IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 July 2024, accepted 24 July 2024, date of publication 29 July 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435376

RESEARCH ARTICLE

Heuristic Algorithm for Obtaining Approximate Optimum Stratification With Mixture of Ratio and Product Estimators

S. E. H. RIZVI[1], FAIZAN DANISH[2], RAFIA JAN[3], ISMAIL A. MAGEED[4], SUMAYA AL MAADEED[5], (Senior Member, IEEE), AND JIHAD MOHAMED ALJA'AM[6]

[1]Division of Statistics and Computer Science, Faculty of Basic Sciences, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Chatha, Jammu and Kashmir 180009, India
[2]Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology-Andhra Pradesh (VIT-AP) University, Amaravati, Andhra Pradesh 522237, India
[3]Department of Statistics, Government Degree College Bejbehara, Anantnag, Jammu and Kashmir 192124, India
[4]University of Bradford, BD7 1DP Bradford, U.K.
[5]Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar
[6]School of Computing and Data Sciences, Liverpool John Moores University (OUC-LJMU), Doha, Qatar

Corresponding authors: Jihad Mohamed Alja'am (j.m.aljaam@ljmu.ac.uk) and Faizan Danish (faizan.danish@vitap.ac.in)

This work was supported by Qatar National Library.

ABSTRACT In this investigation, we examined the impact of employing simple random sampling on the stratification points pertaining to the two independent variables. The study focused on a variable (X) exhibiting a robust correlation, and we employed a combination of ratio and product estimators to select a representative sample and establish the population mean. By maintaining a comprehensive superpopulation framework, we successfully identified concise equations that effectively reduced the overall variability within the dataset. To reveal the underlying nature of these mathematical derivations, we employed the cumulative cube roots rule to determine nearly optimal stratification points for the two research variables. The validity of this suggested rule was assessed through rigorous testing utilizing empirical and simulated data obtained from diverse distributions.

INDEX TERMS Optimum stratification, ratio estimator, product estimator, super-population model.

I. INTRODUCTION

One widely used approach in contemporary survey design is the utilization of stratified random sampling. In this methodology, the division of a population into distinct strata is crucial, emphasizing the need for homogeneity within each stratum. This homogeneity maximizes the accuracy of key population characteristic estimates, such as means and totals. Over time, the evolution of stratified sampling has been significant, with continuous efforts aimed at enhancing the precision of estimates, bringing them closer to true population parameters. Frequently, National Statistical Offices (NSO), government departments, and private organizations request expedited surveys that maintain a fixed cost, minimal effort, and a short timeframe, all while upholding the quality and precision of estimates for effective decision-making. Undoubtedly, the planning and administration of these surveys must be swift and meticulous. Stratified sampling offers numerous advantages over alternative sampling methods, particularly when dealing with diverse populations. Some key benefits include enhanced representativeness for minimized parameter estimation, more precise and efficient estimation compared to simple random sampling, increased reliability, and generalizability of findings, heightened statistical power in hypothesis testing and inferential statistics, assurance of adequate representation for rare subgroups, leading to more robust analyses. Furthermore, the flexibility of stratified sampling allows for its application with various sampling techniques, enabling researchers to tailor the method to specific study requirements. It facilitates meaningful comparative

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Bellan.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

analyses between subgroups and contributes to reduced variability in estimates, thereby improving the precision of statistical analyses. Despite these advantages, it's worth noting that stratified sampling does demand a more extensive initial effort to accurately identify and define strata, especially in diverse populations.

The ratio estimator is a statistical method used to estimate a population parameter by calculating the ratio of two related variables. This technique is particularly useful when there is a strong correlation between these variables. For instance, in survey sampling, the ratio of two auxiliary variables might be employed to estimate a population characteristic, providing a more efficient and accurate estimate than traditional methods. The product estimator is a statistical approach that involves estimating a population parameter by multiplying two related variables. This method is often applied in situations where the joint distribution of these variables is well understood. Further, it can used when there is negative correlation between the two variables. By utilizing the product of these variables, researchers can obtain estimates that capture the intricate relationships within the population, offering a valuable tool for improving the precision of statistical inference in various research domains. In both ratio and product estimation, the key lies in leveraging the relationships between variables to obtain more accurate and efficient estimates of population parameters, contributing to the advancement of statistical methodologies in research and data analysis. However, we may have the circumstances when one variable is a Ratio estimator and other Product which will lead us to use Mixture estimator.

The optimal stratification problem was first raised in [1]. Variable under consideration as a stratification variable under Neyman allocation was investigated in [2], an extension of [1]'s work for the univariate cases. Furthermore, there are cases where numerous features are used for estimation, making straight optimum allocation problematic. The proportional allocation technique has been investigated in the past for two features [3], [4]. References [4], [5], [6], [7], [8], and [9] provide examples of the many distinct contexts for which various techniques have been offered. Using a penalized objective function optimized via the Simulated Annealing technique, an algorithm is proposed in [10] to solve the multivariate stratification problem. Algorithms for stratifying asymmetric populations using power allocation to estimate sample sizes are introduced in [11] and [12] employs Dynamic Programming and Neyman allocation to address the stratification issue under the assumption of a Weibull distribution for the stratification variable. Several R packages, such as GA4 [13] Stratification stratify R (sample on the R CRAN, are available for stratification). To handle inconsistencies between the stratification variable and a study variable, the authors of [14] present a method that uses two models. Methods have been presented over the past few decades, with most falling into the categories of approximation or optimization [15]. In [22], an exact technique is presented for allocating resources, and it is used by the BRKGA (Biassed Random Key Genetic Algorithm) and GRASP (Greedy Randomised Adaptive Search Procedure) algorithms published in [16]. The stratification points for two research variables are calculated using a dynamic programming method [23]. The results refers to the stratification points obtained by the proposed method using different frequency distribution while as accuracy means the percentage relative efficiency which is obtained using the variance estimated from varaious methods.

One of the research factors may have a strong positive association with the stratification variable in real-world scenarios, whereas the other may have a large negative correlation. Using the assumptions of a strong positive correlation between Y and X (the auxiliary variable) and a strong negative correlation between Z (the independent variable) and X (the stratification variable), this study seeks to locate the optimal points at which to apply stratification to two independent variables in a simple random sampling design. The sample is picked so that the mean of the entire population can be estimated by applying ratio and product estimators with the help of the auxiliary variable (X). Minimal equations are obtained in a superpopulation framework by minimizing the total variance with the help of the variables of interest. The conditional variance functions V(y|x) and V(z|x) are also assumed to be known based on prior knowledge of the functional relationship between Y and Z with respect to X.

II. EXPRESSION OF VARIANCE AND COVARIANCE

Assume population of 'N' units be split into 'L' strata. The separate ratio-estimate for population mean in stratified random sampling are given.

$$\bar{Y}_{st.R_1} = \sum_{h=1}^{L} W_h \bar{Y}_{hR_1} \tag{1}$$

where $W_h = \frac{N_h}{N}$, h$^{\text{th}}$ stratum weight

$\bar{Y}_h$ = sample mean of Y

$\bar{Y}_{hR_1} = \left(\bar{Y}_h / \bar{x}_h\right) \bar{X}_h = R_{1h}\bar{X}_h$

$\bar{x}_h$ = sample mean of X

$\bar{X}_h$ = population mean of auxiliary variable X

We can use the combined product estimator if we assume that in each stratum, the regression lines of the stratification variable on the auxiliary variable are linear and pass through the origin and if we consider that characteristic Z fulfils $R_{21} = R_{22} = R_{23} = \ldots = R_{2L}$. For the case of stratified sampling, the combined estimators are provided by:

$$\bar{Z}_{st.P} = \frac{\sum \left(W_h \bar{Z}_h\right) \sum \left(W_h \bar{X}_h\right)}{\bar{X}} \tag{2}$$

where, $\bar{x}_h, \bar{X}$ and $\bar{Z}_h$ denotes sample mean and population means. If the finite population correction (fpc) is neglected, the approximate variances of these estimators under proportional allocation are given by:

$$V(\bar{Y}_{st.R_1})_P = \frac{1}{n} \sum_{h=1}^{L} W_h \left(\sigma_{hy}^2 + R_{1h}^2 \sigma_{hx}^2 - 2R_{1h}\sigma_{hxy}\right) \tag{3}$$

and

$$V(\bar{Z}_{st.P})_P = \frac{1}{n}\sum_{h=1}^{L} W_h\left(\sigma_{hz}^2 + R_2^2\sigma_{hx}^2 - 2R_2\sigma_{hxz}\right) \quad (4)$$

For the covariance expression, we have the following theorem:

Theorem 1: The covariance expression between the estimators $\bar{Y}_{st.R_1}$ and $\bar{Z}_{st.P}$ is given below:

$$Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\sigma_{hyz} + R_2\sigma_{hyz} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hxz}\right) \quad (5)$$

Proof: Using partially the obtained from from [17], we have:

$$Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \frac{1}{\bar{X}}Cov[\{\sum W_h(\epsilon_{1h}\bar{X}_h - \frac{\xi\bar{Y}_h}{\bar{X}_h}\}, \times \bar{X}\sum W_h\epsilon_{2h} + \bar{Z}\sum W_h\xi_h] \quad (6)$$

which, to simplification, results in

$$Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \frac{1}{\bar{X}}Cov[\{\sum W_h^2[\bar{X}Cov\left(\epsilon_{1h}, \epsilon_{2h}\right), \bar{Z}Cov\left(\epsilon_{1h}, \xi_h\right) \times \bar{X}R_{1h}Cov\left(\xi_h, \epsilon_{2h}\right) - \bar{Z}R_{1h}Cov\left(\xi_h, \xi_h\right)] \quad (7)$$

and finally, we have:

$$Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right) \quad (8)$$

and hence, the Lemma is proved.

Under the proportional method of allocating the sample size to different strata, the formula for covariance as given by (5) reduces to

$$Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \frac{1}{n}\sum_{h=1}^{L} W_h\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right) \quad (9)$$

III. MINIMAL EQUATIONS

Let $\{x_h\}$ represent the stratification points in the interval (a, b) of the stratification variable. Then, corresponding to those strata boundaries, the generalized variance G_6, as given by equation (10), is defined as follows:

$$G_6 = \begin{vmatrix} \sigma_y^2 & \sigma_{yz} \\ \sigma_{zy} & \sigma_z^2 \end{vmatrix} = \sigma_y^2\sigma_z^2 - \left(\sigma_{yz}\right)^2 \quad (10)$$

where σ_y^2, σ_z^2 and σ_{yz} denote $V(\bar{Y}_{st.R_1})_P$, $V(\bar{Z}_{st.P})_P$ and $Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right)_P$ respectively.

Differentiating G_6 partially w.rto $\{x_h\}$ and then put its derivative to zero, we get

$$\frac{\partial G_6}{\partial x_h} = \sigma_y^2\frac{\partial\sigma_z^2}{\partial x_h} + \sigma_z^2\frac{\partial\sigma_y^2}{\partial x_h} - 2\sigma_{yz}\frac{\partial\sigma_{yz}}{\partial x_h} = 0, h = 1, 2, \ldots, L-1 \quad (11)$$

Inserting the values of σ_y^2, σ_z^2 and σ_{yz} from (3), (4) and (9) in (11), we have

$$\sigma_y^2\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hz}^2 + R_2^2\sigma_{hx}^2 - 2R_2\sigma_{hxz}\right)\right] + \sigma_z^2\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hy}^2 + R_{1h}^2\sigma_{hx}^2 - 2R_{1h}\sigma_{hxy}\right)\right] - 2\sigma_{yz}\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right)\right] = 0 \quad (12)$$

The approximation regression model can now be written as follows, assuming that the functional relationships between Y and X, Z and X, and each stratum are linear and that the regression lines pass through the origin:

$$Y_1 = R_{1h}X + e_{1i} \quad (13)$$

$$Z_1 = R_2X + e_{2i} \quad (14)$$

Various are the variance expressions for proportional allocation in various models ([18]):

$$\sigma_y^2 = V\left(\bar{Y}_{st.R_1}\right)_P = \frac{\mu_{\eta_1}}{n} \quad (15)$$

$$\sigma_z^2 = V\left(\bar{Y}_{st.P}\right)_P = \frac{4R_2^2}{n}\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \frac{\mu_{\eta_2}}{n} \quad (16)$$

and the covariance term can be obtained as:

$$\sigma_{yz} = Cov\left(\bar{Y}_{st.R_1}, \bar{Z}_{st.P}\right) = \sum_{h=1}^{L} W_h\left(R_{1h}R_2\sigma_{hx}^2 + R_{1h}R_2\sigma_{hx}^2 - R_{1h}R_2\sigma_{hx}^2 - R_{1h}R_2\sigma_{hx}^2\right) = 0 \quad (17)$$

Putting these values in the minimal equation (12), we get

$$\frac{\mu_{\eta_1}}{n}\frac{\partial}{\partial x_h}\left[4R_2^2\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \mu_{\eta_2}\right] + \left[\frac{4R_2^2}{n}\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \frac{\mu_{\eta_2}}{n}\right] \times \frac{\partial}{\partial x_h}\mu_{\eta_1} - 0 = 0$$

Or

$$\frac{4R_2^2}{n}\frac{\mu_{\eta_1}}{n}\frac{\partial}{\partial x_h}\left(\sum_{h=1}^{L} W_h\sigma_{hx}^2\right) = 0$$

which gives:

$$W_h\frac{\partial\sigma_{hx}^2}{\partial x_h} + \sigma_{hx}^2\frac{\partial W_h}{\partial x_h} + W_i\frac{\partial\sigma_{ix}^2}{\partial x_h} + \sigma_{ix}^2\frac{\partial W_i}{\partial x_h} = 0 \quad (18)$$

Putting the values of partial derivatives involved in (18) and then solving, we get

$$x_h - \mu_{hx} = \mu_{ix} - x_i \quad (19)$$

and therefore, the required minimal equations become

$$x_h = \frac{\mu_{hx} + \mu_{ix}}{2}, i = h+1, h = 1, 2, \ldots, L-1 \quad (20)$$

Here, we observe that the minimal equations obtained in this circumstance are the same as obtained in the case where both the variables under consideration are negatively correlated with the auxiliary variable, i.e.., the product method of estimation [19].

Remark 1: It can be easily verified that the expressions for variance and covariance can also be obtained from the corresponding expressions for stratified simple random sampling estimators under proportional allocation, as given in [18], by taking $C_1(x_h) = constant$ (say C) and $C_2(x_h) = 2Rx_h$. Therefore, the minimal equations for the present estimator can also be obtained as a particular case of the minimal equations of simple random sampling estimators with the same substitution.

Theorem 2: The regression models are given by if study variable Y has a positive correlation with stratification variable X and study variable Z has a negative correlation with X.

$$Y = R_{1h}X + e_1$$
$$Y = R_2X + e_2$$

The given assumptions state that the disturbance terms e_1 and e_2 satisfy certain conditions. The conditional expectation of the error terms are assumed to be zero more specifically, $E\left(e_1/X\right) = 0, E\left(e_1, e_1'/X, X'\right) = 0$ for $X \neq X'$ and $V\left(e_1/X\right) > 0$ (j = 1, 2) for all $X \in (a, b)$ with $(b - a) < \infty$, and further if the function $I_4(X)f(x)$ belong to Ω, then the system of equations (20) giving optimum points of strata boundaries $\{x_h\}$, corresponding to lowest variation of G_6 as $K_h^2 \int_{x_{h-1}}^{x_h} I_4(t)f(t)dt$ = constant, h = 1,2,…,L. For a sufficiently large number of strata, where the expressions of $O\left(m^4\right), m = (Sup(K_h))$ can be neglected and the function $I_4(t) = 4R_2^2\mu_{\eta_1}$

IV. SOLUTION OF OBTAINED EQUATION

The results can be stated as follows:

I. If the expression of the obtained equations, we retain only the first term thereby neglecting the rest ones, then the two sides are equalized if

$$K_h = Constant = (b - a)/L, \forall h = 1, \ldots, L \tag{21}$$

and stratification points are

$$x_h = a + h\frac{(b - a)}{L} \; with \; x_0 = a \; and \; x_L = b$$

Now using the system of minimal equations and putting $\lambda = 1, 1/2$ *and* $1/3$, we have following approximate system of equations.

$$\text{II. } K_h^2 \int_{x_{h-1}}^{x_h} I_3(t)f(t)dt = C_1, h = 1, 2, \ldots, L \tag{22}$$

$$\text{III. } K_h \left[\int_{x_{h-1}}^{x_h} \sqrt{I_3(t)f(t)}dt\right]^2 = C_2, h = 1, 2, \ldots, L \tag{23}$$

$$\text{IV. } \left[\int_{x_{h-1}}^{x_h} \sqrt[3]{I_3(t)f(t)}dt\right]^3 = C_3, h = 1, 2, \ldots, L \tag{24}$$

In single study variable, several forms of $Q(x_{h-1}, x_h)$ have been developed [26]. In all the above system of equations, C_i's (i = 1,2,3) are the constraints to be determined. This may be pointed here that the approximate system of equations (24) is more approximate from practical point of view for which the exact value of the constant C_3 is given:

$$C_3 = \frac{1}{L^3}\left[\int_{x_{h-1}}^{x_h} \sqrt[3]{I_3(t)f(t)}dt\right]^3$$

This may be obtained the same fashion as obtained for the stratified random sampling estimate by changing $I_1(t)$ by $I_4(t)$ where $I_4(t) = 4R_2^2\mu_{\eta_1}$. One of the systems of equations, which is most appropriate, is given by:

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{I_4(t)f(t)}dt = C_3^{1/3}, h = 1, 2, \ldots, L \tag{25}$$

The above equation indicates that if the function $R_4(x) = I_4(X)f(x)$, where $I_4(x) = 4R_2^2\mu_{\eta_1}$, is bounded and its first two derivatives exist in the given interval, for particular value of L taking similar intervals on Cum $\sqrt[3]{R_4(x)}$ will give approximately optimum strata boundaries (AOSB) $\{x_h\}$. One of the most suitable systems of equations is given by:

$$\int_{x_{h-1}}^{x_h} \sqrt[3]{f(t)}dt = C_4^{1/3}, h = 1, 2, \ldots, L \tag{26}$$

where $C_4 = \frac{1}{L}\left[\int_a^b \sqrt[3]{f(t)}dt\right]^3$

Thus, for any distribution, the precision of stratification will vary from stratified and product type estimators, having different solutions of conditional variances.

V. PROPOSED RULES

In case of $R_4(x) = I_4(X)f(x)$, where $I_4(x) = 4R_2^2\mu_{\eta_1}$ is bounded and has differentiable first two derivatives in the given interval, then for a fixed L taking equidistant intervals on the Cum $\sqrt[3]{R_4(x)}$ will give AOSB $\{x_h\}$.

Remark 2: Since the function $I_4(x)$ is itself a constant; therefore, the proposed rule reduces to Cum $\sqrt[3]{f(x)}$ rule. The set of AOSB will remain unchanged with respect to the form of conditional variance, viz. $\eta_1(x)$ and $\eta_2(x)$ for given distribution and given L.

VI. NUMERICAL ILLUSTRATION

The response of the developed method for obtaining the stratification points $\{x_h\}$ of AOSB has been demonstrated empirically likewise [20] and [24]. In this regard, we have taken several distribution functions of stratification variable X as follows:

Uniform distribution:

$$f(x) = 1 \quad 1 \leq x \leq 2$$

Exponential distribution:

$$f(x) = e^{x+1} \quad 1 \leq x \leq \infty$$

Right triangular distribution:

$$f(x) = 2(2 - x) \quad 1 \leq x \leq 2$$

Standard Normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad 0 \le x \le \infty$$

A methodology has been used to obtain the AOSB values for this purpose, followed by the other parametric values for each stratum. Through an iterative process, the values of AOSB for various points of L, the number of strata, were calculated with an error of 0.0005 for each distribution. Tables 1, 2, 3 and 4 present the numerical values of AOSB, KG_6 and % R.E. (Percentage Relative Efficiency), where K is given by $K = {}^{n^2}/_{4R_2^2\mu_{\eta_1}}$, have been presented in Tables 1, 2, 3 and 4 for Uniform, right triangular, exponential, and standard normal distributions, respectively.

TABLE 1. Uniform distribution.

L	AOSB				KG_6	%R.E.
1	1.0000	2.0000			0.5249	
2	1.0000	1.4905	2.0000		0.4619	113.6393
3	1.0000	1.3628	1.6389	2.0000	0.2836	162.8702
4	1.0000	1.2567	1.4682	1.7582	0.137	207.0073
	2.0000					
5	1.0000	1.19852	1.3568	1.5782	0.0596	229.8658
	1.7283	2.0000				
6	1.0000	1.1682	1.3782	1.4387	0.0240	248.2299
	1.6734	1.8276	2.0000			

TABLE 2. Right triangular distribution.

L	AOSB				KG_6	%R.E.
1	1.0000	2.0000			0.5286	
2	1.0000	14153	2.0000		0.4083	129.4636
3	1.0000	1.2461	1.5376	2.0000	0.2396	170.3806
4	1.0000	1.2034	1.4183	1.5931	0.1064	225.2256
	2.0000					
5	1.0000	1.1684	1.2961	1.5086	0.0445	238.6721
	1.7164	2.0000				
6	1.0000	1.1385	1.2583	1.4179	0.0178	250.4494
	1.5376	1.7538	2.0000			

TABLE 3. Exceptional distribution.

L	AOSB				KG_6	%R.E.
1	1.0000	6.0000			6.2843	
2	1.0000	2.6731	6.0000		4.9283	127.5146
3	1.0000	1.8375	3.4028	6.0000	3.1761	155.1683
4	1.0000	1.6429	2.6078	3.7928	1.6008	198.4046
	6.0000					
5	1.0000	1.6087	2.1208	3.0839	0.739	216.6198
	4.259	6.0000				
6	1.0000	1.4461	1.8927	2.6091	0.3175	232.7559
	3.4126	4.4261	6.0000			

Table 2 shows the stratification points for the Right Triangular distribution, while Tables 3 and 4 display those for the Exponential and Standard Normal distributions, respectively. Across these Tables 2, 3 and 4, it becomes apparent that the percentage relative efficiency consistently improves with an increasing number of strata.

TABLE 4. Standard normal distribution.

L	AOSB				KG_6	%R.E.
1	0.0000	1.0000			0.5831	
2	0.0000	0.4015	1.0000		0.4193	139.0651
3	0.0000	0.3186	0.6753	1.0000	0.2528	165.8623
4	0.0000	0.2379	0.4528	0.7928	0.1384	182.659
	1.0000					
5	0.0000	0.19852	0.3568	0.5827	0.06375	217.098
	0.7351	1.0000				
6	0.0000	0.1528	0.3829	0.4261	0.0284	224.4718
	0.6429	0.8253	1.0000			

The random numbers were generated using Uniform distribution, Exponential distribution, right triangular distribution, and standard normal distribution, which are presented in Figures 1, 2, 3 and 4 as given below:

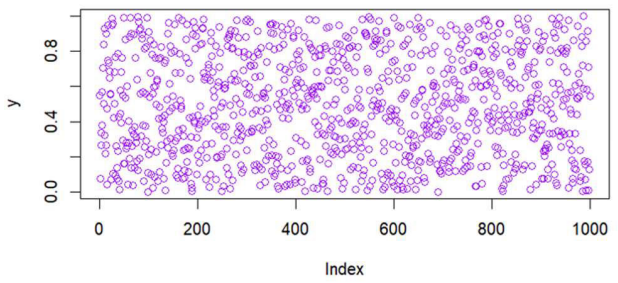

FIGURE 1. Uniform distribution.

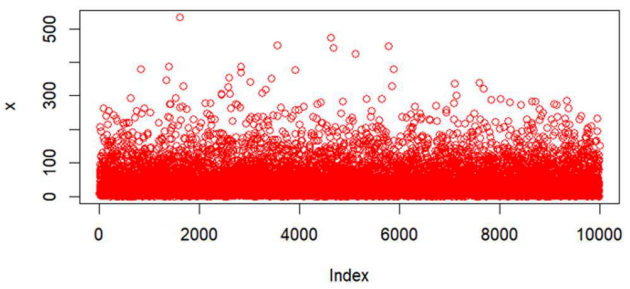

FIGURE 2. Exponential distribution.

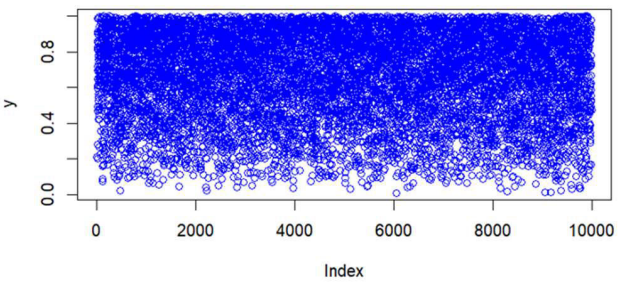

FIGURE 3. Right triangular distribution.

The Figure 1 presents the simulated data generated when the variable has Uniform distribution, Figure 2 presents for

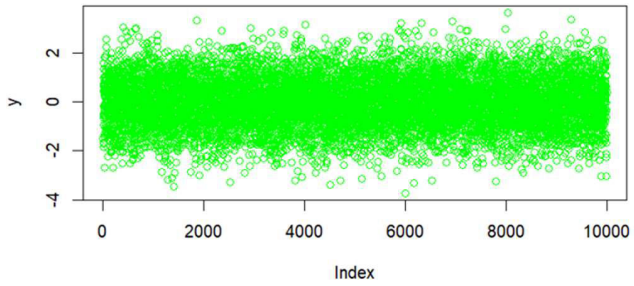

FIGURE 4. Standard normal distribution.

Exponential distribution, Figure 3 presents for Right Triangular distribution and Figure 4 presents the data generated using Standard normal distribution. By their Figures we can see the behaviors of the data generated using simulation. The results obtained in Table 1 can be interpreted easily such as, in case of Uniform distribution and for two strata, the boundary points for the first strata will be from 1.0000 to 1.4905 and for the second strata from 1.4905 to 2.0000 and similarly for three strata, the boundary points will be from 1.0000 to 1.3628, 1.3628 to 1.6389 and 1.6328 to 2.0000 for first, second and third strata respectively. Similarly, the results obtained in rest distribution can be interpreted similarly. All these values depend only on the form of distribution considered. The following tables show that the % R.E. has enhancement with enhancement in the number of subpopulations. However, the R.E. is higher for the present case of mixture estimator than the usual estimator of stratified random sampling. It may be seen that the % R.E. ranges from 113.6393 to 248.2299, 129.4636 to 250.4494, 127.5146 to 232.7559 and 139.0651 to 224.4718 for Uniform, right triangular, exponential distributions, and Standard normal distributions, respectively.

VII. THE DATA SET

In this section, we have utilized a data set on related to the apple production in Jammu and Kashmir, India [27]. In this study, the findings emphasize the effectiveness of using stratified random sampling combined with the ''Equalization of cumulative'' method for estimating apple production in the Shopian district and across Jammu and Kashmir. Neyman allocation surpasses equal and proportional allocation methods, showcasing its potential for greater precision. Additionally, the results highlight the importance of sample size in improving estimation accuracy. Increasing the sample size from 10 to 40 consistently enhances precision, with the ''Equalization of cumulative'' method providing the highest accuracy, followed by the '*Equalization of cumulative of 1/2 {r(y) + f(y)}*' method. Based on these insights, it has been highly recommended that stakeholders adopt stratified random sampling with the ''Equalization of cumulative'' method and Neyman allocation for accurate apple production estimates in both Shopian and Jammu and Kashmir. By implementing these methods and increasing the sample size, stakeholders can significantly improve the precision and reliability of production estimates, aiding in informed decision-making and resource allocation within the apple industry. The data has been utilized and the proposed method has obtained the stratification points given in the following Table 5:

TABLE 5. Percentage increase in efficiency in apple production data set.

Sample Size	$\text{Cum}\sqrt[3]{R_4(x)}$		
	2	**3**	**4**
10	3369.002	8528.391	8408.298
20	1682.521	7671.602	10235.77
30	2236.181	8828.273	12037.85
40	2258.512	8855.346	12761.48

It can be observed from the Table 5 which presents the Percentage Increase in Efficiency in apple production data set, which shows the increase in the gain Iin PRE while utilizing the proposed method and increasing the sample size as well as strata too.

VIII. SIMULATION STUDY

1) A simulation study was conducted to evaluate the validity and relative precision of the proposed method compared to other methods using R statistical software. The following methods were compared:
2) Dalenius and Hodges (1959)cum $\sqrt{f}$ method
3) Gunning and Horgan (2004) geometric method
4) Lavallee-Hidiroglou (1988) method using Kozak's (2004) method
5) Proposed method

In this study, the R software generated a data set of 10,000 observations, assuming a uniform distribution for the auxiliary variable. The minimum and maximum values of the auxiliary variable were found to be 0.000763 and 1.49672, respectively, with a total deviation (k) of 1.495957.

The stratification points were determined using the proposed method, as discussed earlier, along with the comparative methods. The variances obtained by each of these methods, including the proposed method, are presented in Table 6.

TABLE 6. Total variance obtained by different methods.

L	cum $\sqrt{f}$ method	geometric method	Lavallee-Hidiroglou method	Proposed Method
2	0.131	0.13708	0.175627	0.069907
3	0.105667	0.111653	0.098467	0.02376
4	0.080653	0.10124	0.092773	0.0138
5	0.064493	0.08556	0.0818	0.010573
6	0.052987	0.073827	0.08048	0.003173

Based on the Table 6 output, it can be concluded that the proposed method outperformed the existing methods in terms of precision. Therefore, utilizing two study variables in the stratification process leads to a gain in precision compared to using the existing methods.

The simulated data were used in the proposed method, and the variance for each corresponding was noted and plotted in Fig.5 as below:

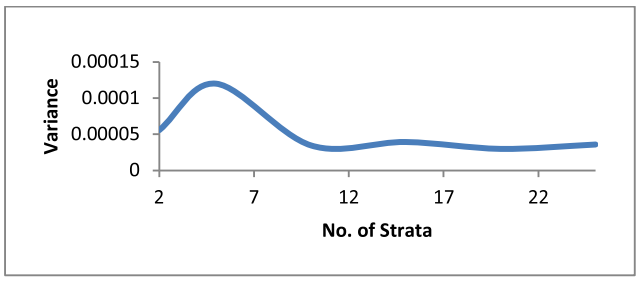

FIGURE 5. Standard normal distribution.

We considered 25 number of strata, but it can be observed that (which displays the graph between number of strata and variance) there is no substantial gain in efficiency for more than 20 strata. Thus, it can be concluded that the number of strata must be chosen very carefully while doing the stratification, and after some number, the variance may have an increasing trend.

IX. CONCLUSION

This research explored the case of combining the ratio and product estimation methods. The proportional allocation method was used to generate mathematical equations that minimized the variance. There has been discussion of using a Cum $\sqrt[3]{R_4(x)}$ Rule to acquire AOSB. Empirical research shows that the efficiency gain is very high, with RE values varying from 113.6393 to 248.2299 for the Uniform distribution, 129.4636 to 250.4494 for the right triangular distribution, 127.5146 to 232.7559 for the exponential distribution, and 139.0651 to 224.4718 for the Standard normal distribution. Asymmetric distributions are more accurately represented by the developed technique, as seen in Tables 1 and 2, which rank the Uniform and right triangular distributions ahead of the exponential and right normal distributions. The suggested method exhibits improvement in the precision of estimates when employing the highly related variable to get the stratification points of populations with Uniform, right triangular exponential distribution, and standard normal distribution. The simulation analysis also reveals an improvement in relative precision with the proposed strategy compared to the state-of-the-art approaches. As a result, the proposed approach has promise for producing reliable estimates of the target variable or feature by mining its frequency distribution. Neutrosophic statistics are chosen over classical statistics when the data originates from a complicated process. References [24] and [25] are examples of studies that have delved into this topic. Neutrosophic statistics, then, may provide a fruitful direction for future study.

REFERENCES

[1] T. Dalenius, "The problem of optimum stratification," *Scandin. Actuarial J.*, vol. 1950, nos. 3–4, pp. 203–213, Jan. 1950, doi: 10.1080/03461238.1950.10432042.
[2] G. Sadasivan and R. Aggarwal, "Optimum points of stratification in bivariate populations," *Sankhya*, vol. 40, pp. 84–97, 1978.
[3] S. P. Ghosh, "Optimum stratification with two characters," *Ann. Math. Statist.*, vol. 34, pp. 866–872, Sep. 1963.
[4] R. Singh, "Approximately optimum stratification on the auxiliary variable," *J. Amer. Stat. Assoc.*, vol. 66, pp. 829–833, Dec. 1971.
[5] T. Dalenius and M. Gurney, "The problem of optimum stratification. II," *Scandin. Actuarial J.*, vol. 1951, nos. 1–2, pp. 133–148, Jan. 1951, doi: 10.1080/03461238.1951.10432134.
[6] F. Danish, S. E. H. Rizvi, M. I. Jeelani, and J. A. Reashi, "Obtaining strata boundaries under proportional allocation with varying cost of every unit," *Pakistan J. Statist. Operation Res.*, vol. 13, no. 3, p. 567, Sep. 2017, doi: 10.18187/pjsor.v13i3.1719.
[7] F. Danish and S. E. H. Rizvi, "Optimum stratification in bivariate auxiliary variables under Neyman allocation," *J. Modern Appl. Stat. Methods*, vol. 17, no. 1, Jun. 2018, doi: 10.22237/jmasm/1529418671.
[8] F. Danish and S. E. H. Rizvi, "Approximately optimum strata boundaries for two concomitant stratification variables under proportional allocation," *Statistics Transition New Series*, vol. 22, no. 4, pp. 19–40, 2021, doi: 10.21307/stattrans-2021-036.
[9] B. K. Gupt and M. I. Ahamed, "Optimum stratification for a generalized auxiliary variable proportional allocation under a superpopulation model," *Commun. Statist. Theory Methods*, vol. 51, no. 10, pp. 3269–3284, May 2022.
[10] M. Kozak and M. R. Verma, "Geometric versus optimization approach to stratification: A comparison of efficiency," *Survey Methodology*, vol. 32, no. 2, pp. 157–163, 2006.
[11] M. Kozak, "Comparison of random search method and genetic algorithm for stratification," *Commun. Statist. - Simul. Comput.*, vol. 43, no. 2, pp. 249–253, Jan. 2014.
[12] K. G. Reddy and M. G. M. Khan, "Optimal stratification in stratified designs using weibull-distributed auxiliary information," *Commun. Statist. Theory Methods*, vol. 48, no. 12, pp. 3136–3152, Jun. 2019.
[13] K. G. Reddy and M. G. M. Khan, "StratifyR: An R package for optimal stratification and sample allocation for univariate populations," *Austral. New Zealand J. Statist.*, vol. 62, no. 3, pp. 383–405, Sep. 2020.
[14] L. P. Rivest, "A generalization of the Lavallée and hidiroglou algorithm for stratification in business surveys," *Surv. Methodol. Statist. Canada*, vol. 28, no. 2, pp. 191–198, 2002.
[15] G. S. Semaan, J. A. de Moura Brito, I. M. Coelho, E. F. Silva, A. C. Fadel, L. S. Ochi, and N. Maculan, "A brief history of heuristics: From bounded rationality to intractability," *IEEE Latin Amer. Trans.*, vol. 18, no. 11, pp. 1975–1986, Nov. 2020.
[16] J. Brito, T. Veiga, and P. Silva, "An optimization algorithm applied to the one dimensional stratification problem," *Survey Methodology*, vol. 45, no. 2, pp. 295–315, 2019.
[17] Rizvi, "Optimum stratification for two study variables using auxiliary information," Ph.D. thesis, Dept. Agricult. Statist., Punjab Agricult. Univ., Punjab, India, 1997.
[18] S. E. H. Rizvi, J. P. Gupta, and M. Bhargava, "Effect of optimum stratification on sampling with varying probabilities under proportional allocation," *Statistica*, vol. 64, no. 4, pp. 721–733, 2004, doi: 10.6092/issn.1973-2201/68.
[19] M. R. Verma, "Approximately optimum stratification for ratio and regression methods of estimation," *Appl. Math. Lett.*, vol. 21, no. 2, pp. 200–207, Feb. 2008.
[20] F. Danish, "Construction of stratification points under optimum allocation using dynamic programming," *Pakistan J. Statist. Operation Res.*, vol. 15, no. 2, pp. 341–355, Jun. 2019.
[21] F. Danish, S. E. H. Rizvi, and C. Bouza, "On approximately optimum strata boundaries using two auxiliary variables," *Revista Investigacion Operacional*, vol. 41, no. 3, pp. 445–460, 2020.
[22] J. André Brito, L. de Lima, P. H. González, B. Oliveira, and N. Maculan, "Heuristic approach applied to the optimum stratification problem," *RAIRO Operations Res.*, vol. 55, no. 2, pp. 979–996, Mar. 2021.

[23] S. E. H. Rizvi and F. Danish, "Approximately optimum strata boundaries under super population model," *Int. J. Math. Oper. Res.*, vol. 1, no. 1, p. 1, 2022, doi: 10.1504/ijmor.2022.10052805.
[24] C. R. Martínez, A. H. German, A. M. Marvelio, and S. Florentin, "Neutrosophy for survey analysis in social sciences," *Neutrosophic Sets Syst.*, vol. 37, p. 49, Jan. 2020.
[25] L. E. Valencia-Cruzaty, M. Reyes-Tomalá, C. M. Castillo-Gallo, and F. Smarandache, "A neutrosophic statistic method to PredictTax time series in Ecuador," *Neutrosophic Sets Syst.*, vol. 34, pp. 33–39, Jan. 2020.
[26] R. Singh and B. V. Sukhatme, "Optimum stratification," *Ann. Inst. Statist. Math.*, vol. 21, pp. 515–528, 1969.
[27] A. A. H. Ahmadini, F. Danish, R. Jan, A. A. Rather, Y. S. Raghav, and I. Ali, "Unlocking the secrets of apple harvests: Advanced stratification techniques in the Himalayan region," *Heliyon*, vol. 10, no. 11, Jun. 2024, Art. no. e31693.

S. E. H. RIZVI started his career as a Lecturer with IAAS, Tribhuvan University, Nepal, in 1987. He is currently a Professor (Statistics)/Dean with the Faculty of Basic Sciences, SKUAST-Jammu, India, where he joined, in 1989, as an Assistant Professor. He has more than 34 years of teaching experience with active involvement in management and planning of various academic activities. He has vast research experience in the field of sampling theory and applied statistics and supervised. He has supervised the M.Sc. and Ph.D. students in statistics. He has a good number of research publications in journals of national and international repute and associated in research projects in different capacities.

FAIZAN DANISH received the Ph.D. degree from the Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Jammu and Kashmir, India.

He is currently an Assistant Professor in statistics with Vellore Institute of Technology-Andhra Pradesh (VIT-AP) University, Vijayawada, Andhra Pradesh, India. Previously, he was a Postdoctoral Fellow with New York University, New York City, USA. With approximately four and half years of combined teaching and research experience, he has authored around 35 research articles in esteemed journals with four books. Additionally, he has was a Biostatistician for Research Consultation Services in Doha, Qatar. His research interests include sampling theory, mathematical programming, applied statistics, and biostatistics. He has proposed innovative methods for determining stratification points using both classical techniques and mathematical programming approaches. He has proficient in various statistical software, including R, STATA, SPSS, OPSTAT, and WINDOSTAT; and has completed several online courses related to software's from prominent global universities. Recognized for his contributions to the field of Statistics, he has received two young scientist awards and a Best Thesis Award. He also serves as a reviewer for several reputable journals.

RAFIA JAN received the Ph.D. degree in statistics from the University for Kashmir, Jammu and Kashmir, India. She is currently a Teaching Faculty Member of the Government Degree College Bejbahara, Jammu and Kashmir, with a more than three years of teaching experience. She has research expertise in sampling theory and applied statistics. She has developed several estimators in different sampling designs.

ISMAIL A. MAGEED received the Doctorate degree in applied probability from The University of Bradford, U.K.

He is currently the U.K. President of the International Society of Fuzzy Set Extensions and Applications (ISFSEA). He has been nominated by numerous high-profile academic institutions to the world prestigious Abel Prize (Noble Prize of Mathematics), in 2025, based on his great services to humanity through revolutionary mathematical applications to advance several scientific disciplines, including engineering and computer science. His current research interests include the unification of queueing theory with information theory and information geometry. He is on the panel board of numerous prominent journals, leading numerous volunteer research teams worldwide. He has been officially honored by Egyptian National Council of Youth, the Arab Human Sciences Organization, and CIDA-PARIS-FRANCE as a governmental recognition of his services to science.

SUMAYA AL MAADEED (Senior Member, IEEE) is currently a Full Professor with the Department of Computer Science and Engineering, Qatar University. She was the Head of the Computer Science Department and the Leader of many committees. She leads a research team in forensics, artificial intelligence applications. She has been awarded several million of USD for research projects in computer sciences and published 100's of papers in top-tier journals and conferences.

JIHAD MOHAMED ALJA'AM received the B.Sc., M.S., and Ph.D. degrees in computing from Southern University [the National Council for Scientific Research (CNRS)], France.

He is currently a Full Professor with the School of Computing and Data Sciences, OUC-Liverpool John Moores University, and the Program Leader of CS Program and the Coordinator of the Research Office. He leads a research team in NLP, knowledge extraction, and multimedia. He was also a Project Manager with IBM, Paris, and an IT Consultant with RTS, France, for several years. He has published more than 210 articles and eight book chapters in computing and information technology. He was a member of the editorial boards of the *Journal of Soft Computing*, *American Journal of Applied Sciences*, the *Journal of Computing and Information Sciences*, the *Journal of Computing and Information Technology*, and the *Journal of Emerging Technologies in Web Intelligence*. He received the Nicolas D. Georganas Best Paper Award from the ACM Transactions on Multimedia Computing, Communications, and Applications, in 2015, the Best Research Paper at the Tenth Annual International Conference on Computer Games Multimedia and Allied Technologies, in 2016, and the Best Research Paper Award at the IEEE ICe3 Conference on eLearning.