Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 3149912, 10 pages
https://doi.org/10.1155/2023/3149912

Research Article

Optimum Stratification Using Dynamic Programming with a Mixture of Ratio and Product Estimators under Super Population Model

Faizan Danish,[1] Rafia Jan,[2] Muhammad Daniyal,[3] and Kassim Tawiah [4,5]

[1]*Department of Mathematics, School of Advanced Sciences, VIT-AP University, Inavolu, Beside AP Secretariat, Amaravati, AP, India*
[2]*Government Degree College for Boys Anantnag, Anantnag, J & K, India*
[3]*Department of Statistics, The Islamia University of Bahawlapur, Bahawlapur, Pakistan*
[4]*Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana*
[5]*Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

Correspondence should be addressed to Kassim Tawiah; kassim.tawiah@uenr.edu.gh

Received 3 November 2022; Revised 4 February 2023; Accepted 10 February 2023; Published 29 April 2023

Academic Editor: Omar-Jacobo Santos

In this study, we have utilized two study variables and one auxiliary variable. The auxiliary variable is used as the stratification variable, and we selected the sample using the stratification variable with a mixture of ratio and product estimators. Under super population set-up, minimal equations have been obtained through minimization of the aggregated variance with the help of the variables under study. The objective function is minimized with respect to the constraints under consideration. The dynamic programming approach has been used to minimize the variance and obtain the optimum strata boundaries. Empirical studies have also been made on the proposed rule utilizing different distributions. A simulation study has been done which shows the gain in precision using the proposed method.

1. Introduction

In stratified random sampling, carefully choosing the optimum strata boundaries would lead to a higher degree of relative precision. Hansen and Hurwitz [1] pioneered the concept of strata boundaries as an extension to Dalenius' [2] work for univariate cases. Sadasivan and Aggarwal [3] studied the variables under consideration as stratification variables under Neyman allocation. For several characteristics under consideration for estimation, it is not possible to utilize direct optimum allocation. Ghosh [4] considered the proportional method of allocation using two stratification variables. Several methods have been proposed in different situations such as Singh [5], Dalenius and Gurney [6], Danish et al. [7], Danish and Rizvi [8], Danish and Rizvi [9], and Gupt and Ahamed [10].

In recent years, there has been an incredible interest in researchers in the area of stratification points. Rizvi et al. [11] used the compromised method and Verma's [12] ratio and regression method for obtaining approximately optimum strata boundaries (AOSB). Danish and Rizvi [8] proposed a method for obtaining stratification points using two highly related variables. Two stratification variables have been used by Danish et al. [13], Danish and Rizvi [9], and Danish and Rizvi [14]. Abo-El Hassan et al. [15] proposed goal programming for obtaining the stratification points. There has been a dramatic increase in studies regarding obtaining strata boundaries, with some of them being the most recent work of Brito et al. [16], Reddy and Khan [17], and Danish et al. [18]. Alshqaq et al. [19] discussed the linear approximation of the multivariate stratified sampling problem with examples. Hamid et al. [20] suggested that the mathematical

goal programming model can determine the optimum strata boundaries by bivariate variables in multiobjective problems with minimum variance.

Brito et al. [21] proposed a hybrid approach that works to identify the cutoff points of the strata through an optimization method and then, an exact method proposed, perform the optimal allocation of the sample to the strata. More specifically, Fadal et al. [22] proposed a heuristic algorithm based on the Biased Random Key Genetic Algorithm (BRKGA) meta-heuristic for the univariate stratification problem for objective (ii). de Moura Brito et al. [23]proposed an exact algorithm based on the concepts of graph theory and the minimization of the expression of variance and application of proportional allocation. Lisic et al. [24], under the hypothesis that the stratification variable has a Weibull distribution, solved the stratification problem using the dynamic programming technique and Neyman allocation. Furthermore, Rizvi and Danish [25] utilized a product estimator for obtaining the stratification points using the classical approach.

In practical situations with two study variables, it may happen that one study variable is highly positively correlated with the stratification variable and the other variables are negatively correlated with the stratification variable. Let us assume that the study variable Y has a high and positive correlation with the auxiliary variable X and that the correlation between another study variable Z and the stratification variable X is negative.

In the present investigation, the issue of stratification points for two study variables is investigated by simple random sampling by selecting a sample for estimating the population means using an auxiliary variable (X) with a mixture of ratio and product estimators implementing the technique of dynamic programming. Under the super population setup, minimal equations have been obtained through minimization of the aggregated variance with the help of the variables under study. Furthermore, past information on the functional relationship of Y and Z on X and the conditional variance functions $V(y|x)$ and $V(z|x)$ are also assumed. The problem is solved as a multistage decision criterion. The auxiliary variable is used as a stratification variable for selecting the sample used as the stratification variable with a mixture of ratio and product estimators. We have utilized dynamic programming for obtaining the stratification points. A simulation study is performed to obtain the relative precision to compare the existing proposed methods.

We present the variance and covariance for the mixture of ratio and product estimators under the superpopulation minimal equations and dynamic programming as a solution procedure, obtaining the optimum sample size, empirical study, simulation study, and conclusions in this paper.

2. Variance and Covariance Expressions under Super-PopulationSet-Up

Let us make L strata from the given population of size N and assume that in each stratum, the regression lines of the two interested variables on the highly related variable are linear and pass through origin.

Let us assume the model as

$$Y_j = C_j(X) + e_j, \quad (j = 1, 2), \tag{1}$$

where $C_j(X)$ is a real function of X, and e_j is disturbance so that $E(e_j/X) = 0$, $E(e_j e_j'/X, X') = 0$, for $x \neq x'$, $V(e_j/X) = \eta_j(x_i) > 0$, $j = 1, 2$, $x \in (a, b)$, $(b - a) < \infty$. $E(e_j c_j) = 0$ but $E(c_1 c_2) \neq 0$. If $f_s(x, y_1, y_2)$ denoted joint density function of (X, Y_1, Y_2) and $f(x)$ marginal of X in the superpopulation model, then we have

$$\begin{aligned} W_h &= \int_{x_{h-1}}^{x_h} f(x)\mathrm{d}x, \\ \mu_{hy_j} &= \mu_{hc_j} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} C_j(x) f(x)\mathrm{d}x, \\ \sigma^2_{hc_j} &= \frac{1}{W_h}\int_{x_{h-1}}^{x_h} C_j^2(x) f(x)\mathrm{d}x - \left(\mu_{hc_j}\right)^2, \\ \sigma^2_{hy_j} &= \sigma^2_{hc_j} + \mu_{h\eta_j}, \end{aligned} \tag{2}$$

where (x_{h-1}, x_h) stratification points, $\mu_{h\eta_j}$ is the average value, and $\eta_j(x_i)$ is the conditional variance of the h^{th} subpopulation.

Let us assume that the population of "N" units are split into "L" strata. The separate ratio estimates for the population mean in stratified random sampling are given by

$$\overline{Y}_{\mathrm{st}.R_1} = \sum_{h=1}^{L} W_h \overline{Y}_{hR_1}, \tag{3}$$

where $W_h = N_h/N$, h^{th} stratum weight, $\overline{Y}_h$ is the mean of Y

$$\begin{aligned} \overline{Y}_{hR_1} &= \left(\frac{\overline{Y}_h}{\overline{x}_h}\right), \\ \overline{X}_h &= R_{1h}\overline{X}_h, \end{aligned} \tag{4}$$

$\overline{x}_h$ is the sample mean of X, $\overline{X}_h$ is the population mean of the auxiliary variable X, $\overline{Y}_{\mathrm{st}.R_1}$ denotes the separate ratio estimates for the population mean in stratified random sampling.

Now, we assume that in each stratum, the regression lines of the stratification variable on the auxiliary variable are linear and pass through the origin. Furthermore, we assume from characteristics Z, $R_{21} = R_{22} = R_{23} = \ldots = R_{2L}$ so that we can use a combined product estimator. The combined estimators in the case of stratified sampling are given by

$$\overline{Z}_{\mathrm{st}.P} = \frac{\sum(W_h\overline{Z}_h)\sum(W_h\overline{X}_h)}{\overline{X}}, \tag{5}$$

where $W_h = N_h/N$, the strata weight, $\overline{x}_h$ is the sample mean of X, $\overline{X}$ is the population mean of X, $\overline{Z}_h$ is the sample mean of Z.

If the finite population correction (FPC) is neglected, the approximate variances of these estimators, under proportional allocation, are given by

$$V\left(\overline{Y}\text{st.}R_1\right)_P = \frac{1}{n}\sum_{h=1}^{L} W_h\left(\sigma_{hy}^2 + R_{1h}^2\sigma_{hx}^2 - 2R_{1h}\sigma_{hxy}\right),$$
$$\left(V\overline{Z}_{\text{st.}P}\right)_P = \frac{1}{n}\sum_{h=1}^{L} W_h\left(\sigma_{hz}^2 + R_2^2\sigma_{hx}^2 - 2R_2\sigma_{hxz}\right). \tag{6}$$

For the covariance expression, we have the following lemma.

Lemma 1. *The covariance expression between the estimators $\overline{Y}_{\text{st.}R_1}$ and $\overline{Z}_{\text{st.}P}$ as defined by equations (3) and (5), respectively, up to the first order of approximation, is given by*

$$\text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{\text{st.}P}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\sigma_{hyz} + R_2\sigma_{hyz} - R_{1h}\sigma_{hxz} - R_{1h}R_2\,\sigma_{hxz}\right). \tag{7}$$

Proof. Using partially the proofs of Lemma 5.1 and Lemma 5.3 from Rizvi [26], we have

$$\text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{\text{st.}P}\right) = \frac{1}{\overline{\overline{X}}}\,\text{Cov}\left[\left\{\sum W_h\left(\epsilon_{1h}\overline{X}_h - \frac{\xi\overline{Y}_h}{\overline{X}_h}\right), \overline{X}\sum W_h\epsilon_{2h} + \overline{Z}\sum W_h\xi_h\right\}\right]. \tag{8}$$

Which by simplification results in

$$\text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{\text{st.}P}\right) = \frac{1}{\overline{\overline{X}}}\,\text{Cov}\left[\left\{\sum W_h^2\left[\overline{X}\text{Cov}\left(\epsilon_{1h}, \epsilon_{2h}\right), \overline{Z}\text{Cov}\left(\epsilon_{1h}, \xi_h\right)\overline{X}R_{1h}\text{Cov}\left(\xi_h, \epsilon_{2h}\right) - \overline{Z}R_{1h}\text{Cov}\left(\xi_h, \xi_h\right)\right]\right\}\right]. \tag{9}$$

Finally, we have

$$\text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{\text{st.}P}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right). \tag{10}$$

Thereby, proving the lemma.

Under the proportional method of allocating the sample size to different strata, the formula for covariance as given by equation (7) reduces to

$$\text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{st.P}\right) = \frac{1}{n}\sum_{h=1}^{L} W_h\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right). \tag{11}$$

□

3. Minimal Equations

Let $\{x_h\}$ represents stratification points in (a, b) of the stratification variable; corresponding to those strata boundaries, the generalized variance G_6 as given by equation.

$$G_6 = \begin{vmatrix} \sigma_y^2 & \sigma_{yz} \\ \sigma_{zy} & \sigma_z^2 \end{vmatrix} = \sigma_y^2\sigma_z^2 - \left(\sigma_{yz}\right)^2, \tag{12}$$

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

Differentiating G_6 partially with respect to $\{x_h\}$ and equating it derivative to zero, we get

$$\frac{\partial G_6}{\partial x_h} = \sigma_y^2\frac{\partial\sigma_z^2}{\partial x_h} + \sigma_z^2\frac{\partial\sigma_y^2}{\partial x_h} - 2\sigma_{yz}\frac{\partial\sigma_{yz}}{\partial x_h} = 0, h = 1, 2, 3 \ldots, L-1. \tag{13}$$

Inserting the values of σ_y^2, σ_z^2 and σ_{yz} from equations (6) and (11) in equation (13), we have

$$\sigma_y^2\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hz}^2 + R_2^2\sigma_{hx}^2 - 2R_2\sigma_{hxz}\right)\right] + \sigma_z^2\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hy}^2 + R_{1h}^2\sigma_{hx}^2 - 2R_{1h}\sigma_{hxy}\right)\right]$$
$$- 2\sigma_{yz}\frac{\partial}{\partial x_h}\left[\sum_{h=1}^{L} W_h\left(\sigma_{hyz} + R_2\sigma_{hxy} - R_{1h}\sigma_{hxz} - R_{1h}R_2\sigma_{hx}^2\right)\right] = 0. \tag{14}$$

Now let us assume that the functional relationship on Y on X and Z on X is linear in each stratum and that the regression lines pass through the origin. Then, the approximate regression model can be given as

$$Y_1 = R_{1h}X + e_{1i}, \tag{15}$$

$$Z_1 = R_2X + e_{2i}, \tag{16}$$

where e_{1i} and e_{2i} represent the error terms in the study variables Y and Z, respectively.

Now, the variance expressions for proportional allocation under these models [27] can be expressed as

$$\sigma_y^2 = V\left(\overline{Y}_{\text{st.}R_1}\right)_P = \frac{\mu_{\eta_1}}{n}, \tag{17}$$

$$\sigma_z^2 = V\left(\overline{Y}_{\text{st.}P}\right)_P = \frac{4R_2^2}{n}\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \frac{\mu_{\eta_2}}{n}, \tag{18}$$

where σ_y^2 and σ_z^2 denotes the general variance of the study variables Y and Z, respectively.

The covariance term can be obtained as

$$\sigma_{yz} = \text{Cov}\left(\overline{Y}_{\text{st.}R_1}, \overline{Z}_{st.P}\right) = \sum_{h=1}^{L} W_h\left(R_{1h}R_2\sigma_{hx}^2 + R_{1h}R_2\sigma_{hx}^2 - R_{1h}R_2\sigma_{hx}^2 - R_{1h}R_2\sigma_{hx}^2\right) = 0. \tag{19}$$

If $f(x)$ is known and is integrable, then, W_h, σ_{hx}^2, and μ_{hx} can be expressed in terms of $(x_{h-1}, x_h,)$ as follows:

$$W_h = \int_{x_{h-1}}^{x_h} f(x)\text{d}x, \tag{20a}$$

$$\sigma_{hx}^2 = \int_{x_{h-1}}^{x_h} x^2 f(x)\text{d}x - \left(\mu_{hx}\right)^2, \tag{20b}$$

$$\mu_{hx} = \int_{x_{h-1}}^{x_h} xf(x)\text{d}x. \tag{20c}$$

Let $f(x_i)$ be the estimated frequency distribution of the variable $x_i\,(i = 1, 2, 3, \ldots, p)$ in the range of (r, s) then we need to find the intermediates points of X to cut up the range (r, s) at $(L-1)$ points $r = x_0 \le x_1 \le x_2 \le \ldots \le x_L = s$ such that the total variance given in equations (17)–(19) is minimum.

$$V\left(\overline{Y}_{\text{st.}R_1}\right)_P = \frac{\mu_{\eta_1}}{n} + \frac{4R_2^2}{n}\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \frac{\mu_{\eta_2}}{n}. \tag{21}$$

This can be written as

$$V\left(\overline{Y}_{\text{st.}R_1}\right)_P = \frac{4R_2^2}{n}\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \frac{\mu_{\eta_1}}{n} + \frac{\mu_{\eta_2}}{n}. \tag{22}$$

For a constant sample size n, reducing the previous variance is equivalent to reducing the variance

$$V\left(\overline{Y}_{\text{st.}R_1}\right)_P = R_2^2\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \mu_{\eta_1} + \mu_{\eta_2}. \tag{23}$$

Thus, the optimization function in equation (23) can be written as a function of the stratification points (x_h, x_{h-1}) only as

$$\gamma_h(x_h, x_{h-1}) = R_2^2\sum_{h=1}^{L} W_h\sigma_{hx}^2 + \mu_{\eta_1} + \mu_{\eta_2}. \tag{24}$$

Thus, the problem of obtaining the stratification points can be expressed as

Minimize $\sum_{h=1}^{L}\gamma(x_h, x_{h-1})$

$$\text{Subject to } r = x_0 \le x_1 \le x_2 \le \ldots \le x_L = s. \tag{25}$$

The length of $X(r, s)$ can be written as $x_L - x_0$. In the same fashion, $u_h = x_h - x_{h-1}$, $(r, s)h = 1, 2, \ldots, L$ where $u_h \ge 0$ indicates the length of h^{th} stratum. Thus, we can write

$$\sum_{h=1}^{L} u_h = \sum_{h=1}^{L}\left(x_h - x_{h-1}\right) = s - r = x_L - x_0. \tag{26}$$

Hence, the last stratification point can be expressed as

$$x_h = x_0 + \sum_{i=1}^{L} u_i = x_{h-1} + u_h. \tag{27}$$

Taking equation (26) as a subject to constraint, the optimization problem can be expressed as

Minimize $\sum_{h=1}^{L}\gamma(q_h, x_{h-1})$

Subject to constraints

$$\sum_{h=1}^{L} u_h = k, \tag{28}$$

$$u_h \ge 0, \quad h = 1, 2, 3, \ldots, L. \tag{29}$$

Obviously, if x_0 is given, then the initial term $\gamma_1(u_1, x_0)$, the objective function of Mathematical Programming Problem (MPP), given in equation (28) is a function of u only. Similarly, if q_1 is given, the second term $\gamma_2(u_2, x_1)$ will be the function of u_2 only and in the same way, the proceeding terms will be expressed as a function of the succeeding terms.

Keeping in view the particular connection between different terms, the optimization problem can be expressed as

Minimize $\sum_{h=1}^{L}\gamma_h(u_h)$

Subject to constraint

$$\sum_{h=1}^{L} u_h = k, \qquad u_h \geq 0, h = 1, 2, 3, \ldots, L. \tag{30}$$

In practical situations, usually, the variable of interest is not known at the initial stage of designing the experiment, thus the highly associated variable is being used for the estimation of stratification points. In the proposed technique, we carry out the optimization technique for the equation (30) on the defined range "k" which is derived from its highly associated variables. It is to be noted here that if the objective function is comprised of any parameters, it should be either fixed or chosen from literature in advance.

4. Dynamic Programming as Solution Procedure

The problem given in equation (28) is a type of problem that can be solved at different stages having a main function along with constraints as separable functions of u_h, which enhances us to utilize the technique of dynamic programming (DP) [28]. Dynamic programming is prominently used in the case of recursion but a plain one and has replicated calls for the same inputs. The approach is to utilize one subproblem's optimal solution as an initial feasible solution in other sub problems to get the optimal solution.

Now, we take a fraction of the problem as

Minimize $\sum_{h=1}^{L} \gamma_h(u_h)$

Subject to constraint

$$\sum_{h=1}^{L} u_h = k_p, \qquad u_h \geq 0, h = 1, 2, 3, \ldots, p, \tag{31}$$

where $k_p < k$

$$k_p = u_1 + u_2 + u_3 + \ldots + u_p, \qquad k_{p-1} = u_1 + u_2 + u_3 + \ldots + u_{p-1} = k_p - u_p. \tag{32}$$

Let $\zeta_p(u_p)$ indicates the lowest value of the MPP equation (30), which means

$$\zeta_p(u_p) = \min\left[\sum_{h=1}^{L} \gamma_h(u_h) \sum_{h=1}^{L} u_h = u_p, u_h \geq 0, h = 1, 2, 3, \ldots, p \,\&\, 1 \leq p \leq L\right]. \tag{33}$$

With this procedure, equation (28) is equal to finding $\zeta_L(u)$ recursively by estimating $\zeta_m(u_m)$ for $m = 1, 2, \ldots, L$ and $0 \leq u_p \leq u$, we have

$$\zeta_p(u_p) = \min\left[\gamma_p(u_h) + \sum_{h=1}^{p-1} \gamma_h(u_h) \sum_{h=1}^{m-1} u_h = k_p - u_p, u_h \geq 0, h = 1, 2, 3, \ldots, p \,\&\, 0 \leq u_p \leq c_p\right]. \tag{34}$$

For the particular value of u_p, we have

$$\zeta_p(u_p) = \gamma_p(u_h) \min\left[\sum_{h=1}^{p-1} \gamma_h(u_h) \sum_{h=1}^{m-1} u_h = k_p - u_p, u_h \geq 0, h = 1, 2, 3, \ldots, p \,\&\, 0 \leq u_p \leq c_p\right]. \tag{35}$$

Thus, we can utilize Bellman's principle of optimality and the recursion equation of the DP for $p \geq 2$

$$\zeta_p(u_p) = \min_{0 \leq u_p \leq c_p} \left[\gamma_p(u_h) + \zeta_{p-1}(k_p - u_p)\right]. \tag{36}$$

If we put $p = 2$, which is for the first stage

$$\zeta_1(u_1) = \gamma_1(u_1) = u_1^* = k_1, \tag{37}$$

where $u_1^* = k_1$ is the total deviation or range of the first stratum. Thus, equations (36) and (37) can be solved in a forward manner for different values of $p = 1, 2, \ldots, L$ to determine the optimum fraction of the problem's objective and then estimate it in a backward manner to estimate the optimum strata boundaries (OSB).

5. Obtaining the Optimum Sample Size

When the stratification points $(x_h - x_{h-1})$ are determined as per the section discussed above, the estimation of the optimum sample size n_h, $h = 1, 2, \ldots, L$ for the h^{th} stratum can be easily determined.

As per the functional relationship defined in equations (16) and (17) for the study variables and auxiliary variable for all strata, we use equation (23) but for the fixed constant sample size "n."

For h^{th} stratum the sample size is

$$n_h = \frac{nW_h\sqrt{{\sigma_{hj(x)}}^2 + \sigma_{he}^2}}{\sum_{h=1}^{L} W_h\sqrt{{\sigma_{hj(x)}}^2 + \sigma_{he}^2}}, \tag{38}$$

where W_h, ${\sigma_{hj(x)}}^2$ andσ_{he}^2 denotes the weight and variance of the h^{th} stratum. ${\sigma_{hj(x)}}^2$ denotes the variance of the functional form of the auxiliary variable and σ_{he}^2 denotes variance of the error term, which can be derived in terms of the stratification points $(x_h - x_{h-1})$. Furthermore, it is to be noted that $1 \le n_h \le N_h$, where N_h denotes the total size of h^{th} stratum.

6. Empirical Study

Let us assume the log-normally distribution auxiliary variable X with probability density function (pdf) as

$$f(x) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-(\log x - \mu)/2\sigma^2}, & x > 0, \sigma > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

Using equations (20a)–(20c), we get

$$W_h = \frac{1}{2}E_1, \tag{40}$$

$$\mu_h = \exp\left(\frac{\sigma^2}{2} + \mu\right)\frac{E_3}{E_1}, \tag{41}$$

$$\sigma_h^2 = \frac{1}{E_1}\left\{\left[\exp\left(2\sigma^2 + 2\mu\right)E_2\right][E_1] - \left[\exp\left(\frac{\sigma^2}{2} + \mu\right)(E_3)\right]^2\right\}, \tag{42}$$

where,

$$\begin{aligned} E_1 &= \operatorname{erf}\left(\frac{\log\left(u_h + x_{h-1}\right) - \mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log\left(x_{h-1}\right) - \mu}{\sigma\sqrt{2}}\right), \\ E_2 &= \operatorname{erf}\left(\frac{\log\left(u_h + x_{h-1}\right) - \mu - 2\sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log\left(x_{h-1}\right) - \mu - 2\sigma^2}{\sigma\sqrt{2}}\right), \\ E_3 &= \operatorname{erf}\left(\frac{\log\left(u_h + x_{h-1}\right) - \mu - \sigma^2}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\log\left(x_{h-1}\right) - \mu - \sigma^2}{\sigma\sqrt{2}}\right), \end{aligned} \tag{43}$$

where

$$\operatorname{erf}(\omega) = \frac{2}{\sqrt{\pi}}\int_0^{\omega} e^{-j^2}\mathrm{d}j, \tag{44}$$

and its properties

$$\begin{aligned} \operatorname{erf}(-\omega) &= -\operatorname{erf}(\omega), \\ \operatorname{erf}(0) &= 0, \\ \operatorname{erf}(\infty) &= 1, \\ \operatorname{erf}(-\infty) &= 1. \end{aligned} \tag{45}$$

Using equations (40)–(42) in equation (30), we get Minimize

$$R_2^2\sum_{h=1}^{L}\frac{1}{2}E_1\frac{1}{E_1}\left\{\left[\exp\left(2\sigma^2 + 2\mu\right)E_2\right][E_1] - \left[\exp\left(\frac{\sigma^2}{2} + \mu\right)(E_3)\right]^2\right\} + \mu_{\eta_1} + \mu_{\eta_2}. \tag{46}$$

Subject to constraint

$$\sum_{h=1}^{L} u_h = k, \tag{47}$$

and $u_h \ge 0, h = 1, 2, 3, \ldots, L$

Let us assume now that the standard log-normal distribution is defined in the interval $x\epsilon[0.000, 30.000]$ that is $x_0 = 0.000$ and $x_L = 30.000$, $\mu = 0, \sigma = 1$. This implies $x_L - x_0 = 30.00 - 0.000 = 30$ and have fixed sample size $n = 200$.

Table 1: Optimum strata boundaries, sample size, and total variance for log-normally distributed auxiliary variables.

L (no. of strata)	OSB	n_h (sample size)	Total variance
2	4.2643	102 98	1.6355
3	6.1785 9.3972	70 64 66	1.0576
4	2.7314 5.3049 11.7431	52 51 47 50	0.8513
5	1.9742 3.1857 7.2533 13.5943	40 42 40 41 37	0.6501
6	1.6476 3.9182 7.5193 11.4136 15.9548	34 36 32 31 35 32	0.4925

Executing the MPP given in equation (47), we get the stratification points along with variance and sample size as presented in Table 1.

Now let us assume the variable X follows gamma distribution with probability density function as

$$f(z) = f(x, s, \theta) = \begin{cases} \frac{1}{\theta^s} x^{s-1} e^{-x/\theta}, & x \geq 0, s, \theta > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{48}$$

where "s" is the slope and "θ" is the scale parameter and Γs is a gamma distribution function defined as

$$\Gamma s = \int_0^\infty e^{-x} x^{s-1} \mathrm{d}z, s > 0. \tag{49}$$

This function is also defined by the upper incomplete gamma function $\Gamma(s, x)$ and a lower incomplete gamma function $\gamma(s, x)$, respectively, as

$$\sqrt{(s, x)} = \int_x^\infty u^{s-1} e^{-u} \mathrm{d}u, \tag{50}$$

and $\gamma(s, x) = \int_0^x u^{s-1} e^{-u} \mathrm{d}u$.

There is also an incomplete gamma function whose values are from 0 to 1 as

$$\begin{aligned} Q(s, x) &= \frac{1}{\Gamma s} \int_x^\infty u^{s-1} e^{-u} \mathrm{d}u, s, x > 0, \\ P(s, x) &= \frac{1}{\Gamma s} \int_0^x u^{s-1} e^{-u} \mathrm{d}u, s, x > 0 \neq 0, \end{aligned} \tag{51}$$

where $Q(s, x)$ and $P(s, x)$ represent upper and lower regularized incomplete gamma function, respectively.

Using these values in equations (20a)–(20c), we get

$$W_h = Q\left(s, \frac{x_{h-1}}{\theta}\right) - Q\left(s, \frac{x_{h-1} + u_h}{\theta}\right), \tag{52}$$

$$\sigma_{hx}^2 = \frac{\theta^2 s(s+1)\left[Q(s+2, x_{h-1}/\theta) - Q(s+2, x_h/\theta)\right]}{\left[Q(s, x_{h-1}/\theta) - Q(s, x_h/\theta)\right]} - \frac{\theta^2 s^2 \left[Q(s+1, x_{h-1}/\theta) - Q(s+1, x_{h-1} + u_h/\theta)\right]^2}{\left[Q(s, x_{h-1}/\theta) - Q(s, x_{h-1} + u_h/\theta)\right]^2}. \tag{53}$$

Using equations (52) and (53) in equation (30), we have

Minimize

$$R_2^2 \sum_{h=1}^{L} \left[Q\left(s, \frac{x_{h-1}}{\theta}\right) - Q\left(s, \frac{x_{h-1} + u_h}{\theta}\right)\right] \left[\frac{\theta^2 s(s+1)\left[Q(s+2, x_{h-1}/\theta) - Q(s+2x_h/\theta,)\right]}{\left[Q(s, x_{h-1}/\theta) - Q(s, x_h/\theta)\right]} - \frac{\theta^2 s^2 \left[Q(s+1, x_{h-1}/\theta) - Q(s+1, x_{h-1} + u_h/\theta)\right]^2}{\left[Q(s, x_{h-1}/\theta) - Q(s, x_{h-1} + u_h/\theta)\right]^2}\right] + \mu_{\eta_1} + \mu_{\eta_2}. \tag{54}$$

Subject to constraint

$$\sum_{h=1}^{L} u_h = k, \tag{55}$$

and $u_h \geq 0, h = 1, 2, 3, \ldots, L$

The maximum likelihood estimate of the parameters for the gamma distribution was found to be $s = 3.836157$ and $\theta = 2.937784$

By assuming the auxiliary variable $x \epsilon [0.0005, 26.000]$ with mean $x_0 = 0.0005$, $x_L = 26.000$ and fixed sample size $n = 300$ and executing the MPP given in equation (47), we get the stratification points presented in Table 2.

7. A Simulation Study

We performed a simulation study to verify the validity of the proposed method by checking its relative precision using the DP technique comparative with the below-mentioned methods utilizing R statistical software.

(i) Dalenius et al. [29] cum $\sqrt{f}$ method

(ii) Gunning and Horgan [30] geometric method

(iii) Lavallée and Hidiroglou [31] approach using Kozak's [32] method

(iv) Khan et al. [33] mathematical programming approach

Table 2: Optimum strata boundaries, sample size, and total variance for gamma distributed auxiliary variables.

L (no. of strata)	OSB	n_h (sample size)	Total variance
2	6.7495	195 105	3.2544
3	5.7664 9.7621	135 119 46	3.0827
4	4.6904 8.3862 12.9146	99 103 78 20	2.9811
5	4.1198 7.6971 11.2892 14.7314	76 92 69 41 22	2.0572
6	6.5385 9.3508 16.2272 19.0509 23.9069	72 87 82 42 32 17	1.7821

Table 3: Total variance obtained by different methods.

L	Cum $\sqrt{f}$ method	Geometric method	Lavallee–Hidiroglou method	Khan et al. [28]	Proposed method
2	0.786	0.82248	1.05376	0.59128	0.41944
3	0.634	0.66992	0.5908	0.22872	0.14256
4	0.48392	0.60744	0.55664	0.16384	0.0828
5	0.38696	0.51336	0.4908	0.06688	0.06344
6	0.31792	0.44296	0.48288	0.02872	0.01904

(v) Proposed method.

We utilized a uniformly distributed auxiliary variable with a data set of 8000, $a = 0.005$, and $b = 1.90$ in R software for our simulation. Our minimum and maximum values came out to be 0.0046 and 1.8842, respectively, with a total deviation $k = 1.877$.

Thus, we have outlined the stratification points using our proposed method as discussed above with the comparative methods. The variance obtained by all these methods along with the proposed method is presented in Table 3. Our proposed method gives a better estimate than the existing methods.

8. Conclusion

In the current investigation, the case of a mixture of ratio and product methods of estimation has been dealt with using mathematical equations obtained after minimizing the variance, which evolved in the estimation. We proposed a method for the estimation of strata boundaries using dynamic programming along with the sample size for each stratum. Through empirical study, it is seen that the gain in efficiency is remarkably high for different distribution functions for the auxiliary variable. Furthermore, Tables 1 to 3 suggest the superiority of our developed method over the existing methods. As a result, our proposed methodology will be useful for obtaining OSB for the variables or characteristics under consideration while using the frequency distribution of the auxiliary variables. When the data are coming from a complex process, neutrosophic statistics is prioritized over classical statistics. Several studies have been done in this regard such as Reddy et al. [34], Martínez et al. [35], Cruzaty et al. [36], and Danish [37]. Thus, the utilization of neutrosophic statistics can be considered in future studies.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Faizan Danish, Rafia Jan, Muhammad Daniyal, and Kassim Tawiah conceptualized the data. Faizan Danish, Rafia Jan, and Muhammad Daniyal did the formal analysis. Faizan Danish, Rafia Jan, Muhammad Daniyal, and KassimTawiah

gave the methodology. Faizan Danish, Rafia Jan, Muhammad Daniyal, and Kassim Tawiah gave the validation. Faizan Danish, Rafia Jan, Muhammad Daniyal, and Kassim Tawiah did the visualization. Faizan Danish, Rafia Jan, and Muhammad Daniyal did the writing–original draft. Muhammad Daniyal and Kassim Tawiah did the writing–editing and reviewing:.

References

[1] M. H. Hansen and W. N. Hurwitz, "On the theory of sampling from finite populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 333–362, 1943.

[2] T. Dalenius, "The problem of optimum stratification," *Scandinavian Actuarial Journal*, vol. 1950, no. 3-4, pp. 203–213, 1950.

[3] G. Sadasivan and R. Aggarwal, "Optimum points of stratification in bi-variate populations," *Sankhya Series C*, vol. 40, pp. 84–97, 1978.

[4] S. P. Ghosh, "Optimum stratification with two characters," *The Annals of Mathematical Statistics*, vol. 34, no. 3, pp. 866–872, 1963.

[5] R. Singh, "Approximately optimum stratification on the auxiliary variable," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 829–833, 1971.

[6] T. Dalenius and M. Gurney, "The problem of optimum stratification. II," *Scandinavian Actuarial Journal*, vol. 1951, no. 1-2, pp. 133–148, 1951.

[7] F. Danish, S. Rizvi, M. I. Jeelani, and J. A. Reashi, "Obtaining strata boundaries under proportional allocation with varying cost of every unit," *Pakistan Journal of Statistics and Operation Research*, vol. 13, no. 3, pp. 567–574, 2017.

[8] F. Danish and S. E. H. Rizvi, "Optimum stratification in Bivariate auxiliary variables under Neyman allocation," *Journal of Modern Applied Statistical Methods*, vol. 17, no. 1, p. 13, 2018.

[9] F. Danish and S. E. H. Rizvi, "Approximately optimum strata boundaries for two concomitant stratification variables under proportional allocation," *Statistics in Transition new series*, vol. 22, no. 4, pp. 19–40, 2021.

[10] B. K. Gupt and M. I. Ahamed, "Optimum stratification for a generalized auxiliary variable proportional allocation under a superpopulation model," *Communications in Statistics - Theory and Methods*, vol. 51, no. 10, pp. 3269–3284, 2022.

[11] S. E. H. Rizvi, J. P. Gupta, and M. Bhargava, "Optimum stratification based on auxiliary variable for compromise allocation," *Metron-international journal of statistics*, vol. 28, no. 1, pp. 201–215, 2002.

[12] M. R. Verma, "Approximately optimum stratification for ratio and regression methods of estimation," *Applied Mathematics Letters*, vol. 21, no. 2, pp. 200–207, 2008.

[13] F. Danish, S. E. H. Rizvi, and C. Bouza, "On approximately optimum strata boundaries using two auxiliary variables," *Operational Research Magazine*, vol. 41, no. 3, pp. 445–461, 2020.

[14] F. Danish and S. E. H. Rizvi, "Optimum stratification by two stratifying variables using mathematical programming," *Pakistan Journal of Statistics*, vol. 35, no. 1, pp. 11–24, 2019.

[15] F. S Abo-El Hassan, R. Hamid, E. A Ismail, and S. M Ezzat, "Behavior of mathematical goal programming for determining optimum stratum boundaries," *Scientific Journal of the faculty of Commerce*, vol. 1, no. 26, pp. 130–143, 2021.

[16] J. A. Brito, L. de Lima, P. Henrique González, B. Oliveira, and N. Maculan, "Heuristic approach applied to the optimum stratification problem," *RAIRO - Operations Research*, vol. 55, no. 2, pp. 979–996, 2021.

[17] K. G. Reddy and M. G. M. Khan, "stratifyR: an R Package for optimal stratification and sample allocation for univariate populations," *Australian & New Zealand Journal of Statistics*, vol. 62, no. 3, pp. 383–405, 2020.

[18] F. Danish, S. Rizvi, M. K. Sharma, S. Dwivedi, B. Kumar, and S. Kumar, "A mathematical programming approach in optimum stratification under neyman allocation for two stratifying variables," *Journal of Reliability and Statistical Studies*, pp. 173–185, 2019.

[19] S. S. A. Alshqaq, A. A. H. Ahmadini, and I. Ali, "Nonlinear stochastic multiobjective optimization problem in multivariate stratified sampling design," *Mathematical Problems in Engineering*, vol. 2022, Article ID 2502346, 16 pages, 2022.

[20] R. Hamid, E. A. Ismail, S. M. Ezzat, and F. S. Abo El Hassan, "Multi-objective mathematical programming model for optimum stratification in multivariate stratified sampling," *Computers and Fluids*, vol. 41, no. 4, pp. 189–208, 2021.

[21] J. Brito, T. Veiga, and P. Silva, "An optimisation algorithm applied to the one-dimensional stratification problem," *Survey Methodology*, vol. 45, no. 2, pp. 295–315, 2019.

[22] A. C. Fadel, L. S. Ochi, J. A. d. M. Brito, and G. S. Semaan, "Micro aggregation heuristic applied to statistical disclosure control," *Information Sciences*, vol. 548, pp. 37–55, 2021.

[23] J. A. de Moura Brito, G. S. Semaan, A. C. Fadel, and L. R. Brito, "An optimization approach applied to the optimal stratification problem," *Communications in Statistics - Simulation and Computation*, vol. 46, no. 6, pp. 4419–4451, 2017.

[24] J. Lisic, H. Sang, Z. Zhu, and S. Zimmer, "Optimal stratification and allocation for the june agricultural survey," *Journal of Official Statistics*, vol. 34, no. 1, pp. 121–148, 2018.

[25] S. Rizvi and F. Danish, "Approximately optimum strata boundaries under super population model," *International Journal of Mathematics in Operational Research*, vol. 1, no. 1, p. 1, 2022.

[26] E. H. Rizvi, "Optimum Stratification for Two Study Variables Using Auxiliary Information," (Unpublished Doctoral Thesis), Punjab Agricultural University-PAU, Punjab, India, 1997.

[27] 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.

[28] M. G. M. Khan, N. Nand, and N. Ahmad, "Determining the optimum strata boundary points using dynamic programming," *Survey Methodology*, vol. 34, no. 2, pp. 205–214, 2008.

[29] T. Dalenius, J. L. Hodges, and J. L. Hodges, "Minimum variance stratification," *Journal of the American Statistical Association*, vol. 54, no. 285, pp. 88–101, 1959.

[30] P. Gunning and J. M. Horgan, "A new algorithm for the construction of stratum boundaries in skewed populations," *Survey Methodology*, vol. 30, no. 2, pp. 159–166, 2004.

[31] P. Lavallée and M. Hidiroglou, "On the stratification of skewed populations," *Survey Methodology*, vol. 14, pp. 33–43, 1988.

[32] M. Kozak, "Optimal stratification using random search method in agricultural surveys," *Stat. Transition*, vol. 6, no. 5, pp. 797–806, 2004.

[33] M. G. M. Khan, D. Rao, A. H. Ansari, and M. J. Ahsan, "Determining optimum strata boundaries and sample sizes

for skewed population with log-normal distribution," *Communications in Statistics - Simulation and Computation*, vol. 44, no. 5, pp. 1364–1387, 2015.

[34] K. G. Reddy, M. G. M. Khan, and S. Khan, "Optimum strata boundaries and sample sizes in health surveys using auxiliary variables," *PLoS One*, vol. 13, no. 4, Article ID e0194787, 2018.

[35] C. R. Martínez, A. H. German, A. M. Marvelio, and S. Florentin, "Neutrosophy for survey analysis in social sciences," *Neutrosophic Sets and Systems*, vol. 37, no. 1, 2020.

[36] L. E. V. Cruzaty, M. R. Tomalá, and C. M. C. Gallo, "A neutrosophic statistic method to predict tax time series in Ecuador," *Neutrosophic Sets and Systems*, vol. 34, pp. 33–39, 2020.

[37] F. Danish, "Construction of stratification points under optimum allocation using dynamic programming," *Pakistan Journal of Statistics and Operation Research*, vol. 15, no. 2, pp. 341–355, 2019.