

A CLASS OF ESTIMATORS OF POPULATION VARIANCE IN STRATIFIED RANDOM SAMPLING

Nursel Koyuncu

Abstract

This study proposes a class of variance estimators for estimating population variance of a study variable using information of auxiliary variable under stratified random sampling scheme. The bias and mean square error of the estimators belonging to class are obtained and the optimum parameters of class are given in stratified random sampling. Efficiency comparison is carried out using a real data set. In this data set, sales profit and waste product of a company are used as a study and auxiliary variable respectively. Moreover we have found that suggested class of estimators are more efficient than classical estimators.

Key words: Ratio estimator; auxiliary information; mean square error; efficiency.

JEL Code: C60, C69

Introduction

Variance estimation is an important issue of statistics. The estimation of variance, when at least one auxiliary variable is available is widely discussed by Garcia and Cebrain (1996), Agrawal and Sthapit (1995), Arcos et al. (2005), Kadilar and Cingi (2006, 2007), Gupta and Shabbir (2008), Shabbir and Gupta (2010), Singh and Solanki (2012). In this study we have defined general class of estimators of variance when one auxiliary variable is available.

Assume that the population of size N is divided into L strata with N_h elements in the h th stratum. Let n_h be the size of the sample drawn from h th stratum of size N_h by using simple random sampling without replacement. The total sample size $\sum_{h=1}^L n_h = n$ and the population size $\sum_{h=1}^L N_h = N$. Let y and x be the study and the auxiliary variables, respectively, assuming values y_{hi} and x_{hi} for the i th unit in h th stratum. Moreover, let $\bar{y}_{(h)} = \sum_{i=1}^{n_h} \frac{y_{(h)i}}{n_h}$,

$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_{(h)}$, and $\bar{Y}_h = \sum_{i=1}^{N_h} \frac{y_{(h)i}}{N_h}$, $\bar{Y} = \bar{Y}_{st} = \sum_{h=1}^L W_h \bar{Y}_{(h)}$ be the sample and population means of y ,

respectively, where $W_h = \frac{N_h}{N}$ is the stratum weight. Similar expressions for x can also be

defined. When the finite population correction $\frac{N_h - n_h}{N_h}$ is ignored, the classical variance of

\bar{y}_{st} is given by $Var(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{S_{y(h)}^2}{n_h} = S_{y(st)}^2$, where $S_{y(h)}^2 = \sum_{i=1}^{N_h} \frac{(y_{(h)i} - \bar{Y}_h)^2}{N_h}$ is the population

variance of y in the h th stratum.

To obtain the bias and mean square error let us define $\delta_{0(h)} = \frac{S_{y(h)}^2 - S_{y(h)}^2}{S_{y(h)}^2}$, $\delta_{1(h)} = \frac{S_{x(h)}^2 - S_{x(h)}^2}{S_{x(h)}^2}$,

$\delta_{2(h)} = \frac{\bar{x}_{(h)} - \bar{X}_{(h)}}{\bar{X}_{(h)}}$. Using these notations we can get the expectations as given by

$$E(\delta_{0(h)}) = E(\delta_{1(h)}) = E(\delta_{2(h)}),$$

$$E(\delta_{0(h)}^2) = \frac{(\lambda_{40(h)} - 1)}{n_h}, \quad E(\delta_{1(h)}^2) = \frac{(\lambda_{04(h)} - 1)}{n_h}, \quad E(\delta_{2(h)}^2) = \frac{C_{x(h)}^2}{n_h}, \quad E(\delta_{0(h)}\delta_{1(h)}) = \frac{(\lambda_{22(h)} - 1)}{n_h},$$

$$E(\delta_{0(h)}\delta_{2(h)}) = \frac{\lambda_{21(h)}C_{x(h)}}{n_h}, \quad E(\delta_{1(h)}\delta_{2(h)}) = \frac{\lambda_{03(h)}C_{x(h)}}{n_h}$$

where $C_{x(h)} = \frac{S_{x(h)}}{\bar{X}_{(h)}}$, $\lambda_{ab(h)} = \frac{\mu_{ab(h)}}{\mu_{20(h)}\mu_{02(h)}}$, $\mu_{ab(h)} = \sum_{i=1}^{N_h} \frac{(y_{(h)i} - \bar{Y}_{(h)})^a (x_{(h)i} - \bar{X}_{(h)})^b}{N_h}$.

General Class of Separate Estimators

Following Koyuncu and Kadilar (2010), a general combined class of variance estimators in stratified random sampling is defined by

$$t_s = \sum_{h=1}^L \frac{W_h^2}{n_h} t_{s(h)} \tag{1}$$

$$t_{s(h)} = H_{(h)}(s_{y(h)}^2, u_{(h)}) \tag{2}$$

where $u_{(h)} = s_{x(h)}^2 / S_{x(h)}^2$ and $H_{(h)}(s_{y(h)}^2, u_{(h)})$ is a function of $s_{y(h)}^2$ and $u_{(h)}$. We can generate many estimators from (2) such as classical ratio, product, regression estimators as given in Table 1. To study the properties of $t_{s(h)}$ we assume following regularity conditions:

1. The point $(s_{y(h)}^2, u_{(h)})$ assumes the value in a closed convex subset R_2 of two dimensional real space containing the point $(S_{y(h)}^2, 1)$,
2. The function $H_{(h)}(s_{y(h)}^2, u_{(h)})$ is continuous and bounded in R_2 ,
3. $H_{(h)}(S_{y(h)}^2, 1) = S_{y(h)}^2$ and $g_{0(h)}(S_{y(h)}^2, 1) = 1$, where $g_{0(h)}(S_{y(h)}^2, 1)$ denotes the first order partial derivative of $g_{0(h)}$ with respect to $s_{y(h)}^2$,
4. The first and second order partial derivatives of $H_{(h)}(s_{y(h)}^2, u_{(h)})$ exist and are continuous and bounded in R_2 .

Expanding $H_{(h)}(s_{y(h)}^2, u_{(h)})$ about the point $(S_{y(h)}^2, 1)$ in a second order Taylor series and using the above regularity conditions, we have

$$t_{s(h)} = H_{(h)}[s_{y(h)}^2 + (s_{y(h)}^2 - S_{y(h)}^2), 1 + (u_{(h)} - 1)] \quad (3)$$

$$t_{s(h)} = H_{(h)}(S_{y(h)}^2, 1) + (s_{y(h)}^2 - S_{y(h)}^2)g_{0(h)} + (u_{(h)} - 1)g_{1(h)} + (u_{(h)} - 1)^2 g_{2(h)} + (s_{y(h)}^2 - S_{y(h)}^2)(u_{(h)} - 1)g_{3(h)} + (s_{y(h)}^2 - S_{y(h)}^2)^2 g_{4(h)} \quad (4)$$

$$t_{s(h)} = s_{y(h)}^2 + (u_{(h)} - 1)g_{1(h)} + (u_{(h)} - 1)^2 g_{2(h)} + (s_{y(h)}^2 - S_{y(h)}^2)(u_{(h)} - 1)g_{3(h)} + (s_{y(h)}^2 - S_{y(h)}^2)^2 g_{4(h)} \quad (5)$$

where

$$g_{1(h)} = \left. \frac{\partial H_{(h)}}{\partial u_{(h)}} \right|_{s_{y(h)}^2 = S_{y(h)}^2, u_{(h)} = 1}, \quad g_{2(h)} = \left. \frac{1}{2} \frac{\partial^2 H_{(h)}}{\partial u_{(h)}^2} \right|_{s_{y(h)}^2 = S_{y(h)}^2, u_{(h)} = 1}, \quad g_{3(h)} = \left. \frac{1}{2} \frac{\partial^2 H_{(h)}}{\partial s_{y(h)}^2 \partial u_{(h)}} \right|_{s_{y(h)}^2 = S_{y(h)}^2, u_{(h)} = 1}$$

$$g_{4(h)} = \left. \frac{1}{2} \frac{\partial^2 H_{(h)}}{\partial s_{y(h)}^4} \right|_{s_{y(h)}^2 = S_{y(h)}^2, u_{(h)} = 1}.$$

To obtain the bias and the *MSE*, let us use the notations $\delta_{0(h)}$ and $\delta_{1(h)}$. Expressing (5) with δs we have

$$t_{s(h)} = S_{y(h)}^2 + S_{y(h)}^2 \delta_{0(h)} + g_{1(h)} \delta_{1(h)} + g_{2(h)} \delta_{1(h)}^2 + S_{y(h)}^2 g_{3(h)} \delta_{0(h)} \delta_{1(h)} + S_{y(h)}^4 g_{4(h)} \delta_{0(h)}^2 \quad (6)$$

$$t_{s(h)} - S_{y(h)}^2 = S_{y(h)}^2 \delta_{0(h)} + g_{1(h)} \delta_{1(h)} + g_{2(h)} \delta_{1(h)}^2 + S_{y(h)}^2 g_{3(h)} \delta_{0(h)} \delta_{1(h)} + S_{y(h)}^4 g_{4(h)} \delta_{0(h)}^2 \quad (7)$$

Taking expectation both sides of (7), we obtain the bias as

$$Bias(t_{s(h)}) = \frac{1}{n_h} \{g_{2(h)}(\lambda_{04(h)} - 1) + S_{y(h)}^2 g_{3(h)}(\lambda_{22(h)} - 1) + S_{y(h)}^4 g_{4(h)}(\lambda_{40(h)} - 1)\} \quad (8)$$

$$Bias(t_s) = \sum_{h=1}^L \frac{W_h^2}{n_h} Bias(t_{s(h)}) \quad (9)$$

Squaring and neglecting higher order terms we have

$$(t_{s(h)} - S_{y(h)}^2)^2 \cong S_{y(h)}^4 \delta_{0(h)}^2 + g_{1(h)}^2 \delta_{1(h)}^2 + 2S_{y(h)}^2 g_{1(h)} \delta_{0(h)} \delta_{1(h)} \quad (10)$$

Taking expectation both sides of (10), we obtain the MSE as

$$MSE(t_{s(h)}) \cong \frac{1}{n_h} [S_{y(h)}^4 (\lambda_{40(h)} - 1) + g_{1(h)}^2 (\lambda_{04(h)} - 1) + 2S_{y(h)}^2 g_{1(h)} (\lambda_{22(h)} - 1)] \quad (11)$$

$$MSE(t_s) = \sum_{h=1}^L \frac{W_h^4}{n_h^2} MSE(t_{s(h)}) \quad (12)$$

On differentiating (11) with respect to $g_{1(h)}$ we obtain optimum value as

$$\frac{\partial MSE(t_{s(h)})}{\partial g_{1(h)}} = 0$$

$$g_{1(h)}^* = -\frac{S_{y(h)}^2 (\lambda_{22(h)} - 1)}{(\lambda_{04(h)} - 1)} \quad (13)$$

Using optimum value in (11) we obtain minimum MSE of $t_{s(h)}$ and t_s as

$$MSE(t_{s(h)})_{min} \cong \frac{S_{y(h)}^4}{n_h} \left[\frac{(\lambda_{40(h)} - 1)(\lambda_{04(h)} - 1) - (\lambda_{22(h)} - 1)^2}{(\lambda_{04(h)} - 1)} \right] \quad (14)$$

$$MSE(t_s)_{min} = \sum_{h=1}^L \frac{W_h^4}{n_h^2} MSE(t_{s(h)})_{min} \quad (15)$$

Now we have considered second class of separate estimators for variance estimation is given by

$$t_k = \sum_{h=1}^L \frac{W_h^2}{n_h} t_{k(h)} \quad (16)$$

$$t_{k(h)} = G_{(h)}(s_{y(h)}^2, m_{(h)}) \quad (17)$$

where $m_{(h)} = \bar{x}_{(h)}/\bar{X}_{(h)}$ and $G_{(h)}(s_{y(h)}^2, m_{(h)})$ is a function of $s_{y(h)}^2$ and $m_{(h)}$. Similarly we can generate many estimators from (17) such as ratio, product, regression estimators as given in Table 1.

To study the properties of $t_{k(h)}$ we assume following regularity conditions:

1. The point $(s_{y(h)}^2, m_{(h)})$ assumes the value in a closed convex subset R_2 of two dimensional real space containing the point $(S_{y(h)}^2, 1)$,
2. The function $G_{(h)}(s_{y(h)}^2, m_{(h)})$ is continuous and bounded in R_2 ,
3. $G_{(h)}(S_{y(h)}^2, 1) = S_{y(h)}^2$ and $\kappa_{0(h)}(S_{y(h)}^2, 1) = 1$, where $\kappa_{0(h)}(S_{y(h)}^2, 1)$ denotes the first order partial derivative of $\kappa_{0(h)}$ with respect to $s_{y(h)}^2$,
4. The first and second order partial derivatives of $G_{(h)}(s_{y(h)}^2, m_{(h)})$ exist and are continuous and bounded in R_2 .

Expanding $G_{(h)}(s_{y(h)}^2, m_{(h)})$ about the point $(S_{y(h)}^2, 1)$ in a second order Taylor series and using the above regularity conditions, we have

$$t_{k(h)} = G_{(h)}[s_{y(h)}^2 + (s_{y(h)}^2 - S_{y(h)}^2), 1 + (m_{(h)} - 1)] \quad (18)$$

$$t_{k(h)} = G_{(h)}(S_{y(h)}^2, 1) + (s_{y(h)}^2 - S_{y(h)}^2)\kappa_{0(h)} + (m_{(h)} - 1)\kappa_{1(h)} + (m_{(h)} - 1)^2\kappa_{2(h)} + (s_{y(h)}^2 - S_{y(h)}^2)(m_{(h)} - 1)\kappa_{3(h)} + (s_{y(h)}^2 - S_{y(h)}^2)^2\kappa_{4(h)} \quad (19)$$

$$t_{k(h)} = s_{y(h)}^2 + (m_{(h)} - 1)\kappa_{1(h)} + (m_{(h)} - 1)^2\kappa_{2(h)} + (s_{y(h)}^2 - S_{y(h)}^2)(m_{(h)} - 1)\kappa_{3(h)} + (s_{y(h)}^2 - S_{y(h)}^2)^2\kappa_{4(h)} \quad (20)$$

where

$$\kappa_{1(h)} = \frac{\partial G_{(h)}}{\partial m_{(h)}} \Big|_{S_{y(h)}^2 = S_{y(h)}^2, m_{(h)} = 1}, \quad \kappa_{2(h)} = \frac{1}{2} \frac{\partial^2 G_{(h)}}{\partial m_{(h)}^2} \Big|_{S_{y(h)}^2 = S_{y(h)}^2, m_{(h)} = 1}, \quad \kappa_{3(h)} = \frac{1}{2} \frac{\partial^2 G_{(h)}}{\partial S_{y(h)}^2 \partial m_{(h)}} \Big|_{S_{y(h)}^2 = S_{y(h)}^2, m_{(h)} = 1},$$

$$\kappa_{4(h)} = \frac{1}{2} \frac{\partial^2 G_{(h)}}{\partial S_{y(h)}^4} \Big|_{S_{y(h)}^2 = S_{y(h)}^2, m_{(h)} = 1}.$$

To obtain the bias and the *MSE*, let us use $\delta_{0(h)}$ and $\delta_{2(h)}$ notations in (20).

$$t_{k(h)} = S_{y(h)}^2 + S_{y(h)}^2 \delta_{0(h)} + \kappa_{1(h)} \delta_{2(h)} + \kappa_{2(h)} \delta_{2(h)}^2 + S_{y(h)}^2 \kappa_{3(h)} \delta_{0(h)} \delta_{2(h)} + S_{y(h)}^4 \kappa_{4(h)} \delta_{0(h)}^2 \quad (21)$$

$$t_{k(h)} - S_{y(h)}^2 = S_{y(h)}^2 \delta_{0(h)} + \kappa_{1(h)} \delta_{2(h)} + \kappa_{2(h)} \delta_{2(h)}^2 + S_{y(h)}^2 \kappa_{3(h)} \delta_{0(h)} \delta_{2(h)} + S_{y(h)}^4 \kappa_{4(h)} \delta_{0(h)}^2 \quad (22)$$

Taking expectation both sides of (22), we obtain the bias as

$$Bias(t_{k(h)}) = \frac{1}{n_h} \left\{ \kappa_{2(h)} C_{x(h)}^2 + S_{y(h)}^2 \kappa_{3(h)} \lambda_{21(h)} C_{x(h)} + \kappa_{4(h)} S_{y(h)}^4 (\lambda_{40(h)} - 1) \right\} \quad (23)$$

$$Bias(t_k) = \sum_{h=1}^L \frac{W_h^2}{n_h} Bias(t_{k(h)}) \quad (24)$$

Squaring and neglecting higher order terms we have

$$(t_{k(h)} - S_{y(h)}^2)^2 = (S_{y(h)}^4 \delta_{0(h)}^2 + \kappa_{1(h)}^2 \delta_{2(h)}^2 + 2S_{y(h)}^2 \kappa_{1(h)} \delta_{0(h)} \delta_{2(h)}) \quad (25)$$

$$MSE(t_{k(h)}) = \frac{1}{n_h} (S_{y(h)}^4 (\lambda_{40(h)} - 1) + \kappa_{1(h)}^2 C_{x(h)}^2 + 2\kappa_{1(h)} S_{y(h)}^2 \lambda_{21(h)} C_{x(h)}) \quad (26)$$

On differentiating (26) with respect to $\kappa_{1(h)}$ we obtain optimum value as

$$\frac{\partial MSE(t_{k(h)})}{\partial \kappa_{1(h)}} = 0$$

$$\kappa_{1(h)}^* = - \frac{S_{y(h)}^2 \lambda_{21(h)} C_{x(h)}}{C_{x(h)}^2} \quad (26)$$

Using optimum value in (26) we obtain minimum *MSE* of $t_{k(h)}$ and t_k as

$$MSE(t_{k(h)})_{\min} = \frac{S_{y(h)}^4}{n_h} ((\lambda_{40(h)} - 1) - \lambda_{21(h)}^2) \quad (27)$$

$$MSE(t_k)_{min} = \sum_{h=1}^L \frac{W_h^4}{n_h^2} MSE(t_{k(h)})_{min} \quad (28)$$

Tab. 1: Some members of t_s and t_k

Some members of t_s	Some members of t_k
$t_{s1} = \sum_{h=1}^L \frac{W_h^2}{n_h} \frac{S_{x(h)}^2}{S_{x(h)}^2} S_{y(h)}^2$	$t_{k1} = \sum_{h=1}^L \frac{W_h^2}{n_h} \frac{\bar{X}_{(h)}}{\bar{X}_{(h)}} S_{y(h)}^2$
$t_{s2} = \sum_{h=1}^L \frac{W_h^2}{n_h} \frac{S_{x(h)}^2}{S_{x(h)}^2} S_{y(h)}^2$	$t_{k2} = \sum_{h=1}^L \frac{W_h^2}{n_h} \frac{\bar{x}_{(h)}}{\bar{X}_{(h)}} S_{y(h)}^2$
$t_{s3} = \sum_{h=1}^L \frac{W_h^2}{n_h} \frac{S_{x(h)}^2}{S_{x(h)}^2 + \alpha_3 (S_{x(h)}^2 - S_{x(h)}^2)} S_{y(h)}^2$	$t_{k3} = \sum_{h=1}^L \frac{W_h^2}{n_h} S_{y(h)}^2 \frac{\bar{X}_{(h)}}{\bar{X}_{(h)} + \alpha_3 (\bar{x}_{(h)} - \bar{X}_{(h)})}$
$t_{s4} = \sum_{h=1}^L \frac{W_h^2}{n_h} (S_{y(h)}^2 + \alpha_3 (S_{x(h)}^2 - S_{x(h)}^2))$	$t_{k4} = \sum_{h=1}^L \frac{W_h^2}{n_h} (S_{y(h)}^2 + \alpha_3 (\bar{x}_{(h)} - \bar{X}_{(h)}))$

Numerical Example

To illustrate the efficiency of suggested estimators in the application, we consider the data concerning the sales profit (y) and waste product (x) of a company's 7634 products are used as a study and auxiliary variable respectively. We have stratified the products as mealy products, vegetables or fruits, meat–fish–chicken, frozen meat. The summary statistics of the data are given in Table 2.

The MSE values of suggested class of estimators have been obtained using (15)-(28) respectively. We have found that minimum mean square error of $MSE(t_s)_{min} = 948201,8$ and $MSE(t_k)_{min} = 1026233$. Note that someone can generate many variance estimators using (2) and (17) as given in Table1. We can say that suggested class of estimators contain most of the

estimators which is defined in literature. In this study for this data set we have found that minimum mean square error of these class of estimators. We can say that member of t_s class of estimator are more efficient than member of t_k class of estimator.

Tab. 2: Data Statistics

	Mealy products	vegetables or fruits	Meat–fish–chicken	Frozen meat
N_h	2692	2683	1714	545
n_h	400	400	250	80
W_h	0.352632958	0.351454021	0.224521876	0.071391
$\bar{Y}_{(h)}$	485.1211631	1166.504962	1053.474624	160.7392
$\bar{X}_{(h)}$	21.20818436	71.1925038	53.47686161	37.23346
$\rho_{xy(h)}$	0.307041472	0.711366419	0.401843163	0.307635
$S_{y(h)}^2$	544292.5448	6889841.622	3981284.264	25916.77
$\mu_{20(h)}$	544090.3559	6889841.622	3978961.461	25869.21
$\mu_{02(h)}$	712.2502101	13938.5986	8002.960644	1732.767
$\mu_{40(h)}$	1.22435E+13	3.2777E+15	3.50587E+15	1.2E+10
$\mu_{04(h)}$	7295541.37	23927277984	3750901117	51289662
$\mu_{22(h)}$	1637043529	4.40552E+12	2.19375E+12	2.11E+08
$\mu_{21(h)}$	14351825.02	3771213648	1876953589	628688.9
$\lambda_{40(h)}$	41.35850434	69.04797147	221.4400999	17.91787
$\lambda_{04(h)}$	14.38110815	123.1558574	58.5644748	17.08242
$\lambda_{22(h)}$	4.224318094	45.87425985	68.89172897	4.70683
$\lambda_{21(h)}$	0.98837058	4.636201918	5.273008371	0.583825

Conclusion

In this study we have suggested general class of estimators of variance when one auxiliary variable is available. We obtain theoretical bias and MSE of class of estimators. For illustration we have used a real data set of a company's sales profit and waste product as study and auxiliary variable respectively. Suggested class of estimators contain many variance estimator and more efficient than many estimator which is defined in the literature

References

Agrawal. M. C.. Sthapit. A. B. "Unbiased Ratio-Type Variance Estimation" *Statistics and Probability Letters* 1995: 25. 361–364.

Arcos. A.. Rueda. M.. Martinez. M. D.. Gonzalez. S.. Roman. Y. "Incorporating the auxiliary information available in variance estimation" *Applied Mathematics and Computation* 2005: 160. 387–399.

Garcia. M. R.. Cebrain. A. A. "Repeated substitution method: The ratio estimator for the population variance" *Metrika* 1996: 43. 101–105.

Gupta. S.. Shabbir. J. "Variance estimation in simple random sampling using auxiliary information" *Hacettepe Journal of Mathematics and Statistics* 2008 : 37. 1. 57-67.

Kadilar. C.. Cingi. H. "Ratio estimators for the population variance in simple and stratified random sampling" *Applied Mathematics and Computation* 2006: 173. 2. 1047-1059.

Kadilar. C.. Cingi. H. "Improvement in Variance Estimation in Simple Random Sampling" *Comm. Statist. Theory Methods* 2007: 36. 11. 2075-2081.

Koyuncu. N.. Kadilar. C. "On improvement in estimating population mean in stratified random sampling" *Journal of Applied Statistics* 2010: 37. 6. 999-1013.

Shabbir. J.. Gupta. S. "Some estimators of finite population variance of stratified sample mean" *Comm. Statist. Theory Methods* 2010: 39. 3001-3008.

Singh. H. P.. Solanki. R. S. "A new procedure for variance estimation in simple random sampling using auxiliary information" *Statistical Papers*. 2012: doi:10.1007/s00362-012-0445-2.

Contact

Nursel Koyuncu

Hacettepe University

Faculty of Science. Statistics Department

06800. Beytepe. Ankara. Turkey

nkoyuncu@hacettepe.edu.tr