ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ON ESTIMATION OF VARIANCE OF MEAN FOR THE REGRESSION ESTIMATOR UNDER STRATIFIED RANDOM SAMPLING

Sheela Misra, Archana Shukla and Ashish Kumar Shukla*
Department of Statistics, University of Lucknow, Lucknow U.P.
Department of Higher Education, Ministry of Human Resource Development, New Delhi
*Department of Statistics, Ramanujan College, University of Delhi, Delhi

DOI: 10.5281/zenodo.221018

ABSTRACT
This chapter deals with the estimation of variance of separate regression type estimator of the population mean in stratified random sampling, its bias and mean square error are obtained and further an optimum class of estimators is obtained having minimum mean square error. Enhancing the practical utility of the optimum estimator, a class of estimators depending upon estimated optimum value based on sample observations is also found. Further comparative study has been done with some earlier estimators.

KEYWORDS: Stratified Random Sampling, Bias, Mean Square Error, Regression Estimator

INTRODUCTION
Let U be a finite population of size N . The study variable and the auxiliary variable are denoted by y and x respectively and the population is partitioned into L non-overlapping strata according to some characteristic. The size of the h^{th} stratum is $N_h\ (h=1,2,.....,L)$ such that $\sum_{h=1}^{L} N_h = N$. A stratified sample of size n is drawn from this population and let n_h be sample size from h^{th} stratum such that $\sum_{h=1}^{L} n_h = n$. The observations on y and x corresponding to i^{th} unit of h^{th} stratum $(h=1,2,.....,L)$ are y_{hi} and x_{hi} respectively. Let $\bar{y}_h$ and $\bar{x}_h$ be sample means and $\bar{Y}_h$ and $\bar{X}_h$ be population means of y and x respectively in h^{th} stratum. Suppose $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ are stratified sample means and $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ and $\bar{X} = \sum_{h=1}^{L} W_h \bar{Y}_h$ are population means of y and x respectively, where $W_h = N_h / N$ is known stratum weight. Let $s_{yh}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_h\right)^2$ and $s_{xh}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(x_{hi} - \bar{x}_h\right)^2$ be sample variances and $S_{yh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \bar{Y}_h\right)^2$ and $S_{xh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(x_{hi} - \bar{X}_h\right)^2$ be population variances of y and x respectively in h^{th} stratum. Finally, let $s_{yxh} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_h\right)\left(x_{hi} - \bar{x}_h\right)$ and

$S_{yxh} = \frac{1}{N_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} - \bar{Y}_h\right)\left(x_{hi} - \bar{X}_h\right)$ be sample and population covariances respectively in h^{th} stratum.
We assume that all parameters corresponding to auxiliary variable x are known and we ignore the finite population correction term $f_h = \left(1 - \frac{n_h}{N_h}\right)$ for simplification. A separate regression-type estimator of population mean $\bar{Y}$ is $\bar{y}_s = \sum_{h=1}^{L} W_h\left\{\bar{y}_h + b_h\left(\bar{X}_h - \bar{x}_h\right)\right\}$, where b_h sample regression coefficient is. Variance of $\bar{y}_s$ is given by

$$V\left(\bar{y}_s\right) = \sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2\left(1 - \rho_h^2\right)}{n_h} \tag{1.1}$$

where $\rho_h = \frac{S_{yxh}}{S_{yh}S_{xh}}$ is population correlation coefficient between y and x in h^{th} stratum.
An estimator of $V\left(\bar{y}_s\right)$ is given by Gupta and Shabbir (2010) is as follows

$$v_s = \sum_{h=1}^{L} W_h^2 \frac{s_{yh}^2\left(1 - r_h^2\right)}{n_h} \tag{1.2}$$

where $r_h = \frac{s_{yxh}}{s_{yh}s_{xh}}$ is sample correlation coefficient between y and x in h^{th} stratum. The mean square error of v_s is

$$MSE\left(v_s\right) = \sum_{h=1}^{L} W_h^4 S_{yh}^4 \frac{\left[\left(\lambda_{40h} - 1\right) + \rho_h^4 B_h + 2\rho_h^2 C_h\right]}{n_h^3} \tag{1.3}$$

where $B_h = \left(\lambda_{04h} - 1\right) + 4\left(\lambda_{22h}/\rho_h^2 - 1\right) - 4\left(\lambda_{13h}/\rho_h - 1\right)$ and $C_h = \left(\lambda_{22h} - 1\right) - 2\left(\lambda_{31h}/\rho_h - 1\right)$
Different authors presented the estimators utilizing auxiliary information and enhanced the efficiency of exiting estimators. These includes Das and Tripathi (1981), Srivastava and Jhajj (1980, 1983), Wu (1985), Prasad and Singh (1990, 1992).

PROPOSED ESIMATOR

Our proposed estimator of $V\left(\bar{y}_s\right)$ using auxiliary information on $\left(\bar{X}_h, S_{xh}^2\right)$. The proposed estimator is given by

$$v_a = \sum_{h=1}^{L} W_h^2\left\{\frac{s_{yh}^2\left(1 - r_h^2\right)}{n_h} + k\left(s_{xh}^2 - S_{xh}^2\right)\right\} \tag{2.1}$$

where k is a characterizing scalar chosen suitably.
We define the following terms:

$s_{yh}^2 = S_{yh}^2\left(1 + e_0\right)$, $s_{xh}^2 = S_{xh}^2\left(1 + e_1\right)$, $s_{yxh} = S_{yxh}\left(1 + e_2\right)$ so that $E\left(e_0\right) = E\left(e_1\right) = E\left(e_2\right)$

and also up to first order of approximation, we have the following expectations that can be derived easily on the lines of Sukhatme et al. (1997):

$E\left(e_0^2\right) = \frac{1}{n_h}\left(\lambda_{40h} - 1\right)$, $E\left(e_1^2\right) = \frac{1}{n_h}\left(\lambda_{04h} - 1\right)$, $E\left(e_2^2\right) = \frac{1}{n_h}\left(\frac{\lambda_{22h}}{\rho_h^2} - 1\right)$ $E\left(e_0e_1\right) = \frac{1}{n_h}\left(\lambda_{22h} - 1\right)$,

$E(e_0e_2)=\frac{1}{n_h}\left(\frac{\lambda_{31h}}{\rho_h}-1\right)$, $E(e_1e_2)=\frac{1}{n_h}\left(\frac{\lambda_{13h}}{\rho_h}-1\right)$ where $\lambda_{pqh}=\frac{\mu_{pqh}}{\mu_{20h}^{p/2}\mu_{02h}^{q/2}}$

and $\mu_{pqh}=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}_h\right)^p\left(x_{hi}-\overline{X}_h\right)^q$

now writing v_a in terms of e_i's, we have

$$v_a=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[S_{yh}^2(1+e_0)-\frac{S_{yxh}^2(1+e_2)^2}{S_{xh}^2(1+e_1)}\right]+k\sum_{h=1}^{L}W_h^2S_{xh}^2(1+e_1-1)$$

$$=\sum_{h=1}^{L}\frac{W_h^2S_{yh}^2}{n_h}\left[(1+e_0)-\rho_h^2\left(1-e_1+2e_2+e_1^2+e_2^2-2e_1e_2+\ldots\right)\right]+k\sum_{h=1}^{L}W_h^2S_{xh}^2e_1$$

$$=\sum_{h=1}^{L}\frac{W_h^2S_{yh}^2}{n_h}\left(1-\rho_h^2\right)\left[1+\frac{e_0+\rho_h^2\left(-e_1+2e_2+e_1^2+e_2^2-2e_1e_2+\ldots\right)}{\left(1-\rho_h^2\right)}\right]+k\sum_{h=1}^{L}W_h^2S_{xh}^2e_1$$

$$v_a-E(v_a)=\sum_{h=1}^{L}W_h^2\frac{S_{yh}^2}{n_h}\left\{e_0+\rho_h^2\left(-e_1+2e_2+e_1^2+e_2^2-2e_1e_2+\ldots\right)\right\}+k\sum_{h=1}^{L}W_h^2S_{xh}^2e_1 \qquad (2.2)$$

Taking expectation on both sides, we have bias up to terms of order O(1/n) to be

$$Bias(v_a)=E[v_a-E(v_a)]$$

$$=\sum_{h=1}^{L}W_h^2\frac{S_{yh}^2}{n_h}\left[E(e_0)+\rho_h^2\left\{\begin{matrix}E(-e_1)+2E(e_2)+\\E(e_1^2)+E(e_2^2)-2E(e_1e_2)\end{matrix}\right\}\right]+k\sum_{h=1}^{L}W_h^2S_{xh}^2E(e_1)$$

$$=\sum_{h=1}^{L}W_h^2\frac{S_{yh}^2}{n_h^2}\left[\rho_h^2\left\{(\lambda_{04h}-1)+\left(\frac{\lambda_{22h}}{\rho_h^2}-1\right)-2\left(\frac{\lambda_{13h}}{\rho_h}-1\right)\right\}\right] \qquad (2.3)$$

Squaring both sides of (2.2) and taking expectation, we have mean square error of v_a up to terms of order O(1/n) to be

$$MSE(v_a)=E[v_a-E(v_a)]^2$$

$$=E\left[\sum_{h=1}^{L}W_h^2\frac{S_{yh}^2}{n_h}\left\{e_0+\rho_h^2(-e_1+2e_2)\right\}+k\sum_{h=1}^{L}W_h^2S_{xh}^2e_1\right]^2$$

$$=\sum_{h=1}^{L}W_h^4\frac{S_{yh}^4}{n_h^2}\left[E(e_0^2)+\rho_h^4\left\{4E(e_2^2)+E(e_1^2)-4E(e_1e_2)\right\}+2\rho_h^2\left\{2E(e_0e_2)-E(e_0e_1)\right\}\right]+$$

$$k^2\sum_{h=1}^{L}W_h^4S_{xh}^4E(e_1^2)+2k\sum_{h=1}^{L}W_h^4\frac{S_{yh}^2S_{xh}^2}{n_h}\left[E(e_0e_1)+\rho_h^2\left\{2E(e_1e_2)-E(e_1^2)\right\}\right]$$

$$=\sum_{h=1}^{L}W_h^4\frac{S_{yh}^4}{n_h^3}\left[(\lambda_{40h}-1)+\rho_h^4B_h+2\rho_h^2C_h\right]+k^2\sum_{h=1}^{L}W_h^4\frac{S_{xh}^4}{n_h}(\lambda_{04h}-1)+$$

$$2k\sum_{h=1}^{L}W_h^4\frac{S_{yh}^2S_{xh}^2}{n_h^2}\left[(\lambda_{22h}-1)+\rho_h^2\left\{2\left(\frac{\lambda_{13h}}{\rho_h}-1\right)-(\lambda_{04h}-1)\right\}\right] \qquad (2.4)$$

The optimum value of k minimizing the mean square error of v_a in (2.4) is given by

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

$$k_o = \frac{S_{yh}^2\left[\rho_h^2\left\{(\lambda_{04h}-1)-2\left(\frac{\lambda_{13h}}{\rho_h}-1\right)\right\}-(\lambda_{22h}-1)\right]}{n_h S_{xh}^2(\lambda_{04h}-1)}$$

$$= \frac{S_{yh}^2 D_h}{n_h S_{xh}^2(\lambda_{04h}-1)} \qquad (2.5)$$

where $D_h = \left[\rho_h^2\left\{(\lambda_{04h}-1)-2\left(\frac{\lambda_{13h}}{\rho_h}-1\right)\right\}-(\lambda_{22h}-1)\right]$

and the minimum mean square error is given by

$$MSE(v_a)_o = \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\left[(\lambda_{40h}-1)+\rho_h^4 B_h + 2\rho_h^2 C_h\right] - \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\frac{D_h^2}{(\lambda_{04h}-1)} \qquad (2.6)$$

Estimator Based on Estimated Optimum $\hat{k}$

For situations where the values of λ_{22h}, λ_{13h}, λ_{04h}, and ρ_h or their good guessed values are not available , the alternative is to replace them by their estimates $\hat{\lambda}_{22h}$, $\hat{\lambda}_{13h}$, $\hat{\lambda}_{04h}$ and r_h based on sample values and get the estimated optimum value of k_o denoted by $\hat{k}$ as

$$\hat{k} = \frac{s_{yh}^2\left[r_h^2\left\{\left(\hat{\lambda}_{04h}-1\right)-2\left(\frac{\hat{\lambda}_{13h}}{r_h}-1\right)\right\}-\left(\hat{\lambda}_{22h}-1\right)\right]}{n_h s_{xh}^2\left(\hat{\lambda}_{04h}-1\right)}$$

$$= \frac{s_{yh}^2\left[r_h^2\left\{\left(\frac{\hat{\mu}_{04h}}{\hat{\mu}_{02h}^2}-1\right)-2\left(\frac{\hat{\mu}_{13h}}{\hat{\mu}_{20h}^{1/2}\hat{\mu}_{02h}^{3/2}}\frac{s_{yh}s_{xh}}{s_{yxh}}-1\right)\right\}-\left(\frac{\hat{\mu}_{22h}}{\hat{\mu}_{20h}^2\hat{\mu}_{02h}^2}-1\right)\right]}{n_h s_{xh}^2\left(\frac{\hat{\mu}_{04h}}{\hat{\mu}_{02h}^2}-1\right)} \qquad (3.1)$$

where $\hat{\lambda}_{22h} = \frac{\hat{\mu}_{22h}}{\hat{\mu}_{20h}^2\hat{\mu}_{02h}^2}$ with $\hat{\mu}_{22h} = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2(x_{hi}-\bar{x}_h)^2$

$\hat{\mu}_{20h} = s_{yh}^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2$ and $\hat{\mu}_{02h} = s_{xh}^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}(x_{hi}-\bar{x}_h)^2$

$\hat{\lambda}_{13h} = \frac{\hat{\mu}_{13h}}{\hat{\mu}_{20h}^{1/2}\hat{\mu}_{02h}^{3/2}}$, $\hat{\lambda}_{04h} = \frac{\hat{\mu}_{04h}}{\hat{\mu}_{02h}^2}$ and $\rho_h = r_h = \frac{s_{yxh}}{s_{yh}s_{xh}}$

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Thus, replacing k_o by estimated optimum $\hat{k}$ in the estimator v_{ae} in (2.1) , we get for wider practical utility of the estimator based on the estimated optimum $\hat{k}$ given by

$$v_{ae} = \sum_{h=1}^{L} W_h^2 \left\{ \frac{s_{yh}^2 \left(1 - r_h^2\right)}{n_h} + \hat{k}\left(s_{xh}^2 - S_{xh}^2\right) \right\} \tag{3.2}$$

To find the mean square error of v_{ae} , let

$$\hat{\mu}_{22h} = \mu_{22h}(1+e_3),\ \hat{\mu}_{13h} = \mu_{13h}(1+e_4),\ \hat{\mu}_{04h} = \mu_{04h}(1+e_5)$$

we have

$$\hat{k} = \frac{S_{yh}^2(1+e_0)\left[\rho_h^2 \frac{(1+e_2)^2}{(1+e_0)(1+e_1)} \left\{ \left(\frac{\mu_{04h}(1+e_5)}{\mu_{02h}^2(1+e_0)^2} - 1 \right) - 2\left(\frac{\mu_{13h}(1+e_4)}{\mu_{20h}^{1/2}(1+e_0)^{1/2}\mu_{02h}^{3/2}(1+e_1)^{3/2}} \rho_h \frac{(1+e_2)}{(1+e_0)^{1/2}(1+e_1)^{1/2}} - 1 \right) \right\} - \left(\frac{\mu_{22h}(1+e_3)}{\mu_{20h}(1+e_0)\mu_{02h}(1+e_1)} - 1 \right) \right]}{n_h S_{xh}^2 (1+e_1) \left\{ \frac{\mu_{04h}(1+e_5)}{\mu_{02h}^2(1+e_0)^2} - 1 \right\}}$$

$$= \frac{S_{yh}^2}{n_h S_{xh}^2} \left[\begin{array}{l} D_h - D_h \dfrac{\lambda_{04h}(-e_1+e_5+\ldots)-e_1}{(\lambda_{04h}-1)} - \rho_h^2\{\lambda_{04h}(-e_1+e_5+\ldots)-(-e_1+e_0+\ldots)\} - \\ \rho_h^2\{\lambda_{04h}(-e_1+e_5+\ldots)-(-e_1+e_0+\ldots)\} \dfrac{\lambda_{04h}(-e_1+e_5+\ldots)-e_1}{(\lambda_{04h}-1)} + \\ 2\rho_h\lambda_{13h}(-e_1+e_2+\ldots) - 2\rho_h\lambda_{13h}(-e_1+e_2+\ldots)\dfrac{\lambda_{04h}(-e_1+e_5+\ldots)-e_1}{(\lambda_{04h}-1)} - \\ \lambda_{22h}(-e_0+e_1+\ldots) + \lambda_{22h}(-e_0+e_1+\ldots)\dfrac{\lambda_{04h}(-e_1+e_5+\ldots)-e_1}{(\lambda_{04h}-1)} \end{array} \right] \tag{3 .3}$$

Substituting $\hat{k}$ from (3.3) in (3.2) and squaring both sides, ignoring terms of ei's greater than two and taking expectation, we have mean square error of v_{ae} to the first degree of approximation, that is up to terms of order O(1/n) to be

$$MSE(v_{ac}) = E\left[\sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} \{e_0 + \rho_h^2(-e_1+2e_2)\} + \hat{k}\sum_{h=1}^{L} W_h^2 \frac{S_{yh}^2}{n_h} D_h e_1 \right]^2$$

$$= \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\left[(\lambda_{40h}-1) + \rho_h^4 B_h + 2\rho_h^2 C_h\right] - \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3} \frac{D_h^2}{(\lambda_{04h}-1)} \tag{3.4}$$

which shows that the estimator v_{ae} in (3.2) based on estimated optimum $\hat{k}$ attains the same minimum mean square error of v_a in (2.6) depending on optimum value k_o in (2.5).

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

CONCLUDING REMARKS

a). From (2.6), for the optimum value of k_o, the estimator v_a attains the minimum mean square error given by

$$MSE(v_a)_o = \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\left[(\lambda_{40h}-1)+\rho_h^4 B_h + 2\rho_h^2 C_h\right] - \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\frac{D_h^2}{(\lambda_{04h}-1)} \quad (4.1)$$

b). From (3.4), the estimator v_{ae} depending upon estimated optimum $\hat{k}$ has the mean square error

$$MSE(v_{ae}) = \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\left[(\lambda_{40h}-1)+\rho_h^4 B_h + 2\rho_h^2 C_h\right] - \sum_{h=1}^{L} W_h^4 \frac{S_{yh}^4}{n_h^3}\frac{D_h^2}{(\lambda_{04h}-1)} \quad (4.2)$$

c). From (4.1) or (4.2), we see that the estimator v_{ae} depending on estimated optimum value is always more efficient than the variance of usual separate regression-type estimator $\bar{y}_s = \sum_{h=1}^{L} W_h \left\{\bar{y}_h + b_h\left(\bar{X}_h - \bar{x}_h\right)\right\}$ for non symmetrical population in the sense of having lesser mean square error.

REFERENCES

[1] Cochran, W.G.(1977): Sampling Techniques. John Wiley and Sons, New York.

[2] Cramer, H. (1946): Mathematical Methods of Statistics, Princeton University Press.

[3] Davies, O.L. and Goldsmith, P.L. (1976). Statistical methods in research and production. *London: Longman Group Ltd.*

[4] Gleser, L.J. and Healy, J.D. (1976). Estimating the mean of a normal distribution with known coefficient of variation, *Journal of the American Statistical Association*, 71, 977-981.

[5] Guilford, J.P. (1975). Psychometric methods, TATA *McGraw Hill Publishing Company Ltd., BomBay.*

[6] Hald, A. (1952). Statistical theory with engineering applications. *John Wiley and Sons, Inc, New York.*

[7] Murthy, M.N. (1977): Sampling Theory & Methods. Statistical Publishing society, Calcutta.

[8] Misra, S., Gupta R.K., Shukla A.K., (2012). Generalized Class of Estimators for Estimation of Finite Population Variance. Int. J. Agricult. Stats. Sci., Vol.8, No.2, pp. 447-458.

[9] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok C.(1984): Sampling Theory of Survey with Application. New Delhi