Chapter

Development of Estimation Procedure of Population Mean in Two-Phase Stratified Sampling

Partha Parichha, Kajla Basu and Arnab Bandyopadhyay

Abstract

This article describes the problem of estimation of finite population mean in two-phase stratified random sampling. Using information on two auxiliary variables, a class of product to regression chain type estimators has been proposed and its characteristic is discussed. The unbiased version of the proposed class of estimators has been constructed and the optimality condition for the proposed class of estimators is derived. The efficacy of the proposed methodology has been justified through empirical investigations carried over the data set of natural population as well as the data set of artificially generated population. The survey statistician may be suggested to use it.

Keywords: stratified random sampling, double sampling, auxiliary variables, chain type estimators, bias, mean square error, efficiency, AMS 2000 Mathematics Subject Classification: 62D05

1. Introduction

In this present paper we have made use of Auxiliary information extracted from the variables having correlation with study variable. Auxiliary information may be utilized at planning, design and estimation stages to develop improved estimation procedures in sample surveys. Sometimes, information on auxiliary variable may be readily available for all the units of population; for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation and number of beds available in different hospitals may be known well in advance in health care surveys. If such information lacks, it is sometimes, relatively cheap to take a large preliminary sample where auxiliary variable alone is measured, such practice is applicable in two-phase (or double) sampling. Two-phase stratified sampling happens to be a powerful and cost effective (economical) technique for obtaining the reliable estimate in first-phase (preliminary) sample for the unknown parameters of the auxiliary variables. For example, Sukhatme [1] mentioned that in a survey to estimate the production of lime crop based on orchards as sampling units, a comparatively larger sample is drawn to determine the acreage under the crop while the yield rate is determined from a sub sample of the orchards selected for determining acreage.

In order to construct an efficient estimator of the population mean of the auxiliary variable in first-phase (preliminary) sample, Chand [2] introduced a technique

of chaining another auxiliary variable with the first auxiliary variable by using the ratio estimator in the first phase sample. The estimator is known as chain-type ratio estimator. This work was further extended by Kiregyera [3, 4], Tracy et al. [5], Singh and Espejo [6], Gupta and Shabbir [7], Shukla et al. [8], Choudhury and Singh [9], Parichha et al. [10] and among others, where they proposed various chain-type ratio and regression estimators.

In practice, the population may often consist of heterogeneous units. For example, in socio-economic surveys, people may live in rural areas, urban localities, ordinary domestic houses, hostels, hospitals and jail, etc. In such a situation one should carefully study the population according to the characteristics of regions and then apply sampling scheme strata wise independently. This procedure is known as stratified random sampling. It may be noted that most of the developments in two-phase sampling scheme are based on simple random sampling only while limited number of attempts are taken to address the problems of two-phase sampling scheme in the platform of stratified random sampling. It may be also noticeable that the most of the research work on two-phase sampling are producing biased estimates. However, biased becomes a serious drawback in sample surveys. A sampling method is called biased if it systematically favors some outcomes over others. It results in a biased sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling. For example, telephone sampling is common in marketing surveys. A simple random sample may be chosen from the sampling frame consisting of a list of telephone numbers of people in the area being surveyed. This method does involve taking a simple random sample, but it is not a simple random sample of the target population (consumers in the area being surveyed). It will miss people who do not have a phone. It may also miss people who only have a cell phone that has an area code not in the region being surveyed. It will also miss people who do not wish to be surveyed, including those who monitor calls on an answering machine and don't answer those from telephone surveyors. Thus the method systematically excludes certain types of consumers in the area. It is obvious that the inferences from a biased sample are not as trustworthy as conclusions from a truly random sample.

Encouraged with the above work, we have proposed a class of product to regression chain type estimators in stratified sampling using two auxiliary variables under double sampling. The unbiased version of the proposed class of estimators has been obtained which make the estimation strategy more practicable. The dominance of the proposed estimation strategy over the conventional ones has been established through empirical investigations carried over the data set of natural as well as artificially generated population.

2. Sampling structures and notations

Consider a finite population U = {1, 2,..., N} of N identifiable units divided into L homogeneous strata with the hth stratum (h = 1, 2,..., L) having N_h. Let y and (x, z) be the study variable and two auxiliary variables respectively taking values y_{ih} and (x_{ih}, z_{ih}), respectively, for the unit $i = 1,2,...N_h$ of the hth stratum. $\overline{Y} = \sum_{h=1}^{L} \overline{Y}_h W_h$, $\overline{X} = \sum_{h=1}^{L} \overline{X}_h W_h$, $\overline{Z} = \sum_{h=1}^{L} \overline{Z}_h W_h$ be population means of the study and the auxiliary variables, and $\overline{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}$, $\overline{X}_h = \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}$, $\overline{Z}_h = \sum_{i=1}^{N_h} \frac{z_{hi}}{N_h}$ be the corresponding stratum means. Here $W_h = \frac{N_h}{N}$ is the known stratum weight.

Let $C_{yh} = \frac{S_{yh}}{\overline{Y}_h}$, $C_{xh} = \frac{S_{xh}}{\overline{X}_h}$ and $C_{zh} = \frac{S_{zh}}{\overline{Z}_h}$ be the coefficients of variation where $S_{y_h} = \sqrt{\frac{\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}_h\right)^2}{N_h-1}}$, $S_{x_h} = \sqrt{\frac{\sum_{i=1}^{N_h}\left(x_{hi}-\overline{X}_h\right)^2}{N_h-1}}$, $S_{z_h} = \sqrt{\frac{\sum_{i=1}^{N_h}\left(z_{hi}-\overline{Z}_h\right)^2}{N_h-1}}$ are the population standard deviations in the *h*th stratum.

Let ρ_{yx_h}, ρ_{yz_h} and ρ_{xz_h} be the correlation coefficients between (y, x), (y, z), and (x, z) respectively in the hth stratum. Chand [2] and Kiregyera [3, 4] discussed a situation in simple random sampling when information on x is unknown but another auxiliary variable z is easily available. It is assumed that population mean of one auxiliary variable z is known in advance and the population mean of the other auxiliary variable x is unknown. We seek to estimate through a two-phase stratified sampling design. Using a simple random sample without replacement (SRSWOR) sampling scheme at each phase, we adopt the double sampling scheme as follows.

i. In the first phase, a preliminary large sample of size n'_h is drawn from the hth stratum of size N_h (h = 1, 2,..., L) and information on the auxiliary variables x and z is observed.

ii. In the second phase, a sub-sample of size n_h is drawn from first phase sample n'_h units from the h th stratum of size N_h and information on both the study variable y and the auxiliary variables x and z is taken.

$\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$, $\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$, $\overline{z}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} z_{hi}$, $\overline{x}'_h = \frac{1}{n'_h}\sum_{i=1}^{n'_h} x_{hi}$, and $\overline{z}'_h = \frac{1}{n'_h}\sum_{i=1}^{n'_h} z_{hi}$ be the corresponding sample means in the hth stratum.

3. Discussion on existing estimation strategies

The usual stratified mean estimator ($\overline{y}_{st}$) for population mean ($\overline{Y}$), is given by

$$\overline{y}_{st} = \sum_{h=1}^{L} w_h \overline{y}_h \tag{1}$$

The mean square error (*MSE*) of $\overline{y}_{st}$, is given by

$$\mathrm{MES}\left(\overline{y}_{st}\right) = \sum_{h=1}^{L} w_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{yh}^2 \tag{2}$$

Motivated with the technique adopted by Chand [2], one may frame the chain ratio-product type estimator in stratified sampling structure as

$$\overline{y}_{RP}^{(h)} = \sum_{h=1}^{L} w_h \overline{y}_h \left(\frac{\overline{X}'_h}{\overline{x}_h}\right)\left(\frac{\overline{Z}_h{}'}{\overline{z}_h}\right) \tag{3}$$

The bias and *MSE* respectively of $\overline{y}^{(h)}{}_{RP}$, to first order of approximation, are obtained as

$$\mathrm{Bias}\left(\overline{y}_{RP}^{(h)}\right) \cong \sum_{h=1}^{L} w_h \overline{y}_h \left[\left(\frac{1}{n_h} - \frac{1}{n'_h}\right) A_{1h} + \left(\frac{1}{n'_h} - \frac{1}{N_h}\right) A_{2h}\right] \tag{4}$$

$$\mathrm{MSE}\left(\overline{y}_{RP}^{(h)}\right) = \sum_{h=1}^{L} w_h^2 s_{yh}^2 \left[\left(\frac{1}{n_h} - \frac{1}{n'_h}\right) A_{3h} + \left(\frac{1}{n'_h} - \frac{1}{N_h}\right) A_{4h} + \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\right] \tag{5}$$

where

$$A_{1h} = C_{xh}^2 - \rho_{yxh} C_{yh} C_{xh} \text{ and } A_{2h} = C_{zh}^2 - \rho_{yzh} C_{yh} C_{zh}$$

$$A_{3h} = \frac{C_{xh}^2}{C_{yh}^2} - 2\rho_{yxh} \frac{C_{xh}}{C_{yh}} \quad \text{and} \quad A_{4h} = \frac{C_{zh}^2}{C_{yh}^2} - 2\rho_{yzh} \frac{C_{zh}}{C_{yh}}$$

Similarly, inspired with the technique adopted by Choudhary and Sing [9], one may frame the two-phase stratified random sampling estimator in stratified sampling as

$$\overline{y}_{cs}^{h} = \sum_{h=1}^{L} w_h \overline{y}_h \left[k_h \left(\frac{\overline{x}_h'}{\overline{x}_h} \right) \left(\frac{\overline{Z}_h{}'}{\overline{z}_h} \right) + (1 - k_h) \left(\frac{\overline{x}_h{}'}{\overline{x}_h} \right) \left(\frac{\overline{z}_h'}{\overline{Z}_h} \right) \right] \tag{6}$$

where $k_h$is constant.

Bias $\left(\overline{y}_{cs}^{h}\right) \cong \sum_{h=1}^{L} w_h \overline{y}_h A_{5h}$

$$A_{5h} = (1 - 2k_h) C_{yh} \left[\left(\frac{1}{n_h} - \frac{1}{n_h'} \right) \rho_{yxh} C_{xh} + \left(\frac{1}{n_h'} - \frac{1}{N_h} \right) \rho_{yzh} C_{zh} \right] + k_h \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) C_{xh}^2 + \left(\frac{1}{n_h'} - \frac{1}{N_h} \right) C_{zh}^2 \right] \tag{7}$$

$$\text{And MSE} \quad \left(\overline{y}_{cs}^{h}\right)_{\min} = \sum_{h=1}^{L} w_h^2 s_{y_h}^2 \times \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) - \frac{\left\{ \left(\frac{1}{n_h} - \frac{1}{n_h'} \right) \rho_{yx_h} C_{x_h} - \left(\frac{1}{n_h'} - \frac{1}{N_h} \right) \rho_{yz_h} C_{z_h} \right\}^2}{\left(\frac{1}{n_h} - \frac{1}{N_h} \right) C_{x_h}^2 + \left(\frac{1}{n_h'} - \frac{1}{N_h} \right) C_{z_h}^2} \right] \tag{8}$$

4. Formulation of proposed estimation strategy

Motivated with the earlier work, discussed above, we have constructed a class of product to regression chain type estimators as

$$t_p = \sum_{h=1}^{L} w_h \overline{y}_h \left\{ k_h \frac{\overline{x}_h'}{\overline{x}_h} + (1 - k_h) \frac{\overline{x}_{id_h}'}{\overline{x}_h} \right\} \tag{9}$$

where $k_h (h = 1, 2, ..., L)$ is a real constant which can be suitably determined by minimizing the M. S. E. of the class of estimator t_p and $\overline{x}_{d_h}' = \overline{x}' + b_{xz_h}(n_h') \left(\overline{Z}_h - \overline{z}_h' \right)$; where $b_{xz_h}(n_h')$ is the regression coefficient between the variables x and z at the hth stratum.

5. Bias and mean square errors of the proposed class of estimator t_p

It can be easily noted that the proposed class of estimators t_p defined in Eqs. (8) is chain product and regression type estimator. Therefore, it is biased estimator for population mean $\overline{Y}$. So, we obtain biases and mean square errors under large sample approximations using the following transformations:

$$\overline{y}_h = \overline{Y}_h(1 + e_1), \overline{x}_h = \overline{X}_h(1 + e_2), \overline{x}_h' = \overline{X}_h(1 + e_3), \overline{z}_h' = \overline{Z}_h(1 + e_4),$$

$$s_{xz_h}' = S_{xz_h}(1 + e_5), \quad s_{z_h'}^2 = S_{z_h}^2(1 + e_6)$$

and $E(e_i) = 0$ for (i = 1, 2,..., 6), e_i for (i = 1, 2,..., 6) are relative error term. Under above transformations the class of estimator t_p may be represented as

$$t_p = \sum_{h=1}^{L} w_h \overline{Y}(1+e_1)\left[(1-k_h)\left\{(1+e_3)(1+e_2)^{-1}\right\} + k_h\left\{(1+e_3) - \frac{\overline{Z}_h}{\overline{X}_h}\beta_{xz_h}(e_4 + e_4e_5 - e_4e_6)\right\}(1+e_2)^{-1}\right] \tag{10}$$

We have the following expectations of the sample statistics of two-phase stratified sampling as

$$\left.\begin{aligned}
&E(e_1^2) = f_1 C_{y_h}^2,\ E(e_2^2) = f_1 C_{x_h}^2,\ E(e_3^2) = f_2 C_{x_h}^2,\ E(e_4^2) = f_2 C_{z_h}^2\\
&E(e_1e_2) = f_1 \rho_{yx_h} C_{y_h} C_{x_h},\ E(e_1e_3) = f_2 \rho_{yx_h} C_{y_h} C_{x_h},\\
&E(e_2e_3) = f_2 C_{x_h}^2,\ E(e_2e_4) = E(e_3e_4) = f_2 \rho_{xz_h} C_{x_h} C_{z_h},\\
&E(e_4e_5) = f_2 \frac{\mu_{102}}{\overline{Z}_h S_{xz_h}},\ E(e_4e_6) = f_2 \frac{\mu_{003}}{\overline{Z}_h S_{z_h}^2},\\
&E(e_2e_5) = f_2 \frac{\mu_{201}}{\overline{X}_h S_{xz_h}},\ E(e_2e_6) = f_2 \frac{\mu_{102}}{\overline{X}_h S_{z_h}^2},\\
&E(e_1e_4) = f_2 \rho_{yz_h} C_{y_h} C_{z_h}.
\end{aligned}\right\} \tag{11}$$

where

$$f_1 = \frac{1}{n_h} - \frac{1}{N_h},\ f_3 = \frac{1}{n_h} - \frac{1}{n_h'},\ f_2 = \frac{1}{n_h'} - \frac{1}{N_h},$$

$$\mu_{pqr} = \frac{1}{N_h}\sum_{i=1}^{N_h}(x_i - \overline{X}_h)^p (y_i - \overline{Y}_h)^q (z_i - \overline{Z}_h)^r; (p, q, r \geq 0)$$

Expanding binomially, using results from Eq. (1) and retaining the terms up to first order of sample size, we have derived the expressions of bias B(.) and mean square error M(.) of the class of estimators t_p as

$$B(t_p) = E(t_p - \overline{Y_h}) = \sum_{h=1}^{L} w_h \overline{Y}\left[(1-k_h) b_{xz_h} \frac{\overline{Z_h}}{\overline{X_h}}\left(f_2 \frac{S_{xz_h}}{\overline{X_h}\,\overline{Z_h}} - f_1 \frac{S_{yz_h}}{\overline{Y_h}\,\overline{Z_h}} - f_2 \frac{\mu_{102}}{S_{xz_h}\overline{Z_h}} - \frac{\mu_{003}}{S_{Z_h}^2\,\overline{Z_h}}\right) + f_3\left(\frac{S_{x_h}^2}{\overline{X_h}^2} - \frac{S_{yx_h}}{\overline{Y}_h\overline{X}_h}\right)\right] \tag{12}$$

$$M(t_p) = E\left[t_p - \overline{Y}_h\right]^2 = \sum_{h=1}^{L} w_h \overline{Y}^2{}_h\left[f_1 C_{y_h}^2 + k_h{}^2 a + 2k_h b + c\right] \tag{13}$$

where $a = \left(f_2 \rho_{xz_h}^2\right) C_{x_h}^2$ and $b = f_2 \rho_{yz_h} \rho_{xz_h} C_{y_h} C_{x_h} - \left(f_2 \rho_{xz_h}^2\right) C_{x_h}^2$

$c = f_3 C_{x_h}^2 - 2 f_3 \rho_{yx_h} C_{y_h} C_{x_h} + \left(f_2 \rho_{xz_h}^2\right) C_{x_h}^2 - 2 f_2 \rho_{yz_h} \rho_{xz_h} C_{y_h} C_{x_h}$.

6. Bias reduction for the proposed class of estimators

In recent time serious drawback is bias of an estimator. Therefore, unbiased versions of the proposed classes of estimators are more desirable. Motivated with

this argument and influenced by the bias correction techniques of Tracy et al. [5] and Bandyopadhyay and Singh [11] we proceed to derive the unbiased version of our proposed class of estimator t_p.

From Eq. (12), we observe that the expression of bias of the estimator t_p contains the population parameters such as μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S^2_{x_h}$, $S^2_{y_h}$, $\overline{Y}_h$, $\overline{X}_h$, S_{yz_h} and $S^2_{z_h}$. Since $S^2_{z_h}$ is known while μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S^2_{x_h}$, $S^2_{y_h}$, $\overline{Y}_h$, $\overline{X}_h$ and S_{yz_h} are unknown, replacing μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S^2_{x_h}$, $S^2_{y_h}$, $\overline{Y}_h$, $\overline{X}_h$, by their respective sample estimator (based on the second phase sample of size m) m_{003}, m_{102}, s_{yz_h}, $s^2_{x_h}$, $s^2_{y_h}$, $\overline{y}_h$, $\overline{x}_h$ and s_{yz_h}, we get an estimator of $B(t_p)$ and

$$b(t_p) = \sum_{h=1}^{L} w_h \overline{y}_h \left[(1-k_h) b_{xz_h} \frac{\overline{z}_h}{\overline{x}_h} \left(f_2 \frac{s_{xz_h}}{\overline{x}_h \overline{z}_h} - f_1 \frac{s_{yz_h}}{\overline{y}_h \overline{z}_h} - f_2 \frac{m_{102}}{s_{xz_h} \overline{z}_h} - \frac{m_{003}}{s^2_{Z_h} \overline{z}_h} \right) + f_3 \left(\frac{s^2_{x_h}}{\overline{x}^2_h} - \frac{s_{yx_h}}{\overline{y}_h \overline{x}_h} \right) \right]. \tag{14}$$

where $m_{pqr} = \frac{1}{m} \sum_{i=1}^{m} (x_{hi} - \overline{x}_h)^p (y_{hi} - \overline{y}_h)^q (z_{hi} - \overline{z}_h)^r$.

Motivating with the bias reduction techniques of Tracy et al. [5] and Bandyopadhyay and Singh [11], we have derived the unbiased version of the proposed class of estimators t_p to the first order of approximations two-phase stratified sampling.

$$t'_p = t_p - b(t_p)$$

which becomes

$$t'_p = \sum_{h=1}^{L} w_h \left[\overline{y}_h \left\{ k_h \frac{\overline{x}'_h}{\overline{x}_h} + (1-k_h) \frac{\overline{x}'_{id_h}}{\overline{x}_h} \right\} - \overline{y_h} \left[(1-k_h) b_{xz_h} \frac{\overline{z_h}}{\overline{x_h}} \left(f_2 \frac{s_{xz_h}}{\overline{x_h}\,\overline{z_h}} - f_1 \frac{s_{yz_h}}{\overline{y_h}\,\overline{z_h}} - f_2 \frac{m_{102}}{s_{xz_h}\overline{z_h}} - \frac{m_{003}}{s^2_{Z_h}\overline{z_h}} \right) + f_3 \left(\frac{s^2_{x_h}}{\overline{x_h}^2} - \frac{s_{yx_h}}{\overline{y_h}\,\overline{x_h}} \right) \right] \right] \tag{15}$$

Thus, the variance of t'_p to the first order of approximation are obtained as

$$V(t'_p) = M(t_p) = \sum_{h=1}^{L} \overline{Y}_h{}^2 \left[f_1 C^2_{y_h} + k_h{}^2 a + 2k_h b + c \right] \tag{16}$$

From Eqs. (10) and (15) it is to be noted that the class of estimators t'_p is preferable over the class of estimators t_p of two –phase sampling set up as t'_p is unbiased (up to first order of sample size) class of estimator of $\overline{Y}_h$ while the class of estimator t_p is biased.

7. Minimum variance of proposed class of estimators

It is obvious from the Eq. (16) that the variances of the proposed class of estimator t'_p depend on the value of the constant k_h. Therefore, we desire to minimize their variances and discussed them below. The optimality condition under which proposed class of estimators t'_p have minimum variance is obtained as

$$k_h = -\frac{b}{a} \tag{17}$$

Substituting the optimum value of the constant k_h in Eq. (19), we have the minimum variance of the class of estimators t'_p as

$$\text{Min.V}\left(t'_p\right) = \sum_{h=1}^{L} W^2{}_h \overline{Y}_h{}^2 \left[f_1 C^2_{y_h} - \frac{b^2}{a} + C \right] \tag{18}$$

8. Efficiency comparison of the proposed strategy

It is important to investigate the performance of the proposed class of estimators with respect to the existing ones. We use the two natural population and one artificially generated population data set to justify the supremacy of the proposed strategy.

8.1 Empirical investigations through natural populations

The data set of two natural populations has been presented below.

- **Population I** (Source: Murthy [12], p. 228)

y: Factory **output** in thousand rupees, *x*: Number of workers in the factory, and z: Fixed capital of factory in thousand rupees.

The data consist of 80 observations which are divided into four strata according to the auxiliary variable z as: (i) $z \leq 500$, (ii) $500 < z \leq 1000$, (iii) $1000 < z \leq 2000$, and $z > 2000$ respectively for allocation of sample size to different strata, Proportional allocation is used.

Stratum 1 $(z \leq 500)$

$$N_1 = 19,\ n'_1 = 11,\ n_1 = 5,\ \overline{Y}_1 = 2669.247,\ \overline{X}_1 = 65.15789$$
$$\overline{Z}_1 = 349.6842,\ C_{y_1} = 0.28363,\ C_{x_1} = 0.17153,\ C_{z_1} = 0.31299$$
$$\rho_{yx_1} = 0.81381,\ \rho_{yz_1} = 0.9364,\ \rho_{xz_1} = 0.9044$$

Stratum 2 $(500 < z \leq 1000)$

$$N_2 = 32,\ n'_2 = 17,\ n_2 = 8,\ \overline{Y}_2 = 4657.625,\ \overline{X}_2 = 139.9668$$
$$\overline{Z}_2 = 706.5938,\ C_{y_2} = 0.14366,\ C_{x_2} = 0.3169,\ C_{z_2} = 0.15457$$
$$\rho_{yx_2} = 0.8883,\ \rho_{yz_2} = 0.9259,\ \rho_{xz_2} = 0.8456$$

Stratum 3 $(1000 < z \leq 2000)$

$$N_3 = 14,\ n'_3 = 8,\ n_3 = 3,\ \overline{Y}_3 = 6537.214,\ \overline{X}_3 = 403.2143$$
$$\overline{Z}_3 = 1539.571,\ C_{y_3} = 0.06365,\ C_{x_3} = 0.20117,\ C_{z_3} = 0.18004$$
$$\rho_{yx_3} = 0.9295,\ \rho_{yz_3} = 0.9835,\ \rho_{xz_3} = 0.9366$$

Stratum 4 $(z > 2000)$

$$N_4 = 15,\ n'_4 = 9,\ n_4 = 4,\ \overline{Y}_4 = 7843.667,\ \overline{X}_4 = 763.2$$
$$\overline{Z}_4 = 2620.533,\ C_{y_4} = 0.08232,\ C_{x_4} = 0.22464,\ C_{z_4} = 0.14156$$
$$\rho_{yx_4} = 0.9787,\ \rho_{yz_4} = 0.9692,\ \rho_{xz_4} = 0.9454$$

- **Population II** (Source: Koyuncu and Kadilar [13]).

y: Number of teachers, *x*: Number of students both primary and secondary schools, and *z*: Number of classes both primary and secondary schools. There are 923 districts in 6 regions (as: (i) Marmara, (ii) Agean, (iii) Mediterranean, (iv) Central Anatolia, (v) Black Sea, (vi): East and Southeast Anatolia) in Turkey in 2007 (source: The Turkish Republic Ministry of Education).

Marmara region

$$N_1 = 127,\ n_1' = 60,\ n_1 = 31,\ \overline{Y}_1 = 703.74,\ \overline{X}_1 = 20804.59$$
$$\overline{Z}_1 = 498.28,\ C_{y_1} = 1.25591,\ C_{x_1} = 1.46538,\ C_{z_1} = 1.115$$
$$\rho_{yx_1} = 0.936,\ \rho_{yz_1} = 0.97891,\ \rho_{xz_1} = 0.93958$$

Agean region

$$N_2 = 117,\ n_2' = 40,\ n_2 = 21,\ \overline{Y}_2 = 413,\ \overline{X}_2 = 9211.79$$
$$\overline{Z}_2 = 318.83,\ C_{y_2} = 1.56155,\ C_{x_2} = 1.64797,\ C_{z_2} = 1.14804$$
$$\rho_{yx_2} = 0.996,\ \rho_{yz_2} = 0.97624,\ \rho_{xz_2} = 0.96958$$

Mediterranean

$$N_3 = 103,\ n_3' = 50,\ n_3 = 29,\ \overline{Y}_3 = 573.17,\ \overline{X}_3 = 14309.3$$
$$\overline{Z}_3 = 431.36,\ C_{y_3} = 1.80307,\ C_{x_3} = 1.9253,\ C_{z_3} = 1.42097$$
$$\rho_{yx_3} = 0.994,\ \rho_{yz_3} = 0.98351,\ \rho_{xz_3} = 0.97655$$

Central Anatolia region

$$N_4 = 170,\ n_4' = 75,\ n_4 = 38,\ \overline{Y}_4 = 424.66,\ \overline{X}_4 = 9478.85$$
$$\overline{Z}_4 = 311.32,\ C_{y_4} = 1.90878,\ C_{x_4} = 1.92206,\ C_{z_4} = 1.47124$$
$$\rho_{yx_4} = 0.983,\ \rho_{yz_4} = 0.98296,\ \rho_{xz_4} = 0.96362$$

Black sea region

$$N_5 = 205,\ n_5' = 40,\ n_5 = 25,\ \overline{Y}_5 = 267.03,\ \overline{X}_5 = 5569.95$$
$$\overline{Z}_5 = 227.20,\ C_{y_5} = 1.51162,\ C_{x_5} = 1.52564,\ C_{z_5} = 1.14811$$
$$\rho_{yx_5} = 0.989,\ \rho_{yz_5} = 0.96434,\ \rho_{xz_5} = 0.96725.$$

The percentage relative efficiencies (PRE) the proposed class of estimators t_p' with respect to different estimators under their respective optimum conditions are shown below.

8.2 Empirical investigations through artificially generated population

An important aspect of simulation is that one builds a simulation model to replicate the actual system. Simulation allows comparison of analytical techniques and helps in concluding whether a newly developed technique is better than the existing ones. Motivated by Singh and Deo [14], Singh et al. [15] and Maji et al. [16] who have been adopted the artificial population generation techniques, we have generated five sets of independent random numbers of size N (N = 100) namely x_{1_k}', y_{1_k}', x_{2_k}', y_{2_k}' and $z_k'(k = 1, 2, 3, \ldots, N)$ from a standard normal distribution with the help of R-software. By varying the correlation coefficients ρ_{yx} and ρ_{xz}, we have

generated the following transformed variables of the population U with the values of $\sigma_y^2 = 50,\ \mu_y = 10,\ \sigma_x^2 = 100,\ \mu_x = 50,\ \sigma_z^2 = 50$ and $\mu_z = 20$ as

$$y_{1_k} = \mu_y + \sigma_y\left[\rho_{xy}x'_{1_k} + \left(\sqrt{1-\rho_{yx}^2}\right)y'_{1_k}\right]$$
$$x_{1_k} = \mu_x + \sigma_x x'_{1_k}$$
$$z_k = \mu_z + \sigma_z\left[\rho_{xz}x'_{1_k} + \left(\sqrt{1-\rho_{xz}^2}\right)z'_k\right]$$
$$y_{2_k} = y_{1_k}$$
$$\text{and } x_{2_k} = x_{1_k}.$$

We have split total population of size N = 100 into 5 strata each of size 20 $[\text{i.e.}, N_h = 20; (h = 1, 2, ..., 5)]$ taking them sequentially and consider $n'_h = 12$ and $n_h = 8;\ (h = 1, 2, ..., 5)$ for the efficiency comparison of the proposed strategy.

The percentage relative efficiencies the proposed class of estimators t'_p with respect to different estimators (under their respective optimum conditions) are derived through the data set of the artificially generated population are obtained as:

9. Conclusion

From the construction of estimation strategy and efficiency comparison of the proposed methodology, following matters are noted.

1. Form **Table 1**, it is clear that the proposed class of estimators is at least 1% better than the existing one in estimating the population mean.

2. Similarly from **Table 2** it is found that the new estimator is at least 28% better than the existing one.

3. It may also be noted from **Tables 1** and **2** that the artificially generated population is homogeneous (the mean and variance of the respective variables are almost same for different strata) where the natural populations are heterogeneous (the mean and variance of the respective variables are different for different strata) in nature. Our suggested estimators performs with equal efficiency for both the types.

Estimator	**PRE**	
	Population I	**Population II**
$(\bar{y}_{st})$	173.3608	192.951
$\bar{y}^{(h)}_{RP}$	101.1429	131.5654
$\bar{y}^{h}_{cs}$	118.3215	172.226

We use following expression to obtain the percent relative efficiency (PRE) of the proposed estimator t'_p *with respect to different estimators as* $\text{PRE} = \frac{V(\bar{y})}{\text{Min.}V(t'_p)} \times 100$.

Table 1.
PRE of the proposed estimator t'_p *with respect to different estimators through data set of natural population.*

Estimator	PRE
	Artificially generated population
$(\bar{y}_{st})$	179.623
$\bar{y}^{(h)}_{RP}$	128.256
$\bar{y}^{h}_{cs}$	154.879

We use following expression to obtain the percent relative efficiency (PRE) of the proposed estimator t'_p *with respect to different estimators as* $\text{PRE} = \frac{V(\bar{y})}{\text{Min.}V(t'_p)} \times 100$.

Table 2.
PRE of the proposed estimator t'_p *with respect to different estimators through data set of artificially generated population.*

4. The unbiased version of the proposed technique has been obtained which make the proposed class of estimators much more practicable.

Thus, it is found that the proposed estimation technique has addressed the problems of estimation through two-phase stratified sampling which may truthful for real life application where population is especially heterogeneous in nature and stratification is essential. Due to the benefits achieved by the new estimator, the survey statistician may be suggested to use it.

Author details

Partha Parichha[1], Kajla Basu[2] and Arnab Bandyopadhyay[1*]

1 Department of Mathematics, Asansol Engineering College, Asansol, India

2 Department of Mathematics, National Institute of Technology, Durgapur, India

*Address all correspondence to: arnabbandyopadhyay4@gmail.com

References

[1] Sukhatme B. Some ratio type estimators in two-phase sampling. Journal of the American Statistics Associations. 1962;**57**:628-632

[2] Chand L. Some ratio type estimators based on two or more auxiliary variables [unpublished PhD thesis]. Ames, Iowa (USA): Iowa State University; 1975

[3] Kiregyera B. A chain ratio type estimators in finite population double sampling using two auxiliary variables. Metrika. 1980;**17**:217-223

[4] Kiregyera B. Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika. 1984;**31**: 215-226

[5] Tracy DS, Singh HP, Singh R. An alternative to the ratio-cum-product estimator in sample surveys. Journal of Statistical Planning and Inference. 1996; **53**:375-387

[6] Singh HP, Espejo MR. Double sampling ratio-product estimator of a finite population mean in sampling surveys. Journal of Applied Statistics. 2007;**34**(1):71-85

[7] Gupta S, Shabbir J. on the use of transformed auxiliary variables in estimating population mean by using two auxiliary variables. Journal of Statistical Planning and Inference. 2007; **137**:1606-1611

[8] Shukla D, Pathak S, Thakur NS. Estimation of population mean using two auxiliary sources in sample surveys. Statistics in Transition. 2012;**13**(1):21-36

[9] Choudhury S, Singh BK. A class of chain ratio–product type estimators with two auxiliary variables under double sampling scheme. Journal of the Korean Statistical Society. 2012;**41**: 247-256

[10] Parichha P, Basu K, Bandyopadhyay A, Mukhopadhyay P. Development of efficient estimation technique for population mean in two phase sampling using fuzzy tools. Journal of Applied Mathematics, Statistics and Informatics. 2017;**13**(2):5-28. DOI: 10.1515/jamsi-2017-0006

[11] Bandyopadhyay A, Singh GN. Predictive estimation of population mean in two-phase sampling. Communications in Statistics: Theory and Methods. 2016;**45**(14):4249-4267. DOI: 10.1080/03610926.2014.919396

[12] Murthy MN. Sampling Theory and Methods. Calcutta: Statistical Publishing Society; 1967

[13] Koyuncu N, Kadilar C. Family of estimators of population mean using two auxiliary variables in stratified sampling. Communications in Statistics: Theory and Methods. 2009;**38**: 2398-2417

[14] Singh S, Deo B. Imputation by power transformation. Statistical Papers. 2003;**4**:555-579

[15] Singh S, Joarder AH, Tracy DS. Median estimation using double sampling. Australian & New Zealand Journal of Statistics. 2001;**43**(1):33-46

[16] Maji R, Singh GN, Bandyopadhyay A. Estimation of population mean in presence of random non-response in two-stage cluster sampling. Communications in Statistics: Theory and Methods, ISSN: 0361-0926. 2018. DOI: 10.1080/03610926.2018.1478101