Estimation of Population Variance using the Knowledge of Kurtosis of an Auxiliary Variable under Simple Random Sampling

Mursala Khan
Department of Mathematics, COMSAT Institute of Information Technology, Abbottabad, Pakistan
Email: mursala.khan@yahoo.com

Ajay Arunachalam
Graduate School of Applied Statistics, NIDA, Thailand
Email: ajay.arunachalam08@gmail.com

Abstract – In this study we have proposed an estimator for population variance \( S^2_y \) of the study variable \( y \) under simple random sampling utilizing the knowledge of the kurtosis of an auxiliary variable \( x \). The properties of the estimators are derived up to first order of expression. We have also derived some efficiency conditions theoretically under which the proposed variance estimator performed better than the usual unbiased estimator, traditional ratio estimator, Upadhyaya and Singh (1999), Tailor and Sharma (2012) and Yadav et al. (2013) estimator. The results have been verified numerically by considering two natural populations from the literature.

Keywords – Simple Random Sampling, Auxiliary Variable, Variance Estimator, Optimum Value, Bias, Mean Squared Error, Efficiency.

I. INTRODUCTION

It is known fact, that proper use of auxiliary information increases the precision of the estimators in such a situations the ratio, product and regression estimators provides more efficient results than the usual simple random sampling. From the estimation of population variance we know about the variability in the units of the population which is essential for future surveys either for stratification or determination of sample size. In order to estimate the finite population mean we use survey data but in many cases the mean is not a appropriate average because it fluctuate from small or large observations or outliers in a set of data so overcome this difficulty we come to variance. The early work on the estimation of population variance was initiated from the work of Evans (1951), Hansen, Hurwitz and Madow (1953). Later on many statistician’s worked on the estimation of population variance using auxiliary information, Isaki (1983) suggested the usual ratio for estimating population variance \( S^2_y \) of the study variable \( y \), using auxiliary information \( S^2_x \) of the auxiliary variable \( x \). Das and Triphati [1978], Srivastava and Jhajj [1980], Upadhyaya and Singh [1983], Upadhyaya and Singh [1999], Singh and Singh (2001, 2003), Upadhyaya et al. (2004), Kadilar and Cingi [2006], Singh et al. (2008), Gupta and Shabbir (2008), Grover (2010), Singh et al. (2011), Khan and Shabbir (2013), Recently Yadav et al. (2013) proposed an estimator for population variance using transformations on both the study variable as well as auxiliary variable when coefficient of variation of an auxiliary variable is known.

Let us consider a finite population of size \( N \) of different units \( U = \{U_1, U_2, U_3, ..., U_N\} \). Let \( y \) and \( x \) be the study and the auxiliary variable with corresponding values \( y_i \) and \( x_i \) respectively for \( i \)-th unit \( i = \{1, 2, 3, ..., N\} \) is defined on a finite population \( U \). Let \( \bar{Y} = (1/N) \sum_{i=1}^{N} y_i \) and \( \bar{X} = (1/N) \sum_{i=1}^{N} x_i \) be the corresponding population means of the study as well as auxiliary variable respectively. Also let \( S^2_y = (1/N - 1) \sum_{i=1}^{N} (y_i - \bar{Y})^2 \) and \( S^2_x = (1/N - 1) \sum_{i=1}^{N} (x_i - \bar{X})^2 \) be the corresponding population variances of the study as well as auxiliary variable respectively and let \( C_y \) and \( C_x \) be the coefficient of variation of the study as well as auxiliary variable respectively, and \( \rho_{yx} \) be the correlation coefficient between \( x \) and \( y \). In order to estimate the unknown population variance by using simple random sampling scheme we take a sample of size \( n \) units from the population \( U \) by using simple random sample without replacement. Let \( y \) and \( x \) be the study and the auxiliary variable with corresponding values \( y_i \) and \( x_i \) respectively for \( i \)-th unit \( i = \{1, 2, 3, ..., n\} \) in the sample.

Let \( \bar{Y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{X} = (1/n) \sum_{i=1}^{n} x_i \) be the corresponding sample means of the study as well as auxiliary variable respectively. Also let \( S^2_y = (1/(n-1)) \sum_{i=1}^{n} (y_i - \bar{Y})^2 \) and \( S^2_x = (1/(n-1)) \sum_{i=1}^{n} (x_i - \bar{X})^2 \) be the corresponding sample variances of the study as well as auxiliary variable respectively. Some notations that we use are, given as below

\[
\beta_{y} = \frac{\mu_{40}}{\mu_{20}^2} \quad \text{and} \quad \beta_{x} = \frac{\mu_{40}}{\mu_{20}^2}
\]

be the coefficient of kurtosis of the study variable \( y \) and the auxiliary variable \( x \) respectively. Further let

\[
\mu_{n,2} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{Y})' (x_i - \bar{X})' \quad \lambda_{22} = \frac{\mu_{22}}{\mu_{20} \mu_{40}}
\]
\[ \lambda_{21} = \frac{\mu_{21}}{\mu_{20} \mu_{02}}, \quad \beta_{0} = (\beta_{1}, -1), \]
\[ \beta_{2} = (\beta_{2}, -1), \quad \lambda_{22} = \lambda_{22} - 1 \text{ and } \theta = \frac{1}{n} - \frac{1}{N}. \]

II. EXISTING ESTIMATORS

In this section we will discuss some of the existing estimators that are available in the literature.

When there is no auxiliary information the usual unbiased to estimate the population variance of the study variable is
\[ t_{0} = s_{y}^{2}. \tag{2.1} \]

The bias and variance, of the estimator \( t_{0} \) up to first order of approximation are given by

\[ \text{Bias}(t_{0}) = 0 \] \tag{2.2}

\[ \text{Var}(t_{0}) = \theta S_{y}^{4} \beta_{2}. \tag{2.3} \]

When the population variance of the auxiliary variable is known the usual ratio estimator to estimate the population variance of the study variable \( S_{y}^{2} \) suggested by Isaki (1983) and is, given by
\[ t_{r} = s_{y}^{2} \frac{S_{y}^{2}}{S_{x}^{2}}. \tag{2.4} \]

The bias and mean squared error, up to first order of approximation are given by

\[ \text{Bias}(t_{r}) = \theta S_{y}^{4} \left[ \beta_{2} - \lambda_{22}^{*} \right] \tag{2.5} \]

\[ \text{MSE}(t_{r}) = \theta S_{y}^{4} \left[ \beta_{2}^{2} + \lambda_{22}^{*} + 2 \lambda_{22}^{*} \right] \tag{2.6} \]

Upadhyaya and Singh (1999), suggested the following estimator for population variance using information on coefficient of kurtosis of the auxiliary variable, given as
\[ t_{us} = s_{y}^{2} \left[ \frac{S_{y}^{2} + \beta_{2}}{S_{y}^{2} + \beta_{2}} \right]. \tag{2.7} \]

The bias and mean squared error, up to first order of approximation are given by

\[ \text{Bias}(t_{us}) = \theta k S_{y}^{4} \left[ k \beta_{2} - \lambda_{22}^{*} \right] \tag{2.8} \]

\[ \text{MSE}(t_{us}) = \theta S_{y}^{4} \left[ \beta_{2} + k^{2} \beta_{2} - 2 k \lambda_{22}^{*} \right] \tag{2.9} \]

Tailor and Sharma (2012) proposed a class of estimators under sampling fraction and in the presence of auxiliary information, as follows
\[ t_{1} = s_{y}^{2} \left[ f + (1-f) \left( \frac{S_{y}^{2}}{S_{x}^{2}} \right) \right] \tag{2.10} \]

\[ t_{2} = s_{y}^{2} \left[ \frac{1-f}{1+2f} \left( \frac{S_{y}^{2}}{S_{x}^{2}} \right) \right] + \frac{3f}{1+2f} \left( \frac{S_{y}^{2}}{s_{y}^{2}} \right) \tag{2.11} \]

\[ t_{3} = s_{y}^{2} \left[ \frac{1-f}{1+3f} \left( \frac{S_{y}^{2}}{S_{x}^{2}} \right) \right] + \frac{4f}{1+3f} \left( \frac{S_{y}^{2}}{S_{y}^{2}} \right) \tag{2.12} \]

The biases and mean squared errors of the estimators, up to first order of approximation are given by

\[ \text{Bias}(t_{1}) = \theta (1-f) S_{y}^{2} \left[ \beta_{2} - \lambda_{22}^{*} \right] \tag{2.13} \]

\[ \text{Bias}(t_{2}) = \theta \left( \frac{3f}{1+2f} \right) S_{y}^{2} \left[ \beta_{2} - \lambda_{22}^{*} \right] \tag{2.14} \]

\[ \text{Bias}(t_{3}) = \theta \left( \frac{4f}{1+3f} \right) S_{y}^{2} \left[ \beta_{2} - \lambda_{22}^{*} \right] \tag{2.15} \]

\[ \text{MSE}(t_{1}) = \theta S_{y}^{4} \left[ \beta_{2} + (1-f) \beta_{2} - 2(1-f) \lambda_{22}^{*} \right] \tag{2.16} \]

\[ \text{MSE}(t_{2}) = \theta S_{y}^{4} \left[ \beta_{2} + \left( \frac{3f}{1+2f} \right) \beta_{2} - 2 \left( \frac{3f}{1+2f} \right) \lambda_{22}^{*} \right] \tag{2.17} \]

\[ \text{MSE}(t_{3}) = \theta S_{y}^{4} \left[ \beta_{2} + \left( \frac{4f}{1+3f} \right) \beta_{2} - 2 \left( \frac{4f}{1+3f} \right) \lambda_{22}^{*} \right] \tag{2.18} \]

Recently Yadav et al. (2013) suggested an estimator for population variance using transformations on both the study as well as the auxiliary variable, given as
\[ t_{s} = s_{y}^{2} \left[ a \left( \frac{C_{2}^{2}}{C_{1}^{2}} \right) + (1-a) \left( \frac{C_{2}^{2}}{C_{1}^{2}} \right) \right] - a \tag{2.19} \]

where \( \alpha, \gamma, a, b, d \) and \( \eta \) are constants.

Also \( s_{y}^{2} = s_{y}^{2} + a \),

\[ C_{1}^{2} = \left( b C_{2}^{2} + d C_{2}^{2} \right) \left( b s_{x}^{2} + d \frac{s_{y}^{2}}{\mu_{x}} \right) \]

\[ C_{2}^{2} = (b + d) C_{2}^{2} = (b + d) \frac{s_{y}^{2}}{\mu_{x}}. \]

The bias and mean squared error, up to first order of approximation are given by

\[ \text{Bias}(t_{s}) = \left( \frac{S_{y}^{2} + \alpha}{2n} \right) \left[ \theta \left( S_{y}^{2} + \alpha \right) + \alpha \left( S_{y}^{2} + \alpha \right) + \alpha \left( S_{y}^{2} + \alpha \right) + \alpha \left( S_{y}^{2} + \alpha \right) \right] \tag{2.20} \]

\[ \text{MSE}(t_{s}) = \theta S_{y}^{4} \left[ \lambda_{10} - \lambda_{22}^{*} C_{1}^{2} \left( \lambda_{22}^{*} - 2 \lambda_{22}^{*} \right) \right] \tag{2.21} \]

III. THE PROPOSED ESTIMATOR

To estimate the population variance, we proposed the following estimator using the knowledge of the coefficient of kurtosis of the auxiliary variable.
\[ S_{y}^{2} = s_{y}^{2} \left[ \lambda_{10} - \lambda_{22}^{*} C_{1}^{2} \left( \lambda_{22}^{*} - 2 \lambda_{22}^{*} \right) \right] \tag{3.1} \]

In order to study large sample properties of \( S_{y}^{2} \), we write
\[ s_{y}^{2} = S_{y}^{2} \left( 1 + \zeta_{y} \right), \quad S_{y}^{2} = S_{y}^{2} \left( 1 + \zeta_{y} \right), \]

Such that \( E(\zeta_{y}) = E(\zeta_{y}) = 0 \)

\[ E(\zeta_{y}) = \theta (\beta_{2}, -1), \quad E(\zeta_{y}) = \theta (\beta_{2}, -1) \text{ and } E(\zeta_{y}) = \theta (\lambda_{22}^{*} - 1) \]

In terms of \( \zeta_{y}^{'}, S_{y}^{2} \), we have

Copyright © 2014 IJASM, All right reserved
\[
\hat{S}_P^2 = \frac{1}{\beta_s^2 + \lambda_2^2} \left[ \alpha \left(1 + \lambda_2^2 \right) - \frac{\lambda_2^2}{\beta_s^2} \right] \]

Expanding the right hand side of above equation, neglecting terms of \( \lambda_2^2 \)'s having power greater than two and subtracting \( \hat{S}_P^2 \) from both sides, we have
\[
\hat{S}_P^2 - \hat{S}_y^2 = \frac{\alpha}{\beta_s^2 + \lambda_2^2} \left[ \lambda_2^2 - k(1 - 2\alpha) \lambda_1^2 - \alpha \lambda_1^2 \right].
\]

Taking expectation on both sides of (3.2), we get bias
\[
\text{Bias} \left( \hat{S}_P^2 \right) = \theta S^2 \left[ \lambda_2^2 \left( k - \frac{\lambda_2^2}{\beta_s^2} \right) \right] \frac{1}{2}
\]

Squaring both sides of (3.2), and expanding up to order 2, we have
\[
\left( \hat{S}_P^2 - \hat{S}_y^2 \right)^2 = S^2 \left[ \lambda_2^2 - k(1 - 2\alpha) \lambda_1^2 \right]^2.
\]

Taking expectation on both sides of (3.2), we get
\[
\text{MSE} \left( \hat{S}_P^2 \right) = \theta S^2 \left[ \beta_s^2 + k^2 - 2k(1 - 2\alpha) \lambda_2^2 \right]
\]

Differentiating equation (3.5), w. r. t. to \( \alpha \) and put it equal to zero, we get the optimum value of \( \alpha \).

\[
\alpha_{opt} = \frac{1}{2} \left( 1 - \frac{\lambda_2^2}{k \beta_s^2} \right)
\]

Substituting the optimum value of \( i.e \alpha_{opt} \) in equation (3.5), we get Minimum Mean Squared Error, given by
\[
\text{MSE} \left( \hat{S}_P^2 \right)_{\text{min}} = \theta S^2 \beta_s^2 \left[ 1 - \frac{\lambda_2^2}{\beta_s^2} \right]
\]

IV. EFFICIENCY COMPARISON

In this section, we have compared mean squared errors of the of the proposed estimator \( \hat{S}_P^2 \) the usual estimator \( t_0 \), \( t_R \), \( t_US \), \( t_1 \), \( t_2 \) and \( t_g \).

(i) By (2.3) and (3.7),
\[
\left[ \text{MSE} \left( t_0 \right) - \text{MSE} \left( \hat{S}_P^2 \right)_{\text{min}} \right] \geq 0
\]

(ii) By (2.6) and (3.7),
\[
\left[ \text{MSE} \left( t_R \right) - \text{MSE} \left( \hat{S}_P^2 \right)_{\text{min}} \right] \geq 0
\]

(iii) By (2.9) and (3.7),
\[
\left[ \text{MSE} \left( t_g \right) - \text{MSE} \left( \hat{S}_P^2 \right)_{\text{min}} \right] \geq 0
\]

We use the following expression for efficiency comparison.
\[
\text{PRE} \left( i, t_i \right) = \frac{\text{MSE} \left( t_i \right)}{\text{MSE} \left( i \right)} \times 100,
\]

where \( i = t_0, t_R, t_US, t_1, t_2, t_g \) and \( t_{GR} \).

To look the estimators closely, we consider two natural populations from the literature of survey, the description of the populations are, given as.

Population-1: [Source: Das (1980)]
It consists of 142 cities of India with population (number of persons) 100,000 and above.

The variates considered are:
\( X \): Census population in the year 1961 and
\( Y \): Census population in the year 1971.
\( N = 142, n = 15, \bar{X} = 4015.2183, \bar{S}_2 = 2900.3872, \beta_s = 40.8536, \beta_s = 48.1567, \lambda_2 = 43.7615, \rho_x = 0.9948, \rho_x = 2.1118, \rho_x = 2.1971, \rho_x = 0.9948. \)

\( X \): Number of rooms per block and
\( Y \): Number of persons per block.

V. EMPIRICAL STUDY
$N = 100$, $n = 10$, $\bar{Y} = 101.10$, $\bar{X} = 58.810$, $S_y = 14.6595$, $S_x = 7.53228$, $C_y = 0.1450$, $C_x = 0.1281$, $\rho_{xy} = 0.6500$, $\beta_2 = 2.2387$, $\beta_3 = 2.3523$, $\lambda_2 = 1.5432$.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Population 1</th>
<th>Population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$MSE(\cdot)$</td>
<td>$PRE(\cdot, s^2)$</td>
</tr>
<tr>
<td>Existing</td>
<td>$\hat{S}^2_y$</td>
<td>1.23x10^{-16}</td>
</tr>
<tr>
<td></td>
<td>$\hat{S}^2_R$</td>
<td>4.58x10^{-14}</td>
</tr>
<tr>
<td></td>
<td>$t_{ss}$</td>
<td>4.58x10^{-14}</td>
</tr>
<tr>
<td></td>
<td>$t_y$</td>
<td>3.34x10^{-14}</td>
</tr>
<tr>
<td></td>
<td>$t_z$</td>
<td>6.38x10^{-14}</td>
</tr>
<tr>
<td></td>
<td>$t_{rp}$</td>
<td>5.32x10^{-14}</td>
</tr>
<tr>
<td>Proposed</td>
<td>$\hat{S}^2_p$</td>
<td>3.32x10^{-14}</td>
</tr>
</tbody>
</table>

VI. Conclusion

Table 1 clearly indicates that the proposed estimator under the optimizing value of the unknown constant is always more efficient than the usual estimator, the ratio estimator suggested by Isaki (1983), Upadhyaya and Singh (1999), Tailor and Sharma (2012) and Yadav et al. (2013) in terms of lesser $MSE$’s and greater $PRE$’s. Having the largest gain in efficiency the proposed estimator appeared to be the best one among all the estimators and would work very well in practical surveys.

ACKNOWLEDGMENT

Authors are very thankful to the anonymous reviewers for their valuable comments that helped in improving the quality of the present work.

REFERENCES

AUTHOR’S PROFILE

Mursala Khanis
Currently working as a lecturer in the Department of Mathematics, COMSATS, Institute of Information Technology Abbottabad, Pakistan. He has completed his Master of Philosophy in Statistics from the Department of Statistics, Quaid-I-Azam University Islamabad, Pakistan, having a research title on “Estimating the Ratio of Two Ratios in Simple and Stratified Random Sampling”.

Ajay Arunachalam
Received his B.E (Computer Engineering) degree from Pune University, India and M.E (Computer Science & Engineering) from Anna University, Chennai, India. He is currently enrolled as a Ph.D. scholar at National Institute of Development Administration (NIDA), Graduate School of Applied Statistics, Thailand. His research interests include area of FOSS in Educational Technology, Algorithms, Wireless and Mobile Networking and P2P networks.