IMPROVING VARIANCE PRECISION IN POPULATION STUDIES: THE ROLE OF POST-STRATIFICATION AND AUXILIARY DATA

M. I. Khan[1], S. Qurat Ul Ain[2], M. Younis Shah[3*]
•
[1]Department of Mathematics, Faculty of Science, Islamic University of Madinah, Madinah 42351, Saudi Arabia
[2]Department of data analytics, Harrisburg University of science and technology
[3*] Division of Statistics & Computer Science, 180009 Jammu, Jammu and Kashmir, India.
[1] khanizhar@iu.edu.sa, [4]andrabiqurat@gmail.com, syeedunis121@gmail.com

Abstract

In this study, we propose an enhanced estimator for the finite population variance in the context of post-stratified sampling, incorporating an auxiliary variable to improve accuracy. We derive expressions for the bias and mean square error (MSE) of the proposed estimator, providing an approximation accurate up to the first order. The theoretical analysis highlights the conditions under which the proposed estimator yields lower bias and reduced MSE, making it a more efficient alternative to traditional methods. To evaluate the practical performance of this estimator, we apply it to two real-world data sets, where our results demonstrate a marked improvement in efficiency over existing estimators. The numerical findings confirm that, in post-stratified sampling, the proposed estimator significantly enhances the precision of variance estimation, especially when the auxiliary variable is highly correlated with the study variable. This work not only contributes a more efficient estimator but also provides valuable insights into the application of auxiliary information in post-stratified sampling designs.

Keywords: Post stratification, Auxiliary variable, Estimation, Population variance and Mean squared error.

I. Introduction

This paper presents an enhanced estimator for population variance under post-stratification by utilizing auxiliary information. The use of auxiliary information in survey sampling has long been recognized for its ability to improve the efficiency of estimators for various population parameters, such as the mean, variance, median, mode, quartiles, interquartile range, percentiles, coefficient of variation, and proportions. Numerous methods for effectively incorporating auxiliary information have been extensively documented in survey sampling literature. Estimators of the ratio, product, and regression types leverage the correlation between the study variable and the auxiliary variable to achieve better precision. In this study, the mean square error (MSE) and bias of the proposed estimator are derived under large sample conditions, accurate to the first order of approximation. Theoretical comparisons with existing estimators are made, and conditions are established under which the proposed estimator is more efficient than those previously developed.

The application of stratified random sampling (STRS) presumes that the sizes and structure of sampling frames for each stratum are already available. Whereas the total population size and the percentage of the unit that belongs to each stratum may be known in many existing system, it is possible that the sample frame for every stratum is neither available or would be costly and difficult to construct. In social surveys relevant census information, where it is necessary to partition the heterogeneous population into different sub-groups, the sampling frame may not be available. In such types of situations, STRS is not applicable as such. In order to resolve these difficulties, post stratification technique is applied, in which a sample of necessary size is first selected from the population employing simple random sampling with or without replacement, and it is then stratified using the stratification variable

The procedure is identical to the one of stratified sampling and the only difference is that the allocation into strata is made ex-post. The gain in precision is related to the sample size in each stratum and (inversely) to the difference between the sample weights and the population ones. The standard error for the post-stratified mean estimator is larger than the stratified sampling one, because additional variability is given by the fact that the sample stratum sizes are themselves the outcome of a random process.

Initially introduced the concept of post-stratification [1]. Later extended this work by [2] investigating [3] classic ratio estimator in the context of post-stratification. They first considered the sample sizes within each stratum as fixed and then accounted for variations across possible stratum sample sizes, drawing on a result from [4]. Several researchers have made notable contributions to the development of post-stratification techniques. Important groundwork in the field laid by [5], followed by [6], who further advanced the methodology. Significant strides in [7] refining the theoretical foundations, while key modifications that improved estimator efficiency introduced by [8]. The scope of post-stratification by applying it to new contexts expanded by [9], and [10] provided valuable insights, enhancing the understanding and application of these methods in finite population estimation. More recently, characterized post-stratification product and ratio-type exponential estimators by [11]. While the regression estimator has generally been shown to outperform ratio and product estimators, this is not the case when the regression line of the primary variable on the auxiliary variable passes through a region near the origin [12].

In the literature, seldom is known about estimation of population variance under post-stratified sampling. A number of estimators for the limited population variance of the post-stratified sample mean utilizing data from the auxiliary variable developed by [13], a new ratio estimators in stratified random sampling using the information of an auxiliary attributes suggested by [14], An exponential estimator in the stratified random sampling taking an auxiliary attribute proposed by [15], An efficient exponential ratio estimator allows estimating the population mean in stratified random sampling using an auxiliary variables developed by [16], memory-type ratio and product estimators for the estimation of population variance based on exponentially weighted moving averages (EWMA) statistic proposed by [17], the generalized estimator of population mean using auxiliary attributes in stratified two- phase sampling introduced by [18], the estimation of rare and clustered population mean using stratified adaptive clustered sampling proposed by [19]. The difficulty of estimating the population mean in the situation of post-stratification discussed by [20].

II. Methods

Let the population of size N that is finite and partitioned into L strata of sizes $N_1, N_2, \dots, N_L$ such that $\sum_{h=1}^{L} N_h = N$. Simple random sampling without replacement (SRSWOR) is used to select a sample of size n from the whole population. Following the method of selection from the population, the number of units falling under the h^{th} stratum is indicated. Let n_h be the size of the sample falling in the h^{th} stratum such that $\sum_{h=1}^{L} n_h = n$, here, it is expected that n is large enough so that the probability of n_h being zero is too low. Let $\bar{y}_{hi}$ *and* $\bar{x}_{hi}$ are the observed values of y and x respectively on the i^{th} unit of the h^{th} stratum. Let $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ and $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ represent the sample means corresponding to the population means $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$ and $\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$ of the study variable (y) and auxiliary variable (x) respectively in the h^{th} stratum.

Let $s_{yh}^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2\, and s_{xh}^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2$represent the sample variances corresponding to the population variances $S_{yh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2 and S_{xh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(x_{hi} - \bar{X}_h)^2$ of the study variable (x) and auxiliary variable (x) respectively in the h^{th} stratum. Also, $c_{yh} = \frac{s_{yh}}{\bar{y}_h}, and\ c_{xh} = \frac{s_{xh}}{\bar{x}_h}$ represent the sample coefficient of variation corresponding to the population coefficient of variation $C_{yh} = \frac{S_{yh}}{\bar{Y}_h}\ and\ C_{xh} = \frac{S_{xh}}{\bar{X}_h}$ respectively. Let $r_{yxh} = \frac{s_{yxh}}{s_{yh}s_{xh}}$ represent the sample correlation coefficient corresponding to the population correlation coefficient $\rho_{yxh} = \frac{S_{yxh}}{S_{yh}S_{xh}}$ between their respective subscripts in the h^{th} stratum.

Further $s_{yxh} = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)$ represents the sample covariance corresponding to population variance $S_{yxh} = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)(x_{hi} - \bar{X}_h)$, the study variable (y) and auxiliary variable (x).

Let $\sigma_{pst}^2 = \sum_{h=1}^{L} w_h^2 \frac{S_{yh}^2}{n_h}$ and $\hat{\sigma}_{pst}^2 = \sum_{h=1}^{L} w_h^2 \frac{s_{yh}^2}{n_h}$ represent the post-stratified population and sample variances of y, where $w_h = \frac{N_h}{N}$ is the h^{th} stratum weight respectively. Ignoring the finite population correction factor to make calculations easier in post-stratified sampling

The variance of the estimator $\hat{\sigma}_{pst}^2$ is provided as

$$Var(\hat{\sigma}_{pst}^2) = \sum_{h=1}^{L} \frac{W_h^4}{n_h^3} S_{yh}^4(\lambda_{400h} - 1) \tag{1}$$

$$where\ \ \lambda_{rsth} = \frac{\mu_{rsh}}{\mu_{200h}^{r/2}\mu_{020h}^{s/2}}\ \ and\ \mu_{rsth} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^r\,(x_{hi} - \bar{X}_h)^s$$

A post-stratified regression estimator $\hat{\sigma}_{pst(reg)}^2$ for the finite population variance of the post-stratified sample mean utilizing data from the study's auxiliary variables developed [13] as:

$$\hat{\sigma}_{pst(reg)}^2 = \sum_{h=1}^{L} \frac{W_h^4}{n_h}\left[(S_{yh}^2 + \hat{\Psi}_{220h}(S_{xh}^2 - s_{xh}^2)\right] \tag{2}$$

Where $\hat{\Psi}_{220h}$ is the sample regression coefficient of y on x with corresponding population regression coefficient $\Psi_{220h} = \frac{S_{yh}^2(\lambda_{220h}-1)}{S_{xh}^2(\lambda_{040h}-1)}$ in h^{th} stratum.

The variance of $\hat{\sigma}_{pst(reg)}^2$ is given by

$$Var(\hat{\sigma}_{pst(reg)}^2) = \sum_{h=1}^{L} \frac{W_h^4}{n_h^3} S_{yh}^4(\lambda_{400h} - 1) - \frac{(\lambda_{220h} - 1)}{(\lambda_{040h} - 1)} \tag{3}$$

A conventional ratio estimator $\sigma_{pst(rat)}^2$ for the finite population variance of the post-stratified sample mean utilizing data from the study's auxiliary variables developed by [13] as:

$$\hat{\sigma}_{pst(rat)}^2 = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left((S_{yh}^2 \frac{S_{xh}^2}{s_{xh}^2}\right) \tag{4}$$

The Bias of $\hat{\sigma}_{pst(rat)}^2$ up to the first order of approximation, is given as

$$Bias\left(\hat{\sigma}^2_{pst(rat)}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} S_{yh}^2 (\lambda_{040h} - 1) - (\lambda_{220h} - 1)$$

The MSE of $\hat{\sigma}^2_{pst(rat)}$ up to the first order of approximation, is given by

$$Var\hat{\sigma}^2_{pst(rat)} = \sum_{h=1}^{L} \frac{W_h^4}{n_h^3} S_{yh}^4 (\lambda_{400h} - 1) + (\lambda_{040h} - 1) - 2(\lambda_{220h} - 1) \tag{5}$$

The new efficient type estimator of population variance developed [21] as follows:

$$\hat{\sigma}^2_{pst(\alpha,\delta)} = \sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{yh}^2 \left[2 - \left(\frac{s_{xh}^2}{S_{xh}^2}\right)^{\alpha} exp\left\{\frac{\delta(s_{xh}^2 - S_{xh}^2)}{s_{xh}^2 - S_{xh}^2}\right\}\right] \tag{6}$$

Where α and δ are unknown constants whose values are to be determined such that the MSE of the proposed estimator is minimum.

The Bias of $\hat{\sigma}^2_{pst(\alpha,\delta)}$ up to the first order of approximation, is given as

$$Bias\left(\hat{\sigma}^2_{pst(\alpha,\delta)}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} S_{yh}^2 \frac{(2\alpha + \delta)(2\alpha + \delta - 2)}{8} (\lambda_{040h} - 1) - \frac{(2\alpha + \delta)}{2} (\lambda_{220h} - 1)$$

The MSE of $\hat{\sigma}^2_{pst(\alpha,\delta)}$ up to the first order of approximation, is given as

$$MSE\left(\hat{\sigma}^2_{pst(reg)}\right) = \sum_{h=1}^{L} \frac{W_h^4}{n_h^3} S_{yh}^4 (\lambda_{400h} - 1) - \frac{(\lambda_{220h} - 1)^2}{(\lambda_{040h} - 1)} \tag{7}$$

A new ratio type estimator under post-stratified sampling developed [20] as follows:

$$\hat{\sigma}^2_{pst(k)} = \sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{yh}^2 \left[k + (1-k) exp\left\{\frac{S_{xh}^2 - s_{xh}^2}{S_{xh}^2 - s_{xh}^2}\right\}\right] \tag{8}$$

The Bias of $\hat{\sigma}^2_{pst(k)}$ up to the first order of approximation, is given by

$$Bias\left(\hat{\sigma}^2_{pst(k)}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} S_{yh}^2 \frac{3}{8} (\lambda_{040h} - 1)(1-k) - \frac{1}{2} (\lambda_{220h} - 1)(1-k)$$

The MSE of $\left(\hat{\sigma}^2_{pst(k)}\right)$ up to the first order of approximation, is given by

$$MSE\left(\hat{\sigma}^2_{pst(reg)}\right) = \sum_{h=1}^{L} \frac{W_h^4}{n_h^3} S_{yh}^4 (\lambda_{400h} - 1) - 2\frac{(\lambda_{220h} - 1)}{(\lambda_{040h} - 1)} \tag{9}$$

III. Proposed Estimator

In the line with the direction of study carried out by [22], we proposed an improved estimator for estimating the finite population variance under the method of post stratified sampling. The bias and MSE of the existing and proposed estimator are derived up to the first order of approximation. The performance of the proposed estimator is the best as compared to existing counterparts in terms of efficiency.

$$\hat{\sigma}^2_{pst(YS)} = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[K_{1h} S_{yh}^2 + K_{2h}(S_{xh}^2 - s_{xh}^2)\right] exp\left(\frac{S_{xh}^2 - s_{xh}^2}{S_{xh}^2 - s_{xh}^2}\right) \tag{10}$$

Where K_{1h} and K_{2h} are unknown constants whose values are to be determined such that the MSE of the proposed estimator is minimum.

For examining the large sample characteristics of the developed estimator, $\hat{\sigma}^2_{pst(YS)}$ we define the random variables up to the first order of approximation as:

$$s_{yh}^2 = S_{yh}^2(1 + \varepsilon_{2h}), \qquad s_{xh}^2 = S_{xh}^2(1 + \varepsilon_{2h}) \quad such\ that\ \ E(\varepsilon_{2h}) = E(\varepsilon_{2h}) = 0$$

Also,

$$E(\varepsilon_{2h}^2)=\frac{1}{n_h}(\lambda_{400h}-1),\;\; E(\varepsilon_{4h}^2)=\frac{1}{n_h}(\lambda_{040h}-1)\quad and\quad E(\varepsilon_{2h}\varepsilon_{4h})=\frac{1}{n_h}(\lambda_{200h}-1)$$

When we use the values of above terms in eq. (10), we have

$$\hat{\sigma}_{pst(YS)}^2=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\left(K_{1h}S_{yh}^2(1+\varepsilon_{2h})-K_{2h}S_{xh}^2-S_{xh}^2(1+\varepsilon_{4h})\right)\right]exp\left(\frac{S_{xh}^2-S_{xh}^2(1+\varepsilon_{4h})}{S_{xh}^2-S_{xh}^2(1+\varepsilon_{4h})}\right)$$

Now expanding the right hand side along with the exponential term of the above equation up to the first degree of approximation, we

$$\hat{\sigma}_{pst(YS)}^2\approx\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\left(K_{1h}S_{yh}^2(1+\varepsilon_{2h})-K_{2h}S_{xh}^2-S_{xh}^2(1+\varepsilon_{4h})\right)\right]\left(1-\frac{1}{2}\varepsilon_{4h}+\frac{3}{8}\varepsilon_{4h}^2\right)\tag{11}$$

Subtracting S_{yh}^2 from both sides of (11), we obtain

$$\begin{aligned}\left(\hat{\sigma}_{pst(YS)}^2-S_{yh}^2\right)=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\Big[\Big(&S_{yh}^2+K_{1h}S_{yh}^2+K_{1h}S_{yh}^2\varepsilon_{2h}-\frac{1}{2}K_{1h}S_{yh}^2\varepsilon_{4h}\\&-K_{2h}S_{xh}^2\varepsilon_{4h}-\frac{1}{2}K_{1h}S_{yh}^2\varepsilon_{2h}\varepsilon_{4h}+\frac{3}{8}K_{1h}S_{yh}^2\varepsilon_{4h}^2-\frac{1}{2}K_{2h}S_{xh}^2\varepsilon_{4h}-S_{xh}^2\Big)\Big]\end{aligned}\tag{12}$$

By applying expectation on both sides of eq. (12), we get the Bias of $\hat{\sigma}_{pst(YS)}^2$ as

$$\begin{aligned}Bias\left(\hat{\sigma}_{pst(YS)}^2\right)=\sum_{h=1}^{L}\frac{W_h^2}{n_h^2}\Big[\Big(&S_{yh}^2\\&+K_{1h}S_{yh}^2\Big\{1+\frac{3}{8}(\lambda_{040h}-1)-\frac{1}{2}(\lambda_{330h}-1)+\frac{1}{2}K_{2h}S_{xh}^2(\lambda_{040h}-1)\Big\}\Big)\Big]\end{aligned}\tag{13}$$

Equation (12) can be squared on both sides we get,

$$\begin{aligned}MSE\left(\hat{\sigma}_{pst(YS)}^2\right)=\sum_{h=1}^{L}\frac{W_h^4}{n_h^2}\Big[&S_{yh}^4+K_{1h}^2S_{yh}^4+K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{400h}-1)-\frac{1}{4}K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{040h}-1)\\&+K_{2h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{040h}-1)+\frac{1}{4}K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{220h}-1)^2+2K_{1h}S_{yh}^4\\&-2K_{1h}S_{yh}^4\frac{1}{n_h}(\lambda_{220h}-1)+\frac{6}{8}K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{040h}-1)+K_{2h}S_{yh}^2S_{yh}^2\frac{1}{n_h}(\lambda_{040h}-1)\\&-K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{220h}-1)+\frac{6}{8}K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{040h}-1)+K_{1h}^2S_{yh}^4\frac{1}{n_h}(\lambda_{220h}-1)\\&-2K_{1h}K_{2h}S_{yh}^2S_{yh}^2\frac{1}{n_h}(\lambda_{220h}-1)\\&+K_{1h}^2S_{yh}^2\frac{1}{n_h}(\lambda_{040h}-1)\Big]\end{aligned}\tag{14}$$

In the above equation, K_{1h} and K_{2h} are unknown constants whose values are to be determined such that the MSE of the proposed estimator is minimum and their optimal values are obtained by differentiating partially equation (14) with respect to K_{1h} and K_{2h} and then equating to zero as:

$$\frac{\partial MSE\left(t_{(YS)}/n_h\right)}{\partial K_{1h}}=0$$

$$K_{1hopt} = \left(\frac{\lambda_{040h}-1}{8}\right)\left[\frac{8-(\lambda_{040h}-1)}{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2}\right]$$

$$\frac{\partial MSE\left(t_{(YS)}/n_h\right)}{\partial K_{2h}} = 0$$

$$K_{2hopt} = \left(\frac{S_{yh}^2}{8S_{xh}^2}\right)\left[\frac{4(\lambda_{220h}-1)^2-(\lambda_{040h}-1)(\lambda_{220h}-1)^2+(\lambda_{040h}-1)}{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2}\right]$$

The reduced MSE of $\hat{\sigma}^2{}_{Pst(YS)}$ at the optimum values of K_{1h} and K_{2h} is obtained as

$$MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min} = \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4 \left\lfloor \frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3 - 16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}} \right\rfloor \tag{15}$$

IV. Efficiency Comparison

The developed estimator has been theoretically compared to the competing estimators of the population variance. As a result, when (1), (3), (5), (7), (9), (11) and (17) are compared, it is clear that the suggested estimator $\hat{\sigma}^2{}_{Pst(YS)}$, would be more efficient than [13] usual unbiased estimator$\hat{\sigma}^2{}_{Pst}$, [13] regression estimator $\hat{\sigma}^2{}_{Pst(reg)}$, [13] regression estimator $\hat{\sigma}^2{}_{Pst(rat)}$, [21] efficient type estimator $\hat{\sigma}^2{}_{Pst(\alpha,\delta)}$, [20] new ratio type estimator $\hat{\sigma}^2{}_{Pst(k)}$.

1. By taking eq. (1) and (15), we get

$$\left\{Var\left(\hat{\sigma}^2{}_{Pst}\right) - MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min}\right\} > 0 \tag{16}$$

$$\hat{\sigma}^2{}_{Pst}\ if:\ \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4 \left\lfloor (\lambda_{400h}-1) - \left\lfloor \frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3 - 16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}} \right\rfloor \right\rfloor > 0$$

2. By taking eq. (3) and (15), we get

$$\left\{Var\left(\hat{\sigma}^2{}_{Pst(reg)}\right) - MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min}\right\} > 0 \tag{17}$$

$$\hat{\sigma}^2{}_{Pst(reg)}\ if\ :\ \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4\left[(\lambda_{400h}-1)-\frac{(\lambda_{220h}-1)^2}{(\lambda_{040h}-1)}\right.$$

$$\left.-\left[\frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3 \\ -16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}\right]\right]>0$$

3. By taking eq. (5) and (15), we get

$$\left\{Var\left(\hat{\sigma}^2{}_{Pst(rat)}\right)-MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min}\right\}>0 \tag{18}$$

$$\hat{\sigma}^2{}_{Pst(reg)}\ if\ :\ \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4\left[\begin{bmatrix}(\lambda_{400h}-1)+(\lambda_{040h}-1)\\ -2(\lambda_{220h}-1)\end{bmatrix}\right.$$

$$\left.-\left[\frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3 \\ -16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}\right]\right]>0$$

4. By taking eq. (7) and (15), we get

5. $$\left\{Var\left(\hat{\sigma}^2{}_{Pst(\alpha,\delta)}\right)-MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min}\right\}>0 \tag{19}$$

$$\hat{\sigma}^2{}_{Pst(\alpha,\delta)}\ if\ :\ \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4\left[(\lambda_{400h}-1)-\frac{(\lambda_{220h}-1)^2}{(\lambda_{040h}-1)}\right.$$

$$\left.-\left[\frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3 \\ -16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}\right]\right]>0$$

5. By taking eq. (9) and (15), we get

$$\left\{Var\left(\hat{\sigma}^2{}_{Pst(k)}\right)-MSE\left(\hat{\sigma}^2{}_{Pst(YS)}\right)_{min}\right\}>0 \tag{20}$$

$$\hat{\sigma}^2_{Pst(\alpha,\delta)}\ if:\ \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h^2}\right)S_{yh}^4\left[(\lambda_{400h}-1)-2\frac{(\lambda_{220h}-1)}{(\lambda_{040h}-1)}-\left[\frac{64\{(\lambda_{400h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}-(\lambda_{040h}-1)^3-16(\lambda_{040h}-1)\{(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}{64\{(\lambda_{040h}-1)+(\lambda_{040h}-1)(\lambda_{040h}-1)-(\lambda_{220h}-1)^2\}}\right]\right]>0$$

As the conditions (16)-(20) are always satisfied, it is inferred that the proposed estimator is more efficient than the other existing estimators under all cases in theory.

V. Empirical Study

5. Numerical Analysis

We will take into account the data set to assess the efficiency of the suggested estimators. The characteristics of the population are described as below:

Population 1:

Let y is the output and x is the fixed capital of 80 factories [23]. The data have classified arbitrarily into four strata as $x \le 500, 500 < x \le 1000$, $1000 < x \le 2000$, and x>2000, respectively.

y: Output
x: Fixed capital

Table 1(a): *Statistical Description of the Population:*

Constants	N_h	n_h	$\bar{Y}_h$	$\bar{X}_h$	s_{yh}^2	λ_{400h}	λ_{040h}	λ_{220h}
Stratum I	20	11	3006.55	65.90	572819.20	3.45	1.55	1.49
Stratum II	31	18	4687.62	141.90	433681.58	1.56	3.09	1.73
Stratum III	13	8	6496.23	392.38	162104.69	1.98	1.49	1.56
Stratum IV	16	8	7795.31	749.50	426528.63	2.35	1.91	2.05

Population 2:

we use the data concerning the number of teachers as study variable and the number of students as auxiliary variable in both primary and secondary schools for 923 districts at six regions (as 1: Marmara 2: Agean 3: Mediterranean 4: Central Anatolia 5: Black Sea 6: East and Southeast Anatolia) in Turkey (Source: Ministry of Education, Republic of Turkey).

Y: number of teachers

X: Number of students

Table 2(a): *Statistical Description of the Population:*

Constants	N_h	n_h	$\bar{Y}_h$	$\bar{X}_h$	s_{yh}^2	λ_{400h}	λ_{040h}	λ_{220h}
Stratum I	127	31	703.740	20804.59	781163.9	3.94783	6.251589	3.720488
Stratum II	117	21	413.000	9211.79	415924.8	17.33181	19.35622	18.35209
Stratum III	103	29	573.174	14309.30	1068054	15.87136	16.3073	16.09088
Stratum IV	170	38	424.664	9478.85	657047.8	13.60375	11.67999	11.65605
Stratum V	205	22	267.029	5569.94	162936.9	22.31908	23.14865	22.30021
Stratum VI	201	39	393.840	12997.59	506549	21.49882	24.26014	21.79386

Table 3: *Conditional values of different estimators using real data sets:*

Conditional values	Population I	Population II
Conditional values I	2964740.76	628422.27
Conditional values II	1651010.71	29677.75
Conditional values III	2308528.95	41662.86
Conditional values IV	1254126.31	29677.75
Conditional values V	1651010.71	224246.86
Conditional values VI	6579180.77	706578.08

Table 4: *Bias values of different estimators using real data sets:*

Estimators	Population I	Population II
$\sigma^2_{(pst)}$	-	-
$\sigma^2_{pst(Reg)}$	-	-
$\sigma^2_{Pst(Rat)}$	248.85	111.00
$\sigma^2_{Pst(\alpha,\delta)}$	-6154.30	-124519.90
$\sigma^2_{Pst(T)}$	-6392.45	-218847.51
$\sigma^2_{Pst(k)}$	-104.74	1024.01
$\sigma^2_{Pst(YS)}$	247404838.9	4.92057E+19

Table 1(b): *MSE and PRE of suggested estimator in relation to $\hat{\sigma}^2{}_{Pst}$*

No.	1	2	3	4	5	6	7
Estimators	$\sigma^2_{(pst)}$	$\sigma^2_{pst(Reg)}$	$\sigma^2_{Pst(Rat)}$	$\sigma^2_{Pst(T)}$	$\sigma^2_{Pst(\alpha,\delta)}$	$\sigma^2_{Pst(k)}$	$\boldsymbol{\sigma^2_{Pst(YS)}}$
MSE	3574695.71	2260965.66	2918483.90	1864081.26	2260965.66	7189135.72	**609954.95**
PRE	100.00	157.40	121.28	191.77	157.40	49.72	**586.06**

Table 2(b): *MSE and PRE of suggested estimator in relation to* $\hat{\sigma}^2{}_{Pst}$

No.	1	2	3	4	5	6	7
Estimators	$\sigma^2_{(pst)}$	$\sigma^2_{pst(Reg)}$	$\sigma^2_{Pst(Rat)}$	$\sigma^2_{Pst(\alpha,\delta)}$	$\sigma^2_{Pst(T)}$	$\sigma^2_{Pst(k)}$	$\boldsymbol{\sigma^2_{Pst(YS)}}$
MSE	651146.08	52401.56	64386.67	52401.56	246970.67	729301.89	**22723.81**
PRE	100.00	1242.61	1011.31	1242.61	263.65	89.28	**2865.48**

Where

$$PRE = \frac{Var\left(\hat{\sigma}^2{}_{Pst}\right)}{MSE\left(\hat{\sigma}^2{}_{Pst(i)}\right)} \times 100 \quad : \quad i = Re, Rat, (\alpha, \delta), T, and\ k.$$

VI. Conclusion

The population variance of the research variable can be effectively estimated using auxiliary data through an improved estimator under post-stratification. The proposed single and combined classes of estimators, such as bias and mean square error (MSE), are derived approximately to the first order of accuracy. Under specific efficiency conditions, the recommended estimator significantly outperforms existing separate and combined estimators. Additionally, empirical research is conducted using both artificially generated symmetric and asymmetric populations, as well as real-world data, to validate the theoretical findings. The results demonstrate that the proposed estimator is more efficient, with a lower MSE and higher percentage relative efficiency (PRE) than the alternatives. This study provides clear evidence supporting the robustness and practicality of the suggested estimator in experimental surveys. Given its superior performance, we strongly recommend its adoption over traditional estimators for post-stratification variance estimation. The integration of theoretical and empirical analyses makes this research highly credible, insightful, and impactful for statistical applications.

VII. Acknowledgement

The author wishes to extent his sincere gratitude to the Deanship of Scientific Research at the Islamic University of Madinah for support provided to the Post- Publication Program (3).

References

[1] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. 1953. Sample Survey Methods and Theory, Vol. 1, New York: John Wiley and Sons.

[2] Ige, A. F. and Tripathi, T. P. (1989). Estimation of population mean using post-stratification and auxiliary information. Abacus, 265-276.

[3] Cochran, W. G. (1940). Sampling Techniques. New York: John Wiley and Sons.

[4] Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian. Annals of Mathematical Statistics, 16: 50-61.

[5] Holt, D., Smith, T. M. F. (1979). Post Stratification. J. Roy. Statistist. Assoc. 61:1172–1183.

[6] Jagers, P. Oden, A., Trulsson, L. (1985). Post-stratification and ratio estimation: usages of auxiliary information in survey sampling and opinion polls. Int. Statist. Rev. 53: 221–238

[7] Vallient, R. (1993). Post stratification and conditional variance estimation. J. Amer. Statist. Assoc., 88:89–96.

[8] Agarwal, M. C. and Panda, K.B. (1993). An efficient estimator in post stratification. Metron 179-188.

[9] Singh, H. P., Ruiz Espejo, M. (2003).Improved post stratified estimation. Bulletin of the International Statistical Institute, 54th session, contributed papers, vol. LX, Two Books, Book 2, 341–342.

[10] Vishwakarma, G. K., Singh, H. P., Singh, S. (2010). A family of estimators of population means using multi-auxiliary variate and post stratification. Nonlinear Anal. Model. Control 15(2):233–253.

[11] Rajesh Tailor & Ritesh Tailor and Sunil Chouhan, (2017). "Improved Ratio- and Product-Type Exponential Estimators for Population Mean in Case of Post-Stratification," Communications in Statistics - Theory and Methods, 46(21), 10387-10393.

[12] Cochran, W. G. (1977). Sampling Techniques 3rd ed. New York: John Wiley and Sons.

[13] Masood, S and Shabbir, J. (2015). On some families of estimators of variance of post-stratified sample mean using two auxiliary variables. Communications in Statistics - Theory and Methods., 2398-2415.

[14] Zaman, T. (2019). Efficient estimators of population mean using auxiliary attribute in stratified random sampling. Advances and Applications in Statistics, 56(2), 153-171.

[15] Zaman, T., & Kadilar, C. (2020). On estimating the population mean using auxiliary character in stratified random sampling. Journal of Statistics and Management Systems, 23(8), 1415-1426.

[16] Zaman, T. (2021). An efficient exponential estimator of the mean under stratified random sampling. Mathematical Population Studies, 28(2), 104-121.

[17] Qureshi, M.N., Kadilar, C. and Hanif, M. (2020). Estimation of rare and clustered population mean using stratified adaptive cluster sampling, Environmental and Ecological Statistics, 27: 151- 170.

[18] Rana, Q., Qureshi, M. N., and Hanif, M. (2022). Generalized estimator for population mean using auxiliary attributes in stratified two- phase sampling. Journal of Statistics, Theory and Applications, 21 (2): 44–57.

[19] Qureshi, M. N., Tariq, M. U., Xiong, Y. and Hanif, M. (2023). Estimation of heterogeneous population variance using Memory-type estimators based on EWMA statistic in the presence of measurement error for time-scaled surveys.

[20] Khalid Ul Islam Rather, M. Iqbal Jeelani, M. Younis Shah, S. E. H. Rizvi and M. K. Sharma (2022). A new ratio type estimator for computation of population mean under post-Stratification. JAMSI, 18 (1).

[21] Yadav, S. K., Mishra, S. S., Mishra, P. P., & Singh, R. S. (2018). New efficient class of estimators for the population variance Int. J. Agricult. Stat. Sci. Vol. 14 (2), 491-496.

[22] Ahmed, S., Sardar, H., Shabbir, J., Zahid, E., Aamir, M and Onyango, R. (2022). Improved Estimation of Finite Population Variance Using Dual Supplementary Information under Stratified Random Sampling. Mathematical Problems in Engineering, 12.

[23] Murthy, M. N. (1967). Sampling Theory and Methods. Calcutta, India: Statistical Publishing Society.