Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 75(2) 2021 147–156

A Revised Calibration Weight based Ratio Estimator in Two-phase Sampling: A Case when Unit Level Auxiliary Information is Available for the First-phase Sample

Sadikul Islam[1], Hukum Chandra[2], U.C. Sud[2], Pradip Basak[3], Nirupam Ghosh[2] and P.R. Ojasvi[1]

[1]*ICAR-Indian Institute of Soil and Water Conservation, Dehra Dun*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[3]*Uttar Banga Krishi Viswavidyalaya, Cooch Behar*

Received 11 February 2021; Revised 06 November 2021; Accepted 04 December 2021

SUMMARY

This paper proposed a revised calibration weight based finite population ratio estimator under two phase sampling design assuming the condition that two auxiliary variable,correlated to numerator and denominator variable of the ratio were available at unit level for the first-phase sample. But, population level auxiliary variables were not available. In this article, two calibration estimator of finite population ratio were discussed, one consists of ratio of two calibration estimator of total and another, consists of combined common calibration weight based ratio estimator under two-phase sampling design. The results of empirical analysis revealed that combined common calibration weight based ratio estimator was best performing and further both the calibration estimator were outperforming over the existing calibration estimator of Islam *et al.* (2019). Hence, combined common calibration weight based ratio estimator declared as a proposed revised calibration weight based finite population ratio estimator under two phase sampling. The theoretical expression of variance estimator as well as optimum sample size for minimum variance for fixed cost were also deliberated for the proposed ratio estimator .

Keywords: Calibration weight, Cost function, Population ratio, Two-phase sampling.

1. INTRODUCTION

In survey sampling, many highly efficient popular estimation methods needs quality dataset of auxiliary variables. For example, calibration estimator of Deville and Särndal (1992) requires population total of auxiliary information, Wu and Luan (2003).Calibration method achieved gain in efficiency through replacing survey weight by a modified weight (calibration weight) using a set of calibration equations related to auxiliary variables, see Särndal *et al.* (1992). Calibration methods are now applied on almost all surveys in official statistics, Devaud and Tille (2019). But, in the beginning of calibration, its application limited mostly to the situation when auxiliary variables were available at population level. In practice, this assumption was very rarely holds or sometimes even if hold the auxiliary variables were outdated and this type of situation is very common in most of the developing nations of the world where data collection methods are not done in regular timely manner. When population level information is not known, in that case two-phase sampling design have considerable attention where it is very expensive to collect data on the variables of interest but, it is relatively inexpensive to collect data on auxiliary variables, correlated to variable of interest, Legg and Fuller (2009). Two-phase sampling have achieved considerable attraction due to two reason, one is efficiency and another, it provides simple mechanism to handle non-response problem, Estavao and Särndal (2002). Hence, to maintain the popularity of calibration approach, there is urgent need to develop calibration estimators under two-phase sampling.

Corresponding author: Sadikul Islam
E-mail address: sadikul.islam@icar.gov.in

Few authors has already explored few applications of calibration under two phase sampling. For example, Estevao and Särndal (2002) proposed two step calibration when auxiliary information is available at two levels (two-phase). Singh (2004) improved the two-phase calibration method discussed by Hidiroglou and Särndal (1998). Estavao and Särndal (2006) proposed ten cases of auxiliary information for calibration in two-phase sampling. It has been observed that these authors were mostly restricted their focus on linear population parameters (For example, population total or mean etc.). Recently, it is found that everybody has interest for in-depth analysis of data,utilizing available high computational facilities, hence cannot be restricted applicability of calibration approach within linear point estimators but need to extend it for the popular non-linear complex estimators too. For example, finite population ratio is a popular non linear parameter which is consists of ratio of population total or mean of two variables and due to practical advantages many times, it is preferred over mean or total estimator, see for example, if the variable of interest is 'number of bullocks per acre of holding' in the population is a ratio of number of bullocks in a holding to area in acre holding (agricultural science), 'per capita monthly income' is ratio of sum of monthly income of house hold to the size of house hold. Similarly 'unemployment rate' is the ratio of number of unemployed individual to the number of individuals in the labour force in the country, see Islam *et al.* (2019).Few authors has discussed calibration approach for different non-linear population parameters under complex sampling design (except two-phase sampling) see, for example, Plikusas and Pumputis (2007, 2010), Farrel and Singh (2005), Kim and Park (2010), Sud *et al.* (2014), Basak *et al.* (2017). Islam *et al.* (2019) proposed calibration estimator of population ratio under two-phase sampling utilizing known ratio of auxiliary variables totals at first phase sample only and the optimization is done through using single calibration equation $\sum_{k=1}^{n_2}\omega_k x_{(y)k} \Big/ \sum_{k=1}^{n_2}\omega_k x_{(z)k} = \sum_{k=1}^{n_1} x_{(y)k} \Big/ \sum_{k=1}^{n_1} x_{(z)k}$, which make it less flexible to capture variation of both y and z variable, simultaneously. In this article,we elaborate two calibration approach of finite population ratio under two-phase sampling design using more than one calibration equation.

The rest of the article is organized as follows. In Section 2, we first introduced revised calibration ratio estimator under two-phase sampling under the situation when auxiliary variables are available at first-phase sample only. The variance and estimator of variance of proposed estimator, in addition optimum sample sizes of first and second sample for minimum variance were discussed in Sub-section 2.2. The empirical performances of different estimators were compared in Section 3, using model-based method with hypothetical population and design-based method with real population data. Finally, concluding remarks were set out in Section 4.

2. CALIBRATION RATIO ESTIMATION UNDER TWO-PHASE SAMPLING

Let Ω be a population of size N where, individual units are indexed by subscript $k = 1, 2...N$. Further, the population has two study variables y and z and its population ratio (denoted as R) y to z is our parameter of interest. The formula of population ratio is $R = t_y / t_z$, where $t_y = \sum_{k\in\Omega} y_k$ be the population total of y and $t_z = \sum_{k\in\Omega} z_k$ be the population total z. Further, consider that $x_{(y)}$ and $x_{(z)}$ be the two auxiliary variable correlated to y and z, respectively. It is assumed that the sampling frame for all the population units are known and per unit data collection cost of $x_{(y)}$ and $x_{(z)}$ is much cheaper compare to the y and z. Here, adopted data collection method was two-phase with a specified probability sampling design. The details of auxiliary variables at different level are as follows:

- *At the Population level: No auxiliary data available*
- *At the first-phase sampling level:* $x_{(y)k}$ *and* $x_{(z)k}$ *are known for all* $k \in s_{(1)}$
- *At the second-phase sampling level:* $x_{(y)k}$ *and* $x_{(z)k}$ *are known for all* $k \in s_{(2)}, \left[s_{(2)} \subset s_{(1)}\right]$

Here, $s_{(1)}$ denotes first-phase sample of size $n_{(1)}$ observing the auxiliary variables [$x_{(y)}$ and $x_{(z)}$] only. Further, $s_{(2)}$ denotes a sub-sample of $s_{(1)}$ of size $n_{(2)}, \left[n_{(2)} < n_{(1)}\right]$ and observed the study variables (y and z) along with auxiliary variables $x_{(y)}$ and $x_{(z)}$. The survey weight for unit k in the first-phase sample $s_{(1)}$ is denoted as $d_{(1)k} = \pi_{(1)k}^{-1}$, where $\pi_{(1)k} = \Pr\left[k \in s_{(1)}\right]$ $ð_{1i} = \Pr\left(i \in s_1\right)$ is the known the first-phase inclusion probability for unit k. Similarly, the survey weight for unit k in the second-phase sample $s_{(2)}$ is denoted as $d_{(2)k} = \pi_{k|s_{(1)}}^{-1}$, where $\pi_{k|s_{(1)}} = \Pr\left[k \in s_{(2)} \mid s_{(1)}\right]$ is the known second-phase conditional inclusion probability for unit *k*. Hence, the overall survey weight (or design weight) for unit *k* can be expressed as $d_k = d_{(1)k} d_{(2)k}$. In addition, $\pi_{(1)kl}$ denotes the joint inclusion probability of k^{th} and l^{th} unit of $s_{(1)}$; $\pi_{kl|s_{(1)}}$ denoted joint conditional inclusion probability of k^{th} and l^{th} unit of $s_{(2)}$. Following the notation of Särndal *et al.*(1992) and Arnab (2017), $\hat{R}_{s_{(2)}}$ be the conventional estimator of population ratio *R* under two-phase sampling design can be defined as

$$\hat{R}_{s_{(2)}} = \frac{\hat{t}_{y|s_{(2)}}}{\hat{t}_{z|s_{(2)}}}, \qquad (1)$$

where,

$$\hat{t}_{y|s_{(2)}} = \sum\nolimits_{k \in s_{(2)}} \left(\frac{y_k}{\pi_{(1)k}} \right) \frac{1}{\pi_{k|s_{(1)}}}, \quad \hat{t}_{z|s_{(2)}} = \sum\nolimits_{k \in s_{(2)}} \left(\frac{z_k}{\pi_{(1)k}} \right) \frac{1}{\pi_{k|s_{(1)}}}$$

and the approximate variance expression of $\hat{R}_{s_{(2)}}$ is

$$V\left(\hat{R}_{Est,R}\right) \simeq E\left[V\left(\hat{R}_{Est,R} \mid s_{(1)}\right)\right] + V\left[E\left(\hat{R}_{Est,R} \mid s_{(1)}\right)\right],$$

$$\simeq E\left[\frac{1}{\left(\hat{t}_{z|s_{(1)}}\right)^2} \sum\sum_{k<l} \left(\pi_{k|s_{(1)}} \pi_{l|s_{(1)}} - \pi_{k,l|s_{(1)}}\right) \left(\frac{u_k / \pi_{(1)k}}{\pi_{k|s_{(1)}}} - \frac{u_l / \pi_{(1)l}}{\pi_{l|s_{(1)}}} \right) + V\left(\frac{\hat{t}_{y|s_{(1)}}}{\hat{t}_{z|s_{(1)}}} \right)\right]$$

$$\simeq E\left[\frac{1}{\left(\hat{t}_{z|s_{(1)}}\right)^2} \sum\sum_{k<l} \left(\pi_{k|s_{(1)}} \pi_{l|s_{(1)}} - \pi_{k,l|s_{(1)}}\right) \left(\frac{u_k / \pi_{(1)k}}{\pi_{k|s_{(1)}}} - \frac{u_l / \pi_{(1)l}}{\pi_{l|s_{(1)}}} \right)\right] + \frac{1}{t_z^2} \sum \sum_{k<l \in \Omega} \left(\pi_{(1)k} \pi_{(1)l} - \pi_{(1)kl}\right) \left(\frac{v_k}{\pi_{(1)k}} - \frac{v_l}{\pi_{(1)l}} \right)^2$$

where $\hat{t}_{y|s_{(1)}} = \sum\nolimits_{k \in s_{(1)}} \frac{y_k}{\pi_{(1)k}}$ and $\hat{t}_{z|s_{(1)}} = \sum\nolimits_{k \in s_{(1)}} \frac{z_k}{\pi_{(1)k}}$,

where

$u_k = y_k - \hat{R}_{s_{(1)}} z_k$, $v_k = y_k - R z_k$, $\hat{R}_{s_{(1)}} = \hat{t}_{y|s_{(1)}} / \hat{t}_{z|s_{(1)}}$ and $k = l = 1, 2 \ldots N$. Here, though $\hat{R}_{s_{(2)}}$ is ratio of two linear unbiased estimator but it is biased as well as non-linear in nature, see Särndal *et al.* (1992). Further, Islam *et al.* (2019) described calibration estimator of population ratio under two-phase sampling utilizing known ratio of auxiliary variables totals at first phase sample only (denoted by $\hat{R}_C$) as

$$\hat{R}_C = \frac{\sum_{k=1}^{n_{(2)}} \omega_k y_k}{\sum_{k=1}^{n_{(2)}} \omega_k z_k} \qquad (2)$$

where,

$$\omega_k = d_k - \left\{ \sum_{k=1}^{n_{(2)}} d_k \left(x_{(y)k} - R^{(1)} x_{(z)k} \right) \Big/ \sum_{k=1}^{n_{(2)}} d_k \left(x_{(y)k} - R^{(1)} x_{(z)k} \right)^2 \right\} d_k \left(x_{(y)k} - R^{(1)} x_{(z)k} \right),$$

$R^{(1)} = \sum\nolimits_{k=1}^{n_{(1)}} x_{(y)k} \Big/ \sum\nolimits_{k=1}^{n_{(1)}} x_{(z)k}$. This calibration estimator was developed through usual approach of distance function minimization using single calibration equation $\sum\nolimits_{k=1}^{n_{(2)}} \omega_k x_{(y)k} \Big/ \sum\nolimits_{k=1}^{n_{(2)}} \omega_k x_{(z)k} = \sum\nolimits_{k=1}^{n_{(1)}} x_{(y)k} \Big/ \sum\nolimits_{k=1}^{n_{(1)}} x_{(z)k}$. In this article, keeping in view to improve the performance in estimation two revised calibration estimator of population ratio were discussed under two phase sampling when auxiliary variables are available for first phase sample only.

2.1 Calibration ratio estimator: Ratio of two calibration total estimator

Following Deville and Särndal (1992) the developed calibration ratio estimator (denoted as $\hat{R}_{RC1}$) under two-phase sampling given as

$$\hat{R}_{RC1} = \frac{\hat{t}_{y_{cal}}}{\hat{t}_{z_{cal}}}, \qquad (2)$$

Here, $\hat{t}_{y_{cal}}$ and $\hat{t}_{z_{cal}}$ are the calibration estimator of population total of *y* and *z* under two-phase sampling, with the formula, $\hat{t}_{y_{cal}} = \sum_{k \in s_{(2)}} w_{(y)k} y_k$ and $\hat{t}_{z_{cal}} = \sum_{k \in s_{(2)}} w_{(z)k} z_k$, respectively where, $w_{(y)k}$ and $w_{(z)k}$ are calibrated weights associated to *y* and *z*, respectively. The calibration weight, $w_{(y)k}$ and $w_{(z)k}$ were obtained by minimizing the distance between the design weights d_k and calibrated weights $w_{(y)k}$ and $w_{(z)k}$ subject to the constraints $\sum_{k=1}^{n_{(2)}} w_{(y)k} x_{(y)k} = \sum_{k=1}^{n_{(1)}} x_{(y)k} = t_{x_{(y)}}^{(1)}$ and $\sum_{k=1}^{n_{(2)}} w_{(z)k} x_{(z)k} = \sum_{k=1}^{n_{(1)}} x_{(z)k} = t_{x_{(z)}}^{(1)}$ individually. Here, loss functions used as $L_1 = \sum_{k=1}^{n_{(2)}} \frac{\left(w_{(y)k} - d_k\right)^2}{d_k}$ and $L_2 = \sum_{k=1}^{n_{(2)}} \frac{\left(w_{(z)k} - d_k\right)^2}{d_k}$. Here, the objective function used as: $\Phi = \sum_{k=1}^{n_{(2)}} \frac{\left(w_{(y)k} - d_k\right)^2}{d_k} + \sum_{k=1}^{n_{(2)}} \frac{\left(w_{(z)k} - d_k\right)^2}{d_k} + 2\lambda_1 \left(\sum_{k=1}^{n_{(2)}} w_{(y)k} x_{(y)k} - t_{x_{(y)}}^{(1)} \right) + 2\lambda_2 \left(\sum_{k=1}^{n_{(2)}} w_{(z)k} x_{(z)k} - t_{x_{(z)}}^{(1)} \right)$. Now, differentiating the function 'Φ' partially with respect to $w_{(y)k}$ and $w_{(z)k}$ independently and equating it to zero, get two function as

$$w_{(y)k} = d_k\left[1-\lambda_1 x_{(y)k}\right] \text{ and} \qquad (3)$$

$$w_{(z)k} = d_k[1-\lambda_2 x_{(z)k}] \qquad (4)$$

Multiplying equation (3) and (4) by $x_{(y)k}$ and $x_{(z)k}$, respectively and summing it over whole range of $k[k=1,2,...,n_{(2)}]$ gives $\lambda_1 = -\left(t^{(1)}_{x_{(y)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}\right)\Big/\sum_{k=1}^{n_{(2)}} d_k x^2_{(y)k}$ and $\lambda_2 = -\left(t^{(1)}_{x_{(z)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}\right)\Big/\sum_{k=1}^{n_{(2)}} d_k x^2_{(z)k}$. Putting λ_1 and λ_2 value to equation (3) and (4), respectively and get the two individual calibration weight for *y* and *z* as

$$w_{(y)k} = d_k + \frac{x_{(y)k} d_k\left(t^{(1)}_{x_{(y)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}\right)}{\sum_{k=1}^{n_{(2)}} d_k x^2_{(y)k}} \text{ and} \qquad (5)$$

$$w_{(z)k} = d_k + \frac{x_{(z)k} d_k\left(t^{(1)}_{x_{(z)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}\right)}{\sum_{k=1}^{n_{(2)}} d_k x^2_{(z)k}} \qquad (6)$$

Hence, putting the value of (5) and (6) to (2) get the revised calibration estimator $\hat{R}_{RC1}$ as

$$\hat{R}_{RC1} = \frac{\sum_{k\in s_{(2)}} d_k y_k \left\{1+\dfrac{x_{(y)k}\left(t^{(1)}_{x_{(y)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}\right)}{\sum_{k=1}^{n_{(2)}} d_k x^2_{(y)k}}\right\}}{\sum_{k\in s_{(2)}} d_k z_k \left\{1+\dfrac{x_{(z)k}\left(t^{(1)}_{x_{(z)}} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}\right)}{\sum_{k=1}^{n_{(2)}} d_k x^2_{(z)k}}\right\}}. \qquad (7)$$

Further, the $\hat{R}_{RC1}$ of the equation (7) re-expressed in simplified form as:

$$\hat{R}_{RC1} = \frac{\hat{t}_y + \left[t^{(1)}_{x_{(y)}} - \hat{t}_{x_{(y)}}\right]\hat{\beta}_y}{\hat{t}_z + \left[t^{(1)}_{x_{(z)}} - \hat{t}_{x_{(z)}}\right]\hat{\beta}_z}, \qquad (8)$$

where, $\hat{t}_y = \sum_{k=1}^{n_{(2)}} d_k y_k$, $\hat{t}_z = \sum_{k=1}^{n_{(2)}} d_k z_k$, $\hat{t}_{x_{(y)}} = \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}$, $\hat{t}_{x_{(z)}} = \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}$, $t^{(1)}_{x_{(y)}} = \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(y)k}$, $t^{(1)}_{x_{(z)}} = \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k}$, $\hat{\beta}_y = \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} y_k \Big/ \sum_{k=1}^{n_{(2)}} d_k x^2_{(y)k}$ and $\hat{\beta}_z = \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} z_k \Big/ \sum_{k=1}^{n_{(2)}} d_k x^2_{(z)k}$.

The advantages of this estimator was, based on two separate calibration weight, but instead has the problem of calculating calibration weights without considering the correlation value between y and z variable.

2.2 Proposed revised calibration weight-based ratio estimator

In this subsection, we proposed combined calibration weight-based ratio estimator which is consists of single common calibration weight for both numerator (*y*) and denominator (*z*) associated to ratio estimator. Following Deville and Särndal (1992) the proposed calibration estimator (denoted as $\hat{R}_{RC2}$) of finite population ratio is defined as

$$\hat{R}_{RC2} = \frac{\sum_{k\in s_{(2)}} w_k y_k}{\sum_{k\in s_{(2)}} w_k z_k}, \qquad (9)$$

where, w_k is the common calibrated weight of *y* and *z*. Here, w_k is determined by minimizing the distance between the original weights d_k and the revised weights w_k, considering the loss function $\sum_{k=1}^{n_{(2)}} \frac{\left(w_k - d_k\right)^2}{d_k}$ subject to the constraints $\sum_{k=1}^{n_{(2)}} w_{(y)k} x_{(y)k} = \sum_{k=1}^{n_{(1)}} x_{(y)k} = t^{(1)}_{x_{(y)}}$ and $\sum_{k=1}^{n_{(2)}} w_{(z)k} x_{(z)k} = \sum_{k=1}^{n_{(1)}} x_{(z)k} = t^{(1)}_{x_{(z)}}$. The objective function used: $\xi = \sum_{k=1}^{n_{(2)}} \frac{\left(w_k - d_k\right)^2}{d_k} + 2\phi_1\left(\sum_{k=1}^{n_{(2)}} w_k x_{(y)k} - t^{(1)}_{x_{(y)}}\right) + 2\phi_2\left(\sum_{k=1}^{n_{(2)}} w_k x_{(z)k} - t^{(1)}_{x_{(z)}}\right)$. Now, differentiating ξ partially with respect to 'w_k' and equating to zero, get the equation

$$w_k = d_k - 2\phi_1 d_k x_{(y)k} - 2\phi_2 d_k x_{(z)k} \qquad (10)$$

Further, multiplying (10) by $x_{(y)k}$ and $x_{(z)k}$ separately and summing it over $k[k=1,2,...,n_{(2)}]$, get the equations

$$\sum_{k=1}^{n_{(2)}} w_k x_{(y)k} = \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} - \phi_1 \sum_{k=1}^{n_{(2)}} d_k x^2_{(y)k} - \phi_2 \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \qquad (11)$$

$$\sum_{k=1}^{n_{(2)}} w_k x_{(z)k} = \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} - \phi_1 \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} - \phi_2 \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \tag{12}$$

By solving the equation (11) and (12), get the solutions as:

$$\phi_1 = \frac{\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right)}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2}$$

and

$$\phi_2 = \frac{\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right)}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2}$$

Putting the value of ϕ_1 and ϕ_2 in equation (10), get the final calibration weight as:

$$w_k = d_k \left(\begin{array}{l} 1 + \dfrac{x_{(y)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} + \\ \dfrac{x_{(z)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} \end{array} \right) \tag{13}$$

Substituating the value of (13) to (9) get the expression of calibration estimator as

$$\hat{R}_{RC.2} = \frac{\sum_{k\in s_{(2)}} d_k y_k \left(\begin{array}{l} 1 + \dfrac{x_{(y)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} + \\ \dfrac{x_{(z)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} \end{array} \right)}{\sum_{k\in s_{(2)}} d_k z_k \left(\begin{array}{l} 1 + \dfrac{x_{(y)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} + \\ \dfrac{x_{(z)k} \left[\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \left(t_{x_{(z)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} \right) - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \left(t_{x_{(y)}}^{(1)} - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} \right) \right]}{\left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2} \end{array} \right)}, \tag{14}$$

Again, $\hat{R}_{RC2}$ (14) can also be re-expressed in simplified form as

$$\hat{R}_{RC2} = \frac{\hat{t}_y + \left(t_{x_{(z)}}^{(1)} - \hat{t}_{x_{(z)}} \right) l_1 + \left(t_{x_{(y)}}^{(1)} - \hat{t}_{x_{(y)}} \right) l_2}{\hat{t}_z + \left(t_{x_{(z)}}^{(1)} - \hat{t}_{x_{(z)}} \right) l_3 + \left(t_{x_{(y)}}^{(1)} - \hat{t}_{x_{(y)}} \right) l_4}, \tag{15}$$

where, $\hat{t}_y = \sum_{k=1}^{n_{(2)}} d_k y_k$, $\hat{t}_z = \sum_{k=1}^{n_{(2)}} d_k z_k$, $\hat{t}_{x_{(y)}} = \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}$, $\hat{t}_{x_{(z)}} = \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}$, $t_{x_{(y)}}^{(1)} = \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(y)k}$, $t_{x_{(z)}}^{(1)} = \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k}$,

$$q = \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 \right)\left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 \right) - \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right)^2 ,$$

$$l_1 = \left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k} y_k \sum_{k=1}^{n_{(2)}} d_k y_k^2 - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} y_k \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} y_k \right) \Big/ q ,$$

$$l_2 = \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} y_k \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} y_k \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right) \Big/ q ,$$

$$l_3 = \left(\sum_{k=1}^{n_{(2)}} d_k x_{(z)k} z_k \sum_{k=1}^{n_{(2)}} d_k x_{(y)k}^2 - \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} z_k \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right) \Big/ q \quad \textbf{and}$$

$$l_4 = \left(\sum_{k=1}^{n_{(2)}} d_k x_{(y)k} z_k \sum_{k=1}^{n_{(2)}} d_k x_{(z)k}^2 - \sum_{k=1}^{n_{(2)}} d_k x_{(z)k} z_k \sum_{k=1}^{n_{(2)}} d_k x_{(y)k} x_{(z)k} \right) \Big/ q .$$

2.2.1 Variance estimation of the proposed $\hat{R}_{RC2}$ estimator

In this sub-subsection, expression of variance and estimator of variance of $\hat{R}_{RC2}$ was discussed. Following, Särndal *et al.* (1992) the expression of approximate variance of $\hat{R}_{RC2}$ given as,

$$\begin{aligned} V\left(\hat{R}_{RC2}\right) &\simeq V_1 E_2\left[\hat{R}_{RC2}\right] + E_1 V_2\left[\hat{R}_{RC2}\right], \\ &\simeq \frac{1}{t_z^2}\left[\sum_{k=1}^{N}\sum_{l=1}^{N} \Delta_{kl}^{(1)} \frac{u_k u_l}{\pi_{(1)k}\pi_{(1)l}}\right] + \\ &\qquad E_1\left[\frac{1}{(t_z^{(1)})^2}\left(\sum_{k=1}^{n_{(1)}}\sum_{l=1}^{n_{(1)}} \Delta_{kl|s_{(1)}} \frac{v_k}{\pi_{(1)k}\pi_{k|s_{(1)}}} \frac{v_l}{\pi_{(1)l}\pi_{l|s_{(1)}}}\right)\right], \end{aligned} \tag{16}$$

where, $t_z = \sum_{k=1}^{N} z_k$, $\Delta_{kl}^{(1)} = \pi_{(1)kl} - \pi_{(1)k}\pi_{(1)l}$,

$u_k = y_k - Rz_k$, $t_z^{(1)} = \sum_{k=1}^{n_{(1)}} d_{(1)k} z_k$, $\Delta_{kl|s_{(1)}} = \pi_{kl|s_{(1)}} - \pi_{k|s_1}\pi_{l|s_{(1)}}$,

$$v_k = y_k - R^{(1)} z_k + \left[l_3^{(1)} R^{(1)} - l_1^{(1)} \right] x_{(z)k} + \left[l_4^{(1)} R^{(1)} - l_2^{(1)} \right] x_{(y)k} ,$$

$$R^{(1)} = \sum_{k=1}^{n_{(1)}} d_{(1)k} y_k \Big/ \sum_{k=1}^{n_{(1)}} d_{(1)k} z_k ,$$

$$l_1^{(1)} = \left(\sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k} y_k \sum_{k=1}^{n_{(1)}} d_{(1)k} y_k^2 - \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(y)k} y_k \sum_{k=1}^{n_{(1)}} d_k x_{(y)k} x_{(z)k} \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k} y_k \right) \Big/ q_{(1)} ,$$

$$l_2^{(1)} = \left(\sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(y)k} y_k \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k}^2 - \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(z)k} y_k \sum_{k=1}^{n_{(1)}} d_{(1)k} x_{(y)k} x_{(z)k} \right) \Big/ q_{(1)} ,$$

$$l_3^{(1)}=\left(\sum_{k=1}^{n_2}d_{(1)k}x_{(z)k}z_k\sum_{k=1}^{n_2}d_{(1)k}x_{(y)k}^2-\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}z_k\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}x_{(z)k}\right)\Big/q_{(1)},$$

$$l_4^{(1)}=\left(\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}z_k\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(z)k}^2-\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(z)k}z_k\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}x_{(z)k}\right)\Big/q_{(1)}$$

and $q_{(1)}=\left(\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}^2\right)\left(\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(z)k}^2\right)-\left(\sum_{k=1}^{n_{(1)}}d_{(1)k}x_{(y)k}x_{(z)k}\right)^2.$

Here, E_1 and E_2 define the expected value over all possible first phase sample $s_{(1)}$ and all possible second phase sample given first phase sample $s_{(1)}$, respectively. V_1 and V_2 define the variance over all possible first phase sample and all possible second phase sample given first phase sample, respectively. Again, following Särndal *et al.* (1992), the approximate estimator of variance of $\hat{R}_{RC2}$ is given as

$$\hat{V}\left(\hat{R}_{RC2}\right)=\frac{1}{\hat{t}_z^2}\left[\sum_{k=1}^{n_{(1)}}\sum_{l=1}^{n_{(1)}}\frac{\Delta_{kl}^{(1)}}{\pi_{(1)kl}}\frac{\hat{u}_k\hat{u}_l}{\pi_{(1)k}\pi_{(1)l}}\right]$$
$$+\left[\frac{1}{\hat{t}_z^2}\left(\sum_{k=1}^{n_{(1)}}\sum_{l=1}^{n_{(1)}}\Delta_{kl|s_1}\frac{\hat{v}_k}{\pi_{(1)k}\pi_{k|s_1}}\frac{\hat{v}_l}{\pi_{(1)l}\pi_{l|s_1}}\right)\right], \qquad (17)$$

where, $\hat{t}_z=\sum_{k=1}^{n_{(2)}}d_kz_k$, $\hat{u}_k=y_k-\hat{R}z_k$,

$\hat{v}_k=y_k-\hat{R}z_k+\left(l_3\hat{R}-l_1\right)b_k+\left(l_4\hat{R}-l_2\right)x_{(y)k}$,

$$l_1=\left(\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}y_k\sum_{k=1}^{n_{(2)}}d_ky_k^2-\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}y_k\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}x_{(z)k}\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}y_k\right)\Big/q,$$

$$l_2=\left(\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}y_k\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}^2-\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}y_k\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}x_{(z)k}\right)\Big/q,$$

$$l_3=\left(\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}z_k\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}^2-\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}z_k\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}x_{(z)k}\right)\Big/q,$$

$$l_4=\left(\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}z_k\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}^2-\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}z_k\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}x_{(z)k}\right)\Big/q$$

and $q=\left(\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}^2\right)\left(\sum_{k=1}^{n_{(2)}}d_kx_{(z)k}^2\right)-\left(\sum_{k=1}^{n_{(2)}}d_kx_{(y)k}x_{(z)k}\right)^2.$

In particular, under simple random sampling (SRS) design, the inclusion probabilities of first and second phase sample were considered as $\pi_{(1)k}=\pi_{(1)l}=\frac{n_{(1)}}{N}$, $\pi_{(1)kl}=\frac{n_{(1)}[n_{(1)}-1]}{N(N-1)}$, $\pi_{kl|s_{(1)}}=\frac{n_{(2)}\left[n_{(2)}-1\right]}{n_{(1)}\left[n_{(1)}-1\right]}$ and $\pi_{k|s_{(1)}}=\pi_{l|s_{(1)}}=n_{(2)}/n_{(1)}$. Hence, the approximate estimator of variance of $\hat{R}_{RC2}$ under SRS design was given as

$$\hat{V}\left(\hat{R}_{RC2}\right)_{SRS}=\left[\frac{N\left(N-n_{(1)}\right)}{\hat{t}_z^2n_{(1)}}\hat{S}_u^2\right]+\left[\frac{1}{\hat{t}_z^2}\frac{N^2\left(n_{(1)}-n_{(2)}\right)}{n_{(1)}n_{(2)}}\hat{S}_v^2\right], \qquad (18)$$

where, $\hat{S}_u^2=\hat{S}_y^2+\hat{R}^2\hat{S}_z^2-2\hat{R}\hat{\rho}\hat{S}_y\hat{S}_z$,

$\hat{v}_k=y_k-\hat{R}z_k+\left(l_3\hat{R}-l_1\right)b_k+\left(l_4\hat{R}-l_2\right)x_{(y)k}$,

$\hat{S}_v^2=\left\{n_{(2)}-1\right\}^{-1}\sum_{k=1}^{n_{(2)}}(\hat{v}_k-\overline{\hat{v}})^2$ and $\overline{\hat{v}}=n_{(2)}^{-1}\sum_{k=1}^{n_{(2)}}\hat{v}_k$. Here, $\hat{S}_y^2$ and $\hat{S}_z^2$ denote the sample mean square of *y* and *z* under SRS and $\hat{\rho}$ is estimate of correlation coefficient between *y* and *z*.

2.2.2 Determination of optimum sample size under fixed cost scenario

Again, the efforts were further given to find out the optimum sample size for first phase and second phase sampling under a fixed overall cost (denoted as C_{fix}) that minimizes the approximate variance (18). The expression of objective function used to minimize (18) as:

$$\delta=\left[\frac{N\left\{N-n_{(1)}\right\}}{t_z^2n_{(1)}}S_u^2\right]+\left[\frac{1}{t_z^2}\frac{N^2\left\{n_{(1)}-n_{(2)}\right\}}{n_{(1)}n_{(2)}}S_v^2\right]+$$
$$\lambda\left[n_{(1)}C_{(1)}+n_{(2)}C_{(2)}-C_{fix}\right]$$

Here, per unit cost of data collection for first-phase and second-phase sample was denoted by notation $C_{(1)}$ and $C_{(2)}$, respectively. First order differentiation of the function δ were done with respect to $n_{(1)}$ and $n_{(2)}$,separately and equating it to zero and we got optimum value of $n_{(1)}$ and $n_{(2)}$ that minimizes the variance estimator (18) for fixed cost C_{fix} as

$$n_{(1)}^{opt}=\frac{C_{fix}\sqrt{S_u^2}}{C_{(2)}\sqrt{S_v^2}+\sqrt{C_{(2)}C_{(1)}\left(S_u^2-S_v^2\right)}}\text{and} \qquad (19)$$

$$n_{(2)}^{opt}=\frac{C_{fix}\sqrt{\left(S_u^2-S_v^2\right)}}{\sqrt{C_{(2)}C_{(1)}S_v^2}+C_{(1)}\sqrt{\left(S_u^2-S_v^2\right)}} \qquad (20)$$

Here, $S_u^2=S_y^2+R^2S_z^2-2R\rho S_yS_z$,

$S_v^2=\frac{1}{(N-1)}\sum_{k=1}^{N}\left(v_k-\overline{v}\right)^2$, $\overline{v}=\sum_{k=1}^{N}v_k$ and $v_k=y_k-R^{(1)}z_k$.

Putting the value of (19) and (20) to (18), achieved minimum variance expression of the proposed estimator $\hat{R}_{RC2}$ for a fixed cost under SRS.

3. EMPIRICAL EVALUATION

In this section, performance evaluation of the developed estimators was done through two types of simulation studies, one was model-based and another was design-based. In model-based simulation studies the population data used were generated through hypothetical population model, in contrary design-based simulation studies based on real population data. Here, we used four estimators for performance comparison are given below:

- Simple ratio estimator $\hat{R}_{s_{(2)}}$ discussed in equation (1)
- $\hat{R}_C$ of Islam *et al.* (2019), discussed in equation (2)
- Calibration estimator $\hat{R}_{RC1}$ described in equation (8)
- Calibration estimator $\hat{R}_{RC2}$ discussed in equation (15)

Two performance evaluation criteria used for comparison were percentage absolute relative bias (ARB) and percentage relative root mean squared error (RRMSE), defined by

$$ARB(\hat{R}) = \frac{1}{H}\sum_{h=1}^{H}\left|\frac{\hat{R}_h - R}{R}\right| \times 100 \text{ and}$$

$$RRMSE(\hat{R}) = \sqrt{H^{-1}\sum_{h=1}^{H}\left(\frac{\hat{R}_h - R}{R}\right)^2} \times 100.$$

Here $\hat{R}_h$ denotes the estimated value of the population ratio for the h^{th} simulation, *R* denotes true population ratio value and *H* denotes the total number of simulation.

3.1 Model-based simulation

In model based simulation, population data were generated through multivariate normal distribution with mean vector, ì $=\left(\mu_y, \mu_z, \mu_{x_{(y)}}, \mu_{x_{(z)}}\right)$=(25,5,40,20) and covariance matrix was

$$\text{Ó} = \begin{bmatrix} \sigma_{yy} & \sigma_{yz} & \sigma_{yx_{(y)}} & \sigma_{yx_{(z)}} \\ \sigma_{zy} & \sigma_{zz} & \sigma_{zx_{(y)}} & \sigma_{zx_{(z)}} \\ \sigma_{x_{(y)}y} & \sigma_{x_{(y)}z} & \sigma_{x_{(y)}x_{(y)}} & \sigma_{x_{(y)}x_{(z)}} \\ \sigma_{x_{(z)}y} & \sigma_{x_{(z)}z} & \sigma_{x_{(z)}x_{(y)}} & \sigma_{x_{(z)}x_{(z)}} \end{bmatrix}$$

. Here, elements of **Ó** represented covariance between the variables and values were taken in such a way that it will provide different correlation value among the variables y, z, $x_{(y)k}$ and $x_{(z)k}$. Here, four set of covariance matrix **Ó** value were considered so that correlation between y and z [denoted as $\rho(y,z)$] are 0.1 and 0.5 within correlation value between y and z with associated auxiliary variables ( ) and $x_{(z)}$ [denoted as $\rho(y,x_{(y)})$ and $\rho(z,x_{(z)})$, respectively] as 0.5 and 0.7, respectively.

Further, for each correlation value four sample size combination of first phase sample and second phase sample were used as $(n_{(1)},n_{(2)})=$ (250, 25), (250, 50), (300, 25) and (300, 50). Here, population size *N* assumed to be 1000. The number of simulation performed *H* =5000.

The values of ARB and RRMSE of different estimators under model based simulations for different correlation between *y*, *z*, $x_{(y)}$ and $x_{(z)}$under different sample size combinations were summarized in Table 1. For (1) 250, $n_{(2)}=50$, $\rho\left(y,x_{(y)}\right)=\rho\left(z,x_{(z)}\right)=0.5$ and $\rho\left(y,z\right)=0.1$ the $\hat{R}_{RC2}$ has minimum ARB and RRMSE value followed by $\hat{R}_{RC1}$ then $\hat{R}_C$ and lastly $\hat{R}_{s_{(2)}}$. Further, fixing the value of $n_{(1)}=250$, $n_{(2)}=50$ and $\rho\left(y,x_{(y)}\right)=\rho\left(z,x_{(z)}\right)=0.5$, with increase in $\rho\left(y,z\right)$ value from 0.1 to 0.5, decreases the ARB and RRMSE value of all the estimators but still $\hat{R}_{RC2}$ is out performer followed by $\hat{R}_{RC1}$ over $\hat{R}_C$ and $\hat{R}_{s_{(2)}}$.

Similarly, for $n_{(1)}=250$, $n_{(2)}=50$, $\rho\left(y,x_{(y)}\right)=\rho\left(z,x_{(z)}\right)=0.7$ and $\rho\left(y,z\right)=0.1$, it was observed that the $\hat{R}_{RC2}$ has minimum ARB and RRMSE value followed by $\hat{R}_{RC1}$ over $\hat{R}_C$ and $\hat{R}_{s_{(2)}}$ and further increase in $\rho\left(y,z\right)$ value from 0.1 to 0.5 decreases the ARB and RRMSE value for all the estimators but still $\hat{R}_{RC2}$ was out performer followed by $\hat{R}_{RC1}$. Further, for a fixed first phase sample value $n_{(1)}=250$ with decrease in second phase sample size from 50 to 25, the ARB and RRMSE value of $\hat{R}_{s_{(2)}}$ was increasing in a faster rate than $\hat{R}_C$ and $\hat{R}_{RC1}$ followed by $\hat{R}_{RC2}$ for all the correlation coefficient combination. Again, for a fixed value of second phase sample size either at $n_{(2)}=25$ or 50, increase in sample size of first phase sample from $n_{(1)}=250$ to 300 for all correlation combinations, it was observed that ARB and RRMSE of .., $\hat{R}_{RC2}$ and $\hat{R}_C$ were decreased further but $\hat{R}_{s_{(2)}}$ maintain the same vale, but still $\hat{R}_{RC2}$ was outperformer over $\hat{R}_{RC1}$. Hence, Table1 overall results show that $\hat{R}_{RC2}$ has minimum ARB and RRMSE value followed by $\hat{R}_{RC1}$ over $\hat{R}_C$ and $\hat{R}_{s_{(2)}}$ for all the sample size combination over different correlation coefficient value.

3.2 Design-based simulation

Here, design-based simulation study was performed to support the validity of the findings of model based simulation. Design-based simulation was performed using real population data MU284 (Särndal *et al.*, 1992). MU284 population has284 municipalities of Sweden. For analysis purpose we considered only four variables of MU284, where 1985 population (P85, in thousands) and revenues from the 1985 Municipal taxation (RMT85, measured in millions of kronor) were assumed to be study variable and 1975 population (P75, in thousands) as the auxiliary variable to P85 and the variable number of municipal employees in 1984 (ME84) as the auxiliary variable to RMT85. Here, our aim was to estimate population ratio of P85 to RMT85. The correlations between the four variables are depicted in Table 2. From the MU284 population data a first phase sample of size $_{(1)}$ =100 was taken using SRS and further second phase sample of three different size $n_{(2)}$ =25, 50and 75were selected from the first-phase sample using SRS. Further, $n_{(2)}$ = 25, 50 and 75 were selected within $n_{(1)}$ =125 and 150, independently, following same the sampling method. The results of design based simulations were depicted in Table 3.

In Table3 for $n_{(1)}$ = 100 and $n_{(2)}$ =75 results found that has minimum ARB and RRMSE value followed by $\hat{R}_{RC1}$ over and $\hat{R}_{s_{(2)}}$. Again for a fixed value of $n_{(1)}$ = 100, with decreasing in second-phase sample size $n_{(2)}$ from 75 to 50 and further reduction to 25, highlighted that ARB and RRMSE value of $\hat{R}_{s_{(2)}}$ estimator, increase drastically compare to the $\hat{R}_{C}$ and $\hat{R}_{RC1}$. Furter, $\hat{R}_{RC2}$ maintained its lower ARB and RRMSE value over the $\hat{R}_{C}$ and $\hat{R}_{RC1}$. Again, for a fixed second phase sample size to 25 and 50, increases in first phase sample size from 100 to 125 and further 150, we observed that both the calibration estimator $\hat{R}_{C}$, $\hat{R}_{RC1}$ and $\hat{R}_{RC2}$ gain in efficiency with diminishing the ARB as well as RRMSE value, though $\hat{R}_{RC2}$ maintain its lowest value over $\hat{R}_{C}$ and $\hat{R}_{RC1}$. Hence, Table3 results showed that $\hat{R}_{RC2}$ has lowest ARB as well as RRMSE value followed by $\hat{R}_{RC1}$ over $\hat{R}_{C}$ and $\hat{R}_{s_{(2)}}$. Hence, we can conclude that design-based simulation results of Table3 supportedthe findings of model-based simulation. Hence, $\hat{R}_{RC2}$ was considered as the best performing estimator.

4. CONCLUDING REMARKS

This research article has proposed the revised calibration weight based estimator of finite population ratio under two-phase sampling when auxiliary variables are available at unit level for first-phase sample and its theoretical approximate variance expression as well as variance estimator were also developed. Optimum sample sizes at first and second-phase sample of the proposed calibration estimator $\hat{R}_{RC2}$ was also achieved that will provide us a minimum variance for a fixed total cost. Finally, our theoretical findings were validated through empirical evaluation studies that confirms that $\hat{R}_{RC2}$ is best performing ratio estimator.

REFERENCES

Arnab, R (2017): Survey Sampling Theory and Applications. Academic Press, Oxford, U.K.

Basak, P., Sud, U.C. and Chandra, H. (2017). Calibration estimation of regression coefficient for

two-stage sampling design. *J. Ind. Soc. Agril. Statist.*, **71(1)**, 1-6.

Devaud, D. and Tille, Y. (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem, (invited paper). doi.org/10.1007/s11749-019-00681-3.

Deville, J.C. and Särndal, C.E. (1992).Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

Estavao, V. M. and Sarndal, C.E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, **74**, 127-147.

Estavao, V.M. and Sarndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling, *J. Off. Statist.*, **18**, 233-255.

Farrell, P.J. and Singh, S (2005). Model-assisted higher order calibration of estimators of variance. *Australia and New Zealand Journal of Statistics*, **47(3)**, 375-383.

Hidiroglou, M.A. and Särndal, C.E. (1998). Use of auxiliary information for two-phases ampling. *Survey Methodology*, **24**, 11-20.

Islam, S., Chandra, H., Sud, U.C., Guha, S and Basak, P. (2019). Calibration approach for estimation of population ratio under double sampling, *J. Ind. Soc. Agril. Statist.*, **73(1)**, 23-29.

Kim, J.K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, **78**, 21-39.

Legg, J.C. and Fuller, W.A. (2009). Handbook of Statistics: Chapter3-Two-Phase Sampling, 29(A), 55-70. doi.org/10.1016/S0169-7161(08)00003-5.

Plikusas, A. and Pumputis, D. (2007). Calibrated estimators of the population covariance. *ActaAppllicationMathematicae*, **97**, 177-187.

Plikusas, A. and Pumputis, D. (2010). Estimation of the finite population covariance using calibration. *Lithuanian Mathematical Journal*, **15(3)**, 325-340.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). Model assisted survey sampling, Springer Verlag, New York.

Singh, S. (2004). Repair of two-phase calibration methodology in survey sampling, *SSC Annual*

Meeting Proceedings of the Survey Methods Section, 115-122.

Sud, U.C., Chandra, H. and Gupta, V.K. (2014). Calibration based product estimator in single and two phase sampling. *Journal of Statistical Theory and Practice*, **8(1)**, 1-11.

Wu, C. and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling, *J. Off. Statist.*, **19(2)**, 119-131.

Table 1. Values of percentage absolute relative biases (ARB) and percentage relative root mean squared errors (RRMSE) of the three estimators under $n_{(2)}=25$ and $n_{(2)}=50$ within $n_{(1)}=250$ and $n_{(1)}=300$, respectively. Each sample size pair contains $\rho(y,z)=0.1$ and 0.5 under $\rho(y,x_{(y)})=\rho(z,x_{(z)})=0.5$ and 0.7. Here, ρ denotes correlation coefficient value

Sample Size	**Estimator**	$\rho(y,x_{(y)})=\rho(z,x_{(z)})=0.5$				$\rho(y,x_{(y)})=\rho(z,x_{(z)})=0.7$			
		$\rho(y,z)=0.1$				$\rho(y,z)=0.1$		$\rho(y,z)=0.5$	
		ARB	**RRMSE**	**ARB**	**RRMSE**	**ARB**	**RRMSE**	**ARB**	**RRMSE**
$n_{(1)}=250$ $n_{(2)}=50$	$\hat{R}_{s_{(2)}}$	2.26	2.82	2.10	2.62	2.96	3.72	2.07	2.57
	$\hat{R}_{CAL}$	2.10	2.60	1.98	2.50	2.55	3.03	1.81	2.27
	$\hat{R}_{RC1}$	2.04	2.56	1.91	2.38	2.36	2.97	1.76	2.19
	$\hat{R}_{RC2}$	1.96	2.48	1.80	2.26	2.02	2.53	1.56	1.94
$n_{(1)}=250$ $n_{(2)}=25$	$\hat{R}_{s_{(2)}}$	3.33	4.12	2.95	3.71	3.21	4.06	3.01	3.74
	$\hat{R}_{CAL}$	3.10	3.82	2.61	3.30	2.79	3.50	2.66	3.33
	$\hat{R}_{RC1}$	3.06	3.77	2.53	3.19	2.70	3.39	2.57	3.18
	$\hat{R}_{RC2}$	2.98	3.65	2.48	3.14	2.46	3.10	2.24	2.78
$n_{(1)}=300$ $n_{(2)}=50$	$\hat{R}_{s_{(2)}}$	2.25	2.83	2.09	2.64	2.10	2.63	2.96	3.72
	$\hat{R}_{CAL}$	2.09	2.63	1.90	2.36	1.79	2.30	2.44	3.12
	$\hat{R}_{RC1}$	2.00	2.51	1.79	2.25	1.72	2.15	2.36	2.97
	$\hat{R}_{RC2}$	1.94	2.43	1.76	2.21	1.52	1.91	2.02	2.53
$n_{(1)}=300$ $n_{(2)}=25$	$\hat{R}_{s_{(2)}}$	3.32	4.23	2.89	3.66	2.87	3.78	2.96	3.72
	$\hat{R}_{CAL}$	2.99	3.81	2.49	3.22	2.55	3.35	2.41	3.09
	$\hat{R}_{RC1}$	2.92	3.71	2.44	3.10	2.47	3.26	2.36	2.97
	$\hat{R}_{RC2}$	2.86	3.62	2.41	3.06	2.41	3.17	2.02	2.53

Table 2. Correlation coefficient value between four variables (RMT85, P85, ME84 and P75) of MU284 (Särndal *et al.*, 1992)

Variables	RMT85	P85	ME84	P75
RMT85	1	0.961	0.999	0.967
P85	0.961	1	0.965	0.998
ME84	0.999	0.965	1	0.971
P75	0.967	0.998	0.971	1

Table 3. Values of percentageabsoluterelative biases (ARB), percentagerelative root mean squared errors (RRMSE) and percentagerelative efficiencies (RE) of the two estimators under $n_{(2)} = 25$, $n_{(2)} = 50$ and $n_{(2)} = 75$ within $n_{(1)} = 100$, $n_{(1)} = 125$ and $n_{(1)} = 150$ respectively from design based simulations using the MU284 population data

First phase sample	Second phase sample	Estimator	ARB	RRMSE
	$_{(2)}$ 75	$\hat{R}_{s_{(2)}}$	7.57	8.71
		$\hat{R}_{CAL}$	7.41	8.66
$n_{(1)} = 100$		$\hat{R}_{RC1}$	7.36	8.62
		$\hat{R}_{RC2}$	6.05	7.05
	$n_{(2)} = 50$	$\hat{R}_{s_{(2)}}$	9.37	10.48
		$\hat{R}_{CAL}$	7.39	8.61
		$\hat{R}_{RC1}$	7.31	8.56
		$\hat{R}_{RC2}$	6.04	7.03
	$n_{(2)} = 25$	$\hat{R}_{s_{(2)}}$	11.62	12.81
		$\hat{R}_{CAL}$	7.39	8.88
		$\hat{R}_{RC1}$	7.11	8.47
		$\hat{R}_{RC2}$	6.29	7.36
$n_{(1)} = 125$	$n_{(2)} = 75$	$\hat{R}_{s_{(2)}}$	7.54	8.66
		$\hat{R}_{CAL}$	6.31	7.40
		$\hat{R}_{RC1}$	6.19	7.34
		$\hat{R}_{RC2}$	5.41	6.28
	$n_{(2)} = 50$	$\hat{R}_{s_{(2)}}$	9.27	10.43
		$\hat{R}_{CAL}$	6.68	7.79
		$\hat{R}_{RC1}$	6.19	7.34
		$\hat{R}_{RC2}$	5.54	6.42
	$n_{(2)} = 25$	$\hat{R}_{s_{(2)}}$	11.79	13.00
		$\hat{R}_{CAL}$	6.50	7.60
		$\hat{R}_{RC1}$	6.19	7.52
		$\hat{R}_{RC2}$	6.03	7.12
$n_{(1)} = 150$	$n_{(2)} = 75$	$\hat{R}_{s_{(2)}}$	7.75	8.83
		$\hat{R}_{CAL}$	5.45	6.50
		$\hat{R}_{RC1}$	5.34	6.35
		$\hat{R}_{RC2}$	5.00	5.76
	$n_{(2)} = 50$	$\hat{R}_{s_{(2)}}$	9.53	10.56
		$\hat{R}_{CAL}$	5.60	6.61
		$\hat{R}_{RC1}$	5.29	6.39
		$\hat{R}_{RC2}$	5.17	5.99
	$n_{(2)} = 25$	$\hat{R}_{s_{(2)}}$	11.78	13.05
		$\hat{R}_{CAL}$	5.80	6.98
		$\hat{R}_{RC1}$	5.36	6.63
		$\hat{R}_{RC2}$	5.69	6.73