliary variable in each enterprise.

Let Y be the monthly salaries amount in 2010 collected by the MES in that year. This is typically a confidential variable for enterprises, only known from business surveys. The auxiliary variable X is the number of employees available from business data registers. The variables Y and X are strongly correlated so we can take advantage of this correlation by using the estimators under study. The MES provided 26980 monthly salary values in 2010, collected for 2316 enterprises which answered this survey in that same year. We take these 26980 values as our population. For the *RRT* part, let S be a normal random variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y+S$ (the salary amount plus a random quantity). The summary statistics about the populations are given below.

<u>Populations Statistics</u>:

$N = 26980$, $\rho_{XY} = 0.8599$

$\mu_X = 113.91$, $\mu_Y = 167.18$ (in thousands of €)

$\sigma_X = 215.8$, $\sigma_Y = 501.4$ and $\sigma_{XY} = 93043$

n = 1000, 2500, 5000 and 10000.

In Tables 4 and 5 below we present the empirical and the theoretical *ARB* results, respectively, for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the proposed estimator ($\hat{\mu}_P$).

The *ARB* results show the good performance for the improved difference-cum-exponential estimator.

The theoretical *MSE* values for both estimators have been obtained using (3.3) and (3.11). These values are given in Table 6.

According to the *MSE* results in Table 6, the proposed estimator ($\hat{\mu}_P$) is considerably better than the difference-cum-exponential estimator ($\hat{\mu}_{DE}$). These results are in line with the theoretical findings and the simulation results. These findings were still valid if we considered two or more auxiliary variables but for our numerical example all the auxiliary variables are strongly correlated, so we opted to only consider an explanatory variable.

6. Conclusion

We can conclude from this study that the estimation of the mean of a sensitive variable can be improved by using correlated non-sensitive auxiliary variables. Our simulation study and the numerical example show that the proposed improved difference-cum-exponential estimator can produce further improvement. We note that the proposed estimator is more efficient than the difference-cum-exponential estimator recently proposed by Koyuncu et al. (2013), which in turn was better than most of the existing estimators of finite population mean. Also there is no additional loss of privacy as compared to what it is for an additively scrambled RRT model.

Acknowledgement

The authors would like to thank the referees for their suggestions that helped improve the presentation of the paper.

 ACCEPTED MANUSCRIPT

References

Eurostat (2008). NACE Rev. 2 - *Statistical Classification of Economic activities in the European Community*, Official Publications of the European Communities, 112-285 and 306-311.

Gupta, S., Shabbir, J., Sousa, R., Corte-Real, P. (2012). Regression Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information, *Communications in Statistics – Theory and Methods,* 41, 2394-2404.

Kadilar, C. and Cingi, H. (2004). Ratio estimators in simple random sampling, *Applied Mathematics and Computation*, 151, 893-902.

Kadilar, C., Candan, M. and Cingi, H. (2007). Ratio estimators using robust regression, *Hacettepe Journal of Mathematics and Statistics*, 36, 2, 81-188.

Koyuncu, N., Gupta, S., and Sousa, R. (2013). Exponential Type Estimators of the Mean of a Sensitive Variable in the Presence of Non-Sensitive Auxiliary Information. To appear in *Communications in Statistics- Simulation and Computation.*

Nangsue, N. (2009). Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing, *World Academy of Science, Engineering and Technology*, 53, 787-790.

Shabbir, J. and Gupta, S. (2007). On improvement in variance estimation using auxiliary information, Communication in Statistics-Theory and Methods, 36(12), 2177-2185.

Shabbir, J. and Gupta, S. (2010). Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute, *Hacettepe Journal of Mathematics and Statistics*, 39(1), 121-129.

Sousa, R., Shabbir, J., Corte-Real, P., Gupta, S. (2010). Ratio Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information, *Journal of Statistical Theory and Practice*, 4(3), 495-507.

Warner, S.L. (1965). Randomized Response: A survey technique for eliminating evasive answer bias, *Journal American Statistical Association*, 76, 916-923.

Table 1 – Empirical absolute relative bias for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$), for the DE estimator with two auxiliary variables ($\hat{\mu}_{DE2}$), for the proposed estimator ($\hat{\mu}_{P}$) and for the proposed estimator with two auxiliary variables ($\hat{\mu}_{P2}$).

Population						
N	ρ_{X1Y} ρ_{X2Y}	**Estimator**	***n* = 50**	***n* = 100**	***n* = 200**	***n* = 300**
1000	**0.6375 0.6773**		0.0093	0.0061	0.0039	0.0041
			0.0012	**0.0009**	**0.0009**	**0.0008**
			0.0041	0.0031	0.0031	0.0036
			0.0021	**0.0014**	**0.0015**	**0.0016**
	0.7056 0.7534		0.0074	0.0054	0.0037	0.0039
			0.0013	**0.0011**	**0.0010**	**0.0009**
			0.0035	0.0032	0.0033	0.0035
			0.0020	**0.0017**	**0.0018**	**0.0018**

Table 2 – Theoretical absolute relative bias for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$), for the DE estimator with two auxiliary variables ($\hat{\mu}_{DE2}$), for the proposed estimator ($\hat{\mu}_{P}$) and for the proposed estimator with two auxiliary variables ($\hat{\mu}_{P2}$).

Population						
N	**ρ_{X1Y} ρ_{X2Y}**	**Estimator**	***n* = 50**	***n* = 100**	***n* = 200**	***n* = 300**
1000	**0.6375 0.6773**		0.0038	0.0018	0.0008	0.0005
			0.0004	**0.0002**	**0.0001**	**0.0001**
			0.0002	0.0001	0.0000	0.0000
			0.0004	**0.0002**	**0.0001**	**0.0000**
	0.7056 0.7534		0.0023	0.0011	0.0005	0.0003
			0.0004	**0.0002**	**0.0001**	**0.0000**
			0.0002	0.0001	0.0000	0.0000
			0.0001	**0.0000**	**0.0000**	**0.0000**

Table 3 – Empirical and theoretical *MSE* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$), for the proposed estimator ($\hat{\mu}_{P}$), for the DE estimator with two auxiliary variables ($\hat{\mu}_{DE2}$) and for the proposed estimator with two auxiliary variables ($\hat{\mu}_{P2}$).

Population				***MSE* Estimation**		***MSE***
N	ρ_{XY}	***n***	**Estimator**	**Empirical**	**Theoretical**	**Condition[1]**
1000	**0.6375 0.6773**	50		0.1174	0.1164	0.0041
				0.0459	0.0437	
				0.0115	0.0112	
				0.0104	0.0097	
		100		0.0561	0.0553	0.0020
				0.0217	0.0207	
				0.0053	0.0053	
				0.0047	0.0046	#REF!
		200		0.0255	0.0246	0.0009
				0.0097	0.0092	
				0.0024	0.0024	
				0.0021	0.0020	
		300		0.0145	0.0144	0.0005
				0.0057	0.0054	
				0.0014	0.0014	
				0.0012	0.0012	
	0.7056 0.7534	50		0.0792	0.0793	0.0027
				0.0077	0.0071	

				0.0097	0.0095		
				0.0022	0.0020		
		100		0.0377	0.0377		0.0013
				0.0037	0.0034		
				0.0045	0.0045		
				0.0010	0.0009		
		200		0.0172	0.0168		0.0006
				0.0018	0.0015		
				0.0020	0.0020		
				0.0005	0.0004		
		300		0.0099	0.0098		0.0003
				0.0012	0.0009		
				0.0012	0.0012		
				0.0003	0.0002		

[1] MSE comparison condition based on expression (3.12).

Table 4 – Empirical absolute relative bias for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the proposed estimator ($\hat{\mu}_{P}$).

Population						
N	$\boldsymbol{\rho_{XY}}$	**Estimator**	***n* = 1000**	***n* = 2500**	***n* = 5000**	***n* = 10000**
26980	0.8599		0.0025	0.0022	0.0016	0.0012
			0.0003	**0.0004**	**0.0004**	**0.0003**

Table 5 – Theoretical absolute relative bias for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the proposed estimator ($\hat{\mu}_P$).

Population						
N	ρ_{XY}	**Estimator**	***n* = 1000**	***n* = 2500**	***n* = 5000**	***n* = 10000**
26980	0.8599		0.0008	0.0003	0.0001	0.0001
			0.0002	**0.0001**	**0.0000**	**0.0000**

Table 6 – Empirical and Theoretical *MSE* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the proposed estimator ($\hat{\mu}_{P}$).

Population				*MSE* Estimation		*MSE*
N	**ρ_{XY}**	***n***	**Estimator**	**Empirical**	**Theoretical**	**Condition** [2]
26980	0.8599	1000		62.58	63.34	0.0205
				6.28	**6.30**	
		2500		24.50	23.92	0.0083
				2.40	**2.38**	
		5000		10.88	10.75	0.0041
				1.09	**1.07**	
		10000		4.19	4.15	0.0020
				0.41	**0.41**	

[2] MSE comparison condition based on expression (3.12).