https://doi.org/10.56093/JISAS.V78I1.4
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 78(1) 2024 29–35

Calibration Estimation of Population Total in Two-Stage Sampling Design under unavailability of Population Level Auxiliary Information for the selected PSUs

Pradip Basak[1], Kaustav Aditya[2] and Deepak Singh[2]

[1]*Uttar Banga Krishi Viswavidyalaya, Cooch Behar*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

Received 05 October 2023; Revised 19 February 2024; Accepted 19 March 2024

SUMMARY

Calibration approach is a popular technique in sample surveys which incorporates auxiliary information in the estimation process assuming that population aggregates of auxiliary variable are available. Many often under two-stage sampling design, such population aggregates of auxiliary variable, i.e., population mean or total are unavailable and under such situations, estimation of population total has been limited to the use of two phase sampling. Therefore, in the present study, efficient estimator of population total is developed under two-stage sampling design when population aggregates of auxiliary variable are unavailable for the selected psu's. The calibrated estimator is developed using the information on known population aggregates of additional auxiliary variable which is less linearly related to the study variable through two step calibration approach. The approximate variance and variance estimator of the proposed calibrated estimator has also been developed. Empirical evaluations using both real and simulated data shows the superior performance of the developed calibrated estimator in comparison to the existing estimators.

Keywords: Auxiliary information; Two step calibration; Two-Stage Sampling.

1. INTRODUCTION

Two stage sampling is the simplest case of multistage sampling which is widely used in large scale surveys since at most of the situations either the sampling frame is unavailable or it could be too expensive to construct such frames. Under two stage sampling, at the first stage, clusters are selected which are referred to as primary stage units or psu's and, at the second stage, a sample of basic elements are drawn from the selected psu's which are called as secondary stage units or ssu's. For example, in consumer expenditure surveys, villages can be considered as psu's and households can be considered as ssu's. Under two-stage sampling design, use of auxiliary information usually improves the estimator of population total (Sukhatme *et al.*, 1984). Calibration is a widely used approach in sample surveys to produce efficient estimators of population parameters at the estimation stage by incorporating axillary information (Deville and Särndal, 1992). Aditya *et al.* (2016) applied calibration approach for estimation of population total in two-stage sampling under the assumption that the population level auxiliary information is available at psu level. Mourya *et al.* (2016) developed calibration estimator for finite population total under two-stage sampling when the auxiliary information is available at the element level for the selected first-stage units in the random sample. Calibration estimation of regression coefficient under two-stage sampling design for different cases of availability of auxiliary information at psu and ssu level has been carried out by Basak *et al.* (2016, 2017, 2018). Biswas *et al.* (2020) developed calibration estimators of the finite population total under two stage sampling design assuming study variable is inversely related to the auxiliary variable. One of the assumptions of calibration approach is that population aggregates are available for the auxiliary variable that is linearly related with the study variable.

Corresponding author: Pradip Basak
E-mail address: pradip@ubkv.ac.in

Many often under two-stage design, such population aggregates of auxiliary variable, i.e., population mean or total are unavailable at the population level and under such situations, estimation of population total has been limited to the use of two phase sampling, see for example, Saini and Bahl (2012), Saini (2013) and references therein. Under such circumstances, it may be assumed that there is availability of additional auxiliary variable which is less closely related to the study variable but its population aggregates are known. Estevao and Särndal (2002) developed estimators of population total under two phase sampling using the information on additional auxiliary variable through two step calibration approach. Guha and Chandra (2020) developed chain ratio type and chain product type estimator of population total under two phase sampling using two step calibration approach assuming that auxiliary information is available for the first variable and unavailable for the second variable. Therefore, the present study considers the problem of estimating the population total under two-stage sampling design when population aggregates of auxiliary variables are unavailable for the selected psu's and develops efficient estimator of population total using the information on known population aggregates of additional auxiliary variable through two step calibration approach.

The rest of this paper is organized as follows. Next Section describes the general notations used for the development of estimators. Section 3 presents the proposed estimators developed using two step calibration approach along with its approximate variance and variance estimator. Section 4 presents the results of the simulation studies to assess the empirical performance of the developed estimators. Finally, Section 5 provides the concluding remarks.

2. NOTATIONS

Let us consider a finite population U of size N which is group into N_{I} clusters each of size N_{i} such that $U = \bigcup_{i=1}^{N_I} U_i$ and $N = \sum_{i=1}^{N_I} N_i$. These clusters are called psu's. At the first stage, n_{I} clusters are selected from N_{I} clusters using a probability sampling design (s_I) such that first and second order inclusion probabilities at the psu level are $\pi_{Ii} = P(i \in s_I)$ and $\pi_{Iij} = P(i, j \in s_I)$. The sampling units within the psu's are called ssu's. At the second stage, n_{i} units are selected from N_{i}, $\forall i \in s_I$ using a probability sampling design (s_i) such that $s = \bigcup_{i=1}^{n_I} s_i$ and $n_s = \sum_{i=1}^{n_I} n_i$. The first and second order inclusion probabilities at the ssu level are $\pi_{k/i} = P(k \in s_i / i \in s_I)$ and $\pi_{kl/i} = P(k,l \in s_i / i \in s_I)$. Let y and x be the study and auxiliary variable respectively. Here, it is assumed that there is availability of additional auxiliary variable z which is less linearly related to the study variable y but its population totals are available for the selected psu's. Let, y_{ik}, x_{ik} and z_{ik}, $\forall i \in s_I$, $k \in s_i$ be values of the variables corresponding to the k^{th} unit of i^{th} selected psu. The population total of y is given by, $t_y = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{N_I} t_{iy}$, where $t_{iy} = \sum_{k=1}^{N_i} y_{ik}$ is the i^{th} psu total of y. Similarly, population total of x is given by $X = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik} = \sum_{i=1}^{N_I} X_i$, where $X_i = \sum_{k=1}^{N_i} x_{ik}$ is the i^{th} psu total of x, and population total of z is, $Z = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} = \sum_{i=1}^{N_I} Z_i$, where $Z_i = \sum_{k=1}^{N_i} z_{ik}$ is the i^{th} psu total of z. Here, it is assumed that population totals of auxiliary variable x is unavailable for the selected psu's, i.e., X_i is unknown $\forall\ i = 1,2,...,n_I$, whereas for additional auxiliary variable z, this information is available, i.e., Z_i is known $\forall\ i = 1,2,...,n_I$. With this, our aim is to estimate the population total, t_y. Following Särndal *et al.* (1992), the π-estimator of population total t_y under two-stage sampling design is given by

$$\hat{t}_{y\pi} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i} y_{ik}, \qquad (1)$$

where, $a_{Ii} = 1/\pi_{Ii}$ and $a_{k/i} = 1/\pi_{k/i}$.

3. PROPOSED CALIBRATION ESTIMATOR

Here, in the first stage, n_I psus are selected. In the second stage, n_i' units are selected out of N_i units at first phase to observe x and z, from each of the n_I selected psus and then at second phase n_i units are drawn from n_i' units.

At first stage: $U_I(N_I)$

$\downarrow$

$s_I(n_I)$

At second stage: $U_i(N_i)\forall i \in s_I$

$\downarrow$

First phase: $s_i'(n_i')$ observe x_{ik}, z_{ik}

$\downarrow$

Second phase: $s_i(n_i)$ observe y_{ik}

Let, a_{Ii} denotes the design weight at the first stage, where $a_{Ii} = 1/\pi_{Ii}$. Let, $a_{1k/i}$ denotes the design weight at the first phase at ssu level and $a_{2k/i}$ denotes the conditional design weight at the second phase at ssu level. Thus, $a_{1k/i} = 1/\pi_{1k/i}$ and $a_{2k/i} = 1/\pi_{2k/i}$, where $\pi_{1k/i}$ is the inclusion probability at first phase and $\pi_{2k/i}$ is the conditional inclusion probability at second phase. Overall design weight corresponding to the k^{th} unit of i^{th} selected psu at the ssu level is given by $a_{k/i} = a_{1k/i}a_{2k/i}$. Let, $w_{1k/i}$ denotes the first phase calibrated weight corresponding to $a_{1k/i}$ at the ssu level and $w_{k/i}$ denotes the overall calibrated weight corresponding to the overall design weight at ssu level, $a_{k/i}$.

First Step Calibration

In the first step calibration, the chi-square distance function measuring the distance between $w_{1k/i}$ and $a_{1k/i}$ is given by

$$\sum_{k=1}^{n_i'} \left(w_{1k/i} - a_{1k/i}\right)^2 / 2a_{1k/i}q_{1k/i}.$$

Here the calibration constraints are,

$$\sum_{k=1}^{n_i'} w_{1k/i}z_{ik} = \sum_{k=1}^{N_i} z_{ik} = Z_i \text{ and } \sum_{k=1}^{n_i'} w_{1k/i} = N_i.$$

First phase calibrated weight, $w_{1k/i}$ are obtained by minimising this objective function subject to the calibration constraints using Lagrangian multiplier approach. The objective function for minimization is given by

$$\phi = \sum_{k=1}^{n_i'} \left(w_{1k/i} - a_{1k/i}\right)^2 / 2a_{1k/i}q_{1k/i} - \lambda_{1/i}'\left(\sum_{k=1}^{n_i'} w_{1k/i}z_{ik} - Z_i\right) - \lambda_{2/i}'\left(\sum_{k=1}^{n_i'} w_{1k/i} - N_i\right).$$

The first step calibrated weights are obtained as

$$w_{1k/i} = a_{1k/i}\left[1 + q_{1k/i}\left(\lambda_{1/i}'z_{ik} + \lambda_{2/i}'\right)\right],$$

where, $\lambda_{1/i}' = \dfrac{\hat{t}_{iq\pi}'\left(Z_i - \hat{t}_{iz\pi}'\right) - \hat{t}_{iqz\pi}'\left(N_i - \hat{t}_{i\pi}'\right)}{\hat{t}_{iq\pi}'\hat{t}_{iqzz\pi}' - \hat{t}_{iqz\pi}'^2}$,

$$\lambda_{2/i}' = \frac{\hat{t}_{iqzz\pi}'\left(N_i - \hat{t}_{i\pi}'\right) - \hat{t}_{iqz\pi}'\left(Z_i - \hat{t}_{iz\pi}'\right)}{\hat{t}_{iq\pi}'\hat{t}_{iqzz\pi}' - \hat{t}_{iqz\pi}'^2},$$

$$\hat{t}_{iqzz\pi}' = \sum_{k=1}^{n_i'} a_{1k/i}q_{1k/i}z_{ik}^2, \qquad \hat{t}_{iqz\pi}' = \sum_{k=1}^{n_i'} a_{1k/i}q_{1k/i}z_{ik},$$

$\hat{t}_{iq\pi}' = \sum_{k=1}^{n_i'} a_{1k/i}q_{1k/i}$, $\hat{t}_{i\pi}' = \sum_{k=1}^{n_i'} a_{1k/i}$ and $\hat{t}_{iz\pi}' = \sum_{k=1}^{n_i'} a_{1k/i}z_{ik}$.

Here, $q_{1k/i}$ is an unknown positive constant. For the particular case when $q_{1k/i} = 1$, the first step calibrated weights are obtained as, $w_{1k/i} = a_{1k/i}\left[1 + \lambda_{1/i}'z_{ik} + \lambda_{2/i}'\right]$. The weights, $w_{1k/i}$ obtained in the first step calibration are used to estimate the psu totals of x for the i^{th} selected psu's as $X_i^* = \sum_{k=1}^{n_i'} w_{1k/i}x_{ik}$, $\forall$ $i=1, 2, \ldots, n_I$, which are required as a constraint in the second step calibration.

Second Step Calibration

In the second step calibration, overall calibration weight at the ssu level, $w_{k/i}$ are obtained. Thus, the chi-square distance function between $w_{k/i}$ and $a_{k/i}$ is given by $\sum_{k=1}^{n_i} \left(w_{k/i} - a_{k/i}\right)^2 / 2a_{k/i}q_{k/i}$ and the calibration constraints are,

$$\sum_{k=1}^{n_i} w_{k/i}z_{ik} = \sum_{k=1}^{n_i'} w_{1k/i}z_{ik} = Z_i,$$

$$\sum_{k=1}^{n_i} w_{k/i}x_{ik} = \sum_{k=1}^{n_i'} w_{1k/i}x_{ik} = X_i^*,$$

$$\sum_{k=1}^{n_i} w_{k/i} = \sum_{k=1}^{n_i'} w_{1k/i} = N_i .$$

The objective function for minimization in the second step calibration is given by

$$\phi = \sum_{k=1}^{n_i} \left(w_{k/i} - a_{k/i}\right)^2 / 2a_{k/i}q_{k/i} - \lambda_{1/i}\left(\sum_{k=1}^{n_i} w_{k/i}z_{ik} - Z_i\right) - \lambda_{2/i}\left(\sum_{k=1}^{n_i} w_{k/i}x_{ik} - X_i^*\right) - \lambda_{3/i}\left(\sum_{k=1}^{n_i} w_{k/i} - N_i\right).$$

Finally, the calibrated weights are obtained as

$$w_{k/i} = a_{k/i}\left[1 + q_{k/i}\left(\lambda_{1/i}z_{ik} + \lambda_{2/i}x_{ik} + \lambda_{3/i}\right)\right],$$

where,

$$\lambda_{1/i} = \frac{\begin{array}{c}\left(X_i^* - \hat{t}_{ix\pi}\right)\left(\hat{t}_{iqx\pi}\hat{t}_{iqz\pi} - \hat{t}_{iqxz\pi}\hat{t}_{iq\pi}\right) + \\ \left(Z_i - \hat{t}_{iz\pi}\right)\left(\hat{t}_{iqxx\pi}\hat{t}_{iq\pi} - \hat{t}_{iqx\pi}^2\right) + \\ \left(N_i - \hat{t}_{i\pi}\right)\left(\hat{t}_{iqxz\pi}\hat{t}_{iqx\pi} - \hat{t}_{iqxx\pi}\hat{t}_{iqz\pi}\right)\end{array}}{\begin{array}{c}\hat{t}_{iqxx\pi}\hat{t}_{iqzz\pi}\hat{t}_{iq\pi} + 2\hat{t}_{iqxz\pi}\hat{t}_{iqx\pi}\hat{t}_{iqz\pi} - \\ \hat{t}_{iqxx\pi}\hat{t}_{iqz\pi}^2 - \hat{t}_{iqzz\pi}\hat{t}_{iqx\pi}^2 - \hat{t}_{iq\pi}\hat{t}_{iqxz\pi}^2\end{array}},$$

$$\lambda_{2/i} = \frac{\begin{array}{c}\left(X_i^* - \hat{t}_{ix\pi}\right)\left(\hat{t}_{iqzz\pi}\hat{t}_{iq\pi} - \hat{t}_{iqz\pi}^2\right) + \\ \left(Z_i - \hat{t}_{iz\pi}\right)\left(\hat{t}_{iqx\pi}\hat{t}_{iqz\pi} - \hat{t}_{iqxz\pi}\hat{t}_{iq\pi}\right) + \\ \left(N_i - \hat{t}_{i\pi}\right)\left(\hat{t}_{iqxz\pi}\hat{t}_{iqz\pi} - \hat{t}_{iqzz\pi}\hat{t}_{iqx\pi}\right)\end{array}}{\begin{array}{c}\hat{t}_{iqxx\pi}\hat{t}_{iqzz\pi}\hat{t}_{iq\pi} + 2\hat{t}_{iqxz\pi}\hat{t}_{iqx\pi}\hat{t}_{iqz\pi} - \\ \hat{t}_{iqxx\pi}\hat{t}_{iqz\pi}^2 - \hat{t}_{iqzz\pi}\hat{t}_{iqx\pi}^2 - \hat{t}_{iq\pi}\hat{t}_{iqxz\pi}^2\end{array}},$$

$$\lambda_{3/i} = \frac{\begin{array}{c}\left(X_i^* - \hat{t}_{ix\pi}\right)\left(\hat{t}_{iqxx\pi}\hat{t}_{iqz\pi} - \hat{t}_{iqzz\pi}\hat{t}_{iqx\pi}\right) + \\ \left(Z_i - \hat{t}_{iz\pi}\right)\left(\hat{t}_{iqxz\pi}\hat{t}_{iqx\pi} - \hat{t}_{iqxx\pi}\hat{t}_{iqz\pi}\right) + \\ \left(N_i - \hat{t}_{i\pi}\right)\left(\hat{t}_{iqzz\pi}\hat{t}_{iqxx\pi} - \hat{t}_{iqxz\pi}^2\right)\end{array}}{\begin{array}{c}\hat{t}_{iqxx\pi}\hat{t}_{iqzz\pi}\hat{t}_{iq\pi} + 2\hat{t}_{iqxz\pi}\hat{t}_{iqx\pi}\hat{t}_{iqz\pi} - \\ \hat{t}_{iqxx\pi}\hat{t}_{iqz\pi}^2 - \hat{t}_{iqzz\pi}\hat{t}_{iqx\pi}^2 - \hat{t}_{iq\pi}\hat{t}_{iqxz\pi}^2\end{array}}.$$

Here,

$$\hat{t}_{iqxx\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i}x_{ik}^2,\ \hat{t}_{iqzz\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i}z_{ik}^2,$$

$$\hat{t}_{iqxz\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i}x_{ik}z_{ik},\ \hat{t}_{iqx\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i}x_{ik},$$

$$\hat{t}_{iqz\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i}z_{ik},\ \hat{t}_{iq\pi} = \sum_{k=1}^{n_i} a_{k/i}q_{k/i},$$

$$\hat{t}_{ix\pi} = \sum_{k=1}^{n_i} a_{k/i}x_{ik},\ \hat{t}_{iz\pi} = \sum_{k=1}^{n_i} a_{k/i}z_{ik},$$

and $\hat{t}_{i\pi} = \sum_{k=1}^{n_i} a_{k/i}$.

Here, $q_{k/i}$ is an unknown positive constant. For the particular case when $q_{k/i} = 1$, the weights are obtained as, $w_{k/i} = a_{k/i}\left[1 + \lambda_{1/i}z_{ik} + \lambda_{2/i}x_{ik} + \lambda_{3/i}\right]$. Finally, the calibrated estimator of population total of *y*, is given by

$$\hat{t}_{y\pi}^c = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} w_{k/i}y_{ik} . \quad (2)$$

Under this situation, the usual double sampling ratio and regression estimator of population total is given by

$$\hat{t}_{y\pi}^{ratio} = \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}}\hat{t}'_{x\pi} , \quad (3)$$

$$\hat{t}_{y\pi}^{reg} = \hat{t}_{y\pi} + b\left(\hat{t}'_{x\pi} - \hat{t}_{x\pi}\right), \quad (4)$$

where,

$$\hat{t}_{y\pi} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}y_{ik} ,\ \hat{t}_{x\pi} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}x_{ik} ,$$

$$\hat{t}'_{x\pi} = \sum_{i=1}^{n_I} a_{Ii}\hat{X}_i ,\ \hat{X}_i = \sum_{k=1}^{n_i'} a_{1k/i}x_{ik} \text{ and}$$

$$b = \frac{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)\left(y_{ik} - \hat{t}_{y\pi}/N\right)}{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)^2} .$$

4. VARIANCE ESTIMATION

The developed calibrated estimators of population total are non-linear in nature. There are two approaches for variance estimation of nonlinear estimator: (i) Analytical approach using the Taylor series linearization and (ii) Resampling approach. In this study, Taylor series linearization technique is used to derive an approximate variance of the estimator as well as the variance estimator. The approximate variance of the

calibrated estimator $\hat{t}_{y\pi}^{c}$ using Taylor series linearization method is obtained as

$$V(\hat{t}_{y\pi}^{c}) = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{Y_i}{\pi_{Ii}}\frac{Y_j}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{E_{k/i}}{\pi_{k/i}}\frac{E_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{X_i}{\pi_{Ii}}\frac{X_j}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta'_{kl/i}\frac{x_k}{\pi'_{k/i}}\frac{x_l}{\pi'_{l/i}}\right]$$

where, $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$, $Y_i = \sum_{k=1}^{N_i} y_{ik}$,

$\Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i}$, $\Delta'_{kl/i} = \pi'_{kl/i} - \pi'_{k/i}\pi'_{l/i}$,

$X_i = \sum_{k=1}^{N_i} x_{ik}$, $E_{k/i} = y_{ik} - B_{ikyx.z}x_{ik} - B_{ikyz.x}z_{ik}$,

$$B_{ikyx.z} = \frac{\left(\sum_{k=1}^{N_i} y_{ik}x_{ik}\right)\left(\sum_{k=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} y_{ik}z_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)}{\left(\sum_{k=1}^{N_i} x_{ik}^2\right)\left(\sum_{k=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)^2},$$

$$B_{ikyz.x} = \frac{\left(\sum_{k=1}^{N_i} y_{ik}z_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}^2\right) - \left(\sum_{k=1}^{N_i} y_{ik}x_{ik}\right)\left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)}{\left(\sum_{k=1}^{N_i} x_{ik}^2\right)\left(\sum_{k=1}^{N_i} z_{ik}^2\right) - \left(\sum_{k=1}^{N_i} x_{ik}z_{ik}\right)^2}.$$

The variance estimator is obtained as,

$$\hat{V}(\hat{t}_{y\pi}^{c}) = \sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{t}_{iy\pi}}{\pi_{Ii}}\frac{\hat{t}_{jy\pi}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\hat{E}_{k/i}}{\pi_{k/i}}\frac{\hat{E}_{l/i}}{\pi_{l/i}} + \left[\sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{t}'_{ix\pi}}{\pi_{Ii}}\frac{\hat{t}'_{jx\pi}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n'_i}\sum_{l=1}^{n'_i}\breve{\Delta}'_{kl/i}\frac{x_k}{\pi'_{k/i}}\frac{x_l}{\pi'_{l/i}}\right].$$

where, $\breve{\Delta}_{Iij} = \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}$, $\breve{\Delta}_{kl/i} = \frac{\pi_{kl/i} - \pi_{k/i}\pi_{l/i}}{\pi_{kl/i}}$,

$\breve{\Delta}'_{kl/i} = \frac{\pi'_{kl/i} - \pi'_{k/i}\pi'_{l/i}}{\pi'_{kl/i}}$, $\hat{t}_{iy\pi} = \sum_{k=1}^{n_i} a_{k/i}y_{ik}$, $\hat{t}'_{ix\pi} = \sum_{k=1}^{n'_i} a_{1k/i}x_{ik}$,

$\hat{E}_{k/i} = y_{ik} - \hat{B}_{ikyx.z}x_{ik} - \hat{B}_{ikyz.x}z_{ik}$,

$$\hat{B}_{ikyx.z} = \frac{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)}{\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)^2},$$

$$\hat{B}_{ikyz.x} = \frac{\left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}z_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}y_{ik}x_{ik}\right)\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)}{\left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}^2\right)\left(\sum_{k=1}^{n_i} a_{k/i}z_{ik}^2\right) - \left(\sum_{k=1}^{n_i} a_{k/i}x_{ik}z_{ik}\right)^2}.$$

5. SIMULATION STUDY

Design based simulation study is conducted to evaluate the empirical performance of the developed estimators. Here, a real survey dataset is considered as a finite population. From this fixed population, repeated random samples are drawn using two-stage sampling design. The survey dataset of 284 municipalities of Sweden popularly referred to as 'MU284 population' is used for the simulation study (Särndal *et al.*, 1992). The 284 municipalities are grouped into 50 clusters and the sizes of the clusters varies from 5 to 9 municipalities. These 50 clusters are psu's and municipalities within the clusters are referred to as ssu's. The dataset contains multiple variables among which three variables are selected for the present study. Here, the variable revenues from the 1985 Municipal taxation (RMT85, measured in millions of kronor, y) is used as study variable. The aim is to estimate total revenues from the 1985 Municipal taxation. Here, 1985 population (P85, in thousands, x) is used as the auxiliary variable and number of seats in the municipal council (S82, z) is used as the additional auxiliary variable. The correlation between study variable, y and additional auxiliary variable, z is found to be 0.58. Therefore, number of Social-Democratic seats in municipal council (SS82) is used as another additional auxiliary variable such that correlation between y and z is 0.40. The correlations among the variables in the population are presented in Table 1.

Table 1. Correlation between variables in MU284 data

Variables	RMT85 (*y*)	P85 (*x*)	S82 (*z*)	SS82 (*z*)
RMT85 (*y*)	1	0.96	0.58	0.40
P85 (*x*)	0.96	1	0.69	0.48
S82 (*z*)	0.58	0.69	1	-

From this population, a two-stage sample is drawn. At the first stage, 20 psu's are drawn and in the second stage, at first phase, 8 units are drawn from each of the selected psu's to observe x only and at second phase, 4 units are drawn. Sample are drawn using simple random sampling without replacement (SRSWOR) at the both stages. The values of different estimators are computed using these sample data. The following estimators of population total under two-stage sampling design are considered in the simulation study.

i) π-estimator, $\hat{t}_{y\pi}$ (denoted as Est- π and given in (1)),

ii) Double sampling ratio estimator, $\hat{t}_{y\pi}^{ratio}$ (denoted as RAT and given in (3)),

iii) Double sampling regression estimator, $\hat{t}_{y\pi}^{reg}$ (denoted as REG and given in (4)),

iv) Developed calibrated estimator, $\hat{t}_{y\pi}^{c}$ using S82 as additional auxiliary variable *z* (denoted as CAL1 and given in (2)).

v) Developed calibrated estimator, $\hat{t}_{y\pi}^{c}$ using SS82 as additional auxiliary variable *z* (denoted as CAL2 and given in (2)).

The simulation is repeated to a total number of *M*=5000 times. The performance of the estimators are evaluated by percentage absolute relative bias (ARB, %) and percentage relative root mean squared error (RRMSE, %), defined by

$$ARB(\hat{T}) = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{\hat{T}_i - T}{T}\right| \times 100 \text{ and}$$

$$RRMSE(\hat{T}) = \sqrt{M^{-1}\sum_{i=1}^{M}\left(\frac{\hat{T}_i - T}{T}\right)^2} \times 100,$$

where $\hat{T}_i$ denotes the estimated value of population total at simulation run *i*, with true value *T* and *M* denotes the number of simulations run. The values of percentage absolute relative bias and percentage relative root mean square error of different estimators are reported in Table 2.

Table 2. Percentage absolute relative bias (ARB, %) and percentage relative root mean square error (RRMSE, %) of different estimators in design based simulation

Estimator	ARB, %	RRMSE, %
Est- π	27.63	33.40
RAT	16.57	19.68
REG	16.25	18.74
CAL1	13.36	15.59
CAL2	16.53	19.31

The results in Table 2 show that the values of both percentage absolute relative bias and relative root mean square error are higher for π-estimator as compared to the other estimators. The developed calibrated estimator has minimum percentage absolute relative bias and relative root mean square error among all the estimators when correlation between *y* and *z* is high. However, for moderate level of correlation between *y* and *z*, percentage absolute relative bias and relative root mean square error of the developed calibrated estimator is lower than double sampling ratio estimator but higher than double sampling regression estimator.

6. CONCLUSIONS

Calibration estimators of the population total have been developed under two-stage sampling design based on the unavailability of auxiliary information for the selected psu's. Monte Carlo simulations based on both simulated and real dataset show the superiority of the proposed calibration estimators of the population total in comparison to the existing estimators such as Horvitz-Thompson, double sampling ratio and regression estimators. Therefore, the developed calibration estimators will produce reliable estimate of population parameters from the two-stage survey data in the situations of unavailability of auxiliary information for the selected psu's.

ACKNOWLEDGEMENTS

The authors also gratefully acknowledge the valuable contributions made by Late Dr. Hukum Chandra, National Fellow and Principal Scientist, ICAR-IASRI, New Delhi.

REFERENCES

Aditya, K., Sud, U.C., Chandra, H. and Biswas, A. (2016). Calibration based regression type estimator of the population total under two stage sampling design. *Journal of the Indian Society of Agricultural Statistics*, **70(1)**, 19-24.

Basak, P., Sud, U.C., and Chandra, H. (2016). Calibration Approach Based Estimator of Finite Population Regression Coefficient under Two-stage Sampling Design. *International Journal of Agricultural and Statistical Sciences*, **12(2)**, 415-422.

Basak, P., Sud, U.C. and Chandra, H. (2017). Calibration Estimation of Regression Coefficient for Two-stage Sampling Design. *Journal of the Indian Society of Agricultural Statistics*, **71(1)**, 1-6.

Basak, P., Sud, U.C., and Chandra, H. (2018). Calibration Estimation of Regression Coefficient for Two-stage Sampling Design using Single Auxiliary Variable. *Journal of the Indian Society of Agricultural Statistics*, **72 (1)**, 1-6.

Biswas, A., Aditya, K., Sud, U. C. and Basak, P. (2020). Product type calibration estimators of finite population total under two stage sampling. *Journal of the Indian Society of Agricultural Statistics*, **74(1)**, 23-32.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* **87**, 376–382.

Estevao, V.M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official* Statistics, **18(2)**, 233–255.

Guha, S. and Chandra, H. (2020). Improved chain-ratio type estimator for population total in double sampling. *Mathematical Population Studies*, **27(4)**, 216-231.

Mourya, K.K., Sisodia, B.V.S. and Chandra, H. (2016). Calibration approach for estimating finite population parameters in two-stage sampling. *Journal of Statistical Theory and Practice*, **10 (3)**, 550-562.

Saini, M. (2013). A class of predictive estimators in two-stage sampling when auxiliary character is estimated. *Indian Journal of Science and Technology*, **6(3)**, 4213-4218.

Saini, M. and Bahl, S. (2012). Estimation of population mean in two stage design using double sampling for stratification and multi-auxiliary information. *International Journal of Computer Applications*, **47(9)**, 17-21.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press. (USA).