American Journal of Operational Research 2016, 6(1): 9-15
DOI: 10.5923/j.ajor.20160601.02

Efficient Estimator for Population Variance Using Auxiliary Variable

Subhash Kumar Yadav[1], Sheela Misra[2], S. S. Mishra[1,*]

[1]Department of Mathematics and Statistics (A Centre of Excellence), Dr. RML Avadh University, Faizabad, U.P., India
[2]Department of Statistics, University of Lucknow, Lucknow, U.P., India

Abstract Population variance is one of the important measures of dispersion. For example one is interested in knowing the estimate of variance of a particular crop, blood pressure, temperature etc. This paper deals with the estimation of population variance using auxiliary information under simple random sampling scheme. In the present paper, we have proposed an improved estimator through well known kappa technique using Yadav et al (2014) paper. The large sample properties of the estimator have been studied up to the first order of approximation that is its bias and mean square error have been obtained up to the first order of approximation. The optimum value of the characterizing scalar kappa has been obtained and for this optimum value of the kappa the minimum mean squared error has been obtained. A comparison has been made with the existing estimators of population variance using secondary data. An improvement of the proposed estimator has been shown over all existing mentioned estimators as it has lesser mean square error as compared to other estimators.

Keywords Ratio estimator, Quartiles, Bias, Mean squared error, Efficiency

1. Introduction

In the theory of survey sampling, the auxiliary information plays paramount role in developing and searching improved estimators of population parameters of the study variable. The auxiliary information is used at both the stages of designing and estimation. Here we have used this information at estimation stage only. The auxiliary variable (*X*) and the main variable (*Y*) under study are highly closely related with each other. When there is a close positive association between the study variable and the auxiliary variable and the line of regression of the study variable *Y* on the auxiliary variable *X* passes through origin, then the ratio type estimator is used for improvement over the parameters of the population under consideration. On the other hand the product type estimators are used for improved estimation of parameters when the auxiliary variable *X* and the study variable *Y* have negative correlation between them. While the regression type estimators are used for the improved estimation of population parameters, when the line of regression does not pass through the origin.

Let the population under investigation is finite and it consists of *N* distinct and identifiable units. Let $(x_i, y_i),\ i = 1, 2, ..., n$ be a random sample of size n from above bivariate population (*X,* Y) of size N using a *SRSWOR* scheme. Let $\overline{X}$ and $\overline{Y}$ respectively are the population means of the auxiliary and the study variables, and let $\overline{x}$ and $\overline{y}$ are the corresponding sample means which are unbiased estimators of population means $\overline{X}$ and $\overline{Y}$ respectively. Let ρ denote the correlation coefficient between the variables *X* and *Y* and Q_r is the inter-quartile range of the auxiliary variable *X*. In this manuscript, we have proposed an improved ratio type estimator of population variance of study variable by suitably using the correlation coefficient ρ between the two variables and Q_r, inter-quartile range of the auxiliary variable *X*. Further we assume that a reliable estimate of the correlation coefficient ρ is available in advance from pilot surveys etc.

* Corresponding author:
sant_x2003@yahoo.co.in (S. S. Mishra)
Published online at http://journal.sapub.org/ajor

2. Variance Estimators in Literature

The sample variance is the most appropriate estimator of population variance and is given by:

$$t_0 = s_y^2, \tag{2.1}$$

This estimator of population variance is unbiased, and it has the variance up to the first degree of approximation as:

$$V(t_0) = \gamma S_y^4 (\lambda_{40} - 1) \tag{2.2}$$

Isaki (1983) proposed the following ratio estimator of population variance using auxiliary information as:

$$t_R = s_y^2\left(\frac{S_x^2}{s_x^2}\right), \tag{2.3}$$

where

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2,\ s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,\ S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2,$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N}X_i,\ \bar{Y} = \frac{1}{N}\sum_{i=1}^{N}Y_i,\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

The first order of approximations for the *Bias* and Mean Square Error (*MSE*) respectively are given by

$$B(t_R) = \gamma S_y^2[(\lambda_{04}-1)-(\lambda_{22}-1)], \tag{2.4}$$

$$MSE(t_R) = \gamma S_y^4[(\lambda_{40}-1)+(\lambda_{04}-1)-2(\lambda_{22}-1)], \tag{2.5}$$

where $\lambda_{rs} = \dfrac{\mu_{rs}}{\mu_{20}^{r/2}\mu_{02}^{s/2}}$, $\mu_{rs} = \dfrac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^r(X_i-\bar{X})^s$, $\gamma = \dfrac{1-f}{n}$ and $f = \dfrac{n}{N}$.

Several authors proposed different estimators by utilizing auxiliary information in different forms. They used it in the form of different parameters of auxiliary variable for estimating the population variance of the main variable under study. Some of them from the literature are as follows,

Upadhyaya and Singh (1999) utilized coefficient of kurtosis $\beta_{2(x)}$ of auxiliary variable and proposed the following estimator of population variance as,

$$\hat{S}_1^2 = s_y^2\left[\frac{S_x^2+\beta_{2(x)}}{s_x^2+\beta_{2(x)}}\right] \tag{2.6}$$

The bias and Mean Squared Error of above estimator up to the first order of approximations respectively are,

$$B(\hat{S}_1^2) = \gamma S_y^2 R_1\left[R_1(\lambda_{04}-1)-(\lambda_{22}-1)\right]$$

$$MSE(\hat{S}_1^2) = \gamma S_y^4\left[(\lambda_{40}-1)+R_1^2(\lambda_{04}-1)-2R_1(\lambda_{22}-1)\right] \tag{2.7}$$

Where, $R_1 = \dfrac{S_x^2}{S_x^2+\beta_{2(x)}}$

Kadilar and Cingi (2006) proposed the following estimators using different parameters of auxiliary information as,

$$\hat{S}_2^2 = s_y^2\left[\frac{S_x^2+C_x}{s_x^2+C_x}\right],\ \hat{S}_3^2 = s_y^2\left[\frac{S_x^2\beta_{2(x)}+C_x}{s_x^2\beta_{2(x)}+C_x}\right],\ \hat{S}_4^2 = s_y^2\left[\frac{S_x^2C_x+\beta_{2(x)}}{s_x^2C_x+\beta_{2(x)}}\right]$$

The bias and Mean Squared Error of above estimators up to the first order of approximations respectively are,

$$B(\hat{S}_i^2) = \gamma S_y^2 R_i\left[R_i(\lambda_{04}-1)\ \ -(\lambda_{22}-1)\right]$$

$$MSE(\hat{S}_i^2) = \gamma S_y^4\left[(\lambda_{40}-1)+R_i^2(\lambda_{04}-1)-2R_i(\lambda_{22}-1)\right]\quad i=2,3,4 \tag{2.8}$$

Where, $R_2 = \dfrac{S_x^2}{S_x^2+C_x}$, $R_3 = \dfrac{S_x^2\beta_{2(x)}}{S_x^2\beta_{2(x)}+C_x}$, $R_4 = \dfrac{S_x^2C_x}{S_x^2C_x+\beta_{2(x)}}$

Subramani and Kumarpandiyan (2012), utilizing various population parameters of auxiliary variable proposed the following estimators of population variance as,

$$\hat{S}_5^2 = s_y^2\left[\frac{S_x^2+Q_1}{s_x^2+Q_1}\right],\ \hat{S}_6^2 = s_y^2\left[\frac{S_x^2+Q_3}{s_x^2+Q_3}\right],\ \hat{S}_7^2 = s_y^2\left[\frac{S_x^2+Q_r}{s_x^2+Q_r}\right],\ \hat{S}_8^2 = s_y^2\left[\frac{S_x^2+Q_d}{s_x^2+Q_d}\right],\ \hat{S}_9^2 = s_y^2\left[\frac{S_x^2+Q_a}{s_x^2+Q_a}\right]$$

The expressions for the bias and Mean Squared Error of above estimators up to the first order of approximations respectively are,

$$B(\hat{S}_i^2) = \gamma S_y^2 R_i\left[R_i(\lambda_{04}-1) \quad -(\lambda_{22}-1)\right]$$

$$MSE(\hat{S}_i^2) = \gamma S_y^4\left[(\lambda_{40}-1)+R_i^2(\lambda_{04}-1)-2R_i(\lambda_{22}-1)\right] \quad i=5,6,7,8,9 \tag{2.9}$$

Where,

$$R_5 = \frac{S_x^2}{S_x^2+Q_1},\ R_6 = \frac{S_x^2}{S_x^2+Q_3},\ R_7 = \frac{S_x^2}{S_x^2+Q_r},\ R_8 = \frac{S_x^2}{S_x^2+Q_d},\ R_9 = \frac{S_x^2}{S_x^2+Q_a}$$

Where $Q_i\ (i=1,2,3)$ are the quartiles, the three points dividing the whole distribution into four equal parts. Further the functions of quartiles are, the inter quartile range, $Q_r = Q_3 - Q_1$, the semi-quartile range $Q_d = \frac{Q_3-Q_1}{2}$ and the quartile average $Q_a = \frac{Q_3+Q_1}{2}$.

Khan and Shabbir (2013) proposed the following estimator using correlation coefficient and the third quartile of the auxiliary variable as,

$$\hat{S}_{10}^2 = s_y^2\left[\frac{S_x^2\rho+Q_3}{s_x^2\rho+Q_3}\right]$$

The expressions for the bias and Mean Squared Error of the estimator up to the first order of approximations respectively are,

$$B(\hat{S}_{10}^2) = \gamma S_y^2 R_{10}\left[R_{10}(\lambda_{04}-1)-(\lambda_{22}-1)\right]$$

$$MSE(\hat{S}_{10}^2) = \gamma S_y^4\left[(\lambda_{40}-1)+R_{10}^2(\lambda_{04}-1)-2R_{10}(\lambda_{22}-1)\right] \tag{2.10}$$

Where, $R_{10} = \frac{S_x^2\rho}{S_x^2\rho+Q_3}$

Yadav et al. (2014), utilizing the correlation coefficient of the inter-quartile range of auxiliary variable proposed the following estimator as,

$$\hat{S}_{11}^2 = s_y^2\left[\frac{S_x^2\rho+Q_r}{s_x^2\rho+Q_r}\right]$$

The bias and Mean Squared Error of the above estimator up to the first order of approximations respectively are,

$$B(\hat{S}_{11}^2) = \gamma S_y^2 R_{11}\left[R_{11}(\lambda_{04}-1)-(\lambda_{22}-1)\right]$$

$$MSE(\hat{S}_{11}^2) = \gamma S_y^4\left[(\lambda_{40}-1)+R_{11}^2(\lambda_{04}-1)-2R_{11}(\lambda_{22}-1)\right] \tag{2.11}$$

Where, $R_{11} = \frac{S_x^2\rho}{S_x^2\rho+Q_r}$.

3. Proposed Estimator

Here, an improved ratio estimator of population variance is being suggested in the light of estimators proposed by Yadav et al. (2014) and Prasad (1989) as,

$$t = \kappa s_y^2 \left[\frac{S_x^2 \rho + Q_r}{s_x^2 \rho + Q_r} \right], \tag{3.1}$$

where κ is a characterizing scalar to be determined such that the MSE of the proposed estimator t is minimized.

To obtain the bias and Mean squared error of the proposed estimator, we wish to define

$s_y^2 = S_y^2(1+e_0)$ and $s_x^2 = S_x^2(1+e_1)$ such that $E(e_i) = 0$ for $(i = 0,1)$ and $E(e_0^2) = \frac{1-f}{n}(\lambda_{40} - 1)$, $E(e_0^2) = \frac{1-f}{n}(\lambda_{04} - 1)$, $E(e_0 e_1) = \frac{1-f}{n}(\lambda_{22} - 1)$.

The proposed estimator t can be written in terms of ε_i's ($i = 0,1$), as

$$t = \kappa S_y^2 (1+e_0)(1+R_{11}e_1)^{-1}$$

Expanding the right hand side of above equation and considering the terms in ε_i's up to the first degree of approximation, we get:

$$t = \kappa S_y^2 (1 + e_0 - R_{11}e_1 - R_{11}e_0 e_1 + R_{11}^2 e_1^2)$$

After subtracting the population variance S_y^2 of study variable on both the sides of above equation, we have,

$$t - S_y^2 = \kappa S_y^2 (1 + e_0 - R_{11}e_1 - R_{11}e_0 e_1 + R_{11}^2 e_1^2) - S_y^2 \tag{3.2}$$

The bias of proposed estimator t is obtained by taking expectations on both sides of (3.2) and putting the values of different expectations, as:

$$B(t) = \lambda\kappa S_y^2 [R_{11}^2(\lambda_{04} - 1) - R_{11}(\lambda_{22} - 1)] + S_y^2(\kappa - 1) \tag{3.3}$$

where $\lambda = \frac{(1-f)}{n}$.

The mean squared error of the proposed estimator t is obtained by squaring both sides of (3.2), simplifying and taking expectation on both sides, up to the first order of approximation as,

$$MSE(t) = S_y^4[\kappa^2\lambda(\lambda_{40} - 1) + (3\kappa^2 - 2\kappa)R_{11}^2\lambda(\lambda_{04} - 1) - 2(2\kappa^2 - \kappa)R_{11}\lambda(\lambda_{22} - 1) + (\kappa - 1)^2] \tag{3.4}$$

$MSE(t)$ is minimum for,

$$\kappa = \frac{A}{B} \tag{3.5}$$

where,

$A = 1 + R_{11}^2\lambda(\lambda_{04} - 1) - R_{11}\lambda(\lambda_{22} - 1)$ and

$B = 1 + \lambda(\lambda_{40} - 1) + 3R_{11}^2\lambda(\lambda_{04} - 1) - 4R_{11}\lambda(\lambda_{22} - 1)$

The minimum MSE of the estimator, t, for this optimum value of κ, is:

$$MSE_{\min}(t) = S_y^4\left[1 - \frac{A^2}{B}\right] \tag{3.6}$$

4. Efficiency Comparison

The proposed estimator t performs better than the estimator t_0 in the sense having lesser mean squared error under the condition:

$$MSE_{\min}(t) - V(t_0) = S_y^2\left[1 - \frac{A^2}{B} - \lambda(\lambda_{40} - 1)\right] < 0, \text{ if } \quad \frac{A^2}{B} + \lambda(\lambda_{40} - 1) > 1 \tag{4.1}$$

The proposed estimator in (3.1) will perform better than the estimator (2.3), under the condition if:

$$MSE_{\min}(t) - MSE(t_R) = S_y^2\left[1 - \frac{A^2}{B} - \lambda\{(\lambda_{40} - 1) + (\lambda_{04} - 1) - 2(\lambda_{22} - 1)\}\right] < 0, \text{ if}$$

$$\frac{A^2}{B} + \lambda\{(\lambda_{40} - 1) + (\lambda_{04} - 1) - 2(\lambda_{22} - 1)\} > 1 \tag{4.2}$$

The proposed estimator t has more efficiency as compared to all other estimators $\hat{S}_i^2\,(i = 1, 2, ..., 11)$ mentioned in this manuscript under the condition if:

$$MSE_{\min}(t) - MSE(\hat{S}_i^2) = S_y^2\left[1 - \frac{A^2}{B} - \lambda\{(\lambda_{40} - 1) + R_i^2(\lambda_{04} - 1) - 2R_i(\lambda_{22} - 1)\}\right] < 0, \; (i = 1, 2, ..., 11)$$

$$\text{if } \frac{A^2}{B} + \lambda\{(\lambda_{40} - 1) + R_i^2(\lambda_{04} - 1) - 2R_i(\lambda_{22} - 1)\} > 1, \tag{4.3}$$

5. Numerical Illustration

Following populations have been considered to examine the performances of different estimators of population variance,

Population I: Italian bureau for the environment protection-APAT Waste 2004

Y: Total amount (tons) of recyclable-waste collection in Italy in 2003.

X: Total amount (tons) of recyclable-waste collection in Italy in 2002.

$N = 103$, $n = 40$, $\bar{Y} = 626.2123$, $\bar{X} = 557.1909$, $\rho = 0.9936$, $S_y = 913.5498$, $C_y = 1.4588$, $S_x = 818.1117$, $C_x = 1.4683$, $\lambda_{04} = 37.3216$, $\lambda_{40} = 37.1279$, $\lambda_{22} = 37.2055$, $Q_1 = 142.9950$, $Q_3 = 665.6250$, $Q_r = 522.6300$, $Q_d = 261.3150$, $Q_a = 404.3100$.

Population II: Italian bureau for the environment protection-APAT Waste 2004

Y: Total amount (tons) of recyclable-waste collection in Italy in 2003.

X: Number of inhabitants in 2003.

$N = 103$, $n = 40$, $\bar{Y} = 62.6212$, $\bar{X} = 556.5541$, $\rho = 0.7298$, $S_y = 91.3549$, $C_y = 1.4588$, $S_x = 610.1643$, $C_x = 1.0963$, $\lambda_{04} = 17.8738$, $\lambda_{40} = 37.1279$, $\lambda_{22} = 17.2220$, $Q_1 = 259.3830$, $Q_3 = 628.0235$, $Q_r = 368.6405$, $Q_d = 184.3293$, $Q_a = 443.7033$.

Population III: Murthy (1967)

Y: Output for 80 factories in a region.

X: Fixed capital.

$N = 80$, $n = 20$, $\bar{Y} = 51.8264$, $\bar{X} = 11.2646$, $\rho = 0.9413$, $S_y = 18.3549$, $C_y = 0.3542$, $S_x = 8.4563$, $C_x = 0.7507$, $\lambda_{04} = 2.8664$, $\lambda_{40} = 2.2667$, $\lambda_{22} = 2.2209$, $Q_1 = 5.1500$, $Q_3 = 16.975$, $Q_r = 11.825$, $Q_d = 5.9125$, $Q_a = 11.0625$.

Population IV: Singh and Cahudhary

The population consists of 70 wheat farms in 70 villages in certain region of India and the variables under considerations are defined as:

Y = area under wheat crop (in acres) during 1974,

X = area under wheat crop (in acres) during 1973,

$N = 70$, $n = 25$, $\bar{Y} = 96.7000$, $\bar{X} = 175.2671$, $\rho = 0.7293$, $S_y = 60.7140$, $C_y = 0.6254$, $S_x = 140.8572$, $C_x = 0.8037$, $\lambda_{04} = 7.0952$, $\lambda_{40} = 4.7596$, $\lambda_{22} = 4.6038$, $Q_1 = 80.1500$, $Q_3 = 225.0250$, $Q_r = 144.8750$, $Q_d = 72.4375$, $Q_a = 152.5875$.

Table 1. Comparison of Bias and Mean square error of different estimators

Estimator	*Bias*				*MSE*			
Population	I	II	III	IV	I	II	III	IV
$\hat{S}_1^2$	2420.6810	135.9827	10.4399	364.3702	67038384403	35796605	3850.1552	1415839
$\hat{S}_2^2$	2379.9609	135.8179	9.2918	363.9722	670169790	35796503	3658.4051	1414994
$\hat{S}_3^2$	2422.3041	135.9929	10.7222	364.4139	670393032	35796611	3898.5560	1415931
$\hat{S}_4^2$	2393.4791	135.8334	8.8117	363.8627	670240637	35796512	3580.8342	1414762
$\hat{S}_5^2$	2259.9938	133.4494	8.1749	359.3822	669558483	35795045	3480.5516	1427990
$\hat{S}_6^2$	1667.7818	129.8456	3.9142	350.4482	667000531	35792955	2908.6518	1408858
$\hat{S}_7^2$	1829.6315	132.3799	5.5038	355.3634	667623576	35794395	3098.4067	1419946
$\hat{S}_8^2$	2125.7591	134.1848	7.8275	359.8641	668911625	35795495	3427.1850	1429077
$\hat{S}_9^2$	1963.6570	131.6458	5.7705	354.8875	668182833	35793951	3133.3256	1418424
$\hat{S}_{10}^2$	1663.3086	127.6040	3.6276	348.1975	666910707	35791562	2878.5603	1398150
$\hat{S}_{11}^2$	1114.184	80.1433	3.9396	228.1034	645476858	21882440	2240.9762	905945
t (Proposed)	-1034.243	-76.2341	-2.7865	-203.321	534782692	16409261	2080.1004	718745

6. Results and Conclusions

This paper deals with the estimation of population variance using improved ratio type estimator. An efficient estimator of population variance using coefficient of correlation and the inter quartile range of the auxiliary variable has been proposed. Up to the first degree of approximation, the expressions for the bias and mean square error of the proposed estimator have been obtained. The optimum value of the characterizing scalar kappa, which minimizes the mean squared error of the proposed estimator, is also obtained. Further the minimum value of the mean square error for this optimum value of kappa has also been obtained. It has been proved theoretically as well as empirically that the proposed estimator performs much better than all of the other mentioned estimators of population variance in the sense of having lesser *Bias* and *MSE*. It is of worth to be mentioned that the knowledge regarding the correlation coefficient ρ should be available in advance. This knowledge of correlation coefficient is either available in advance (generally) or it is obtained from prior studies like pilot surveys etc. In case if we do not have prior knowledge of correlation coefficient, then it in the expression of estimator is replaced by its estimate and there is no effect on the mean squared error the estimator. Therefore it is strongly recommended that the proposed estimator should be preferred over the estimators mentioned in this manuscript for the estimation of population variance under simple random sampling scheme.

ACKNOWLEDGEMENTS

The authors are very much thankful to the editor in chief of AJOR and the anonymous learned referees for critically examining the manuscript and giving the valuable suggestions for further improvement in the earlier draft.

REFERENCES

[1] Isaki, C, T., Variance estimation using auxiliary information, *Journal of American Statistical Association*, 78, 117- 123 (1983).

[2] Kadilar, C. and Cingi, H., Improvement in variance estimation using auxiliary information, *Hacettepe Journal of mathematics and Statistics*, 35, 111-15 (2006).

[3] Kadilar, C. and Cingi, H., Ratio estimators for population variance in simple and Stratified sampling, *Applied Mathematics and Computation,* 173, 1047-1058 (2006).

[4] Khan, M. and Shabbir, J., A Ratio Type Estimator for the Estimation of Population Variance using Quartiles of an Auxiliary Variable, *Journal of Statistics Applications & Probability*, 2, 3, 319-325 (2013).

[5] Murthy, M. N., Sampling Theory and Methods, Statistical Publishing Society Calcutta, India, (1967).

[6] Singh, D. and Chaudhary, F. S., Theory and analysis of sample survey designs, New-Age International Publisher, (1986).

[7] Subramani, J. and Kumarapandiyan, G., Variance estimation using quartiles and their functions of an auxiliary variable, *International Journal of Statistics and Applications,* 2, 67-72 (2012).

[8] Upadhyaya, L. N. and Singh, H. P., Use of auxiliary information in the estimation of population variance, mathematical forum, 4, 33-36 (1983).

[9] http://www.osservatorionalerifiuti.it/ElencoDocPub.asp?A_TipoDoc=6.

[10] Yadav, S.K., Mishra, S.S. and Gupta, S., Improved Variance Estimation Utilizing Correlation Coefficient and Quartiles of an Auxiliary Variable, Communicated to American Journal of Mathematics and Mathematical Sciences, (2014).