ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 41 (2021), 123-132

Hierarchic Estimation for Population Variance

K.B. Panda[1] and P. Das[2]
[Received on May, 2020. Accepted on July, 2021]

ABSTRACT

We, in this paper, extending the fabric of hierarchic estimation for mean introduced by Agrawal and Sthapit (1997) and carried forward successively by Panda and Sahoo (2015) and Panda and Das (2018) to the dimension of variance estimation, propose a new ratio estimator for the estimation of population variance using simple random sampling without replacement (SRSWOR) scheme. The novelty of the proposed estimator of order k is that it is predictive in form under the assumption of balanced sampling. Conditions under which the proposed estimator with optimal k fares better than the usual variance estimator and the ratio estimator due to Isaki (1983) have been arrived at. In addition to this, the supremacy of the proposed estimator over the one due to Kadilar and Cingi (2006) is established numerically. Theoretical findings are supported by numerical illustration.

1. Introduction

The need of estimating population variance with greater accuracy has led to the emergence of ratio, product and regression estimators of population variance. By different authors, there are several estimators constructed resorting to ratio, product and regression methods of estimation in the literature to address the problem of estimating population variance. Kadilar and Cingi (2006), Gupta, and Shabbir (2008), Adichwal, Raghav and Singh (2015), Adichwal, Kumar and Singh (2018) have proposed estimators of population variance using auxiliary information. We, in this paper, following the structural framework which resembles hierarchic estimation, construct a new ratio estimator for estimating population variance.

✉ : P. Das
Email: prabhatadas91@gmail.com

Extended author information available after reference list of the article.

Consider a population $U = \{U_1, U_2 \ldots\ldots U_i, \ldots\ldots, U_N\}$ consisting of N units. Let y and x be the study and auxiliary variables with population means $\bar{Y}$ and $\bar{X}$, respectively. Let a sample of size n be drawn from this population according to simple random sampling without replacement (SRSWOR).

Let $s_y^2 = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 / n$ and $s_x^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 / n$ be the sample variances and $s_y^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 / n$ and $s_x^2 = \sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2 / N$ be the population variances of y and x, respectively. Let $C_y = S_y / \bar{Y}$ and $C_x = S_x / \bar{X}$ be the coefficients of variation of y and x, respectively, and ρ_{xy} be the correlation coefficient between y and x. We assume that parameters of x such as $\bar{X}, S_x^2, C_x$ etc. are known.

Let $e_0 = \left(s_y^2 - S_y^2\right)$ and $e_1 = \left(s_x^2 - S_x^2\right) / S_x^2$ be such that $E(e_i) = 0$, $(i = 0, 1)$. Moreover, to the first order of approximation, we have $E\left(e_0^2\right) = \theta\left(\beta_2(y) - 1\right), E\left(e_1^2\right) = \theta\left(\beta_2(x) - 1\right)$ and $E\left(e_0 e_1\right) = \theta\left(\delta_{22} - 1\right)$,

where $\theta = \left(\frac{1}{n} - \frac{1}{N}\right), \delta_{pq} / \left(\mu_{20}^{p/2}\mu_{02}^{q/2}\right), \mu_{pq} = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^p \left(x_i - \bar{X}\right)^q / N$, and $\beta_2(y) = \mu_{40} / \mu_{20}^2, \beta_2(x) = \mu_{40} / \mu_{02}^2$ are the coefficients of kurtosis of y and x, respectively.

The traditional ratio estimator for population variance due to Isaki (1983) is given by

$$S_R^2 = s_y^2\left(\frac{S_x^2}{s_x^2}\right) \tag{1.1}$$

The bias and MSE of this estimator, to the first order of approximation, are given by

$$Bias\left(S_R^2\right) = \theta S_y^2\left[\beta_2(x) - \delta_{22}\right], \tag{1.2}$$

and $$MSE\left(s_R^2\right) = \theta S_y^4\left[\left(\beta_2(y) - 1\right) + \left(\beta_2(x) - 1\right) - 2\left(\delta_{22} - 1\right)\right] \tag{1.3}$$

2. The Proposed Estimator and Justification of its Predictive Character

We propose the following ratio estimator for estimating S_y^2 as

$$s_R^{2(k)} = \left(1-\lambda^k\right)s_y^2 + \lambda^k s_R^2 \tag{2.1}$$

where λ = 1 – n / N and k is a positive number greater than 0.

Following the predictive approach due to Basu (1971) and Smith (1976), Agrawal and Panda (1999) split the population variance as

$$S_y^2 = \frac{n}{N}S_y^2 + \frac{N-n}{N}S_{ry}^2 + \frac{n(n-N)}{N^2}\left(\bar{y}-y_r\right)^2. \tag{2.2}$$

Thus the predictive format for the estimation of S_y^2 assumes the following form:

$$\hat{S}_y^2 = \frac{n}{N}s_y^2 + \frac{N-n}{N}\hat{S}_{ry}^2 + \frac{n(n-N)}{N^2}\left(\bar{y}-\hat{\bar{y}}_r\right)^2, \tag{2.3}$$

Where s_{ry}^2 and $\bar{y}_r$ are the variance and the mean in respect to the unobserved segment of the population and $\hat{s}_{ry}^2$ and $\hat{\bar{y}}_r$ are their respective predictors. Now if we use $\left(1-\lambda^k\right)s_y^2 + \lambda^k s_y^2\left(\frac{s_{rx}^2}{s_x^2}\right)$ and $\bar{y}\left(\frac{\bar{x}_r}{\bar{x}}\right)$ as implied predictors of $\hat{s}_{ry}^2$ and $\hat{\bar{y}}_r$ respectively where s_{rx}^2 and $\bar{x}_r$ that are analogous to s_{ry}^2 and $\bar{y}_r$ relate to x character, then we have, under balanced sampling, the predictor of S_y^2 as

$$\hat{S}_y^2 = \left(1-\lambda^k\right)s_y^2 + \lambda^k s_R^2$$

which points out to the fact that the proposed estimator $s_R^{2(k)}$, under balanced sampling, is endowed with predictive character.

3. Comparison of Bias and Mean Square Error of the Proposed Estimator Vis-a-Vis the Competing Estimators

The bias of the estimator $s_R^{2(k)}$, to the first order of approximation, can be worked out as

$$B\left(s_R^{2(k)}\right) = \lambda^k \theta S_y^2\left[\beta_2(x)-\delta_{22}\right]. \tag{3.1}$$

It is clear from (3.1) that the absolute value of the bias obtained above is, for $k \geq 1$, invariably less than that of the conventional ratio estimator for population variance given in (1.2).

The MSE of $s_R^{2(k)}$, to the first order of approximation, can be found as

$$MSE\left(s_R^{2(k)}\right) = \theta S_y^4 \left[\left(\beta_2(y)-1\right) + \lambda^{2k}\left(\beta_2(x)-1\right) - 2\lambda^k\left(\delta_{22}-1\right)\right] \qquad (3.2)$$

To determine k optimally in order to minimize (3.2), we have

$$\frac{\partial MSE\left(s_R^{2(k)}\right)}{\partial k} = 0$$

$$\Rightarrow \left(\lambda^{2k}\log\lambda^2\right)\left(\beta_2(x)-1\right) - 2\left(\lambda^k \log\lambda\right)\left(\delta_{22}-1\right) = 0$$

$$\Rightarrow \left(\lambda^{2k}\log\lambda^2\right)\left(\beta_2(x)-1\right) - 2\left(\lambda^k \log\lambda\right)\left(\delta_{22}-1\right)$$

$$\Rightarrow \lambda^k = \frac{\delta_{22}-1}{\beta_2(x)-1} \qquad (3.3)$$

$$\Rightarrow \lambda^{2k} = \frac{\left(\delta_{22}-1\right)^2}{\left(\beta_2(x)-1\right)^2} \qquad (3.4)$$

Now putting the values of λ^k and λ^{2k} given respectively in (3.3) and (3.4) in (3.2) we can get the minimum mean square error with optimal k which can be denoted as $MSE\left(s_R^{2(k)}\right)_{opt}$.

The proposed estimator fares better than the traditional ratio estimator for population variance due to Isaki (1983) if

$$\mathrm{MSE}\left(s_R^2\right) - \mathrm{MSE}\left(s_R^{2(k)}\right)_{opt} > 0$$

$$\Rightarrow \theta s_y^4 + \left[\beta_2(y)-1+\left(\beta_2(x)-1\right)-2\left(\delta_{22}-1\right)\right]$$
$$+ - \theta s_y^4\left[\beta_2(y)-1+\lambda^{2k}\left(\beta_2(x)-1\right)-2\lambda^k\left(\delta_{22}-1\right)\right] > 0$$

$$\Rightarrow \left[\left(\beta_2(y)-1\right)+\left(\beta_2(x)-1\right)-2\left(\delta_{22}-1\right)\right] - \left[\left(\beta_2(y)-1\right)\left(\beta_2(x)-1\right)-2\lambda^k\left(\delta_{22}-1\right)\right] > 0$$

$$\Rightarrow \left(1-\lambda^{2k}\right)\left(\beta_2(x)-1\right) - 2\left(\delta_{22}-1\right)\left(1-\lambda^k\right) > 0$$

$$\Rightarrow \left(1-\lambda^{2k}\right)\left(\beta_2(x)-1\right)-2\left(\delta_{22}-1\right)\left(1-\lambda^{k}\right)$$

$$\Rightarrow \left(1-\lambda^{2k}\right)\left(1+\lambda^{k}\right)/\left(1-\lambda^{k}\right) > 2\left(\delta_{22}-1\right)/\left(\beta_2(x)-1\right)$$

$$\Rightarrow \frac{1}{2}\left(1+\lambda^{k}\right) > \left(\delta_{22}-1\right)/\left(\beta_2(x)-1\right), \tag{3.5}$$

and it fares better than s_y^2 if

$$\mathrm{MSE}\left(s_y^2\right)-\mathrm{MSE}\left(s_R^{2(k)}\right)_{opt} > 0$$

$$\left(\delta_{22}-1\right)/\left(\beta_2(x)-1\right) > \frac{1}{2}\lambda^{k}. \tag{3.6}$$

Thus, $s_R^{2(k)}$ will perform better than both s_R^2 and s_y^2 when

$$\frac{1}{2}\lambda^{k} < \left(\delta_{22}-1\right)/\left(\beta_2(x)-1\right) < \frac{1}{2}\left(1+\lambda^{k}\right) \tag{3.7}$$

a condition which holds good in practice quite often. Under optimality of k, i.e., when (3.3) holds, the above condition reduces to

$$\frac{1}{2}\lambda^{k} < \lambda^{k} < \frac{1}{2}\left(1+\lambda^{k}\right), \tag{3.8}$$

which is invariably true as $\lambda < 1$ and $k \geq 1$, indicating the supremacy of the proposed estimator over its competitors. The bounds given in (3.6) are called the efficiency bounds, the term in the middle of (3.6) being treated as a pivotal quantity. By choosing values of the sampling fraction $f\left(=\frac{n}{N}\right)$ and hence $\lambda(=1-f)$, we have computed the following table which gives the bounds of $\left(\delta_{22}-1\right)/\left(\beta_2(x)-1\right)$ within which $s_R^{2(k)}$ (for various values of k) will be more efficient than s_R^2 and s_y^2.

Table 1: Efficiency bounds of $(\delta_{22} - 1) / (\beta_2 (x) - 1)$ for various values of f and k.

	K					
F	1	2	5	8	10	50
0.05	(0.475,0.975)	(0.451,0.951)	(0.387,0.587)	(0.332,0.532)	(0.299,0.799)	(0.038,0.538)
0.10	(0.450,0.950)	(0.405,0.905)	(0.295,0.795)	(0.215,0.715)	(0.174,0.674)	(0.003,0.503)
0.20	(0.400,0.900)	(0.320,0.820)	(0.164,0.664)	(0.084,0.584)	(0.054,0.554)	(0.000,0.500)
0.25	(0.375,0.875)	(0.281,0.781)	(0.118,0.618)	(0.050,0.550)	(0.028,0.528)	(0.000,0.500)
0.30	(0.350,0.850)	(0.245,0.745)	(0.084,0.584)	(0.028,0.528)	(0.014,0.514)	(0.000,0.500)
0.40	(0.300,0.800)	(0.180,0.680)	(0.038,0.538)	(0.008,0.508)	(0.003,0.503)	(0.000,0.500)
0.50	(0.250,0.750)	(0.125,0.625)	(0.016,0.516)	(0.002,0.502)	(0.001,0.501)	(0.000,0.500)
0.60	(0.200,0.700)	(0.080,0.580)	(0.005,0.505)	(0.000,0.500)	(0.000,0.500)	(0.000,0.500)
0.70	(0.150,0.650)	(0.045,0.545)	(0.001,0.501)	(0.000,0.500)	(0.000,0.500)	(0.000,0.500)
0.80	(0.100,0.600)	(0.020,0.520)	(0.000,0.500)	(0.000,0.500)	(0.000,0.500)	(0.000,0.500)

Table 1 can be acted as a device to locate a suitable value of k for given values of the pivotal quantity and f. Knowledge of the pivotal quantity consisting of various population parameters such as the population δ_{22} and coefficients of kurtosis, as they remain stable over a period of time, can be gathered from past survey, pilot survey, educated guess etc. For a specified value of the pivotal quantity, Table 1 provides more than one value of k which ensures better performance of $s_R^{2(k)}$ vis $-$ a′-viss$_R^2$ and s_y^2. However the optimal value of k can be arrived at from equation (3.3) provided $(\delta_{22} - 1) / (\beta_2 (x) - 1) < 1$. When an optimum value of k is not obtainable, a suitable value of k that renders $s_R^{2(k)}$ superior to s_R^2 and s_y^2 might still be found from the above table.

Here attention must be paid to the fact that when k will be zero the proposed estimator for population variance is no different from the one due to Isaki (1983).

4. Empirical Investigation

For the purpose of empirical investigation, we have considered the following three sets of data which are taken from various sources followed by Gupta and Shabbir (2008).

DATA 1: $Source: Kadilar\ and\ Cingi\ (2006)$

The data is consisted of 104 villages in the East Anatolia Region of Turkey in 1999.The variables of interest are as:

Y: The level of apple production (in 100 tones)

and

X: The number of apple trees.

For this data, we have

$N = 104$, $n = 20$, $\theta = 0.04038$,

$$\bar{Y} = 6.254, \bar{X} = 13931.683, S_y = 11.67, S_x = 23029.072,$$

$$\beta_2(y) = 16.523, \beta_2(x) = 17.516, \delta_{22} = 14.398, \lambda^k = 0.8112.$$

DATA 2: $Source: Das,\ 1988$

The data is consisted of 278 villages/towns/wards under Gajole Police Station of Malda district of West Bengal, India. The variables of interest are as:

Y: The number of agricultural laborers in 1971

and

X: The number of agricultural laborers in 1961.

For this data, we have

$N = 278$, $n = 30$, $\theta = 0.02974$, $\bar{Y} = 39.068$, $\bar{X} = 25.111$, $S_y = 56.457167$,

$S_x = 40.674797$, $\beta_2(y) = 25.8969$, $\beta_2(x) = 38.8898$, $\delta_{22} = 26.8142$, $\lambda^k = 0.6812$.

DATA 3: *Source: Cochran* , p. 325

The data is consisted of 100 blocks in a large city. The variables of interest are as:

Y : The number of persons per block

and

X : The number of rooms per block .

For this data, we have

$$N = 100, n = 10, \theta = 0.09, \bar{Y} = 101.1, \bar{X} = 58.8,$$

$$S_y = 14.6595, S_x = 7.53228, \beta_2(y) = 2.3523, \beta_2(x) = 2.2387, \delta_{22} = 1.5432, \lambda^k = 0.4385.$$

We provide the Percentage Gain in Efficiency of the proposed estimator with respect to its competitors in the following Table 1.

Table 1: Percentage gain in efficiency of the proposed estimator $s_R^{2(k)}$ with respect to s_R^2 ,s_y^2 and $\hat{S}_{KC1}^2$

Data set	**Percentage gain in efficiency of $s_R^{2(k)}$ with respect to s_y^2**	**Percentage gain in efficiency of $s_R^{2(k)}$ with respect to s_R^2**	**Percentage gain in efficiency of $s_R^{2(k)}$ with respect to $\hat{S}_{KC1}^2$**
Data 1	233.56	12.66	12.64
Data 2	240.65	52.67	52.32
Data 3	21.39	35.07	34.78

The above Table shows the supremacy of the proposed estimator $s_R^{2(k)}$ over $s_R^2, \hat{S}_{KC1}^2$ and

5. Conclusion

The proposed estimator of order k, as established by the theoretical findings along with numerical illustrations, is superior to the one due to Isaki (1983) and s_y^2 under conditions which hold good in a wide-ranging practical situations very often. There is a fairly good amount of gain in efficiency of the proposed estimator over the one due to Kadilar and Cingi (2006) as has been evident from the table 1.The second column of Table 1 suggests that there exist cases where the estimator due to Isaki (1983) is less efficient than the customary estimator of

population variance, our newly proposed estimator performs better thus being recommended for use in practice.

Acknowledgement

We thank the referee for suggestions leading to the improvement of the paper.

References

Adichwal, N.K., Raghav, Y.S. and Singh, R. (2015): A new exponential ratio-type estimator for population variance with linear combination of two auxiliary attributes, *Elixir international journal*, **89**, 36451-36457.

Adichwal, N.K., Kumar, J. and Singh, R. (2018): An improved generalized class of estimators for population variance using auxiliary variables, *Congent Mathematics & Statistics*, 5:1.

Agrawal, M.C. and Sthapit, A. B. (1997): Hierarchic predictive ratio-based & product-based estimators and their efficiencies. *Journal of Applied Statistics,* **24(1)**, 97-104.

Agrawal, M.C. and Panda, K.B. (1999): A predictive justification for variance estimation using auxiliary information. *Jour. Ind. Ag. Statistics*, **52(2)**, 192-200.

Basu, D. (1971): An essay on the logical foundations of statistical inference, Part I, Foundations of Statistical Inference, Ed. By V.P. Godambe and D.A. Sportt, New York, 203-233.

Cochran, W. G. (1977): Sampling Techniques, 3rdedn. (Wiley & Sons).

Das, A. K. (1988): Contribution to the theory of sampling strategies based on auxiliary information (Ph.D. thesis submitted to Bidhan Chandra Krishi Vishwavidyalaya, Mohanpur, Nadia, West Bengal, India).

Kadilar, C. and Cingi, H. (2006): Improvement in variance estimation using auxiliary information, *Hacettepe Journal of Mathematics and Statistics* **35(1)**, 111-115.

Panda, K.B. and Sahoo, N. (2015): Systems of exponential ratio-based and exponential product-based estimators with their efficiency. *ISOR Journal of Mathematics,* **11(3)**, PP 73-77.

Panda, K.B. and Das, P. (2018): Efficient hierarchic multivariate product-based estimator. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, **5(1)**, pp.65-69.

Panda, K.B. and Das, P. (2018): Efficient hierarchic predictive multivariate product estimator based on harmonic mean. *International Journal of Mathematics Trends and Technoloy (IJMTT)* 56(6), pp.14-18.

Gupta, Sat and Shabbir, Javid (2008): Variance estimation in simple random sampling using auxiliary information. *Hacettepe Journal of Mathematics and Statistics,* **37(1)**, 57-67.

Authors and Affiliations

K.B. Panda[1] and P. Das[2]

K.B.Panda
kunja.st@utkaluniversity.ac.in

[1,2]Department of Statistics, Utkal University, Bhubaneswar, India.