Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/040102

An Improved Estimator of Population Variance using known Coefficient of Variation

Sheela Misra, Dipika Kumari and Dharmendra Kumar Yadav*

Department of Statistics, University of Lucknow, Lucknow, India 226007.

Received: 27 jun. 2016, Revised: 15 Nov. 2016, Accepted: 22 Nov. 2016
Published online: 1 Jan. 2017

Abstract: In the present article, an improved estimator $(s^2_{y_k})$ over usual unbiased estimator of population variance (s^2_y) is proposed by using known coefficient of variation (C_y) of the study variable y. Asymptotic expression for its bias and mean square error (MSE) have been obtained. For more practical utility the study of proposed estimator under estimated optimum value of k has also been carried out. A comparative study has been made between the proposed estimator and the conventional estimator. Numerical illustration is also given in support of the present study.

Keywords: Bias, Coefficient of Variation, Efficiency, Mean Square Error, Simple Random Sampling.

1 Introduction

The problem of estimating population variance arises in many practical situations like agricultural, biological and medical studies [Bland and Altman (1986)] [3]. The problem has been well dealt in literature in simple random sampling. This problem is considered by Wakimoto (1971) [20] in stratified sampling. Variance estimation in PPS and general sampling design was also considered by Das and Tripathi(1977) [6], Liu (1974) [13], Chaudhury (1978) [5], Mukhopadhyay (1978) [14], Swain and Mishra(1994) [18]. Estimation of population variance under super-population models has been carried out by Mukhopadhyay (1982) [15], Padmawar and Mukhopadyay (1981) [16]. Taking advantage of high correlation between study and auxiliary variables, Isaki (1983) [10] proposed ratio and regression type estimators of population variance. Biradar and Singh (1998) [2], Agrawal and Panda (1999) [1] explored their discussion under prediction approach. The assumption of a known coefficient of variation is actually common in many agricultural, biological and industrial applications. If the situation arises that the population mean is proportional to the population standard deviation, then knowing the proportionality constant is equivalent to knowing the population coefficient of variation. For a more thorough discussion of this concept we suggest Gleser and Healy(1976) [8]. For estimation of finite population variance we assume that the finite population consists of N identifiable units $(U_1, U_2, U_3, ..., U_N)$ taking the values $(Y_1, Y_2, Y_3, ..., Y_N)$ on study variable y. Let

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i,\ \sigma_y{}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

and

$$C_y = \frac{\sigma_y}{Y}$$

be the population mean, variance and coefficient of variation of y respectively. Similarly

$$\mu_r = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^r,$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}\ ,\ \gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}},\ \beta_2 = \frac{\mu_4}{\mu_2^2}\ ,\ \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

* Corresponding author e-mail: dipikascholar@gmail.com

Let,

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

be the sample mean and variance of y based on a sample s =(1,2,..,n) taken from U by simple random sampling.

The proposed estimator, using known coefficient of variation of study variable y for the estimation of population variance σ_y^2 is

$$s_{y_k}^2 = s_y^2\left(\bar{y}^2\frac{C_y^2}{s_y^2}\right)^k \tag{1}$$

where k is the characterizing scalar to be chosen suitably.

2 Bias and mean square error of Proposed Estimator

For the sake of simplicity we are assuming that the population size N is large as compared to sample size n so that finite population correction (fpc) is ignored.

Let, $\bar{y} = \bar{Y}(1+e_0)$, $s_y^2 = \sigma_y^2(1+e_1)$, $E(e_0) = E(e_1) = 0$, $E(e_0^2) = \frac{C_y^2}{n}$, $E(e_1^2) = \frac{\gamma_{2y}+2}{n}$ $E(e_0e_1) = \frac{\gamma_{1y}C_y}{n}$ From (1) we have

$s_{y_k}^2 = s_y^2\left(\bar{y}^2\frac{C_y^2}{s_y^2}\right)$

Now expressing proposed estimator in terms of ei's

$s_{y_k}^2 = \sigma_y^2(1+e_1)\left(\bar{Y}^2(1+e_0)^2\right)\left(\frac{C_y^2}{\sigma_y^2(1+e_1)}\right)^k$

$= \sigma_y^2[1+2ke_0-(k-1)e_1+k(2k-1)e_0^2-2k(k-1)e_0e_1+\frac{k(k-1)}{2}e_1^2+.........]$

$$(s_{y_k}^2-\sigma_y^2) = \sigma_y^2[2ke_0-(k-1)e_1+k(2k-1)e_0^2-2k(k-1)e_0e_1+\frac{k(k-1)}{2}e_1^2+.......] \tag{2}$$

On taking expectation on both the sides of (2) and using first order of approximation, we get the bias of proposed estimator s_y^2 as

$$Bias(s_{y_k}^2) = (s_{y_k}^2-\sigma_y^2) = \frac{\sigma_y^2}{n}[2k(k-1)C_y^2-4k(k-1)\gamma_1C_y+k(k-1)(\gamma_2y+2)] \tag{3}$$

Again, squaring (2) both sides and taking expectation, we have the mean square error of $s_{y_k}^2$ up to first order of approximation to be

$MSE(s_{y_k}^2) = E(s_{y_k}^2-\sigma_y^2)^2$

$= \sigma_y^4[4kE(e_0^2)+(k-1)^2E(e_1^2)-4k(k-1)E(e_0e_1)]$

$$= \frac{\sigma_y^4}{n}(\gamma_2y+2)+\frac{\sigma_y^4}{n}[k^2(4C_y^2-4\gamma_1yC_y+\gamma_2y+2)+2k(2\gamma_1yC_y-\gamma_2y-2)] \tag{4}$$

The optimum value of k which minimizes the mean square error of $s_{y_k}^2$ in (4) is given by

$$k_0 = \frac{-(2\gamma_{1y}C_y-\gamma_{2y}-2)}{(4C_y^2-4\gamma_{1y}C_y+\gamma_{2y}+2)} \tag{5}$$

The minimum value of mean square error of proposed estimator $s_{y_k}^2$ for k_0 is given by

$$MSE(s_{y_k}^2)_{min} = \frac{\sigma_y^4}{n}(\gamma_{2y}+2)-\frac{\sigma_y^4}{n}\left(\frac{(2\gamma_{1y}C_y-\gamma_{2y}-2)^2}{4C_y^2-4\gamma_{1y}C_y+\gamma_{2y}+2}\right) \tag{6}$$

3 Estimator with Estimated optimum Value of k

An alternative procedure for calculating mean square error when values of $\gamma_1 y$ and $\gamma_2 y$ or their good guessed values are not available is to replace these values involved in the optimum k by their estimates $\hat{\gamma}_{1y}$ and $\hat{\gamma}_{2y}$ based on sample values and get the estimated optimum value of k denoted by $\hat{k}$ as

$$\hat{k} = -\frac{(2\hat{\gamma}_{1y}C_y - \hat{\gamma}_{2y} - 2)}{(4C_y^2 - 4\hat{\gamma}_{1y}C_y + \hat{\gamma}_{2y} + 2)} \tag{7}$$

Where, $\hat{\gamma}_{1y} = \frac{\hat{\mu}_3}{\hat{\mu}_2^{\frac{3}{2}}}$ and $\hat{\gamma}_{2y} = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} - 3$ with $\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}), \hat{\mu}_2 = s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4$

Thus, replacing k by estimated optimum value of k in the estimator $s^2_{y_k}$ in (1),we get for wider practical utility of the estimator based on the estimated optimum value $\hat{k}$ is given as

$$s^2_{y_{\hat{k}}} = s_{y^2}(\bar{y}^2 \frac{C_y{}^2}{s_y^2})^{\hat{k}} \tag{8}$$

To find the bias and mean square error of $s^2_{y_{\hat{k}}}$, let

$\hat{\mu}_3 = \mu_3(1+e_2), \hat{\mu}_4 = \mu_4(1+e_3)$
Along with

$$\bar{y} = \bar{Y}(1+e_0), s_y^2 = \sigma_y^2(1+e_1)$$

$$\hat{k} = -\frac{2\frac{\mu_3(1+e_2)}{\sigma_y^3(1+e_1)^{\frac{3}{2}}}C_y - \frac{\mu_4(1+e_3)}{\sigma_y^4(1+e_1)^2} + 1}{4C_y^2 - 4\frac{\mu_3(1+e_2)}{\sigma_y^3(1+e_1)^{\frac{3}{2}}}C_y + \frac{\mu_4(1+e_3)}{\sigma_y^4(1+e_1)^2} - 1}$$

$$= \frac{2\gamma_1 yC_y(1+e_2-\frac{3}{2}e_1-\frac{3}{2}e_1e_2+\frac{15}{2}e_1^2-.....)-(\gamma_2 y+3)(1+e_3-2e_1-2e_1e_3+3e_1^2-.......)+1}{4C_y^2-4\gamma_1 yC_y(1+e_2-\frac{3}{2}e_1-\frac{3}{2}e_1e_2+\frac{15}{2}e_1^2-......)+(\gamma_2 y+3)(1+e_3-2e_1-2e_1e_3+3e_1^2-)-1}$$

$$= \left(\frac{2\gamma_1 yC_y - \gamma_2 y - 2}{4C_y^2 - 4\gamma_1 yC_y + \gamma_2 y + 2}\right)\left(1 + \frac{2\gamma_1 yC_y(e_2 - \frac{3}{2}e_1 - \frac{3}{2}e_1e_2 + \frac{15}{2}e_1^2 - ......) - (\gamma_2 y+3)(e_3 - 2e_1 - 2e_1e_3 + 3e_1^2 - ....)}{2\gamma_1 yC_y - \gamma_2 y - 2}\right)$$
$$* \left(1 + \frac{(\gamma_2 y+3)(e_3 - 2e_1 - 2e_1e_3 + 3e_1^2 - ....) - 4\gamma_1 yC_y(e_2 - \frac{3}{2}e_1 - \frac{3}{2}e_1e_2 + \frac{15}{2}e_1^2 - ....)}{4C_y^2 - 4\gamma_1 yC_y + \gamma_2 y + 2}\right)^{-1}$$
$$= -\left(\frac{2\gamma_1 yC_y - \gamma_2 y - 2}{4C_y^2 - 4\gamma_1 yC_y + \gamma_2 y + 2}\right)\left\{1 + \frac{2\gamma_1 yC_y(e_2 - \frac{3}{2}e_1 - \frac{3}{2}e_1e_2 + \frac{15}{2}e_1^2 - ......) - (\gamma_2 y+3)(e_3 - 2e_1 - 2e_1e_3 + 3e_1^2 - ......)}{2\gamma_1 yC_y - \gamma_2 y - 2}\right\}$$
$$+ \frac{2\gamma_1 yC_y - \gamma_2 y - 2}{4C_y^2 - 4\gamma_1 yC_y + \gamma_2 y + 2}\left\{\frac{(\gamma_2 y+3)(e_3 - 2e_1 - 2e_1e_3 + 3e_1^2 - .....) - 4\gamma_1 yC_y(e_2 - \frac{3}{2}e_1 - 3/2e_1e_2 + \frac{15}{2}e_1^2 - ......)}{4C_y^2 - 4\gamma_1 yC_y + \gamma_2 y + 2}\right\} \tag{9}$$

Substituting $\bar{y} = \bar{Y}(1+e_0)$, $s_y^2 = \sigma_y^2(1+e_1)$ and $\hat{k}$ from (9) in (2), we have

$$\left(s^2_{y_{\hat{k}}} - \sigma_y^2\right) = \sigma_y^2\left[e_1 - \frac{2\gamma_{1y}C_y - \gamma_{2y} - 2}{4C_y^2 - 4\gamma_1 yC_y + \gamma_{2y} + 2}\left\{2e_0 - e_1 - e_0^2 + 2e_0e_1 - \frac{e_1^2}{2} + ...\right\}\right] \tag{10}$$

Taking expectation of (10) and ignoring terms of $e_i's$ greater than power two, we can easily check that the bias of $s^2_{y_k}$ is of $O(\frac{1}{n})$, hence the bias of $s^2_{y_k}$ negligible for large value of n, that is the estimator $s^2_{y_k}$ is approximately unbiased estimator of population variance.

Now squaring and taking expectation of (10), we have
$MSE(s^2_{y_{\hat{k}}}) = \sigma_y^4[e_1 - \frac{2\gamma_1 yC_y - \gamma_{2y} - 2}{(4C_y^2 - 4\gamma_1 yC_y + \gamma_{2y} + 2)}2e_0 - e_1]^2$

$$= \frac{\sigma_y^4}{n}(\gamma_{2y} + 2) - \frac{\sigma_y^4}{n}\left\{\frac{(2\sigma_{1y}C_y - \gamma_{2y} - 2)^2}{(4C_y^2 - 4\gamma_1 yC_y + \gamma_{2y} + 2)}\right\} \tag{11}$$

Which is same as mean square error for the optimum value of k that is estimator $s^2_{y_k}$ based on estimated value of optimum k also has same mean square error as that of the estimator $s^2_{y_k}$ based on optimum k.

4 Theoretical Efficiency Comparison

We compare the proposed estimator $s^2_{y_k}$ with respect to usual unbiased estimator of population variance ${s_y}^2$ and the condition for which the proposed estimator will efficient is given by

$MSE(s^2_{y_k}) - MSE(s^2_y) < 0$

$$2\gamma_1 y C_y < \gamma_{2y} + 2 \tag{12}$$

5 Numerical Illustration

For numerical illustration, we consider the two data as

(1) Data given in Cochran(1977, pg34) dealing with the weekly expenditure of family on food(y) of 33 low-income families, the required values are calculated from data are

$n = 33, \bar{y} = 27.49, \sigma^2_y = 99.613033, C_y = 0.363036, \gamma_{1y} = 1.4651, \gamma_{2y} = 2.7146$

Using above values, we have

$$MSE(s^2_y) = 1417.63112 \tag{13}$$

$$MSE\left(s^2_{y_k}\right)_{min} = 1065.13127 \tag{14}$$

The percent relative efficiency (PRE) of the proposed estimator over the usual unbiased estimator for population variance is 133%.

(2) Generated population from normal distribution by using simulation technique through R software. The Description of this data is as follows $Y = N(5,10), n = 5000, \bar{y} = 4.95, \sigma^2_y = 99.38, C_y = 2.014, \gamma_1 y = 0.039, \gamma_2 y = -0.041$

Using above values, we have

$$MSE(s^2_y) = 3.87 \tag{15}$$

$$MSE(s^2_{y_k})_{min} = 3.52 \tag{16}$$

The percent relative efficiency (PRE) of the proposed estimator over the usual unbiased estimator for population variance is 109%. Hence from both the data set we can conclude that proposed estimator is better the usual unbiased estimator for population variance.

6 Concluding Remarks

(a) From (6) it is observed that the proposed estimator will perform better than usual unbiased estimator of population variance.

(b) The estimator $s^2_{y_k}$ with optimum value k_0 and the estimator based on estimated optimum $\hat{k}$ have same mean square error given by

$MSE(s^2_{y_{\hat{k}}}) = MSE({s_y}^2)_{min} = \frac{\sigma_{y^4}}{n}(\gamma_2 y + 2) - \frac{\sigma^4_y}{n}\frac{(2\gamma_1 y C_y - \gamma_2 y - 2)^2}{(4C^2_y - 4\gamma_1 y C_y + \gamma_2 y + 2)}$

(c) For normal population (,i.e.for $\gamma_{1y} = 0$ and $\beta_2 y = 3$),the optimum value of k from(5),reduces to

$k = \frac{1}{1+2{C_y}^2}$

For which mean square error of proposed estimator becomes

$MSE(s^2_{y_k}) = \frac{\sigma_y^4}{n}\frac{(2C_y^2)}{(1+C_y^2)}$

showing that the proposed estimator is more efficient than usual unbiased estimator in normal parent population also.

(d) If for any dataset (12) holds then proposed estimator will be better than the usual unbiased estimator of population variance.

(e) From numerical illustration (1) it is observed that proposed estimator is 133% more efficient than the usual unbiased estimator for population variance.

(f) From simulation data analysis it is observed that proposed estimator is 109% more efficient than the usual unbiased estimator for population variance.

Acknowledgement

The authors are grateful to the anonymous referee for the valuable suggestion and helpful comments that improved this paper. Authors are also thankful to editor in chief for his/her support and suggestions.

References

[1] Agrawal, M. C. and Panda, K. B., A predictive justification for variance estimation using auxiliary information, Jour. Ind. Soc. Ag. Stat., 1999, 2, 52, 192200.

[2] Biradar, R. S. and Singh, H. P. , Predictive estimation of finite population variance, Cal. Statist. Assoc. Bull., 1998, 48, 229-235.

[3] Bland, J. M. and Altman, D. G. , Statistical method for assessing agreement between two methods of clinical measurement , Lance, 1986, 1, 8476, 307310.

[4] Cochran, W.G., Sampling Techniques , Wiley Eastern Private Limited, New Delhi., 1963, Second Edition , 307-310.

[5] Chaudhury, A., On estimating the variance of a finite population. , Metrika , 1978, 25, 66-67..

[6] Das, A. K. and Tripathi, T. P. , Admissible estimators for quadratic forms in finite populations., Bull. Inter. Stat. Inst., 1977, Second Edition , 47,4, 132-135.

[7] Das, A. K. and Tripathi, T. P. , Use of auxiliary information in estimating the finite population variance., Sankhya, 1978, 4, c , 139-148.

[8] Gleser, L. J., and Healy, J. D. , Estimating the Mean of a Normal Distribution with Known Coefficient of Variation , Journal of the American Statistical Association, 1976, 71,977-981.

[9] Gupta, S. and Shabbir, J. , Variance estimation in simple random sampling using auxiliary information, Hacettepe Journal of Mathematics and Statistics, 2008, 37, 57-67.

[10] Isaki, C.T., Variance estimation using Auxiliary Information , Jour. Amer. Statist. Asssoct , 1983, 78,117-123.

[11] Kadilar, C. and Cingi, H. , Improvement in variance estimation using auxiliary information, Hacettepe Journal of Mathematics and Statistics, 2006a , 35, 1, 111-115.

[12] Kadilar, C. and Cingi, H. , Ratio estimators for population variance in simple and stratified sampling , Applied Mathematics and Computation, 2006b , 1731047-1058.

[13] Liu, T. P., A generalized unbiased estimator for the variance of a finite population, Sankhya, 1974 , c,3623-32.

[14] Mukhopadhyay, P., Estimating a finite population variance under a super population model , Metrika , 1978, 25115-122.

[15] Mukhopadhyay, P., Optimum Strategies for estimating the variance of a finite population under a super population model , Metrika , 1982, 29143158.

[16] Padmwar, V. R. and Mukhopadhya, Y. P., Estimation of symmetric functions of a finite population , Metrika , 1981, 3131-97.

[17] Sukhatme P.V., Sukhatme B.V., Sukhatme, S. and Ashok, C., Sampling Theory of Surveys with Applications , Iowa State University Press, Ams , 1984.

[18] Swain, A. K. P. C. and Mishra, G., Estimation of population variance under unequal probability sampling , Sankhya , 1994, B, 56374-384.

[19] Tripathi, T. P., Singh, H. P. and Upadhyaya, L. N., A general method of estimation and its application to the estimation of coefficient of variation , Statistics in Transition , 2002, 5, 61081-1102.

[20] Wakimoto, K. , Stratified random sampling (I): Estimation of Population variance, Ann. Inst. Stat. Math. , 1971, 23, 233-252.

[21] Wolter, K. M. , Introduction to variance estimation , New York, NY: Springer- Verlag .

Sheela Misra is Professor and Head of the Department of Statistics, University of Lucknow, Lucknow, India. She is Ph.D. Supervisor of Dipika and Dharmendra Kumar Yadav. She has a lot of contribution in the field of Sampling Theory, Gender Statistics and Biostatistics. She has successfully organized many National and International Conferences and many Training Programmes in the Department.

Dipika Kumari is working as Research Scholar at the Department of Statistics, University of Lucknow. She has presented many research papers in National and International Conferences. Her area of research is Sampling theory

Dharmendra Kumar Yadav is Research Scholar at Department of Statistics, University of Lucknow. He has presented many research papers in National and International Conferences. His area of research is Non-Sampling Errors.