scientific reports

Check for updates

OPEN

Efficient class of estimators for finite population mean using auxiliary attribute in stratified random sampling

Housila P. Singh[1], Anurag Gupta[2]✉ & Rajesh Tailor[1]

The aim of this paper is to develop more effective methods for estimating population means in sample surveys using auxiliary attributes. To achieve this goal, we introduce a modified version of the estimators proposed by Koyuncu (2013b) and Shahzad et al. (2019), as well as a new class of estimators. We derive expressions for the bias and mean squared error of these new estimators up to the first degree of approximation. Our results show that the suggested classes of estimators perform better than other existing methods, with the lowest mean squared error under optimal conditions. We also conduct an empirical investigation to support our findings.

The use of auxiliary attribute is a well-known method for improving the efficiency of an estimator in estimating population parameters. Auxiliary attributes (say ϕ), which are highly correlated with the study variable (y), are commonly encountered in practice. Examples include a person's height (y), the amount of milk produced by a cow (y), and the yield of a particular variety of wheat (y), which may respectively depend on factors such as gender, breed of the cow, and type of wheat. For more examples, see Kendale and Stuart[1], Shabbir and Gupta[2], and Sharma and Singh[3], among others. The estimation of the population mean of the study variable (y) using an auxiliary attribute (ϕ) under simple random sampling without replacement has been extensively studied. See, for instance, Naik and Gupta[4], Jhajj et al.[5], Solanki and Singh[6], Singh et al.[7], Gupta and Tailor[8], and other relevant literature.

The simple random sampling scheme is commonly used when the population units are homogeneous. However, in many practical situations, the population units are heterogeneous, and to obtain a better estimate of the population parameters, we use stratified random sampling. Therefore, our objective is to estimate the population mean of the study variable (y) using information on an auxiliary attribute (ϕ) under stratified random sampling. Various authors, including Sharma and Singh[3], Koyuncu[9,10], Shahzad et al.[11], Zaman[12,13], Hussain et al.[14,15], Zaman et al.[16,17], and Ahmad et al.[18,19], have discussed the problem of estimating the population mean of the study variable using an auxiliary attribute under stratified random sampling.

Shahzad et al.[20,21] used a calibration approach in stratified random sampling. They also proposed an estimator to estimate the coefficient of variation using calibrated estimators in stratified random sampling Shahzad et al.[22].

Notations. Consider a population size N unit, is divided into L strata units hth stratum containing N_h units, where $h = 1, 2, \ldots, L$ such that $\sum_{h=1}^{L} N_h = N$. A simple random sample of size n_h is drawn without replacement the hth stratum such that $\sum_{h=1}^{L} n_h = n$. Let (y_{hi}, ϕ_{hi}) be observed value of study variable y and the auxiliary attribute ϕ on the ith unit of the hth stratum, respectively, where $i = 1, 2, ..., N_h$ and $h = 1, 2, ..., L$.

Further, let $\overline{y}_h = \sum_{h=1}^{n_h} \frac{y_{hi}}{n_h}$ and $\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$ be unbiased estimators of population means $\overline{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}$ and $\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$, where $W_h = \frac{N_h}{N}$ is the stratum weight.

We also assume that

$$\phi_{hi} = \begin{cases} 1, & i^{\text{th}} \text{ unit of the } h^{\text{th}} \text{ stratum possesses the attribute } \phi, \\ 0, & \text{otherwise.} \end{cases}$$

[1]School of Studies in Statistics, Vikram University, Ujjain, M.P. 456010, India. [2]Indian Agricultural Statistics Research Institute, ICAR, New Delhi 110012, India. ✉email: gupta.glg@gmail.com

Let $s_{yh}^2=\sum_{i=1}^{n_h}\frac{(y_{hi}-\overline{y}_h)^2}{(n_h-1)}$ and $s_{\phi h}^2=\sum_{i=1}^{n_h}\frac{(\phi_{hi}-p_h)^2}{(n_h-1)}$ be the hth sample variances and $S_{yh}^2=\sum_{i=1}^{N_h}\frac{(y_{hi}-\overline{Y}_h)^2}{(N_h-1)}$ and $S_{\phi h}^2=\sum_{i=1}^{N_h}\frac{(\phi_{hi}-P_h)^2}{(N_h-1)}$ be the hth population variances of the study variable y and the auxiliary attribute ϕ, respectively.

Further, $s_{y\phi h}=\sum_{i=1}^{n_h}\frac{(y_{hi}-\overline{y}_h)(\phi_{hi}-p_h)}{(n_h-1)}$ and $\widehat{\rho}_{y\phi h}=\frac{s_{y\phi h}}{s_{yh}s_{\phi h}}$ are the hth sample covariance and point bi-serial correlation and $S_{y\phi h}=\sum_{i=1}^{N_h}\frac{(y_{hi}-\overline{Y}_h)(\phi_{hi}-P_h)}{(N_h-1)}$ and $\widehat{\rho}_{y\phi h}=\frac{S_{y\phi h}}{S_{yh}S_{\phi h}}$ are population covariance and point bi-serial correlation between the study variable y and the auxiliary attribute ϕ, respectively.

To derive the bias and mean squared error (MSE) of the estimators, we write

$$\overline{y}_{st}=\overline{Y}(1+e_o),\ p_{st}=P(1+e_1)$$

such that $E(e_o)=E(e_1)=0$ and

$$E\left(e_0^2\right)=\frac{1}{\overline{Y}^2}\sum_{h=1}^{L}W_h^2\gamma_h S_{yh}^2=A_y,$$

$$E\left(e_1^2\right)=\frac{1}{P^2}\sum_{h=1}^{L}W_h^2\gamma_h S_{\phi h}^2=A_\phi,$$

$$E(e_0e_1)=\frac{1}{\overline{Y}P}\sum_{h=1}^{L}W_h^2\gamma_h S_{y\phi h}=A_{y\phi},$$

where $p_{st}=\sum_{h=1}^{L}W_hp_h$ such that $E\left(p_{st}\right)=P=\sum_{h=1}^{L}W_hP_h$ and $\gamma_h=\frac{N_h-n_h}{n_hN_h}$.

Reviewing some existing estimators. The conventional unbiased estimators for population mean $\overline{Y}$ of the study variable y under stratified random sampling is given by

$$t_0=\overline{y}_{st}=\sum_{h=1}^{L}W_h\overline{y}_h \tag{1}$$

The variance/MSE of the estimator t_0 is given by

$$MSE(t_0)=\sum_{h=1}^{L}W_h^2\gamma_h S_{yh}^2=\overline{Y}^2A_y. \tag{2}$$

The ratio estimator for population mean $\overline{Y}$ using auxiliary attribute ϕ in stratified random sampling due to Naik and Gupta[4] is given by

$$t_1=\overline{y}_{st}\left(\frac{P}{p_{st}}\right). \tag{3}$$

The MSE of the estimator t_1 up to the first degree of approximation (fda), is given by

$$MSE(t_1)=\overline{Y}^2\left[A_y+A_\phi(1-2C)\right], \tag{4}$$

where $C=\frac{A_{y\phi}}{A_\phi}$.

The stratified version of ordinary product estimator for population mean $\overline{Y}$ is defined by

$$t_2=\overline{y}_{st}\left(\frac{p_{st}}{P}\right). \tag{5}$$

The MSE of t_2 to the fda, is given by

$$MSE(t_2)=\overline{Y}^2\left[A_y+A_\phi(1+2C)\right]. \tag{6}$$

The usual regression estimator for $\overline{Y}$ is given by

$$t_3=\overline{y}_{st}+b_{st}\left(P-p_{st}\right), \tag{7}$$

where b_{st} is the sample regression coefficient of y on ϕ.

To the fda the MSE of t_3 is given by

$$MSE(t_3)=\overline{Y}^2A_y\left(1-\rho_{y\phi}^2\right), \tag{8}$$

where $\rho_{y\phi} = \frac{A_{y\phi}}{\sqrt{A_y A_\phi}}$.

Koyuncu[9] suggested a class of estimators for $\overline{Y}$ is given by

$$t_4 = \left[w_1\overline{y}_{st} + w_2(P - p_{st})\right]\left\{\frac{a_{st}P + b_{st}}{a_{st}p_{st} + b_{st}}\right\}, \tag{9}$$

where (w_1, w_2) are suitable constants, $a_{st}(\neq 0)$ and b_{st} are either real numbers or the functions of the known parameters for the hth stratum of the auxiliary attribute ϕ, such as standard deviation $S_{\phi(st)} = \sum_{h=1}^{L} W_h\sigma_{\phi h}$, coefficient of variation $C_{\phi(st)} = \sum_{h=1}^{L} W_h C_{\phi h}$ with $C_{\phi h} = \frac{S_{\phi h}}{P_h}$, skewness $\beta_{1(\phi)st} = \sum_{h=1}^{L} W_h\beta_{1h}(\phi)$ and kurtosis $\beta_{2(\phi)st} = \sum_{h=1}^{L} W_h\beta_{2h}(\phi)$ and correlation coefficient $\rho_{(y\phi)st} = \sum_{h=1}^{L} W_h\rho_{y\phi h}$ where $\beta_{1h(\phi)} = \frac{\mu_{3h}^2(\phi)}{\mu_{2h}^3(\phi)}$, $\beta_{2h(\phi)} = \frac{\mu_{4h}(\phi)}{\mu_{2h}^2(\phi)}$, $\sigma_{\phi h}^2 = \mu_{2h}(\phi) = \frac{1}{N_h}\sum_{i=1}^{N_h}(\phi_{hi} - P_h)^2$, $\mu_{3h}(\phi) = \frac{1}{N_h}\sum_{i=1}^{N_h}(\phi_{hi} - P_h)^3$ and $\mu_{4h}(\phi) = \frac{1}{N_h}\sum_{i=1}^{N_h}(\phi_{hi} - P_h)^4$.

The *MSE* of the estimator t_4 to the *fda* is given by

$$MSE(t_4) = \overline{Y}^2\left[1 + w_1^2 B_1 + w_2^2 B_2 + 2w_1 w_2 B_3 - 2w_1 B_4 - 2w_2 B_5\right], \tag{10}$$

where $B_1 = \left[1 + A_y + \upsilon_{st}A_\phi(3\upsilon_{st} - 4C)\right]$, $B_2 = \frac{A_\phi}{R^2}$, $B_3 = \left(\frac{A_\phi}{R}\right)(2\upsilon_{st} - C)$, $B_4 = \left[1 + \upsilon_{st}A_\phi(\upsilon_{st} - C)\right]$, $B_5 = \left(\frac{A_\phi}{R}\right)\upsilon_{st}$, $R = \frac{\overline{Y}}{P}$ and $\upsilon_{st} = \frac{a_{st}P}{(a_{st}P + b_{st})}$.

The *MSE* t_4 at (10) is minimized for

$$\left.\begin{array}{l} w_1 = \frac{(B_2B_4 - B_3B_5)}{(B_1B_2 - B_3^2)} \\ w_2 = \frac{(B_1B_5 - B_3B_4)}{(B_1B_2 - B_3^2)} \end{array}\right\}. \tag{11}$$

Therefore, the resulting minimum *MSE* of t_4 is given by

$$\begin{aligned} MSE_{\min}(t_4) &= \overline{Y}^2\left[1 - \frac{\left(B_2B_4^2 - 2B_3B_4B_5 + B_1B_5^2\right)}{\left(B_1B_2 - B_3^2\right)}\right] \\ &= \overline{Y}^2\left[1 - \frac{\left(A_\phi - \upsilon_{st}^2 A_\phi A_{y\phi}^2 - \upsilon_{st}^2 A_\phi^2 + \upsilon_{st}^2 A_\phi^2 A_y\right)}{\left(A_\phi + A_yA_\phi - \upsilon_{st}^2 A_\phi^2 - A_{y\phi}^2\right)}\right]. \end{aligned} \tag{12}$$

Using information on auxiliary attribute ϕ, Sharma and Singh[3] proposed the following exponential type estimators for $\overline{Y}$ as

$$t_{1e} = \overline{y}_{st}exp\left(\frac{P - p_{st}}{P + p_{st}}\right), \text{ (ratio type exponential estimator)} \tag{13}$$

$$t_{2e} = \overline{y}_{st}exp\left(\frac{p_{st} - P}{p_{st} + P}\right), \text{ (product-type exponential estimator)} \tag{14}$$

$$t_{\alpha e} = \overline{y}_{st}exp\left\{\frac{\alpha(P - p_{st})}{(P + p_{st})}\right\}, \tag{15}$$

where α being a suitable chosen constant.

To the *fda* the *MSE*s of the estimators t_{1e}, t_{2e} and $t_{\alpha e}$, are respectively given by

$$MSE(t_{1e}) = \overline{Y}^2\left[A_y + \frac{A_\phi}{4}(1 - 4C)\right], \tag{16}$$

$$MSE(t_{2e}) = \overline{Y}^2\left[A_y + \frac{A_\phi}{4}(1 + 4C)\right], \tag{17}$$

$$MSE(t_{\alpha e}) = \overline{Y}^2\left[A_y + \frac{\alpha A_\phi}{4}(\alpha - 4C)\right]. \tag{18}$$

The $MSE(t_{\alpha e})$ is minimum when

$$\alpha = 2C. \tag{19}$$

This yields the minimum *MSE* of $t_{\alpha e}$ is given by

$$MSE_{\min}(t_{\alpha e}) = \overline{Y}^2 A_y\left(1-\rho_{y\phi}^2\right). \tag{20}$$

Sharma and Singh[3] proposed the following class of estimators for population mean $\overline{Y}$ as

$$t_5 = \left[w_1\overline{y}_{st} + w_2(P-p_{st})\right] exp\left\{\frac{a_{st}(P-p_{st})}{a_{st}(P+p_{st})+2b_{st}}\right\}, \tag{21}$$

where (a_{st}, b_{st}) are same as defined for the class of estimators t_4 at (9) and (w_1, w_2) are suitable chosen constants to be determined such that MSE of t_5 is minimum.

If we set $a_{st} = 1$ and $b_{st} = NP$ in (21), then the class of estimators t_5 reduces to the Shahzad et al.[11] class of estimators for $\overline{Y}$ as

$$t_6 = \left[w_1\overline{y}_{st} + w_2(P-p_{st})\right] exp\left\{\frac{(P-p_{st})}{P+p_{st}+2NP}\right\}. \tag{22}$$

We note that the expressions of bias and MSE of the class of estimators t_5 derived by Sharma and Singh[3] [Eqs. (4.6) and (4.7), p. 1789] are not correct. The correct expressions of bias and MSE of the estimator t_5 to the *fda* are respectively given by

$$B(t_5) = \overline{Y}(w_1A_4 + w_2A_5 - 1), \tag{23}$$

and

$$MSE(t_5) = \overline{Y}^2\left[1 + w_1^2A_1 + w_2^2A_2 + 2w_1w_2A_3 - 2w_1A_4 - 2w_2A_5\right], \tag{24}$$

where $A_1 = \left[1 + A_y + \upsilon_{st}A_\phi(\upsilon_{st} - 2C)\right]$, $A_2 = \frac{A_\phi}{R^2}$, $A_3 = \left(\frac{1}{R}\right)A_\phi(\upsilon_{st} - C)$, $A_4 = \left[1 + \frac{\upsilon_{st}A_\phi}{8}(3\upsilon_{st} - 4C)\right]$, $A_5 = \frac{\upsilon_{st}A_\phi}{2R}$.

The MSE t_5 at (24) is minimized for

$$\left.\begin{aligned} w_1 &= \frac{A_2A_4 - A_3A_5}{A_1A_2 - A_3^2} \\ w_2 &= \frac{A_1A_5 - A_3A_4}{A_1A_2 - A_3^2} \end{aligned}\right\}. \tag{25}$$

Therefore, the minimum MSE of t_5 is given by

$$MSE_{\min}(t_5) = \overline{Y}^2\left[1 - \frac{A_2A_4^2 - 2A_3A_4A_5 + A_1A_5^2}{A_1A_2 - A_3^2}\right] \tag{26}$$

Putting $a_{st} = 1$ and $b_{st} = NP$ in (24), we get the MSE of t_6 to the *fda* is given by

$$MSE(t_6) = \overline{Y}^2\left[1 + w_1^2A_{1(1)} + w_2^2A_{2(1)} + 2w_1w_2A_{3(1)} - 2w_1A_{4(1)} - 2w_2A_{5(1)}\right], \tag{27}$$

where $A_{1(1)} = \left[1 + A_y + \upsilon_{st(1)}A_\phi\left(\upsilon_{st(1)} - 2C\right)\right]$, $A_{2(1)} = \frac{A_\phi}{R^2}$, $A_{3(1)} = \left(\frac{1}{R}\right)A_\phi\left(\upsilon_{st(1)} - C\right)$, $A_{4(1)} = \left[1 + \frac{\upsilon_{st(1)}A_\phi}{8}\left(3\upsilon_{st(1)} - 4C\right)\right]$, $A_{5(1)} = \frac{\upsilon_{st(1)}A_\phi}{2R}$, $\upsilon_{st(1)} = \frac{1}{(N+1)}$.

The MSE t_6 at (27) is minimum when

$$\left.\begin{aligned} w_1 &= \frac{A_{2(1)}A_{4(1)} - A_{3(1)}A_{5(1)}}{A_{1(1)}A_{2(1)} - A_{3(1)}^2} \\ w_2 &= \frac{A_{1(1)}A_{5(1)} - A_{3(1)}A_{4(1)}}{A_{1(1)}A_{2(1)} - A_{3(1)}^2} \end{aligned}\right\}. \tag{28}$$

Therefore, the minimum MSE of t_6 is given by

$$MSE_{\min}(t_6) = \overline{Y}^2\left[1 - \frac{A_{2(1)}A_{4(1)}^2 - 2A_{3(1)}A_{4(1)}A_{5(1)} + A_{1(1)}A_{5(1)}^2}{A_{1(1)}A_{2(1)} - A_{3(1)}^2}\right]. \tag{29}$$

Koyuncu[10] and Shahzad et al.[11] proposed the following class of estimators for $\overline{Y}$ as

$$t_7 = \left[w_1\overline{y}_{st} + w_2\left(\frac{p_{st}}{P}\right)^\gamma\right] exp\left\{\frac{a_{st}(P-p_{st})}{a_{st}(P+p_{st})+2b_{st}}\right\}, \tag{30}$$

where (w_1, w_2, γ) are suitable chosen constants and (a_{st}, b_{st}) are same as defined earlier.

To the *fda*, the MSE of t_7 is given by

$$MSE(t_7) = \overline{Y}^2\left[1 + w_1^2C_1 + w_2^2C_2 + 2w_1w_2C_3 - 2w_1C_4 - 2w_2C_5\right], \tag{31}$$

where $C_1 = \left[1 + A_y + \upsilon_{st} A_\phi (\upsilon_{st} - 2C)\right]$, $C_2 = \frac{1}{P^2 R^2}\left[1 + A_\phi\left\{\gamma^2 + \upsilon_{st}^2 - 2\gamma\upsilon_{st} + \gamma(\gamma - 1)\right\}\right]$, $C_3 = \left(\frac{1}{PR}\right)[1 + A_\phi\left\{\left(\upsilon_{st}^2 + \frac{\gamma(\gamma-1)}{2} - \upsilon_{st}\gamma\right) + (\gamma - \upsilon_{st})C\right\}]$, $C_4 = \left[1 + \frac{\upsilon_{st} A_\phi}{8}(3\upsilon_{st} - 4C)\right]$, $C_5 = \frac{1}{PR}\left[1 + \left\{\frac{\gamma(\gamma-1)}{2} - \frac{\gamma\upsilon_{st}}{2} + \frac{3}{8}\upsilon_{st}^2\right\}A_\phi\right]$.

$MSE(t_7)$ at (31) is minimized for

$$\left.\begin{aligned} w_1 &= \frac{C_2C_4 - C_3C_5}{C_1C_2 - C_3^2} \\ w_2 &= \frac{C_1C_5 - C_3C_4}{C_1C_2 - C_3^2} \end{aligned}\right\}. \tag{32}$$

Therefore, the minimum MSE of t_7 is given by

$$MSE_{\min}(t_7) = \overline{Y}^2\left[1 - \frac{C_2C_4^2 - 2C_3C_4C_5 + C_1C_5^2}{C_1C_2 - C_3^2}\right] \tag{33}$$

In this paper we have suggested a class of estimators for population mean $\overline{Y}$ of the study variable y using auxiliary attribute ϕ. Expressions of bias and MSE of the proposed class of estimators are obtained up to terms of order 0 (n^{-1}).

We have obtained the optimum condition under which the MSE of the proposed class of estimators is minimum. We have derived the conditions under which the suggested class of estimators is more efficient than the conventional estimator and the estimators due to Naik and Gupta[4], Koyuncu[9], Sharma and Singh[3] and Shahzad et al.[11]. Numerical illustration is given in support of the proposed study.

Suggested class of estimators

We note that the exponent part of (30) is obtained on using the transformation $\left(a_{st}p_{st} + b_{st}\right)$ such that $E\left\{a_{st}p_{st} + b_{st}\right\} = (a_{st}P + b_{st})$, in the first bracket coefficient of w_2 is $\left(\frac{p_{st}}{P}\right)^\gamma$ which does not use the transformation $\left(a_{st}p_{st} + b_{st}\right)$. Thus, authors are in opinion that coefficient of w_2 should be $\left(\frac{a_{st}p_{st}+b_{st}}{a_{st}P+b_{st}}\right)^\gamma$. Hence the modified suggested class of estimators for $\overline{Y}$ is given by

$$t_{7(m)} = \left[w_1\overline{y}_{st} + w_2\left(\frac{a_{st}p_{st} + b_{st}}{a_{st}P + b_{st}}\right)^\gamma\right]\exp\left\{\frac{a_{st}\left(P - p_{st}\right)}{a_{st}\left(P + p_{st}\right) + 2b_{st}}\right\}. \tag{34}$$

where (w_1, w_2) are suitably chosen constants to be determined such that MSE of $t_{7(m)}$ is minimum; and (a_{st}, b_{st}, γ) are same as defined earlier.

To the *fda*, the bias and MSE of $t_{7(m)}$ are respectively given by

$$B\left(t_{7(m)}\right) = \overline{Y}(w_1D_4 + w_2D_5 - 1), \tag{35}$$

and

$$MSE\left(t_{7(m)}\right) = \overline{Y}^2\left[1 + w_1^2D_1 + w_2^2D_2 + 2w_1w_2D_3 - 2w_1D_4 - 2w_2D_5\right], \tag{36}$$

where $D_1 = \left[1 + A_y + \upsilon_{st}A_\phi(\upsilon_{st} - 2C)\right]$, $D_2 = \frac{1}{R^2P^2}\left[1 + \upsilon_{st}^2\theta(2\theta - 1)A_\phi\right]$, $D_3 = \left(\frac{1}{RP}\right)\left[1 + \frac{\upsilon_{st}(2\theta-1)}{8}A_\phi (2\theta + 4C - 3)\right]$, $D_4 = \left[1 + \frac{\upsilon_{st}}{8}A_\phi\left(3\upsilon_{st} - 4C\right)\right]$, $D_5 = \frac{1}{RP}\left[1 + \frac{\upsilon_{st}^2\theta(\theta-1)}{2}A_\phi\right]$, $\theta = \frac{(2\gamma-1)}{2}$.

The $MSE\left(t_{7(m)}\right)$ at (36) is minimized for

$$\left.\begin{aligned} w_1 &= \frac{(D_2D_4 - D_3D_5)}{\left(D_1D_2 - D_3^2\right)} \\ w_2 &= \frac{(D_1D_5 - D_3D_4)}{\left(D_1D_2 - D_3^2\right)} \end{aligned}\right\}. \tag{37}$$

Therefore, the minimum MSE of $t_{7(m)}$ is given by

$$MSE_{\min}\left(t_{7(m)}\right) = \overline{Y}^2\left[1 - \frac{\left(D_2D_4^2 - 2D_3D_4D_5 + D_1D_5^2\right)}{\left(D_1D_2 - D_3^2\right)}\right] \tag{38}$$

Now we can conclude this as a theorem given below.

Theorem 2.1 *The MSE of $t_{7(m)}$ is greater than or equal to the minimum MSE of $t_{7(m)}$.*

$$MSE\left(t_{7(m)}\right) \geq MSE_{\min}\left(t_{7(m)}\right)$$

An alternative class of estimators

We propose another class of estimators for population mean as $\overline{Y}$ as

$$t_8 = \left[w_1 \overline{y}_{st} + w_2 \exp\left\{ \frac{\delta a_{st}(P - p_{st})}{a_{st}(P + p_{st}) + 2b_{st}} \right\} \right] \left(\frac{a_{st}P + b_{st}}{a_{st}p_{st} + b_{st}} \right)^{\eta}. \tag{39}$$

where $(w_1, w_2, a_{st}, b_{st})$ are same as defined earlier and (δ, η) are constants which take real numbers like (– 1,0,1).

To the *fda*, the bias and *MSE* of t_8 are respectively given by

$$B(t_8) = \overline{Y}(w_1 E_4 + w_2 E_5 - 1), \tag{40}$$

and

$$MSE(t_8) = \overline{Y}^2\left[1 + w_1^2 E_1 + w_2^2 E_2 + 2w_1 w_2 E_3 - 2w_1 E_4 - 2w_2 E_5\right], \tag{41}$$

where $E_1 = \left[1 + A_y - 4\eta \upsilon_{st} A_{y\phi} + \eta(2\eta + 1)\upsilon_{st}^2 A_\phi\right]$, $E_2 = \frac{1}{R^2P^2}\left[1 + \theta(2\theta + 1)\upsilon_{st}^2 A_\phi\right]$, $E_3 = \left(\frac{1}{RP}\right)\left[1 + \frac{(\eta+\theta)(\eta+\theta+1)}{2}\upsilon_{st}^2 A_\phi - (\eta + \theta)\upsilon_{st} A_{y\phi}\right]$, $E_4 = \left[1 + \frac{\eta \upsilon_{st}}{2}\left\{\frac{(\eta+1)}{2}\upsilon_{st} A_\phi - 2A_{y\phi}\right\}\right]$, $E_5 = \frac{1}{RP}\left[1 + \frac{\theta(\theta+1)}{2}\upsilon_{st}^2 A_\phi\right]$.

The $MSE(t_8)$ at (41) is minimized for

$$\left.\begin{aligned} w_1 &= \frac{(E_2E_4 - E_3E_5)}{(E_1E_2 - E_3^2)} \\ w_2 &= \frac{(E_1E_5 - E_3E_4)}{(E_1E_2 - E_3^2)} \end{aligned}\right\}. \tag{42}$$

Substitution of (42) in (41) provides the minimum *MSE* of t_8 is given by

$$MSE_{\min}(t_8) = \overline{Y}^2\left[1 - \frac{\left(E_2E_4^2 - 2E_3E_4E_5 + E_1E_5^2\right)}{\left(E_1E_2 - E_3^2\right)}\right] \tag{43}$$

Now we have the following theorem.

Theorem 3.1 *The MSE of t_8 is greater than or equal to the minimum MSE of t_8.*

$$MSE(t_8) \geq MSE_{\min}(t_8)$$

Efficiency comparison

From (2), (4), (6), (8), (16) and (17) we have

$$MSE\left(t_0 = \overline{y}_{st}\right) - MSE(t_3) = \overline{Y}^2 A_y \rho_{y\phi}^2 \geq 0, \tag{44}$$

$$MSE(t_1) - MSE(t_3) = \overline{Y}^2 A_\phi (1 - C)^2 \geq 0, \tag{45}$$

$$MSE(t_2) - MSE(t_3) = \overline{Y}^2 A_\phi (1 + C)^2 \geq 0, \tag{46}$$

$$MSE(t_{1e}) - MSE(t_3) = \overline{Y}^2 \frac{A_\phi}{4}(1 - 2C)^2 \geq 0, \tag{47}$$

$$MSE(t_{2e}) - MSE(t_3) = \overline{Y}^2 \frac{A_\phi}{4}(1 + 2C)^2 \geq 0, \tag{48}$$

It follows from (44) to (46) that the regression estimator t_3 is more efficient than $\overline{y}_{st}$, t_1, t_2, t_{1e} and t_{2e}.

From (8), (12), (22), (26), (29), (33), (38) and (43) we have

$$MSE(t_3) - MSE_{\min}(t_4) = \overline{Y}^2\left[A_y\left(1 - \rho_{y\phi}^2\right) + \frac{\left(B_2B_4^2 - 2B_3B_4B_5 + B_1B_5^2\right)}{\left(B_1B_2 - B_3^2\right)} - 1\right] \geq 0 \tag{49}$$

$$MSE(t_3) - MSE_{\min}(t_5) = \overline{Y}^2\left[A_y\left(1 - \rho_{y\phi}^2\right) + \frac{\left(A_2A_4^2 - 2A_3A_4A_5 + A_1A_5^2\right)}{\left(A_1A_2 - A_3^2\right)} - 1\right] \geq 0 \tag{50}$$

$$MSE(t_3) - MSE_{\min}(t_6) = \overline{Y}^2\left[A_y\left(1 - \rho_{y\phi}^2\right) + \frac{\left(A_{2(1)}A_{4(1)}^2 - 2A_{3(1)}A_{4(1)}A_{5(1)} + A_{1(1)}A_{5(1)}^2\right)}{\left(A_{1(1)}A_{2(1)} - A_{3(1)}^2\right)} - 1\right] \geq 0 \tag{51}$$

$$MSE(t_3) - MSE_{\min}(t_7) = \overline{Y}^2\left[A_y\left(1-\rho_{y\phi}^2\right) + \frac{\left(C_2C_4^2 - 2C_3C_4C_5 + C_1C_5^2\right)}{\left(C_1C_2 - C_3^2\right)} - 1\right] \geq 0 \quad (52)$$

$$MSE(t_3) - MSE_{\min}\left(t_{7(m)}\right) = \overline{Y}^2\left[A_y\left(1-\rho_{y\phi}^2\right) + \frac{\left(D_2D_4^2 - 2D_3D_4D_5 + D_1D_5^2\right)}{\left(D_1D_2 - D_3^2\right)} - 1\right] \geq 0 \quad (53)$$

$$MSE(t_3) - MSE_{\min}(t_8) = \overline{Y}^2\left[A_y\left(1-\rho_{y\phi}^2\right) + \frac{\left(E_2E_4^2 - 2E_3E_4E_5 + E_1E_5^2\right)}{\left(E_1E_2 - E_3^2\right)} - 1\right] \geq 0 \quad (54)$$

It follows from (49) to (54) that the classes of estimators $t_4, t_5, t_6, t_7, t_{7(m)}$ and t_8 are more efficient than the regression estimator t_3.

Further from (12), (26), (29), (33), (38) and (43) we have

$$MSE_{\min}(t_8) < MSE_{\min}(t_4) \text{ if } M_2 < M_1, \quad (55)$$

$$MSE_{\min}(t_8) < MSE_{\min}(t_5) \text{ if } M_3 < M_1, \quad (56)$$

$$MSE_{\min}(t_8) < MSE_{\min}(t_6) \text{ if } M_4 < M_1, \quad (57)$$

$$MSE_{\min}(t_8) < MSE_{\min}(t_7) \text{ if } M_5 < M_1, \quad (58)$$

$$MSE_{\min}(t_8) < MSE_{\min}\left(t_{7(m)}\right) \text{ if } M_6 < M_1, \quad (59)$$

where $M_1 = \frac{\left(E_2E_4^2-2E_3E_4E_5+E_1E_5^2\right)}{\left(E_1E_2-E_3^2\right)}$, $M_2 = \frac{\left(A_y - \upsilon_{st}^2A_\phi A_{y\phi}^2 - \upsilon_{st}^2A_\phi^2 + \upsilon_{st}^2A_\phi^2A_y\right)}{\left(A_\phi + A_yA_\phi - \upsilon_{st}^2A_\phi^2 - A_{y\phi}^2\right)}$, $M_3 = \frac{\left(A_2A_4^2-2A_3A_4A_5+A_1A_5^2\right)}{\left(A_1A_2-A_3^2\right)}$, $M_4 = \frac{\left(A_{2(1)}A_{4(1)}^2-2A_{3(1)}A_{4(1)}A_{5(1)}+A_{1(1)}A_{5(1)}^2\right)}{\left(A_{1(1)}A_{2(1)}-A_{3(1)}^2\right)}$, $M_5 = \frac{\left(C_2C_4^2-2C_3C_4C_5+C_1C_5^2\right)}{\left(C_1C_2-C_3^2\right)}$, $M_6 = \frac{\left(D_2D_4^2-2D_3D_4D_5+D_1D_5^2\right)}{\left(D_1D_2-D_3^2\right)}$.

Therefore we can say that the proposed class of estimators t_8 is more efficient than the estimators t_4, t_5, t_6, t_7 and $t_{7(m)}$ as long as the conditions (55), (56), (57), (58), and (59) respectively are satisfied.

Numerical illustration

To judge the merits of the suggested class of estimators t_8 over other existing estimators, we have computed the percent relative efficiency (PRE) of different estimators with respect to usual unbiased estimator $\overline{y}_{st}$ by using the following formulae:

$$PRE\left(t_1, \overline{y}_{st}\right) = \frac{A_y}{\left[A_y + A_\phi(1-2C)\right]} \times 100, \quad (60)$$

$$PRE\left(t_2, \overline{y}_{st}\right) = \frac{A_y}{\left[A_y + A_\phi(1+2C)\right]} \times 100, \quad (61)$$

$$PRE\left(t_3 \text{ or } t_{\alpha e}, \overline{y}_{st}\right) = \frac{1}{\left[1-\rho_{y\phi}^2\right]} \times 100, \quad (62)$$

$$PRE\left(t_4, \overline{y}_{st}\right) = \frac{A_y}{[1-M_2]} \times 100, \quad (63)$$

$$PRE\left(t_{1e}, \overline{y}_{st}\right) = \frac{A_y}{\left[A_y + \frac{A_\phi}{4}(1-4C)\right]} \times 100, \quad (64)$$

$$PRE\left(t_{2e}, \overline{y}_{st}\right) = \frac{A_y}{\left[A_y + \frac{A_\phi}{4}(1+4C)\right]} \times 100, \quad (65)$$

$$PRE\left(t_5, \overline{y}_{st}\right) = \frac{A_y}{[1-M_3]} \times 100, \quad (66)$$

$$PRE\left(t_6,\overline{y}_{st}\right)=\frac{A_y}{[1-M_4]}\times 100, \tag{67}$$

$$PRE\left(t_7,\overline{y}_{st}\right)=\frac{A_y}{[1-M_5]}\times 100, \tag{68}$$

$$PRE\left(t_{7(m)},\overline{y}_{st}\right)=\frac{A_y}{[1-M_6]}\times 100, \tag{69}$$

$$PRE\left(t_8,\overline{y}_{st}\right)=\frac{A_y}{[1-M_1]}\times 100, \tag{70}$$

To demonstrate the effectiveness of the proposed class of estimators t_8, we utilise data on the number of teachers as the study variable (y), and the number of students classified as more or less than 750 in both primary and secondary schools as the auxiliary attribute ϕ for 923 districts across six regions (as 1: Marmara, 2: Agean, 3: Mediterranean, 4: Central Anatolia, 5: Black Sea, 6: East and Southest Anatolia) in Turkey in 2007 (Source: The Turkish Republic Ministry of Education).

The summary statistics of the data are given in Table 1. We applied Neyman[23] allocation for allocating the samples to various strata[24]. Source: Koyuncu[9].

It is observed from Table 2 that the estimators t_1 and t_{1e} are more efficient than the usual unbiased estimator $\overline{y}$ (which does not utilize the auxiliary attribute). The product estimator t_2 and product-type exponential estimator $t_{2(e)}$ perform poor than $\overline{y}$ (due to positive correlation between y and φ). The t_3 is more efficient than t_1, t_{1e}, t_2 and t_{2e}. The performance of the estimators (t_4, t_5, t_6) are almost same but marginally better than estimators t_1, t_{1e}, t_2, t_{2e} and t_3.

Values	Stratum (h)					
	1	2	3	4	5	6
N_h	127	117	103	170	205	201
n_h	31	21	29	38	22	39
$\overline{Y}_h$	703.74	413.00	573.17	424.66	267.03	393.84
S_{yh}	883.83	644.92	1033.46	810.58	403.65	711.72
$S_{\phi h}$	0.213	0.159	0.25	0.31	0.28	0.21
P_h	0.952	0.97	0.93	0.88	0.91	0.95
$S_{\phi yh}$	25.26	9.98	37.45	44.62	21.04	18.66
$\beta_{2(\phi h)}$	16.92	35.57	10.34	4.23	6.67	15.56

Table 1. Data statistics.

Different values of scalars		Estimators										
a_{st}	b_{st}	t_1	$t_{1(e)}$	t_2	$t_{2(e)}$	t_3	t_4	t_5	t_6	$t_7(\gamma=-1)$	$t_{7(m)}(\gamma=-1)$	$t_8(\delta=-1, \eta=1)$
$C_{\phi(st)}$	NP	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	3366.42	1.71E+10	1.49E+11
1	NP	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	3363.75	1.21E+09	1.06E+10
$\beta_{2\phi(st)}$	NP	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	3318.90	6.68E+06	5.84E+07
$C_{\phi(st)}$	$\beta_{2\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	3308.06	4.44E+06	3.88E+07
1	$\beta_{2\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	3162.20	3.43E+05	3.02E+06
$C_{\phi(st)}$	1	101.82	101.75	92.28	96.75	102.00	103.17	103.17	103.17	2786.33	3.47E+04	3.10E+05
1	$C_{\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1747.40	2.33E+03	2.02E+04
$\beta_{2\phi(st)}$	1	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1573.86	1.70E+03	1.43E+04
$\beta_{2\phi(st)}$	$C_{\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1519.10	1.55E+03	1.29E+04
NP	$\beta_{2\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1515.26	1.54E+03	1.28E+04
NP	1	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1499.67	1.50E+03	1.24E+04
NP	$C_{\phi(st)}$	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1498.76	1.50E+03	1.24E+04
1	0	101.82	101.75	92.28	96.75	102.00	103.17	103.18	103.17	1498.43	1.50E+03	1.24E+04

Table 2. *PRE*s values of different estimators of $\overline{Y}$ with respect to $\overline{y}$.

Table 2 also demonstrates that the proposed estimator t_8 with $\left(\delta = -1, \eta = 1, a_{st} = C_{\phi(st)}, b_{st} = NP\right)$ has the largest PRE(= 1.49E+11) followed by $t_{7(m)}$ with $\left(\lambda = -1, a_{st} = C_{\phi(st)}, b_{st} = NP\right)$. It is further observed that the proposed classes of estimators $t_{7(m)}$ and t_8 are always better than the classes of estimators t_1[4], t_{1e}, t_2 and t_{2e}, t_3 (difference estimator), t_4[9], t_5[3], t_6[11], t_7[10] for all choices of (a_{st}, b_{st}). The proposed class of estimators t_8 is the best among the estimators closed in Table 2.

Thus, our recommendation is to use the suggested class of estimators $t_{7(m)}$ and t_8 in practice.

Conclusion

In this article, we propose two classes of estimators for estimating the population mean $\overline{Y}$ of the study variable y using information on an auxiliary attribute (ϕ). The suggested classes of estimators are wide-ranging. The biases and mean squared errors of the proposed classes of estimators $t_{7(m)}$ and t_8 are derived up to the first degree of approximation. The optimum estimators in the classes of estimators $t_{7(m)}$ and t_8 are investigated using the minimum mean squared error formulae. An empirical study is conducted to evaluate the efficiency of the proposed classes of estimators $t_{7(m)}$ and t_8 and the findings are presented in Table 2. The results of Table 2 demonstrate that the suggested classes of estimators $t_{7(m)}$ and t_8 are more efficient than the recently developed classes of estimators t_4, t_5,t_6, t_7 and t_3 by Koyuncu[9], Sharma and Singh[3], Shahzad et al.[11], Koyuncu[10], and the difference estimator, with a considerable gain in efficiency. Therefore, we conclude that the proposed classes of estimators and are justified and can be used in practice.

One potential direction for future research is the application of advanced statistical techniques, such as machine learning and artificial intelligence, can be explored to improve the accuracy and efficiency of the estimators. These techniques can also help in identifying relevant auxiliary variables for improving the estimation process. The impact of various sampling designs on the estimation process can be investigated. For example, the effect of unequal sample sizes in different strata, non-response rates, and measurement errors on the accuracy and efficiency of the estimators can be studied. Finally, the extension of the current research to other types of population parameters, such as variance and quantiles, can also be explored. This can lead to the development of new classes of estimators and further improve the accuracy and efficiency of the estimation process.

Data availability

All the necessary data generated and/or analysed during the current study are included in this published article.

Received: 21 December 2022; Accepted: 4 May 2023
Published online: 24 June 2023

References

1. Kendall, M. G. & Stuart, A. *The Advanced Theory of Statistics* 2nd edn. (Charles Griffin and Company Limited, 1967).
2. Shabbir, J. & Gupta, S. On estimating the finite population mean with known population proportion of an auxiliary variable. *Pak. J. Stat.* **23**(1), 1–9 (2007).
3. Sharma, P. & Singh, R. Efficient estimator of population mean in stratified random sampling using auxiliary attribute. *World Appl. Sci. J.* **27**(12), 1786–1791 (2013).
4. Naik, V. D. & Gupta, P. C. A note on estimation of mean with known population proportion of an auxiliary character. *J. Indian Soc. Agric. Stat.* **48**(2), 151–158 (1996).
5. Jhajj, H. S., Sharma, M. K. & Grover, L. K. A family of estimators of population mean using information on auxiliary attribute. *Pak. J. Stat.* **22**(1), 43–50 (2006).
6. Solanki, R. S. & Singh, H. P. Improved estimation of population mean using population proportion of an auxiliary character. *Chilean J. Stat.* **4**(1), 3–17 (2013).
7. Singh, H. P., Gupta, A. & Tailor, R. Estimation of population mean using a difference-type exponential imputation method. *J. Stat. Theory Pract.* **15**, 1–43 (2021).
8. Gupta, A. & Tailor, R. Ratio in ratio type exponential strategy for the estimation of population mean. *J. Reliability Stat. Stud.* 551–564 (2021).
9. Koyuncu, N. Improved estimation of population mean in stratified random sampling using information on auxiliary attribute. In *Proceeding of 59th ISI World Statistics Congress, Hong Kong, China, 25–30* (2013).
10. Koyuncu, N. Efficient combined estimators of population mean using auxiliary attribute under stratified random sampling. In *Proceeding of 11th International Conference of Numerical Analysis and Applied Mathematics*, vol. 1558, 1466–1469 (2013).
11. Shahzad, U., Hanif, M., Koyuncu, N. & Garcia, A. V. A family of ratio estimators in stratified random sampling utilizing auxiliary attribute alongside the nonresponse. *J. Stat. Theory Appl.* **18**(1), 12–25 (2019).
12. Zaman, T. Efficient estimators of population mean using auxiliary attribute in stratified random sampling. *Adv. Appl. Stat.* **56**(2), 153–171 (2019).
13. Zaman, T. An efficient exponential estimator of the mean under stratified random sampling. *Math. Popul. Stud.* **28**(2), 104–121 (2021).
14. Hussain, S., Ahmad, S., Saleem, M. & Akhtar, S. Finite population distribution function estimation with dual use of auxiliary information under simple and stratified random sampling. *PLoS ONE* **15**(9), e0239098 (2020).
15. Hussain, S., Akhtar, S. & El-Morshedy, M. Modified estimators of finite population distribution function based on dual use of auxiliary information under stratified random sampling. *Sci. Prog.* **105**(3), 00368504221128486 (2022).
16. Zaman, T. & Bulut, H. An efficient family of robust-type estimators for the population variance in simple and stratified random sampling. *Commun. Stat.-Theory Methods* (2021).
17. Zaman, T. & Kadilar, C. Exponential ratio and product type estimators of the mean in stratified two-phase sampling. *AIMS Math.* **6**(5), 4265–4279 (2021).
18. Ahmad, S. *et al.* Dual use of auxiliary information for estimating the finite population mean under the stratified random sampling scheme. *J. Math.* 1–12 (2021).
19. Ahmad, S. *et al.* Improved estimation of finite population variance using dual supplementary information under stratified random sampling. *Math. Probl. Eng.* **2022** 12. https://doi.org/10.1155/2022/3813952 (2022).

20. Shahzad, U., Ahmad, I., Almanjahie, I. M., Al-Noor, N. H., & Hanif, M. A novel family of variance estimators based on L-moments and calibration approach under stratified random sampling. *Commun. Stat.-Simul. Comput.* 1–14. https://doi.org/10.1080/03610918.2021.1945629 (2021).
21. Shahzad, U., Ahmad, I., Almanjahie, I. & Al-Noor, N. H. L-Moments based calibrated variance estimators using double stratified sampling. *Comput. Mater. Continua* **68**(3), 3411–3430 (2021).
22. Shahzad, U., Ahmad, I., García-Luengo, A. V., Zaman, T. & Al-Noor, N. H. Kumar, A estimation of coefficient of variation using calibrated estimators in double stratified random sampling. *Mathematics* **11**, 252 (2023).
23. Neyman, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selecting. *J. R. Stat. Soc.* **97**, 558–606 (1934).
24. Cochran, W. G. *Sampling Techniques* 3rd edn. (Wiley, 1977).

Acknowledgements
We would like to express our sincere gratitude to the anonymous reviewers and the editor for their valuable comments, feedback, and suggestions, which greatly improved the quality and impact of this manuscript. Their thoughtful and constructive criticisms and insights have been immensely helpful in shaping the final version of this paper. We appreciate their time, effort, and expertise in reviewing and editing this work. I also acknowledged to Dhanashree Bhure, Editorial Support at Scientific Reports for assisting and guiding us for the proper submission of the manuscript.

Author contributions
Idea of the estimator generation is of H.P.S. Theoretical study and comparison have been carried out by R.T. A.G. has carried out the empirical study of the estimator and drafted the whole research paper in article form. All authors read and approved the final study manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.G.