An improved generalized class of estimators for population variance using auxiliary variables

Mr. Nitesh K. Adichwal¹, Dr. Jitendra Kumar¹ and Prof. Rajesh Singh¹*

Abstract: This paper proposed an improved generalized class of estimator for estimating population variance using auxiliary variables based on simple random sampling without replacement. The expression of mean square error of the proposed estimator is obtained up to the first order of approximation. We have derived the conditions for the parameters under which the proposed estimator performs better compared to the usual estimator and other existing estimators. An empirical study and simulation study are also carried out with the support of theoretical results.

Subjects: Science; Mathematics & Statistics; Statistics & Probability

Keywords: population variance; auxiliary information; simple random sampling; mean square error; bias; simulation

1. Introduction

Through sampling literature it is well established that the use of Auxiliary information in Sampling Survey at the estimation stage improve the efficiency of the estimators of the population parameters. viz. population mean, variance, median, population correlation coefficient, etc. Ratio, regression and product method of estimation are good examples in this context. Several author including Ahmed, Raman, and Hossain (2000), Gupta and Shabbir (2008), Singh and Solanki (2009–2010, 2013), Yadav and Kadilar (2013), Solanki and Singh (2013) proposed estimators of population variance using information of auxiliary variable under different sampling schemes. Asghar, Sanaullah, and Hanif (2014) proposed generalized exponential-type estimators using information on auxiliary variable for estimating the population variance \( S_y^2 \). Singh, Chauhan, Sawan, and Smarandache (2009)

ABOUT THE AUTHOR

Mr. Nitesh K. Adichwal is a research fellow at the Banaras Hindu University in Varanasi. His research interest is in the area of sampling survey and demography. He has completed his MSc in Statistics from Purvanchal University, Jaunpur in 2010. He has also completed his MSc in Population Studies from International institute for population sciences, Mumbai in 2014. This paper has also published research papers in the area of sampling survey and demography in various national and international journals.

PUBLIC INTEREST STATEMENT

Variance estimation plays an important role in survey sampling. It is used for two purposes. One is the analytic purpose such as for constructing confidence intervals or performing hypothesis testing and another one is for descriptive purpose such as to evaluate the efficiency of the survey designs or to provide estimates for planning surveys. Several authors have paid their attention towards formation of estimators for the estimation of population variance using auxiliary information. Motivated by their works, we have proposed an improved generalized class of estimators for estimation of population variance using auxiliary information. The expression of mean square error is derived up to the first order of approximation. We have compared our proposed estimator with the other existing estimators. The efficiencies of the estimators are validated using real population data-sets and simulation study.
and Adichwal, Sharma, and Singh (2015) proposed estimator of population variance using two auxiliary variables and suggested to use two auxiliary variables, if they are made available. In this paper, we have also proposed an improved estimator for the population variance $S^2_y$ based on simple random sampling without replacement utilizing information of two auxiliary variables X and Z.

Let us consider a simple random sample (SRS) of size $n$ is drawn from the given population of $N$ units. Let the value of the study variable $Y$ and the auxiliary variables $X$ and $Z$ for the $i$th units ($i = 1, 2, 3, ..., N$) of the population be denoted by $y_i$, $x_i$ and $z_i$ for the $i$th unit in the sample ($i = 1, 2, 3, ..., n$) by $y_i$, $x_i$ and $z_i$ respectively. From the sample observation we have

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i, \quad s^2_y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y})^2, \quad s^2_x = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad s^2_z = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2$$

Let us define

$$\epsilon_y = \frac{s^2_y - s^2_z}{s^2_y}, \quad \epsilon_x = \frac{s^2_x - s^2_z}{s^2_x}, \quad \epsilon_z = \frac{s^2_z - s^2_x}{s^2_z}$$

such that $E(\epsilon_y) = E(\epsilon_x) = E(\epsilon_z) = 0$

$$E(\epsilon^2_y) = \frac{1}{n}(\sigma^2_{\epsilon_y} - 1), \quad E(\epsilon^2_x) = \frac{1}{n}(\sigma^2_{\epsilon_x} - 1), \quad E(\epsilon^2_z) = \frac{1}{n}(\sigma^2_{\epsilon_z} - 1)$$

$$E(\epsilon_y \epsilon_x) = \frac{1}{n}(\sigma_{\epsilon_y \epsilon_x} - 1), \quad E(\epsilon_y \epsilon_z) = \frac{1}{n}(\sigma_{\epsilon_y \epsilon_z} - 1), \quad E(\epsilon_x \epsilon_z) = \frac{1}{n}(\sigma_{\epsilon_x \epsilon_z} - 1)$$

where, $\sigma_{pqr} = \frac{\mu_{pqr}}{\mu_{pqr,pq,pr}}$, $\mu_{pqr} = \frac{1}{n} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})(z_i - \bar{z})$; $p$, $q$, $r$ being the non-negative integers.

### 2. Estimators in literature

In order to estimate population variance of the study variable $Y$, using the information of two auxiliary variables $X$ and $Z$, Singh et al. (2009) proposed a general class of exponential estimator given as

$$t_1 = s^2_y \left[ k \exp \left( \frac{s^2_x - s^2_z}{s^2_x + s^2_z} \right) + (1-k) \exp \left( \frac{s^2_z - s^2_y}{s^2_z + s^2_y} \right) \right]$$

Adichwal et al. (2015) proposed the following two generalized class of estimators for estimation of population variance using two auxiliary variables as

$$t_w = ws^2_y + (1-w)s^2_y \exp \left[ \alpha \frac{s^2_y - s^2_z}{s^2_y + s^2_z} \right]$$

$$t_w = \delta s^2_y \left( s^2_z \right)^{\beta} \exp \left[ \gamma \frac{s^2_z - s^2_y}{s^2_z + s^2_y} \right] + (1-\delta) s^2_y \left( \frac{s^2_z}{s^2_y} \right)^{\beta} \exp \left[ \gamma \frac{s^2_z - s^2_y}{s^2_z + s^2_y} \right]$$

The mean square error (MSE) expressions of the estimators $t_1$, $t_w$ and $t_w$ are, respectively, given by

$$\text{MSE}(t_1) = \frac{s^2_y}{n} \left[ \sigma^2_{\epsilon_y} + \frac{k^2}{4}(\sigma^2_{\epsilon_y} - 1) + \frac{(1-k)^2}{4}(\sigma^2_{\epsilon_y} - 1) \right]$$

$$\text{MSE}(t_w) = \frac{s^2_y}{n} \left[ \sigma^2_{\epsilon_y} + \frac{(1-k)^2}{4}(\sigma^2_{\epsilon_y} - 1) + \frac{k^2(1-k)}{2}(\sigma^2_{\epsilon_y} - 1) \right]$$

$$\text{MSE}(t_w) = \frac{s^2_y}{n} \left[ \sigma^2_{\epsilon_y} + \frac{(1-k)^2}{4}(\sigma^2_{\epsilon_y} - 1) + \frac{k^2(1-k)}{2}(\sigma^2_{\epsilon_y} - 1) \right]$$
where, \( k = \frac{\left( \delta_{004} + 2(\delta_{020} + \delta_{002}) + \delta_{022} - 4 \right)}{(\delta_{040} + \delta_{020} + 2\delta_{022} - 4)} = k_0 \)

\[
\text{MSE}(t_0) = \frac{1}{n} \left[ \frac{s^2}{4} (w_0 - 1)^2 g^2(\delta_{004} - 1) + (\delta_{040} - 1) + (w_0 - 1)^2 g^2(\delta_{004} - 1) + 2g(w_0 - 1)(\delta_{220} - 1)
+ \alpha g(w_0 - 1)(\delta_{232} - 1) + \alpha g^2(w_0 - 1)^2(\delta_{022} - 1) \right]
\]

(2.5)

where, \( w_0 = \left[ 1 - \frac{2(g(\delta_{004} - 1) + g(\delta_{040} - 1))}{2(\delta_{004} - 1) + 2\delta_{020} - 1 + 2g(\delta_{020} - 1)} \right] \)

\[
\text{MSE}(t_0) = \frac{S^2_y}{n} \left[ \rho^2(\delta_0 g - g - \delta_0)^2(\delta_{040} - 1) + (\delta_{040} - 1) + \rho^2(\delta_0 g - g - \delta_0)^2(\delta_{040} - 1) + 2\beta(\delta_0 g - g - \delta_0)(\delta_{220} - 1)
+ \gamma(\delta_0 g - g - \delta_0)(\delta_{232} - 1) + \gamma(\delta_0 g - g - \delta_0)^2(\delta_{022} - 1) \right]
\]

(2.6)

where, \( \delta_0 = \left[ \frac{4\rho^2(\delta_{004} - 1) + 4\rho(\delta_{040} - 1) - 4\beta(\delta_{004} - 1) - 1 + 4\gamma(\delta_{020} - 1) + 4\gamma(\delta_{022} - 1)}{(g - 1)\left[ 4\rho^2(\delta_{004} - 1) + 4\rho(\delta_{040} - 1) + 4\gamma(\delta_{020} - 1) + 4\gamma(\delta_{022} - 1) \right]} \]

3. Proposed estimators

In this paper, we have proposed a generalized exponential-type estimator for population variance for the study variable \( Y \) based on simple random sampling without replacement using information of two auxiliary variable \( X \) and \( Z \) given by

\[
t = \lambda S^2_Y \exp \left[ \eta \left\{ \frac{S^2_y - s^2_x}{S^2_x + (a - 1)s^2_x} \right\} + \psi \left\{ \frac{S^2_z - s^2_x}{S^2_z + (b - 1)s^2_z} \right\} \right]
\]

(3.1)

where \( \eta, \psi, a, \) and \( b \) are suitable chosen constants to be determined such that the MSE of \( t \) is minimum.

Expanding Equation (3.1) in terms of \( e's \) up to the first order of approximation, we have

\[
t = \lambda S^2_Y (1 + e_\lambda) \exp \left[ \frac{-\eta e_x}{a + (a - 1)e_x} + \frac{-\psi e_z}{b + (b - 1)e_z} \right]
\]

(3.2)

or, \( t - S^2_Y = (\lambda - 1)S^2_Y + \lambda S^2_Y \left\{ \frac{-\eta e_x}{a} - \frac{\psi e_z}{b} - \frac{\eta e_x e_z}{a b} - \frac{\eta e_x e_y}{a} \right\} \)

(3.3)

or, \( t - S^2_Y = (\lambda - 1)S^2_Y + \lambda S^2_Y \left\{ \frac{-\psi e_x}{a} - \frac{\eta e_z}{b} - \frac{\eta e_x}{a} - \frac{\psi e_z}{b} - \frac{\eta e_x e_z}{a b} - \frac{\eta e_x e_y}{a} \right\} \)

(3.4)

where, \( \frac{a}{\eta} = \eta^* \) and \( \frac{b}{\psi} = \psi^* \)

Squaring both sides of Equation (3.4) and taking its expectations we get the MSE expressions of the estimator \( t \) as,

\[
\text{MSE}(t) = (\lambda - 1)^2 S^4_Y + \frac{2}{n} S^4_Y \left[ (\delta_{000} - 1) + \psi^2(\delta_{004} - 1)
+ \eta^2(\delta_{004} - 1) - 2\psi^*(\delta_{020} - 1) - 2\eta^*(\delta_{220} - 1) + 2\eta^*\psi^*(\delta_{022} - 1) \right]
\]

(3.5)

Partially differentiating Equation (3.5) with respect to \( \lambda, \eta^*, \) and \( \psi^* \) and equating it to zero, we get the optimum value of \( \lambda, \eta^* \) and \( \psi^* \) as
\[ \lambda_{\text{opt}} = \frac{1}{1 + \frac{1}{n} \left( \sigma_{400}^2 + \psi^2(\sigma_{004}^2 - 1) + \eta_2^2(\sigma_{040}^2 - 1) - 2\psi(\sigma_{022} - 1) - 2\eta_2(\sigma_{220} - 1) + 2\eta_2\psi(\sigma_{022} - 1) \right) } \]

\[ \eta^*_{\text{opt}} = \frac{(\sigma_{220} - 1) - \psi(\sigma_{022} - 1)}{(\sigma_{040} - 1)} \]

\[ \psi^*_{\text{opt}} = \frac{(\sigma_{202} - 1)(\sigma_{040} - 1) - (\sigma_{220} - 1)(\sigma_{022} - 1)}{(\sigma_{040} - 1)(\sigma_{004} - 1) - (\sigma_{022} - 1)^2} \]

Substituting the optimum value of \( \lambda_{\text{opt}}, \eta^*_{\text{opt}}, \) and \( \psi^*_{\text{opt}} \) in Equation (3.5), we obtain the minimum MSE associated with the estimators \( \tilde{t} \) as,

\[ \text{MSE}(\tilde{t})_{\min} = (\lambda_{\text{opt}} - 1)^2 S^2_y + \frac{\lambda^2_{\text{opt}}}{n} S^2_y \left[ (\sigma_{400}^2 - 1) + \psi^2(\sigma_{004}^2 - 1) \right. \]

\[ \left. + \eta^2(\sigma_{040} - 1) - 2\psi(\sigma_{022} - 1) - 2\eta_2(\sigma_{220} - 1) + 2\eta_2\psi(\sigma_{022} - 1) \right] \]

### 4. Efficiency comparison

In this section, we are comparing the minimum MSE of the proposed estimator \( \tilde{t} \) with usual estimator \( S^2_y \) and other existing estimators.

The variance of the usual estimator \( S^2_y \) under SRSWOR is given by

\[ V(S^2_y) = \frac{S^2_y}{n} \left[ \sigma_{400}^2 - 1 \right] \]  \[ (4.1) \]

\[ V(S^2_y) - \text{MSE}(\tilde{t})_{\min} = \frac{1 - \lambda^2_{\text{opt}}}{n} S^2_y \left( \sigma_{400}^2 - 1 \right) - \left( \lambda_{\text{opt}} - 1 \right)^2 S^2_y - \frac{\lambda^2_{\text{opt}}}{n} \left[ \psi^2(\sigma_{004}^2 - 1) \right. \]

\[ \left. + \eta^2(\sigma_{040} - 1) - 2\psi(\sigma_{022} - 1) - 2\eta_2(\sigma_{220} - 1) + 2\eta_2\psi(\sigma_{022} - 1) \right] \geq 0 \]  \[ (4.2) \]

\[ \text{MSE}(t_n) - \text{MSE}(\tilde{t})_{\min} = \frac{1 - \lambda^2_{\text{opt}}}{n} S^2_y \left[ \frac{\pi^2}{4} (w_0 - 1)^2 g^2(\sigma_{004}^2 - 1) + (w_0 - 1) g^2(\sigma_{040}^2 - 1) \right. \]

\[ \left. + 2g(w_0 - 1)(\sigma_{022} - 1) + ag(w_0 - 1)(\sigma_{220} - 1) + ag^2(w_0 - 1)^2(\sigma_{022} - 1) \right] \]

\[ - \left[ (\lambda_{\text{opt}} - 1)^2 S^2_y + \frac{\lambda^2_{\text{opt}}}{n} S^2_y \left( \sigma_{400}^2 - 1 \right) + \psi^2(\sigma_{004}^2 - 1) \right. \]

\[ \left. + \eta^2(\sigma_{040} - 1) - 2\psi(\sigma_{022} - 1) - 2\eta_2(\sigma_{220} - 1) + 2\eta_2\psi(\sigma_{022} - 1) \right] \geq 0 \]  \[ (4.3) \]
When the conditions (4.2) to (4.4) are satisfied, our suggested estimator \( t \) will be more efficient as compared to \( \sigma^2 \), \( t_1 \), \( t_n \) and \( t_M \) respectively.

5. Empirical study
To illustrate the performance of various estimators of \( \sigma^2 \), we consider the following data sets

**Population I** [Source: Sarjinder Singh (2003, p. 1116)].

- \( y \): Fish caught in year 1995, \( x \): Fish caught in year 1993, \( z \): Fish caught in year 1994,
- \( N = 69, \ n = 25 \).

\[
\begin{align*}
&\partial_{400} = 7.7685, \ \partial_{040} = 9.9860, \ \partial_{004} = 9.9851, \ \partial_{202} = 8.3107, \ \partial_{202} = 8.1715, \ \partial_{022} = 9.6631, \\
&S_y^2 = 37199578.
\end{align*}
\]

**Population II** [Source: Murthy (1967, p. 399)].

- \( y \): Area under wheat in 1964, \( x \): Area under wheat in 1963, \( z \): Cultivated area in 1961,
- \( N = 34, \ n = 15 \).

\[
\begin{align*}
&\partial_{400} = 3.7879, \ \partial_{040} = 2.9123, \ \partial_{004} = 2.8082, \ \partial_{220} = 3.1046, \ \partial_{202} = 2.9790, \ \partial_{022} = 2.7379, \\
&S_y^2 = 22564.55704.
\end{align*}
\]

From Table 1, it is clear that the MSE of the estimator “\( t \)” is less as compared to the estimators \( S_y^2 \), \( t_1 \), \( t_n \) and \( t_M \). In terms of PRE’s, it is observed that the PRE of the proposed estimator \( t \) is higher as

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( \text{PREs} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_y^2 )</td>
<td>Population I: 100.0000, Population II: 100.0000</td>
</tr>
<tr>
<td>( t_1 )</td>
<td>Population I: 848.7573, Population II: 698.2471</td>
</tr>
<tr>
<td>( t_n )</td>
<td>Population I: 848.7573, Population II: 698.2471</td>
</tr>
<tr>
<td>( t_M )</td>
<td>Population I: 848.7573, Population II: 698.2471</td>
</tr>
<tr>
<td>( t )</td>
<td>Population I: 876.6598, Population II: 716.6599</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{MSE}(t_n) - \text{MSE}(t)_{\text{min}} & = \frac{S_y^4}{n} \left[ \beta^2(\partial_0 g - g - \delta_0)^2(\partial_{040} - 1) + (\partial_{400} - 1) \right. \\
& + \frac{\gamma^2}{4}(\partial_0 g - g - \delta_0)^2(\partial_{004} - 1) + 2\beta(\partial_0 g - g - \delta_0)(\partial_{202} - 1) \\
& + \gamma(\partial_0 g - g - \delta_0)(\partial_{202} - 1) + \gamma(\partial_0 g - g - \delta_0)^2(\partial_{022} - 1) \\
& - \left[ (\partial_{\text{opt}} - 1)S_y^4 + \frac{2\partial_{\text{opt}}}{n}S_y^4 \left( (\partial_{040} - 1) + \psi^2(\partial_{040} - 1) \right. \\
& + \eta(\partial_{040} - 1) - 2\psi(\partial_{202} - 1) - 2\eta(\partial_{202} - 1) \\
& + 2\eta(\partial_{022} - 1) \right] \right] \geq 0
\end{align*}
\]
compared to PRE’s of the estimators $s_y^2$, $t_w$, $t_1$, and $t_2$ for both the given data-sets. So, the estimator $t$ is more efficient than the estimator $s_y^2$ and other existing estimators.

6. Simulation study

This section presents the computational procedure for the comparison of proposed estimator with other existing estimators. The simulation study is based on the algorithm proposed by Reddy, Rao, and Boiroju (2010) to illustrate the performance of various estimators of $S_y^2$. The following algorithm explain the simulation procedure used in this paper.

**Step-1:** Generate two independent random variable $X$ from $N\left(\mu, \sigma^2\right)$ and $X_1$ from $N\left(\mu_1, \sigma_1^2\right)$ using box-Muller method (Johnson, 1987).

**Step-2:** Set $Y = \rho X + \sqrt{1 - \rho^2}X_1$ where $0 < \rho = 0.4, 0.6, 0.8 < 1$.

**Step-3:** return the pair $(Y, X)$.

**Step-4:** Consider the population-I with the parameters $\mu = 5, \sigma = 3, \mu_1 = 5$ and $\sigma_1 = 3$ in step-1 and repeat the steps 1 to 3 for 1000 times. This population will contain the same variance for the variable $Y$ and $X$.

**Step-5:** Similarly, generate the population-II with the parameters $\mu = 3, \sigma = 2, \mu_1 = 5$ and $\sigma_1 = 3$ in step-1 and repeat the steps 1 to 3 for 1000 times. This population will have different variances for the variable $Y$ and $X$.

**Step-6:** From the population of size $N = 1000$, draw 2000 SRSs $(y_i, x_i)$ $(i = 1, 2, \ldots, n)$ without replacement of size $n = 40, 50$ and $60$.

**Step-7:** The Average mean squared error (MSE) of the estimators are defined by

$$\text{Average MSE}(t) = \frac{1}{2000} \sum_{k=1}^{2000} E\left(t_k - S_y^2\right)^2$$

The PRE of an estimator $t$ with respect to the usual estimator $s_y^2$ is defined by

$$\text{PRE}(t) = \frac{\text{MSE}\left(s_y^2\right) \times 100}{\text{MSE}(t)}$$

The obtained results of the simulation study are as follows.

From Tables 2 and 3 we observe that as the sample size and the value of correlation coefficient increases, the average PRE’s of the estimators also increases for both the population I and II. The average PRE of the estimator “$t$” is higher in all cases indicating that the proposed estimator $t$ is more efficient as compared to usual estimator $s_y^2$ and other existing estimators $t_w$, $t_1$, and $t_2$. 


7. Conclusion
This paper proposed an improved generalized class of estimator for estimating population variance based on simple random sampling without replacement using information of two auxiliary variables. The performance of the proposed estimator is verified by using two real population data-sets and by simulation study. Tables 1-3 clearly show that the proposed estimator \(t\) is more efficient as compared to the usual estimator and other existing estimators. Hence, it is recommended for use in practice.

Funding
The authors received no direct funding for this research.

Author details
Mr. Nitesh K. Adichwal
E-mail: nnitesh139@gmail.com
Dr. Jitendra Kumar
E-mail: statsjkumar@gmail.com
Prof. Rajesh Singh
E-mail: rsinghstat@gmail.com

1 Department of Statistics, Banaras Hindu University, Varanasi, Uttar Pradesh, India.

Citation information
Cite this article as: An improved generalized class of estimators for population variance using auxiliary variables, Mr. Nitesh K. Adichwal, Dr. Jitendra Kumar & Prof. Rajesh Singh, Cogent Mathematics & Statistics (2018), 5: 1454579.

References