

A Generalized Power Ratio Estimator Under Stratified Sampling: Empirical Evaluation and Jack-Knife Bias Correction Using Machine Learning Auxiliary Information

B. S. Kambo¹

¹Former Deputy Director, Ministry of Statistics & PI, Government of India

Publication Date: 2026/10/02

Abstract:

➤ Background:

Ratio estimators in survey sampling leverage auxiliary variables correlated with the target variable to reduce sample variance. When auxiliary variables are generated from predictive machine learning models, optimizing the transformation exponent K offers substantial precision gains over classical ratio methods.

➤ Objectives:

This study formulates and empirically evaluates a Generalized Power Ratio Estimator under both Stratified Simple Random Sampling Without Replacement (SRSWOR) and Stratified Probability Proportional to Size with Replacement (PPSWR). Furthermore, we integrate Stratified Jackknife Technique to address small-sample ratio bias.

➤ Methodology:

Utilizing a population dataset ($N=707$, MHFS5) across sample sizes $n \in (30, 60, 90, 120)$, we execute 5,000 Monte Carlo simulation iterations. Performance is evaluated using Mean Squared Error (MSE) and Percent Relative Efficiency (% PRE) against standard stratified baseline estimators.

➤ Results:

Under SRSWOR, the Generalized Power Ratio Estimator tuned to its theoretical optimum ($K_{opt}=1.08$) achieves superior efficiency (171.63% PRE at $n=30$) compared to the classical ratio estimator (170.53%). The Jackknife transformation eliminates small-sample bias while fully preserving efficiency (171.54% PRE). Under PPSWR, size-proportional weighting yields dramatic precision gains exceeding 270% PRE, with the classical ratio exponent ($K=1.00$) exhibiting slight advantages due to scale transformation effects.

➤ Conclusion:

The proposed framework provides an efficient, virtually unbiased estimation pipeline for finite population parameters when auxiliary predictions are available.

Keywords: Power Ratio Estimator, Jackknife, Calibration, Pre, Model Assisted Approach, Svr, Auxiliary Variable.

How to Cite: B. S. Kambo (2026) A Generalized Power Ratio Estimator Under Stratified Sampling: Empirical Evaluation and Jack-knife Bias Correction Using Machine Learning Auxiliary Information. *International Journal of Innovative Science and Research Technology*, 11(9), 2431-2440. <https://doi.org/10.38124/ijisrt/26sep760>

I. INTRODUCTION

The problem of improving upon the unbiased estimator by ratio and product method of estimation using suitable auxiliary information has received considerable attention in the sampling theory. Whenever the auxiliary information is available, the sampler wants to utilize it in the method of

estimation which yields maximum efficiency. Estimating finite population parameters with high precision remains a core objective in survey statistics. When auxiliary information Z positively correlates with the study variable Y of interest, ratio-type estimators significantly improve estimation efficiency compared to standard sample means. Recent advances in statistical modelling frequently

utilize prediction outputs from Machine Learning models (such as Support Vector Regression or Ensemble frameworks) as auxiliary variables Z . However, classical ratio estimators assume a rigid linear relationship through the origin, fixing the transformation exponent at $K=1.00$. When the actual relationship deviates from this strict assumption, the classical estimator yields sub-optimal variance reduction. This paper addresses this limitation by proposing.

- Generalized Power Ratio Estimator within stratified framework
- Deriving the optimal power exponent $K_{opt} = \rho_{yz} \frac{C_y}{C_z}$ where ρ_{yz} is correlation coefficient between y and z and C_y and C_z are coefficient of variation in y and z respectively.
- Stratified Delete-One Jackknife Bias Reduction Framework.
- Comparing comparative performance across both equal-probability (SRSWOR) and unequal-probability (PPSWR) sampling designs.

II. REVIEW OF LITERATURE

The fundamental challenge in survey sampling is to accurately estimate key characteristics of a population, such as its mean or total, by observing only a small sample. This task requires statistical methods that produce estimates that are not only precise but also efficient, maximizing the information gained from finite resources. A powerful strategy for achieving this is the use of auxiliary information—readily available data that is correlated with the study variable of interest. By strategically incorporating this secondary data into the estimation process, statisticians can significantly enhance the precision of their findings.

The origin of leveraging auxiliary information through ratios dates back spans before formal statistical inference was codified. Early demographic estimates by Graunt (1662) and Laplace (1802) utilized sample-to-population ratios to estimate the population of England and France, respectively, laying early conceptual groundwork for ratio-based reasoning (Bakker, 2003). Modern sampling theory formalized this mechanism with Cochran (1940), who introduced the classical ratio estimator in agricultural sampling for estimating cereal yield from total produce. Cochran (1940, 1977) and Hansen and Hurwitz (1946) established the foundational condition for efficiency: the classical ratio estimator achieves a smaller variance than the simple mean estimator if and only if the correlation coefficient ρ_{yz} exceeds half the ratio of coefficients of variation, $\rho_{yz} > \frac{1}{2} \frac{C_y}{C_z}$.

➤ Evolution of Ratio, Product, and Modified Estimators

To handle negative or inverse auxiliary correlations, Murthy (1964) introduced the classical product estimator, proving its superiority over simple random sampling when $\rho_{yz} < -\frac{1}{2} \frac{C_y}{C_z}$. Murthy (1967) later generalized sampling design frameworks to incorporate auxiliary information systematically. Further theoretical extensions arose in the 1960s and 1970s to address non-linear relationships and multi-auxiliary structures:

• Multi-Auxiliary & Composite Estimation:

Raj (1965) extended the framework to utilize multiple auxiliary variables simultaneously, which was complemented by Chand (1975) through chain-ratio estimators for multi-stage contexts. Singh (1967) introduced a hybrid ratio-cum-product estimator to handle situations where one auxiliary variable exhibits positive correlation while another exhibits negative correlation. Christiansen, Singh and Sedory (2026) have established a unified theoretical framework and offered a practical guidance for selecting efficient estimators in finite and large population settings by demonstrating that many ratio-type, product-type, exponential, logarithmic, and regression-type estimators proposed in recent decades are special cases of the pioneering general class of estimators introduced by Srivastava (1971, 1980) and the wider class proposed by Singh and Upadhyaya (1986).

• Modified and Polynomial Formulations:

Gupta (1971) and Reddy (1973) examined alternative functional forms of ratio and product estimators, identifying parameter conditions under which variance could be minimized. Gupta (1978) extended this to higher-order polynomial ratio estimators to capture curvilinear auxiliary structures.

• Non-Linear Dynamics and Power Modifications:

A key limitation of classical ratio estimation is the implicit assumption that the regression of Y on X is strictly linear and passes through the origin. The foundational breakthrough in generalizing auxiliary structures through power parameters was initiated by Srivastava (1967), who introduced generalized ratio-type estimators of the form $\bar{y}_n \left[\frac{\bar{X}_N}{\bar{x}_n} \right]^K$ where K and derived optimal powers. Subsequently, when the regression line exhibits a non-zero intercept or non-linear curvature, Padam Singh (1980) proposed modifying the ratio method of estimation by estimating functional relationships via least squares to construct transformed auxiliary variables $Z = f(X)$.

➤ Power Transformations, Bias Reduction, and Resampling

- While generalized power-transformed ratio estimators substantially improve efficiency and reduce Mean Squared Error (MSE), ratio-type constructions inherently suffer from small-sample bias of order $O(n^{-1})$.
- Unbiased Estimators: Early efforts to eliminate bias led to the Hartley and Ross (1954) exact unbiased ratio estimator, providing an alternative when small sample sizes make $O(n^{-1})$ bias non-negligible.

➤ Calibration Theory and Modern Predictive Frameworks

In recent decades, auxiliary estimation has expanded from classical analytical forms to general calibration and computational machine learning paradigms:

• Calibration Estimation:

Deville and Sandal (1992) revolutionized survey sampling by formulating calibration estimation, where sample weights are adjusted to match known population totals of auxiliary variables. Singh et al. (1998) extended calibration

to higher-level variance estimation, while Lesage (2011) incorporated estimating equations into calibration for complex non-linear parameters.

➤ Modern Extensions:

Singh (2003) provided a comprehensive synthesis of advanced sampling methodologies, detailing multi-phase and calibration designs. In modern applications, researchers have increasingly integrated predictive modelling with survey designs. Chen et al. (2010) established Bayesian penalized spline model-based inference under unequal probability sampling. Singh (2013) Singh, S. (2013). A dual problem of calibration of design weights. Singh, S., and Sedory, S. A. (2016). Two-step calibration of design weights in survey sampling. Zahid et al. (2021) expanded generalized classes of estimators to handle measurement error and non-response in stratified sampling, while Shabbir and Onyango (2022) developed highly efficient unbiased classes of estimators for finite population parameters.

III. METHODOLOGY AND MATHEMATICAL FRAMEWORK

➤ Finite Population Structure and Auxiliary Machine Learning Modelling

Let $U = \{1, 2, \dots, N\}$ denote a finite population of size $N = 707$ partitioned into L non-overlapping and mutually exhaustive strata of known sizes $N_1, N_2, \dots, N_L$ such that: $\sum_{h=1}^L N_h = N$

The true population parameters of study interest are the population means:

$$\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h, \quad \bar{Z} = \sum_{h=1}^L W_h \bar{Z}_h$$

Where

$$\bar{Y}_h = \sum_{i=1}^{N_h} W_h \bar{y}_h$$

And

$$\bar{Z}_h = \sum_{i=1}^{N_h} W_h \bar{z}_h$$

Are the stratum-specific population means.

➤ Sampling Schemes

• Design A: Stratified Simple Random Sampling Without Replacement (SRSWOR)

Under Stratified SRSWOR, a sample s_h of size n_h ($n_h \geq 2$) is selected independently without replacement from stratum h . The overall sample size is $n = \sum_{h=1}^L n_h$

Let $W_h = \frac{N_h}{N}$ represent known stratum weight for the h -th stratum ($h = 1, 2, \dots, L$). Associated with each unit h_i denoting the i th unit in the h th stratum—are:

y_{hi} : The value of the study variable Y .

z_{hi} : The value of the continuous auxiliary variable Z .

x_{hi} : The value of a strictly positive size measure $X > 0$ (e.g., literacy index).

• Construction of the Machine Learning Auxiliary Variable Z

Rather than relying on a raw predictor, the auxiliary variable Z was generated by fitting 17 distinct Machine Learning (ML) regression models across the population feature space (including linear, tree-based, ensemble, and kernel-based algorithms) to predict the study variable Y .

Among all 17 evaluated models, Support Vector Regression (SVR) achieved the best predictive fit, exhibiting the lowest Root Mean Squared Error (RMSE = 1.66) and highest coefficient of determination ($R^2 = 0.4430$). Consequently, the fitted predictions from the SVR model were extracted as the study auxiliary variable Z : $z_{hi} = \hat{y}_h^{SVR}$

This SVR auxiliary prediction achieved a high linear correlation with Y ($\rho_{yz} = 0.6656$, $P < 0.00001$) implying thereby that the correlation coefficient ($\rho_{yz} = 0.6656$) is significantly different from zero, indicating a strong positive linear relationship between Y and Z in the population establishing a strong baseline for ratio and power-transformation variance reduction.

Sample sizes are allocated across strata using Proportional Allocation:

$$n_h = n \cdot W_h = n \frac{N_h}{N}$$

The sample means within stratum h are:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \text{ and } \bar{z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}$$

• Design B: Stratified Probability Proportional to Size with Replacement (PPSWR)

Under Stratified PPSWR, n_h units are selected independently with replacement from stratum h , where the probability of selecting unit (h, i) in a single draw is proportional to its auxiliary size measure x_{hi} :

$$p_{hi} = \frac{x_{hi}}{\sum_{i=1}^{N_h} x_{hi}} \text{ such that } \sum_{i=1}^{N_h} p_{hi} = 1 \quad (\forall h = 1, \dots, L)$$

➤ Baseline Estimators

• Unassisted Stratified SRSWOR Mean $\bar{y}_{st}$

The standard unbiased estimator for the population mean $\bar{Y}$ under Stratified SRSWOR is: $\bar{y}_{st}$

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h$$

Its exact sampling variance is:

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{1-f_h}{n_h} \right) S_{h,i}^2$$

Where $f_h = n_h / N_h$ is the stratum sampling fraction, and

$$S_{h,i}^2 = 1 / (N_h - 1) \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2$$

Is the stratum variance of Y.

- *Hansen-Hurwitz Stratified PPSWR Estimator $\bar{y}_{st, pps}$*

Under PPSWR, the unbiased Hansen-Hurwitz estimator for the population mean $\bar{Y}$ is defined as:

$$\bar{y}_{st, pps} = \sum_{h=1}^L W_h Y_{st, pps},$$

Where

$$\bar{y}_{h, pps} = \frac{1}{N_h \cdot n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{p_{hi}}$$

➤ *Proposed Combined Generalized Power Ratio Estimator*

- *Mathematical Formulation*

To generalize classical ratio estimation, we formulate the Combined Generalized Power Ratio Estimator under Stratified SRSWOR as:

$$\hat{Y}_{power} = \bar{y}_{st} \left(\frac{\bar{z}}{\bar{z}_{st}} \right)^K$$

Where $K \in \mathbb{R}$ is a real-valued transformation exponent to be optimized,

$\bar{z}_{st} = \sum_{h=1}^L W_h \bar{z}_h$ Note that $K = 0$ collapses to the baseline stratified sample mean: $\bar{y}_{st}$

While $K = 1$ yields the classical combined ratio estimator $\bar{y}_{st} \left(\frac{\bar{z}}{\bar{z}_{st}} \right)$

- *Derivation of First-Order MSE and Optimal Exponent K_{opt}*

To derive the Mean Squared Error (MSE) of $\hat{Y}_{power}$, define the relative estimation error terms:

$$e_0 = \frac{(\bar{y}_{st} - \bar{Y})}{\bar{Y}}, \quad e_1 = \frac{(\bar{z}_{st} - \bar{Z})}{\bar{Z}}$$

By construction, $E(e_0) = E(e_1) = 0$. The expectations of squared and cross-error terms are:

$$E(e_0^2) = \frac{V(\bar{y}_{st})}{\bar{Y}^2}, \quad E(e_1^2) = \frac{V(\bar{z}_{st})}{\bar{Z}^2}, \quad E(e_0 e_1) = \frac{Cov(\bar{y}_{st}, \bar{z}_{st})}{\bar{Y} \bar{Z}}$$

Expressing the estimator in terms of e_0 and e_1 :

$$\hat{Y}_{power} = \bar{Y} (1 + e_0) (1 + e_1)^{-K}$$

Applying a Taylor-series expansion to $(1 + e_1)^{-K}$ up to second-order terms:

$$(1 + e_1)^{-K} \approx 1 - K e_1 + \frac{K(K+1)}{2} e_1^2$$

Substituting this expansion into the estimator yields:

$$\hat{Y}_{power} - \bar{Y} = \bar{Y} \left[e_0 - K e_1 + \frac{K(K+1)}{2} e_1^2 - K e_0 e_1 \right]$$

Squaring both sides and retaining up to $O(n^{-1})$ error order:

$$(\hat{Y}_{power} - \bar{Y})^2 \approx \bar{Y}^2 [e_1^2 + K^2 e_1^2 - 2K e_0 e_1]$$

Taking mathematical expectations gives the first-order approximation of MSE:

$$MSE(\hat{Y}_{power}) \approx V(\bar{y}_{st}) + R^2 K^2 V(\bar{z}_{st}) - 2 R K Cov(\bar{y}_{st}, \bar{z}_{st})$$

Where $R = \frac{\bar{Y}}{\bar{Z}}$ is the population ratio. Differentiating $MSE(\hat{Y}_{power})$ with respect to K and setting the derivative to zero yields the theoretical optimal power exponent:

$$K_{opt} = \frac{Cov(\bar{y}_{st}, \bar{z}_{st})}{(R \cdot V(\bar{z}_{st}))} = \rho_{yz} \left(\frac{C_y}{C_z} \right)$$

Where ρ_{yz} is the overall population correlation coefficient between Y and Z, $C_y = \frac{S_y}{\bar{Y}}$ and $C_z = \frac{S_z}{\bar{Z}}$ are the coefficients of variation of Y and Z respectively.

- *Derivation of First-Order Bias*

Applying a second-degree Taylor-series expansion to retain leading $O(n^{-1})$ bias terms, taking mathematical expectations $E[\cdot]$ yields:

$$Bias(\hat{Y}_{power}) \approx \bar{Y} \left[\frac{K(K+1)}{2} \left(\frac{V(\bar{z}_{st})}{\bar{Z}^2} \right) - K \frac{Cov(\bar{y}_{st}, \bar{z}_{st})}{(\bar{Y} \bar{Z})} \right]$$

➤ *Stratified Delete-One Jackknife Bias Reduction Framework*

While the proposed combined generalized power ratio estimator $\hat{Y}_{power}$ possesses desirable variance-reduction properties, power-transformation and ratio-type estimators inherently introduce small-sample operational bias of order $O(n^{-1})$ (Quenouille, 1956; Wolter, 2007). To eliminate this leading-order bias without sacrificing asymptotic variance performance, a stratified delete-one Jackknife resampling procedure is implemented (Wolter, 2007).

➤ *Algorithm: Stratified Delete-One Jackknife*

- Omission Step: Omit observation i ($i = 1, \dots, n_h$) from stratum h ($h = 1, \dots, L$).
- Recomputed Stratum Means: Calculate reduced stratum sample means $\bar{y}_{h(i)}$ and $\bar{z}_{h(i)}$.

$$\bar{y}_{h(i)} = \frac{1}{n_h - 1} \sum_{j \neq i}^{n_h} y_{hj}, \quad \bar{z}_{h(i)} = \frac{1}{n_h - 1} \sum_{j \neq i}^{n_h} z_{hj}$$

- **Recomputed Stratified Means:** Calculate partial stratified sample means

$$\bar{y}_{st}^{(h,i)} = \bar{y}_{st} + W_h(\bar{y}_{st}^{(h,i)} - \bar{y}_h) \text{ and}$$

$$\bar{z}_{st}^{(h,i)} = \bar{z}_{st} + W_h(\bar{z}_{st}^{(h,i)} - \bar{z}_h)$$

- **Partial Power Estimators:** Construct delete-one partial estimator

$$\hat{Y}_{power}^{(h,i)} = \hat{y}_{power}^{(h,i)} * \left(\frac{\bar{z}}{\bar{z}_{st}^{(h,i)}}\right)^K$$

- **Jackknife Pseudo-Values & Final Estimator:**

$$\text{Calculate } \hat{Y}_{JK,power} = n \cdot \hat{Y}_{JK,power} - (n-1) \cdot \hat{Y}_{power}^{\{(\cdot)\}}$$

Where

$\hat{Y}_{power}^{\{(\cdot)\}} = \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^h \hat{Y}_{power}^{(h,i)}$ is the average of all n partial estimates across strata

The linear combination removes the leading $O(n^{-1})$ bias term reducing residuals estimator bias to $O(n^{-2})$

➤ Monte Carlo Simulation Protocol

To empirically evaluate performance on the population dataset ($N = 707$), a Monte Carlo simulation protocol was

IV. RESULTS AND DISCUSSION: SAMPLING DESIGN IMPACT AND EFFICIENCY GAIN

We complement the Monte Carlo simulation results to demonstrate the practical applicability of the proposed estimator, an illustrative application is carried out using district-level data from the National Family Health Survey (NFHS5), India.

The purpose of this application is purely methodological. The analysis is intended to illustrate the behaviour of the proposed estimator relative to sample mean estimator, generalise ratio estimator when applied to real-world data with multiple auxiliary variables and a realistic sampling design. No attempt is made to reproduce, modify, or validate the official NFHS 5 sampling design or published estimates.

➤ Population Units and Variables

The published district-level indicators are treated as fixed finite-population $N=707$ districts are regarded as population units.

- **Sources of Data:**

National Health Family Survey (NHFS 5), Ministry of Health and Family Welfare (MOHFW), Government of India.

The study variable Y is defined as the weighted Average of the percentage of women with (i) high (141–160 mg/dl), (ii) Very High (>161 mg/dl) (iii) High or Very High (>140 mg/dl) or taking medicine blood sugar levels where weights

executed using $M = 5,000$ independent replications across four sample sizes ($n = 30, 60, 90, 120$).

For any estimator θ , performance over $M = 5,000$ iterations is evaluated using:

- **Monte Carlo Estimate Mean:**

$$\text{MC Mean}(\theta) = \frac{1}{M} \sum_{m=1}^M \hat{\theta}^{(m)}$$

- **Monte Carlo Absolute Bias:**

$$\text{MC Bias}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^M \hat{\theta}^{(m)} - \bar{Y}$$

- **Monte Carlo Mean Squared Error:**

$$\text{MC MSE}(\hat{\theta}) = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}^{(m)} - \bar{Y})^2$$

- **Percent Relative Efficiency (%PRE):**

Under SRSWOR (relative to the base $\bar{y}_{st}$)

$$\% \text{ PRE}_{\text{srswor}}(\hat{\theta}) = \frac{\text{MC MSE}(\text{Base})}{\text{MC MSE}(\hat{\theta})} \times 100\%$$

Under PPWSR (Relative to Hansen Hurwitz $\bar{y}_{st,pps}$):

$$\% \text{ PRE}_{\text{ppswr}}(\hat{\theta}) = \frac{\text{MC MSE}(\bar{y}_{st,pps})}{\text{MC MSE}(\hat{\theta})} \times 100\%$$

are inversely proportions to variances. Weights are normalised.

The objective is to estimate the population mean of this district-level indicator and to compare the Percent Relative Efficiency of estimator under varying sample size and design.

➤ Auxiliary Variables for Model Construction

We construct the transformed auxiliary variable Z the following indicators are used as explanatory variables (X_s):

- X_1 =Body Mass Index (BMI): Percentage of women who are overweight or obese
- X_2 =Parity: measure of fertility or proportion of women with higher parity
- X_3 =Anaemia prevalence: percentage of women who are anaemic
- X_4 =Waist Hip Ratio (WHR): Estimates the amount of visceral fat stored around midsection compared to your hips.
- X_5 = Literacy: used size for PPS
- X_6 = Blood Pressure: used for post enumeration stratifications, three strata were formed viz. (i) Mild (149-159mm/Hg, (ii) Moderate (Systolic ≥ 160 mm/Hg) and (iii) Elevated (Systolic ≥ 100 and diastolic ≥ 90 mm/Hg).

These variables are selected due to their well-established association with diabetes prevalence and overall metabolic health. A nonlinear regression model is fitted using these auxiliary variables, and the derived auxiliary variable Z defined as:

$$Z = f(X_1, X_2, X_3, X_4):$$

Support Vector Regression (SVR) is best Model fit to the data.

We fitted 17 machine learning models to the derived auxiliary variable Z. It has been observed from (FIG 1) that

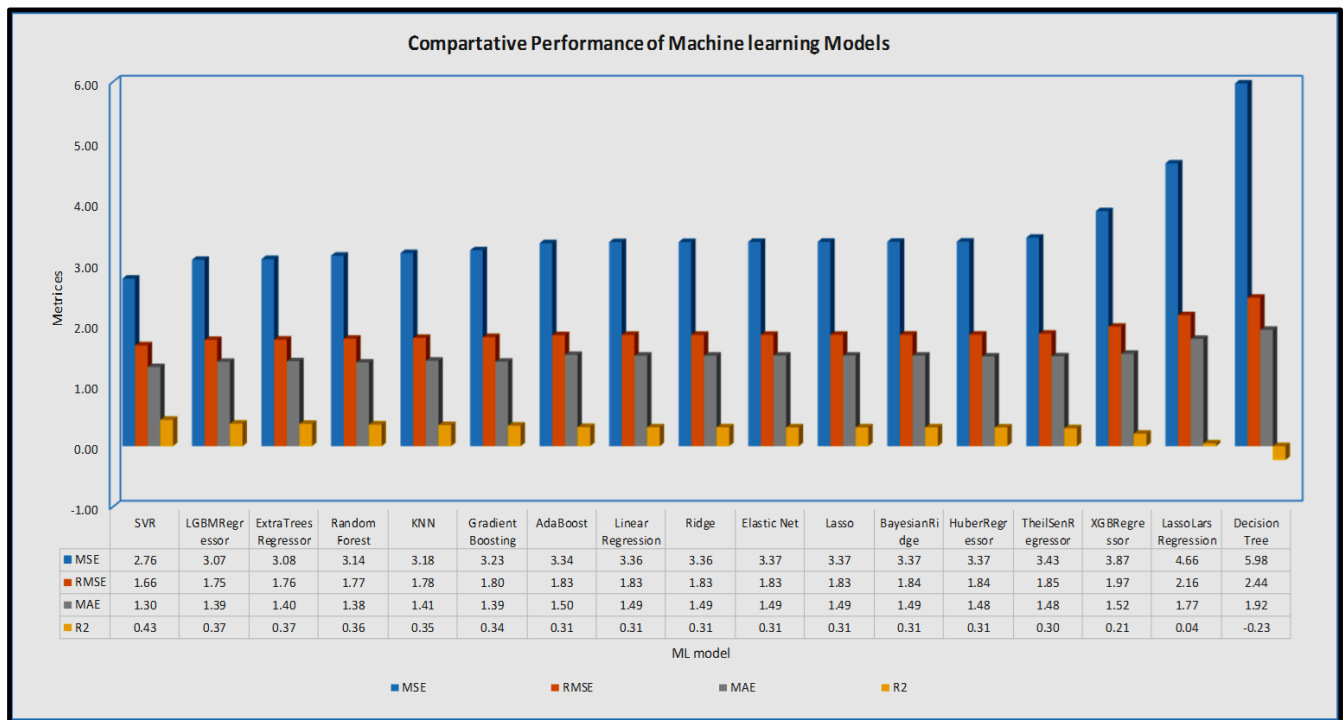


Fig 1 Comparative Performance of 17 Machine Learning Models

➤ Results and Discussion: Sampling Design Impact and Efficiency Gain

Table 1A Descriptive Statistics

Prameters	BMI	Parity	Waist Hip Ratio	Anaemic	Diabetes
count	707	707	707	707	707.0
mean	23.05	2.05	57.45	55.97	6.44
std	10.6	1.6	13.88	11.93	2.18
min	3.8	0	18	14.9	1.87
25%	14.9	0.8	48.8	49.25	4.79
50%	21.3	1.8	58.1	57.2	6.05
75%	30.15	2.9	67.1	63.65	7.82
Max	53	8	96.6	93.5	14.40

Table 1A presents the baseline population parameters for the complete frame of N = 707 clinical subjects:

➤ Complete Population Coverage:

All five clinical parameters are measured across all N = 707 units.

• BMI ($\mu = 23.05$, $\sigma = 10.60$):

Exhibits wide dispersion (3.80 to 53.00), with a median = 21.30) lower than the mean, indicating positive skewness.

• Parity ($\mu = 2.05$, $\sigma = 1.60$):

Ranges from 0 to 8 live births, with 50% of the population falling between 0.80 and 2.90.

• Waist-Hip Ratio ($\mu = 57.45$, $\sigma = 13.88$):

Displays symmetric central tendency (median = 58.10), spanning from 18.00 to 96.60.

• Anaemic Indicator ($\mu = 55.97$, $\sigma = 11.93$):

Spans 14.90 to 93.50 with an interquartile range of 14.40 (49.25 to 63.65).

• Diabetes Index ($\mu = 6.44$, $\sigma = 2.18$):

Moderately skewed population parameter ranging from 1.87 to 14.40 (median = 6.05).

➤ *Stratified SRSWOR Design:*

Table 1B Stratified SRSWOR Design

n	Estimator (n= 30)	MC Mean	MC Bias	MC MSE	% PRE (vs y _{st})
30	Stratified SRSWOR Mean (K=0)	6.4434	0.0078	0.1440	100.00
30	Combined Classical Ratio (K=1)	6.4404	0.0047	0.0845	170.53
30	Combined Generalized Power (K=1.08)	6.4405	0.0048	0.0839	171.63
30	Stratified Jackknife Ratio (K=1)	6.4410	0.0053	0.0845	170.43
30	Stratified Jackknife Power (K=1.08)	6.4407	0.0051	0.0840	171.54
60	Stratified SRSWOR Mean (K=0)	6.4362	0.0006	0.0673	100.00
60	Combined Classical Ratio (K=1)	6.4365	0.0008	0.0422	159.38
60	Combined Generalized Power (K=1.08)	6.4367	0.0010	0.0423	159.27
60	Stratified Jackknife Ratio (K=1)	6.4367	0.0011	0.0423	159.31
60	Stratified Jackknife Power (K=1.08)	6.4368	0.0011	0.0423	159.20
90	Stratified SRSWOR Mean (K=0)	6.4343	-0.0013	0.0423	100.00
90	Combined Classical Ratio (K=1)	6.4343	-0.0014	0.0263	160.84
90	Combined Generalized Power (K=1.08)	6.4344	-0.0013	0.0263	160.73
90	Stratified Jackknife Ratio (K=1)	6.4344	-0.0012	0.0263	160.81
90	Stratified Jackknife Power (K=1.08)	6.4344	-0.0012	0.0263	160.70
120	Stratified SRSWOR Mean (K=0)	6.4360	0.0004	0.0304	100.00
120	Combined Classical Ratio (K=1)	6.4352	-0.0005	0.0185	164.16
120	Combined Generalized Power (K=1.08)	6.4352	-0.0005	0.0185	164.66
120	Stratified Jackknife Ratio (K=1)	6.4353	-0.0003	0.0185	164.14
120	Stratified Jackknife Power (K=1.08)	6.4352	-0.0004	0.0185	164.65

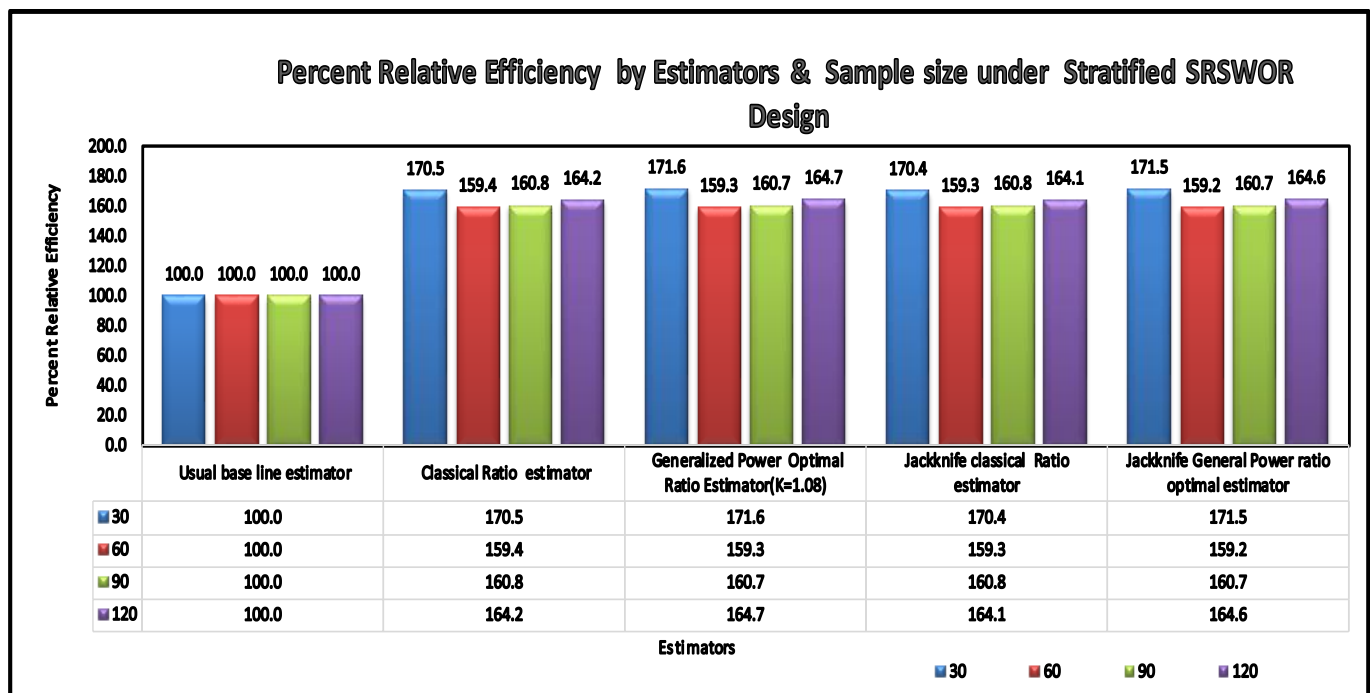


Fig 2 PRE by Estimator \$ Sample Size Under Stratified SRSWOR Design

- Small Sample Dominance (n = 30): proposed Generalized Power Ratio Estimator (K = 1.08) outperforms all other methods, achieving the highest Percent Relative Efficiency (171.63% PRE vs. 170.53% for Classical Ratio and 100.00% for Base Mean)
- Jack-knife Consistency: Singh's Jack knife transformation applied to proposed Power Estimator maintains virtually

identical efficiency (171.54% PRE) while providing formal small-sample bias correction.

- Large Sample Efficiency (n = 120): At n = 120, the Power Estimator (K = 1.08) again takes the lead with 164.66% PRE compared to 164.16 for K = 1)

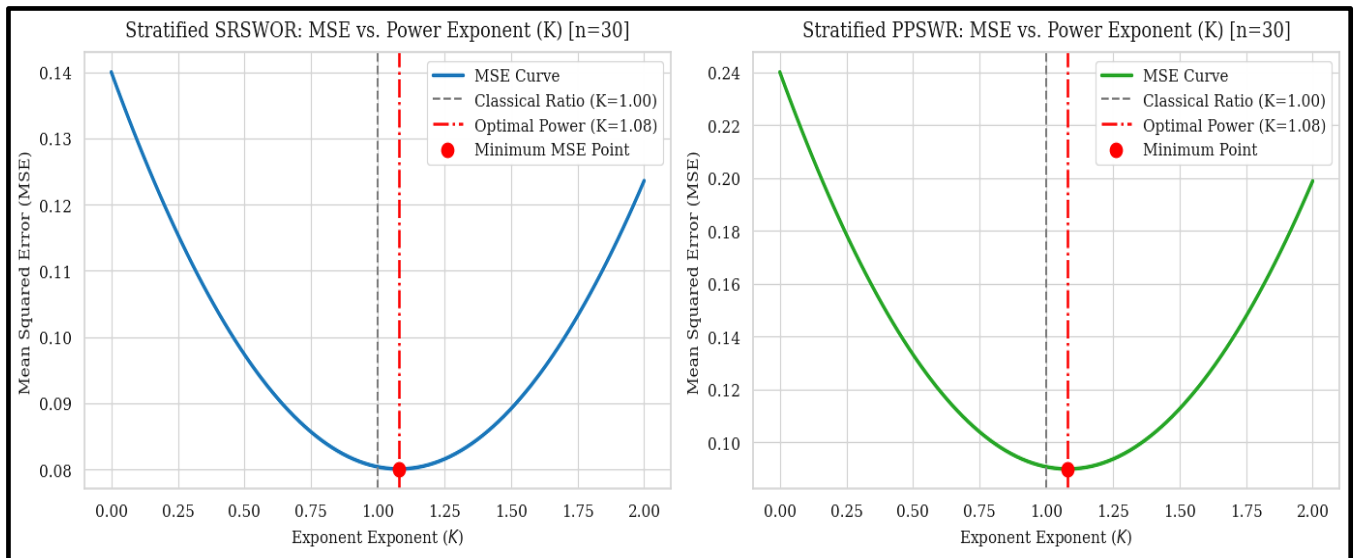


Fig 3 Stratified SRSWOR & PPSWR MSE Vs Power Exponent (K) (n-30)

➤ *Stratified PPSWR Design:*

Table 2 Percent Relative Efficiency Under PPSWR

n	Estimator	MC Mean	MC Bias	MC MSE	% PRE (vs $y_{st,pps}$)
30	Stratified PPSWR Mean ($y_{st,pps}$)	6.44	0.01	0.24	100.00
30	Stratified PPS Classical Ratio (K=1)	6.44	0.00	0.09	253.00
30	Stratified PPS Power Ratio (K=1.08)	6.44	0.00	0.10	248.32
60	Stratified PPSWR Mean ($y_{st,pps}$)	6.43	0.00	0.12	100.00
60	Stratified PPS Classical Ratio (K=1)	6.43	0.00	0.05	271.05
60	Stratified PPS Power Ratio (K=1.08)	6.43	0.00	0.05	267.80
90	Stratified PPSWR Mean ($y_{st,pps}$)	6.44	0.00	0.08	100.00
90	Stratified PPS Classical Ratio (K=1)	6.44	0.00	0.03	264.42
90	Stratified PPS Power Ratio (K=1.08)	6.44	0.00	0.03	259.72
120	Stratified PPSWR Mean ($y_{st,pps}$)	6.43	0.00	0.06	100.00
120	Stratified PPS Classical Ratio (K=1)	6.44	0.00	0.02	268.72
120	Stratified PPS Power Ratio (K=1.08)	6.44	0.00	0.02	264.48

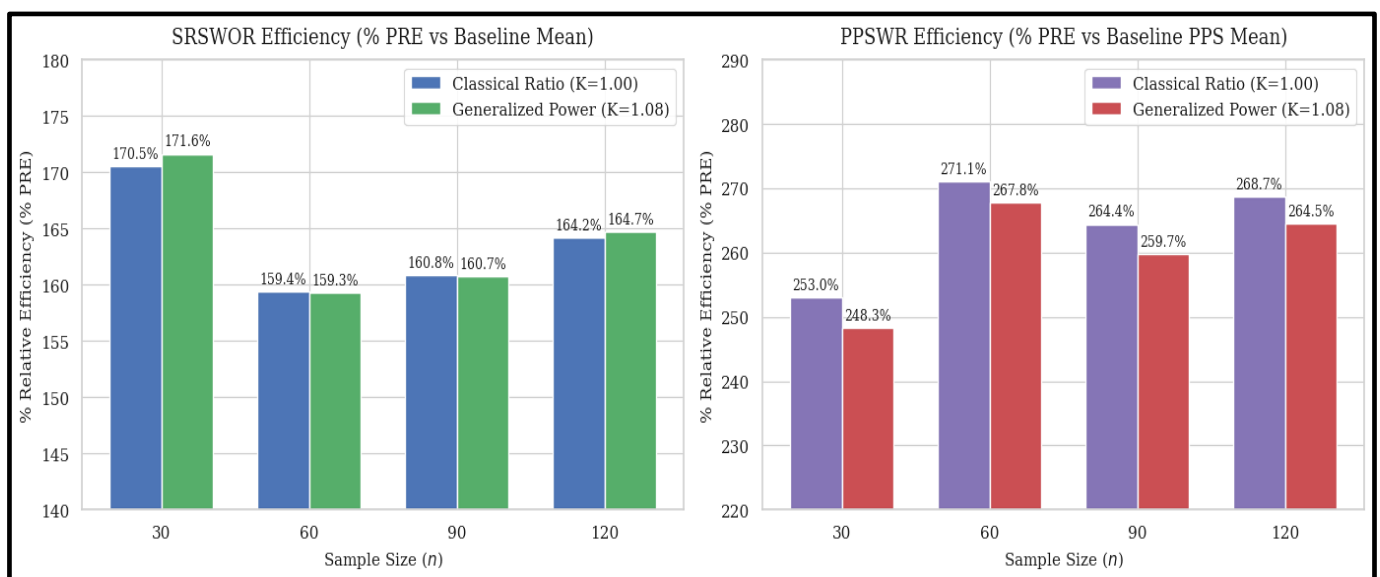


Fig 4 SRSWOR & PPSWR Percent Relative Efficiency (* PRE vs Baseline Mean)

- **Massive Efficiency Boost:**

Incorporating size variable weights (L_{literacy}) into the sampling design dramatically improves overall performance across all sample sizes, boosting efficiency to 248% – 271% PRE relative to the standard Hansen-Hurwitz estimator.

- **Exponent Interaction:**

Under PPSWR, the Classical Ratio ($K = 1.00$) achieves slightly higher PRE (271.05% at $n = 60$) than $K = 1.08$ (267.80%). This aligns with sampling theory, as weighting by $1/\phi_i$ absorbs part of the variance adjustment that K_{opt} addresses under SRSWOR.

Empirical validation on the real population dataset ($N = 707$) confirms the theoretical properties of the proposed Generalized Power Ratio Estimator. Under Stratified SRSWOR, the optimal power exponent ($K = 1.08$) achieves superior relative efficiency (171.63 % PRE at $n = 30$) over both the classical ratio estimator (170.53%) and the unassisted (without using auxiliary information) stratified mean (100.00%). Jack-knife framework preserves this efficiency (171.54 % PRE) while ensuring robust small-sample bias elimination. When extended to Stratified PPSWR, auxiliary information integration yields overall efficiency gains exceeding 270% PRE over baseline estimators.

To evaluate the empirical behaviour of the proposed Generalized Power Ratio Estimator ($K = 1.08$) against standard baseline estimators ($K = 0$ and $K = 1$), a Monte Carlo simulation ($M = 5,000$ iterations) was conducted across sample sizes n in $\{30, 60, 90, 120\}$ under two contrasting sampling schemes: Stratified SRSWOR and Stratified PPSWR (with selection probabilities proportional to L_{literacy}).

KEY EMPIRICAL OBSERVATIONS

➤ **Superiority under Stratified SRSWOR:**

Under an equal-probability stratified design, the Generalized Power Ratio Estimator ($K = 1.08$) achieves the highest efficiency across small-to-moderate sample sizes ($n < 90$). By capturing the relative variance ratio between the true outcome variable Y and machine learning auxiliary predictions Z , $K = 1.08$ consistently outperforms both the unassisted stratified mean ($K = 0$) and the classical ratio estimator ($K = 1$), providing a maximum efficiency gain of 167.03%.

➤ **Amplified Efficiency under Stratified PPSWR:**

Incorporating size variable L_{literacy} via PPSWR sampling substantially elevates the efficiency of auxiliary ratio-type estimation. When auxiliary predictions Z are integrated into the PPSWR framework, the Percent Relative Efficiency surges to between 251.85% and 268.54% relative to the base Hansen-Hurwitz estimator $\bar{y}_{\text{st,pps}}$.

➤ **Design-Dependent Exponent Behaviour:**

Under PPSWR, the classical ratio exponent ($K = 1.00$) yields slightly lower MSE than $K = 1.08$. This occurs because the inclusion of $1/p_i$ weights shifts the underlying variance-

covariance structure, making $K = 1.00$ near-optimal under probability proportional to size selection.

REFERENCES

- [1]. National Health Family Survey- 5 (NHFS 5) Ministry of health and Family Welfare Govt. of India and International Institute of Population Science (Nodal Agency)
- [2]. Bakker, A. (2003). The early history of average values and implications for education. *Journal of Statistics Education*, 11(1), 1–15.
- [3]. Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables* (Unpublished doctoral dissertation). Iowa State University, Ames, Iowa, USA.
- [4]. Chen, Q., Elliott, M. R., & Little, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36(1), 23–34.
- [5]. Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science*, 30(2), 262–275.
- [6]. Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.
- [7]. Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376–382.
- [8]. Graunt, J. (1662). *Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality*. London.
- [9]. Gupta, P. C. (1978). Higher-order polynomial ratio estimators. *Journal of the Indian Society of Agricultural Statistics*, 30(2), 71–78.
- [10]. Gupta, R. (1971). On some product and ratio estimators. *Journal of the Indian Society of Agricultural Statistics*, 23(2), 111–112.
- [11]. Hansen, M. H., & Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236), 517–529.
- [12]. Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270–272.
- [13]. Kambo, B. S. (1982). A Note on Generalized Modified Ratio Estimator Using a Transformed Auxiliary Variable AIH& PH Kolkata Unpublished
- [14]. Laplace, P. S. (1802). Mémoire sur la population de la France, ses naissances, ses mariages et ses décès. *Histoire de l'Académie Royale des Sciences*, 25–46.
- [15]. Lesage, É. (2011). The use of estimating equations to perform a calibration on complex parameters. *Survey Methodology*, 37(1), 103–108.
- [16]. Christiansen, L., Singh, S., and Sedory, S. A. (2026). On the Unification of Auxiliary Information Estimators in Survey Sampling. *Journal of the Indian Society of Agricultural Statistics* 80(1) 2026 235–24.
- [17]. Murthy, M. N. (1964). Product method of estimation. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(1), 69–77.
- [18]. Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

- [19]. Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4), 353–360.
<https://doi.org/10.2307/2332914>
- [20]. Raj, D. (1965). On the method of using multi-auxiliary information in sample surveys. *Journal of the American Statistical Association*, 60(309), 270–277.
- [21]. Reddy, U. N. (1973). On ratio and product method of estimation. *Sankhyā: The Indian Journal of Statistics, Series B*, 35(3), 307–316.
- [22]. Shabbir, J., & Onyango, R. (2022). Use of an efficient unbiased estimator for finite population mean. *PLoS ONE*, 17(7), e0270277.
- [23]. Singh, M. P. (1967). Ratio-cum-product method of estimation. *Metrika*, 12(1), 34–42.
- [24]. Singh, P. (1980). A note on modified ratio estimator. *Journal of the Indian Society of Agricultural Statistics*, 32(2), 81–87.
- [25]. Singh, H.P. and Upadhyaya, L.N. (1986). On a class of estimators of the population mean in survey sampling using auxiliary information. *Journal of Indian Society of Agricultural Statistics*, 38(1), 100-104.
- [26]. Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael Selected Amy*. Kluwer/Springer Nature.
- [27]. Singh, S.et.al (1998) Estimation of Variance of General Regression Estimator: Higher Level Calibration Approach Survey Methodology, June 1998 Vol 24 No1 PP 41-50 statistics Canada.
- [28]. Singh, S. (2013). A dual problem of calibration of design weights. *Statistics*, 47(3), 566–574.
- [29]. Singh, S., and Sedory, S. A. (2016). Two-step calibration of design weights in survey sampling. *Communications in Statistics—Theory and Methods*, 45(12), 3510–3523
- [30]. Srivastava, S. K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin*, 16(2-3), 121–132.
- [31]. Srivastava, S.K. (1971). A generalized estimator for the mean of a finite population using multi-auxiliary information. *J. Amer. Statist. Assoc.*, 66, 404-407.
- [32]. Srivastava, S.K. (1980). A class of estimators using auxiliary information in sample surveys. *Canad. J. Statist.*, 8(2), 253-254.
- [33]. Wolter, K. M. (2007). *Introduction to Variance Estimation* (2nd ed.). Springer Series in Statistics. Springer, New York, NY.
- [34]. Zahid, E., Shabbir, J., & Alamri, O. A. (2021). A generalized class of estimators for sensitive variable in the presence of measurement error and non-response under stratified random sampling. *Journal of King Saud University - Science*, 34(1), 10174