STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 637–650, DOI 10.21307/stattrans-2017-004

A NEW ESTIMATOR OF MEAN USING DOUBLE SAMPLING

Kalyan Rao Vadlamudi[1], **Stephen A. Sedory, Sarjinder Singh**

ABSTRACT

In this paper, we consider the problem of estimation of population mean of a study variable by making use of first-phase sample mean and first-phase sample median of the auxiliary variable at the estimation stage. The proposed new estimator of the population mean is compared to the sample mean estimator, ratio estimator and the difference type estimator for the fixed cost of the survey by using the concept of two-phase sampling. The magnitude of the relative efficiency of the proposed new estimator has been investigated through simulation study.

Key words: Two-phase sampling, relative efficiency, analytical and empirical comparison.

1.Introduction

Consider a population Ω consisting of N units. Let (y_i, x_i), $i = 1,2,..., N$ be the values of the study variable Y and auxiliary variable X for the ith unit in the population.

Let

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.1}$$

and

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2}$$

be the population means of the study and auxiliary variables respectively. Survey statisticians are often interested in estimating the population mean $\bar{Y}$ of the study variable. It is also well known that if the population mean $\bar{X}$ of an auxiliary variable is known then it can be used to improve estimation strategies in survey sampling. Examples of such estimators are the ratio estimator due to Cochran

[1] Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX 78363, USA.

(1940) and the linear regression estimator due to Hansen, Hurwitz and Madow (1953).

If such auxiliary information of the population mean $\bar{X}$ is not known or complete auxiliary information is not known, then it can be relatively cheaper to obtain information on the auxiliary variable by taking a large preliminary sample for estimating population mean of the auxiliary variable to be used at the estimation stage. In other words, in the case of single auxiliary variable X, if the population mean $\bar{X}$ of the auxiliary variable is unknown then we consider taking a preliminary large sample of m units by using simple random and without replacement sampling (SRSWOR) from the population of N units. In the sample s_1 of m units, we observe only the auxiliary variable x_i , $i = 1,2,...,m$. From the given first-phase sample s_1 of m units, we select another subsample s_2 of n units by using SRSWOR sample. In the sample s_2, we measure the ordered pairs (y_i, x_i), $i = 1,2,...,n$ of the study and auxiliary variables. Then an unbiased estimator of the population mean $\bar{X}$ based on the first-phase sample information as the sample mean is given by:

$$\bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1.3}$$

The unbiased estimators of the population means $\bar{Y}$ and $\bar{X}$ of the study and auxiliary variables based on the second-phase sample information are, respectively, given by:

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.4}$$

and

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.5}$$

Neyman (1938) invented this sampling technique called double sampling or two-phase sampling, and later work related to this scheme is extensively reviewed in Singh (2003). It leads to ratio and regression type estimators of the population mean $\bar{Y}$ in two-phase sampling as:

$$\bar{y}_{\text{rat(d)}} = \bar{y}_n\left(\frac{\bar{x}_m}{\bar{x}_n}\right) \tag{1.6}$$

and

$$\bar{y}_{\text{reg(d)}} = \bar{y}_n + \beta\left(\bar{x}_m - \bar{x}_n\right) \tag{1.7}$$

The variances of the sample mean, ratio and regression type estimators are, respectively, given by;

$$V(\bar{y}_{\mathrm{n}}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$$

$$V(\bar{y}_{\mathrm{rat(d)}}) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right] \tag{1.8}$$

and

$$V(\bar{y}_{\mathrm{reg(d)}}) = \left(\frac{1}{m} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right) S_y^2\left[1 - \rho_{xy}^2\right] \tag{1.9}$$

where

$$R = \frac{\bar{Y}}{\bar{X}};\quad \rho_{xy} = \frac{S_{xy}}{S_x S_y};\quad S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X});$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2, \text{ and}$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$

To our knowledge, the pioneer contributors, to the problem of estimating median, are Kuk and Mak (1989) by proposing very clear estimators of median in the presence of auxiliary information. Singh, Joarder and Tracy (2001) extended their idea to the situation of median estimation in two-phase sampling. The importance of double sampling and improvements on the estimation of population mean can also be seen in several publications by Vishwakarma and Kumar (2015), Vishwakarma and Gangele (2014), Vishwakarma and Singh (2011), Amin *et al*. (2016), and Sanaullah *et al*. (2014). However, none of these papers deal with the situation of making use of estimator of median of the auxiliary variable at the estimation stage of population mean of the study variable in two-phase sampling. This motivated the authors to think on these lines if some improvements can be seen by making the use of first-phase median of the auxiliary variable.

In the next section, we introduce a new estimator of the population mean in two-phase sampling which makes use of first-phase sample mean and sample median of the auxiliary variable.

2. Estimator

Let $\hat{M}_x^*$ be the median of the auxiliary variable X based on the first phase sample s_1 of m units. Let $\hat{M}_x$ be the median of the auxiliary variable X based on the second phase sample s_2 of n units.

Suppose $x_{(1)}, x_{(2)},..., x_{(m)}$ are the x values of first-phase sample units in the ascending order. Further, let t_1 be an integer such that $X_{(t_1)} \le M_x \le X_{(t_1+1)}$ and let $p_1 = t_1/n$ be the proportion of X values in the first phase sample that are less than or equal to the median value M_x, an unknown population parameter. If $\hat{p}_1$ is a predictor of p_1, the first-phase sample median $\hat{M}_x^*$ can be written in terms of quintiles as $\hat{Q}_x(\hat{p}_1)$, where $\hat{p}_1 = 0.5$.

Suppose $x_{(1)}, x_{(2)},..., x_{(n)}$ are the x values of second-phase sample units in the ascending order. Further, let t_2 be an integer such that $X_{(t_2)} \le M_x \le X_{(t_2+1)}$ and let $p_2 = t_2/n$ be the proportion of X values in the second phase sample that are less than or equal to the median value M_x, an unknown population parameter. If $\hat{p}_2$ is a predictor of p_2, the first-phase sample median $\hat{M}_x$ can be written in terms of quintiles as $\hat{Q}_x(\hat{p}_2)$, where $\hat{p}_2 = 0.5$.

Now we define a new estimator of the population mean $\bar{Y}$ in two-phase sampling as:

$$\bar{y}_{kal} = \bar{y}_n + \beta_1^*(\bar{x}_m - \bar{x}_n) + \beta_2^*(\hat{M}_x^* - \hat{M}_x) \tag{2.1}$$

where β_1^* and β_2^* are unknown partial regression coefficients to be determined such that the variance of the estimator is minimum.

It may be worth pointing out that the proposed estimator $\bar{y}_{kal}$ is an extension of the recent estimator of due to Lamichhane, Singh and Diawara (2015) from single-phase sampling to two-phase sampling.

To study the asymptotic properties of the proposed estimator, $\bar{y}_{kal}$, let us define the following error terms:

$$\varepsilon_0 = \frac{\bar{y}_n}{\bar{Y}} - 1;\ \ \varepsilon_1 = \frac{\bar{x}_n}{\bar{X}} - 1;\ \ \varepsilon_2 = \frac{\hat{M}_x}{M_x} - 1;\ \ \varepsilon_3 = \frac{\bar{x}_m}{\bar{X}} - 1 \text{ and } \varepsilon_4 = \frac{\hat{M}_x^*}{M_x} - 1$$

such that

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0;\ E(\varepsilon_3) \approx E(\varepsilon_4) \approx 0$$

$$E(\varepsilon_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_y^2;\ \ E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2;$$

$$E(\varepsilon_2^2) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2 M_x^2}$$

$$E(\varepsilon_3^2)=\left(\frac{1}{m}-\frac{1}{N}\right)C_x^2;\;\; E(\varepsilon_4^2)=\left(\frac{1}{m}-\frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2};$$

$$E(\varepsilon_0\varepsilon_1)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xy}C_xC_y;\;\; E(\varepsilon_1\varepsilon_3)=\left(\frac{1}{m}-\frac{1}{N}\right)C_x^2;$$

$$E(\varepsilon_0\varepsilon_2)=-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{YMx}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x};\; E(\varepsilon_0\varepsilon_3)=\left(\frac{1}{m}-\frac{1}{N}\right)\rho_{xy}C_xC_y;$$

$$E(\varepsilon_0\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{YM_x}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x};$$

$$E(\varepsilon_1\varepsilon_2)=-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{XMx}\{f_x(M_x)\}^{-1}}{\bar{X}M_x}$$

$$E(\varepsilon_1\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XM_x}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x};$$

$$E(\varepsilon_2\varepsilon_3)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XM_x}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x}$$

$$E(\varepsilon_2\varepsilon_4)=\left(\frac{1}{m}-\frac{1}{N}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2};\;\text{ and}$$

$$E(\varepsilon_3\varepsilon_4)=-\left(\frac{1}{m}-\frac{1}{N}\right)\frac{S_{XMx}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x}$$

where

$$S_{XM_x}=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(I_{x_i}-0.5\right),$$

$$f=n/N,$$

$$S_{YM_x}=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\left(I_{x_i}-0.5\right)$$

and

$$I_{x_i}=\begin{cases}1, & \text{if} \quad X_i\le M_x\\ 0, & \text{otherwise}\end{cases}$$

Note that we used the following main result from Kuk and Mak (1989) in deriving the variance and co-variance expressions for the sample means and

sample median, that is, if F_x be the cumulative distribution function of X, then the sample median can be approximated as:

$$\hat{M}_x = M_x + (0.5 - p_x) f_x^{-1}(M_x) + ...$$

where p_x be the proportion of I_{x_i} values taking a value of 1.

The new estimator $\bar{y}_{kal}$ in terms of ε_i, $i = 0,1,2,3,4$ can be written as

$$\begin{aligned}\bar{y}_{kal} &= \bar{y}_n + \beta_1^*(\bar{x}_m - \bar{x}_n) + \beta_2^*(\hat{M}_x^{|} - \hat{M}_x) \\ &= \bar{Y}(1+\varepsilon_0) + \beta_1^*[\bar{X}(1+\varepsilon_3) - \bar{X}(1+\varepsilon_1)] + \beta_2^*[M_x(1+\varepsilon_4) - M_x(1+\varepsilon_2)] \\ &= \bar{Y} + \bar{Y}\varepsilon_0 + \beta_1^*\bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2) \qquad (2.2)\end{aligned}$$

Now we have the following theorems:

Theorem 2.1. The new proposed estimator $\bar{y}_{kal}$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof. Taking expected value on both sides of (2.2), we get

$$\begin{aligned}E(\bar{y}_{kal}) &= E\left[\bar{Y} + \bar{Y}\ \varepsilon_0 + \beta_1^*\bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2)\right] \\ &= \bar{Y} + \bar{Y}\ E(\varepsilon_0) + \beta_1^*\bar{X}(E(\varepsilon_3) - E(\varepsilon_1)) + \beta_2^* M_x(E(\varepsilon_4) - E(\varepsilon_2)) \\ &= \bar{Y} + 0 \\ &= \bar{Y} \qquad (2.3)\end{aligned}$$

Thus the new estimator $\bar{y}_{kal}$ is an unbiased estimator of $\bar{Y}$ and it proves the theorem.

Theorem 2.2. The minimum variance, to the first order of approximation, of the new proposed estimator $\bar{y}_{kal}$ is given by

$$Min.V(\bar{y}_{kal}) = \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy}S_{XM_x})^2}{S_y^2 S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right) \qquad (2.4)$$

Proof. By the definition of variance, the variance of the unbiased estimator $\bar{y}_{kal}$ is given by

$$V(\bar{y}_{kal}) = E\left[\bar{y}_{kal} - \bar{Y}\right]^2$$

$$= E\left[\bar{Y}\varepsilon_0 + \beta_1^* \bar{X}(\varepsilon_3 - \varepsilon_1) + \beta_2^* M_x(\varepsilon_4 - \varepsilon_2)\right]^2$$

$$= E\Bigg[\bar{Y}^2\varepsilon_0^2 + \beta_1^{*2}\bar{X}^2(\varepsilon_3 - \varepsilon_1)^2 + \beta_2^{*2}M_x^2(\varepsilon_4 - \varepsilon_2)^2$$

$$-2\beta_1^*\bar{X}\bar{Y}\varepsilon_0(\varepsilon_3 - \varepsilon_1) - 2\beta_2^* M_x\bar{Y}\varepsilon_0(\varepsilon_4 - \varepsilon_2) + 2\beta_1^*\beta_2^*\bar{X}M_x(\varepsilon_3 - \varepsilon_1)(\varepsilon_4 - \varepsilon_2)\Bigg] \tag{2.5}$$

Further note that

$$E(\varepsilon_3 - \varepsilon_1)^2 = \left(\frac{1}{n} - \frac{1}{m}\right)C_x^2 \tag{2.6}$$

$$E(\varepsilon_4 - \varepsilon_2)^2 = \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{4\{f_x(M_x)\}^2 M_x^2} \tag{2.7}$$

$$E(\varepsilon_0\varepsilon_3 - \varepsilon_0\varepsilon_1) = -\left(\frac{1}{n} - \frac{1}{m}\right)\rho_{xy}C_xC_y \tag{2.8}$$

$$E(\varepsilon_0\varepsilon_4 - \varepsilon_0\varepsilon_2) = \left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{YM_X}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x} \tag{2.9}$$

and

$$E(\varepsilon_3\varepsilon_4 - \varepsilon_1\varepsilon_4 - \varepsilon_3\varepsilon_2 + \varepsilon_1\varepsilon_2) = -\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{XM_X}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x} \tag{2.10}$$

On substituting (2.6) – (2.10) into (2.5), we have

$$V(\bar{y}_{kal}) = \bar{Y}^2\varepsilon_0^2 + \beta_1^{*2}\bar{X}^2E(\varepsilon_3 - \varepsilon_1)^2 + \beta_2^{*2}M_x^2E(\varepsilon_4 - \varepsilon_2)^2$$

$$-2\beta_1^*\bar{X}\bar{Y}E(\varepsilon_0\varepsilon_3 - \varepsilon_0\varepsilon_1) - 2\beta_2^*M_x\bar{Y}E(\varepsilon_0\varepsilon_4 - \varepsilon_0\varepsilon_2) + 2\beta_1^*\beta_2^*\bar{X}M_xE(\varepsilon_3 - \varepsilon_1)(\varepsilon_4 - \varepsilon_2)$$

$$= \bar{Y}^2\left(\frac{1}{n} - \frac{1}{N}\right)C_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\beta_1^{*2}\bar{X}^2C_x^2 + \beta_2^{*2}M_x^2\left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{4\{f_x(M_x)\}^2M_x^2}$$

$$+2\beta_1^*\bar{X}\,\bar{Y}\left(\left(\frac{1}{n} - \frac{1}{m}\right)\rho_{xy}C_xC_y\right) - 2\beta_2^*\bar{Y}M_x\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{YM_X}\{f_x(M_x)\}^{-1}}{\bar{Y}\,M_x}$$

$$-2\beta_1^*\beta_2^*\bar{X}M_x\left(\frac{1}{n} - \frac{1}{m}\right)\frac{S_{XM_X}\{f_x(M_x)\}^{-1}}{\bar{X}\,M_x} \tag{2.11}$$

On differentiating $V(\bar{y}_{kal})$ with respect to β_1^* and β_2^* and each equating to zero, we have,

$$\frac{\partial V(\bar{y}_{kal})}{\partial \beta_1^*} = 0 \text{, and } \frac{\partial V(\bar{y}_{kal})}{\partial \beta_2^*} = 0$$

On solving the system of linear equations, the optimum values of β_1^* and β_2^* are given by

$$\beta_1^* = \frac{\{f_x(M_x)\}^{-2} S_{YM_x} S_{XM_x} - S_{xy} S_y^2}{\{f_x(M_x)\}^{-2}\left(\frac{1}{4}S_x^2 - S_{XM_x}^2\right)} \tag{2.12}$$

and

$$\beta_2^* = \frac{S_x^2 S_{YM_x} - S_{xy} S_{XM_x}}{\{f_x(M_x)\}^{-1}\left(\frac{1}{4}S_x^2 - S_{XM_x}^2\right)} \tag{2.13}$$

On substituting the optimum values of β_1^* and β_2^*, the minimum variance is given by

$$Min.V(\bar{y}_{kal}) = \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)\left(S_y^2 - \frac{S_{xy}^2}{S_x^2} - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right)$$

$$= \left(\frac{1}{m} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{m}\right)S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25S_x^2 - S_{XM_x}^2\right)}\right) \tag{2.14}$$

which proves the theorem.

Remark: It may be worth pointing out that the replacement of β_1^* and β_2^* with their consistent estimates lead to a new estimator of the population mean in two two-phase sampling which has same mean square error up to the first order of approximation as the minimum variance in (2.14). Such changes do not affect the results up to the first order of approximation. ((6.1) and (6.2) in Singh, Singh, and Upadhayaya (2007)).

3. Comparison of different estimators

Note that

$$0.25S_x^2 - S_{XM_x}^2 > 0$$

$$\frac{S_{XM_x}^2}{S_x^2} < \frac{1}{4} \tag{3.1}$$

Thus the proposed new estimator $\bar{y}_{kal}$ is always more efficient than the sample mean, ratio and the regression type estimator in two-phase sampling. It may be worth pointing out that the final minimum variance of the proposed estimator is free from the value of $f_x(M_x)$, hence from computational point.

In the next section, we focus on the cost analysis in two-phase sampling because it is considered as one among the list of cost effective sampling schemes in survey sampling.

4. Cost Analysis

In this section we consider comparison of different estimators with cost aspects. Let C_0 be the overhead cost, C_1 be the cost of information from one unit in the first phase sample; and C_2 be the cost of information from one unit in the second phase sample. Note that the value of C_1 is always smaller than that of C_2 . Thus the total cost function is given by

$$C = C_0 + mC_1 + nC_2 \tag{4.1}$$

Now we have the following results

For the fixed cost C of the survey, the minimum variance of the proposed estimator $\bar{y}_{kal}$ is given by

$$Min.V(\bar{y}_{kal})_{fixed\_cost} =$$

$$= \frac{\left(\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left[1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right]}\right)^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.2}$$

The optimum values of m and n are given by

$$m = m_{kal} = \frac{(C - C_0)\sqrt{S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2\left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}\right]} \tag{4.3}$$

and

$$n = n_{kal} = \frac{(C - C_0)\sqrt{S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2\left(\rho_{xy}^2 + \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)} + \sqrt{C_2 S_y^2\left(1 - \rho_{xy}^2 - \frac{(S_x^2 S_{YM_x} - S_{xy} S_{XM_x})^2}{S_y^2 S_x^2 \left(0.25 S_x^2 - S_{XM_x}^2\right)}\right)}\right]} \tag{4.4}$$

which proves the theorem.

In case of single-phase sampling, the total cost function is given by

$$C = C_0 + nC_2 \tag{4.5}$$

From (4.11), we have the optimum sample size as:

$$n = \frac{C - C_0}{C_2} = n_u \text{ (say)} \tag{4.6}$$

The variance of the sample mean estimator is given by

$$V(\bar{y}_{\mathrm{n}})_{fixed\_\mathrm{cost}} = \left(\frac{C_2}{C - C_0} - \frac{1}{N}\right) S_y^2 \tag{4.7}$$

The optimum values of m and n for the ratio estimator are given by

$$m = \frac{(C - C_0)\sqrt{2RS_{xy} - R^2 S_x^2}}{\sqrt{C_1}\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy})}\right]} = m_{rat} \tag{4.8}$$

and

$$n = \frac{(C - C_0)\sqrt{(S_y^2 + R^2 S_x^2 - 2RS_{xy})}}{\sqrt{C_1}\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy})}\right]} = n_{rat} \tag{4.9}$$

The minimum variance of the ratio estimator for the fixed cost is given by:

$$Min.V(\bar{y}_{rat})_{fixt\_\mathrm{cost}} = \frac{\left[\sqrt{C_1(2RS_{xy} - R^2 S_x^2)} + \sqrt{C_2(S_y^2 + R^2 S_x^2 - 2RS_{xy}}\right]^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.10}$$

The optimum values of m and n for the regression type estimator $\bar{y}_{reg(d)}$ are given by

$$m = \frac{(C - C_0)\sqrt{S_y^2 \rho_{xy}^2}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 \left(1 - \rho_{xy}^2\right)}\right]} = m_{reg} \tag{4.11}$$

and

$$n = \frac{(C - C_0)\sqrt{S_y^2 \left(1 - \rho_{xy}^2\right)}}{\sqrt{C_1}\left[\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 \left(1 - \rho_{xy}^2\right)}\right]} = n_{reg} \tag{4.12}$$

Note that for the fixed cost of the survey, the variances of the regression type estimator $\bar{y}_{reg(d)}$ is given by

$$Min.V(\bar{y}_{\text{reg}})_{fixed\_\text{cost}} = \frac{\left(\sqrt{C_1 S_y^2 \rho_{xy}^2} + \sqrt{C_2 S_y^2 (1 - \rho_{xy}^2)}\right)^2}{C - C_0} - \frac{S_y^2}{N} \tag{4.13}$$

The percent relative efficiency of the new proposed estimator $\bar{y}_{kal}$ with respect to the sample mean estimator $(\bar{y}_n)$, ratio estimator $(\bar{y}_{\text{rat(d)}})$, and regression type estimators $(\bar{y}_{\text{reg(d)}})$:

$$RE(0) = RE(\text{sample mean}) = \frac{Min.V(\bar{y}_n)_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.14}$$

$$RE(1) = RE(\text{ratio}) = \frac{Min.V(\bar{y}_{rat})_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.15}$$

$$RE(2) = RE(\text{reg}) = \frac{Min.V(\bar{y}_{reg})_{fixed\_\text{cost}}}{Min.V(\bar{y}_{kal})_{fixed\_\text{cost}}} \times 100\% \tag{4.16}$$

In order to see the magnitude of the relative efficiency of the proposed estimator $\bar{y}_{kal}$ over the mean, ratio and the regression type estimator, we did simulation study.

5. Simulation Study

From (3.1), it is clear that the maximum value of $S_{XM_x} < \frac{S_x}{2}$. We consider many populations of size $N = 50{,}000$, $\bar{X} = 50$, $\bar{Y} = 20$, $S_y^2 = 5$, $S_x^2 = 10$ and

different values of the correlation coefficient ρ_{xy}. Note that it is very likely that the possible value of $0 < S_{XM_X} < 0.5$ and $0 < S_{YM_X} < 0.5$. Consider a situation of having total cost $C = \$5000$, overhead cost $C_0 = \$500$, cost of collecting information from one unit in the first-phase sample $C_1 = \$4$, and the cost of collecting information from one unit in the sample $C_2 = \$10$. We wrote R-codes to compute the percent relative efficiency values and the optimum sample sizes for the four estimators. There are many situations where the proposed estimator performs better than the existing estimators, and in the simulation study we stored only those combinations of ρ_{xy}, S_{XM_X} and S_{YM_X} where the value of $RE(2)$ is greater than 105%. In other words, the proposed estimator is at least 105% more efficient than the linear regression type estimator in double sampling. The results so obtained are presented in Table 5.1

Table 5.1. Percent relative efficiency of the proposed new estimator over the three estimators and optimum sample sizes

ρ_{xy}	S_{XM_X}	S_{YM_X}	n_u	m_{rat}	n_{rat}	m_{reg}	n_{reg}	m_{kal}	n_{kal}	$RE(0)$	$RE(1)$	$RE(2)$
0.80	0.05	0.40	450	483	406	515	386	540	370	107.7	108.3	105.4
0.80	0.05	0.45	450	483	406	515	386	550	364	110.1	110.6	107.8
0.80	0.05	0.50	450	483	406	515	386	565	354	113.4	114.0	111.0
0.80	0.10	0.45	450	483	406	515	386	544	368	108.7	109.2	106.4
0.80	0.10	0.50	450	483	406	515	386	556	360	111.4	112.0	109.1
0.80	0.15	0.45	450	483	406	515	386	539	371	107.5	108.1	105.3
0.80	0.15	0.50	450	483	406	515	386	549	364	109.8	110.4	107.5
0.80	0.20	0.50	450	483	406	515	386	543	368	108.5	109.0	106.2
0.80	0.25	0.50	450	483	406	515	386	538	371	107.4	107.9	105.1
0.85	0.05	0.35	450	516	385	568	352	600	332	115.4	112.8	106.8
0.85	0.05	0.40	450	516	385	568	352	616	322	118.9	116.3	110.1
0.85	0.05	0.45	450	516	385	568	352	637	309	123.9	121.2	114.8
0.85	0.05	0.50	450	516	385	568	352	670	288	131.7	128.8	122.0
0.85	0.10	0.35	450	516	385	568	352	594	336	113.8	111.3	105.4
0.85	0.10	0.40	450	516	385	568	352	606	328	116.6	114.1	108.0
0.85	0.10	0.45	450	516	385	568	352	623	317	120.7	118.1	111.8
0.85	0.10	0.50	450	516	385	568	352	649	301	126.7	123.9	117.3
0.85	0.15	0.40	450	516	385	568	352	598	333	114.8	112.3	106.4
0.85	0.15	0.45	450	516	385	568	352	613	324	118.2	115.6	109.4
0.85	0.15	0.50	450	516	385	568	352	633	311	122.9	120.2	113.8
0.85	0.20	0.45	450	516	385	568	352	604	330	116.1	113.6	107.5
0.85	0.20	0.50	450	516	385	568	352	620	319	120.0	117.3	111.1
0.85	0.25	0.45	450	516	385	568	352	596	334	114.4	111.9	106.0
0.85	0.25	0.50	450	516	385	568	352	610	326	117.6	115.0	108.9
0.85	0.30	0.50	450	516	385	568	352	602	331	115.7	113.2	107.1

Discussion: From the simulation study, it is observed that a high value of correlation coefficient ρ_{xy} and a high value of S_{YM_X} is required for the proposed new estimator to show percent relative efficiency of at least 105%. For example, if $\rho_{xy} = 0.80$, $S_{XM_X} = 0.05$, $S_{YM_X} = 0.40$, with a total cost of \$5000, if instead of we use only single phase sample mean estimator with optimum sample size $n_u = 450$, we should use the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$, then the percent relative efficiency value is $RE(0) = 107.7\%$. Instead of using the ratio estimator with optimum sample sizes $m_{rat} = 483$ and second-phase sample size $n_{rat} = 406$, if one uses the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$ then the percent relative efficiency value is $RE(1) = 108.3\%$. In the same way, if we use the proposed estimator with optimum sample sizes $m_{kal} = 540$ and $n_{kal} = 370$, then the relative efficiency of the proposed estimator over the regression method of estimation with optimum sample sizes $m_{reg} = 515$ and $n_{reg} = 386$ is $RE(2) = 105.4\%$.

Note that $n_{kal}(= 370) < n_{reg}(= 386) < n_{rat}(= 406) < n_u(= 450)$. The optimum second-phase sample size remains lowest in case of the proposed estimator, thus the proposed estimator reduces efforts for collecting data on the second-phase of the sample for the fixed cost of the survey and provides efficient results. The rest of the results in Table 5.1 can also be interpreted in the same way. We conclude that there exist several situations where the proposed new estimator can be used more efficiently for a fixed cost of the survey.

Acknowledgements

The authors are grateful to the Admin: Patryk Barszcz, Editorial Office, Statistics in Transition new series, and a referee for bringing the original manuscript in the presnt form.

REFERENCES

AMIN, M. N. U., SHAHBAZ, M. Q., KADILAR, C. (2016). Ratio estimators for population mean using robust regression in double sampling, Gael University Journal of Science, 29 (4), pp. 793–798.

COCHRAN, W. G., (1940). Some properties of estimators based on sampling scheme with varying probabilities, Austral. J. Statist., 17, pp. 22–28.

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953). Sample survey methods and theory, New York, John Wiley and Sons, pp. 456–464.

KUK, A. Y. C., MAK, T. K. (1989). Median estimation in the presence of auxiliary information, J. R. Statist. Soc., B, 51, pp. 261–269.

LAMICHHANE, R., SINGH, S., DIAWARA, N., (2015). Improved Estimation of Population Mean Using Known Median of Auxiliary Variable, Communications in Statistics-Simulation and Computation, DOI: 10.1080/03610918.2015.1062102.

NEYMAN, J., (1938). Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc., 33, pp. 101–116.

SANAULLAH, A., ALI, H. A., AMIN, M. N. U., HANIF, M., (2014). Generalized exponential chain ratio estimator under stratified two-phase random sampling. Applied Mathematics and Computation, 226, pp. 541–547.

SINGH, S., (2003). Advanced sampling theory with applications: How Michael selected Amy. Kluwer: The Academic Publisher, The Netherlands.

SINGH, S., JOARDER A. H., TRACY D. S., (2001). Median estimation using double sampling. Australian & New Zealand J. Statist., pp. 43, 33–46.

SINGH, S., SINGH, H. P., UPADHAYAYA, L. N., (2007). Chain ratio and regression type estimators for median estimation in survey sampling. Statistical Papers, 48 (1), pp. 23–46.

VISHWAKARMA, G. K., SINGH, H. P., (2011). Separate ratio-product estimator for estimating population mean using auxiliary information. Journal of Statistical Theory and Applications, 10 (4), pp. 653–664.

VISHWAKARMA, G. K., GANGELE, R. K., (2014). A class of chain ratio-type exponential estimators in double sampling using two auxiliary variates. Applied Mathematics and Computation, 227, pp. 171–175.

VISHWAKARMA, G. K., KUMAR, M., (2015). An efficient class of estimators for the mean of a finite population in two-phase sampling using multi-auxiliary variates, Commun. Math. Stat., 3, pp. 477–489.