STATISTICS IN TRANSITION-new series, July 2009
Vol. 10, No. 1, pp. 75—84

A CHAIN-TYPE ESTIMATOR FOR POPULATION VARIANCE USING TWO AUXILIARY VARIABLES IN TWO-PHASE SAMPLING

Housila P. Singh, Nidhi Mathur and Prem Chandra[1]

ABSTRACT

In this paper we have suggested a chain-type estimator of population variance by making use of a known coefficient of kurtosis of second auxiliary variable in two-phase sampling. It is shown that the proposed estimator is better than usual estimators under some realistic conditions.

Key words: Two-phase sampling, Chain ratio-type estimator, Study variable, Auxiliary variable, Mean square error.

1. Introduction

In many sample surveys studies, the information on more than single auxiliary variables correlated with study variable is either readily known or may be made known by diverting a part of the survey resources. This information is used at the estimation stage, thereby leading to a better choice of estimators. Out of many, ratio and regression methods of estimation are good examples in this context.

Consider a finite population $U = (U_1, U_2, ..., U_N)$. Let y and x be the variable under study and auxiliary variable, taking values y_i and x_i respectively for the i^{th} unit. Define

$$S_y^2 = \frac{N}{(N-1)}\sigma_y^2 \text{ with } \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 ,$$

$$S_x^2 = \frac{N}{(N-1)}\sigma_x^2 \text{ with } \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 ,$$

[1] Department of Biostatistics, All India Institute of Medical Sciences, New Delhi, India.

where $\bar{Y} = \sum_{i=1}^{N} y_i \Big/ N$ and $\bar{X} = \sum_{i=1}^{N} x_i \Big/ N$ are the population means of y and x respectively.

For large N,

$$S_y^2 \cong \sigma_y^2 \text{ and } S_x^2 \cong \sigma_x^2 .$$

Isaki (1983) suggested a ratio estimator for S_y^2 of y assuming that the population variance S_x^2 of x is known. When the two variables y and x are closely related but no information is available on the population variance S_x^2 of x, we seek to estimate of S_y^2 of y from a sample 's', obtained through a two-phase (or double) selection. Allowing simple random sampling without replacement (SRSWOR) in each phase, the two phase sampling will be as follows:

(a) The first phase sample $s^1 \left(s^1 \subset U\right)$ of fixed size n is drawn to measure only x in order to obtain a good estimate S_x^2.

(b) Given s^1, the second-phase sample s $\left(s \subset s^1\right)$ of fixed size m is selected to measure y only.

Let $s_{ym}^2 = \sum_{j \in s} \left(y_j - \bar{y}_m\right)^2 \Big/ (m-1)$, $\bar{y}_m = \sum_{j \in s} y_j \Big/ m$,

$s_{xm}^2 = \sum_{j \in s} \left(x_j - \bar{x}_m\right)^2 \Big/ (m-1)$, $\bar{x}_m = \sum_{j \in s} x_j \Big/ m$, and

$s_{xn}^2 = \sum_{j \in s^1} \left(x_j - \bar{x}_n\right)^2 \Big/ (n-1)$, $\bar{x}_n = \sum_{j \in s^1} x_j \Big/ n$.

The two-phase sampling version of Isaki's (1983) estimator for S_y^2 is defined by

$$t_1 = s_{ym}^2 \left(s_{xn}^2 \big/ s_{xm}^2\right) \tag{1.1}$$

Sometimes even if S_x^2 is unknown, information on a cheaply ascertainable variable z is available on all units of the population with known population variance

$$S_z^2 = \frac{N}{(N-1)} \sigma_z^2 \text{ with } \sigma_z^2 = \frac{1}{N} \sum_{j=1}^{N} \left(z_j - \bar{Z}\right)^2 \text{ and } \bar{Z} = \sum_{j=1}^{N} z_j \Big/ N .$$

For large N, $S_z^2 \cong \sigma_z^2$.

Motivated by Chand (1975) one may suggest a chain ratio-type estimator for S_y^2 as

$$t_{R1} = s_{ym}^2 \left(\frac{s_{xn}^2}{s_{xm}^2} \right) \left(\frac{S_z^2}{s_{zn}^2} \right) \tag{1.2}$$

where $s_{zn}^2 = \sum_{j \in s^1} \left(z_j - \bar{z}_n \right)^2 \Big/ (n-1)$ with $\bar{z}_n = \sum_{j \in s^1} z_j \Big/ n$.

To the first degree of approximation, the biases of t_{Rd} and t_{R1}, the variance of s_y^2 and mean square errors (MSEs) of t_{Rd} and t_{R1} are respectively given by

$$B(t_{Rd}) = \left(\frac{1}{m} - \frac{1}{n} \right) S_y^2 \, C_1^2 \left(1 - \kappa_{01} \right) \tag{1.3}$$

$$B(t_{R1}) = S_y^2 \left[\left(\frac{1}{m} - \frac{1}{n} \right) C_1^2 \left(1 - \kappa_{01} \right) + \left(\frac{1}{n} - \frac{1}{N} \right) C_2^2 \left(1 - \kappa_{02} \right) \right] , \tag{1.4}$$

$$Var\left(s_{ym}^2 \right) = S_y^4 \left(\frac{1}{m} - \frac{1}{n} \right) C_0^2 \tag{1.5}$$

$$MSE(t_{Rd}) = S_y^4 \left[\left(\frac{1}{m} - \frac{1}{n} \right) C_0^2 + \left(\frac{1}{m} - \frac{1}{n} \right) C_1^2 \left(1 - 2\kappa_{01} \right) \right] , \tag{1.6}$$

$$MSE(t_{R1}) = MSE(t_{Rd}) + S_y^4 \left(\frac{1}{n} - \frac{1}{N} \right) C_2^2 \left(1 - 2\kappa_{02} \right) , \tag{1.7}$$

where $C_0 = \sqrt{(\lambda_{400} - 1)}$, $C_1 = \sqrt{(\lambda_{040} - 1)}$, $C_2 = \sqrt{(\lambda_{004} - 1)}$,
$\rho_{01} = (\lambda_{220} - 1) \Big/ \sqrt{(\lambda_{400} - 1)(\lambda_{040} - 1)}$,
$\rho_{02} = (\lambda_{202} - 1) \Big/ \sqrt{(\lambda_{400} - 1)(\lambda_{004} - 1)}$, $\rho_{12} = (\lambda_{022} - 1) \Big/ \sqrt{(\lambda_{040} - 1)(\lambda_{004} - 1)}$,
$\kappa_{01} = \rho_{01} C_0 / C_1$,
$\kappa_{02} = \rho_{02} C_0 / C_2$, $\kappa_{12} = \rho_{12} C_1 / C_2$, $\kappa_{21} = \rho_{12} C_2 / C_1$,
$\lambda_{pqr} = \mu_{pqr} \Big/ \left\{ \mu_{200}^{p/2} \, \mu_{020}^{q/2} \, \mu_{002}^{r/2} \right\}$,
$\mu_{pqr} = (1/N) \sum_{i=1}^{N} \left(y_i - \bar{Y} \right)^p \left(x_i - \bar{X} \right)^q \left(z_i - \bar{Z} \right)^r$, (p, q, r) being non-negative integers.

In the present investigation we have suggested a modified chain ratio-type estimator for S_y^2 and discussed its properties. Further generalization of the suggested estimator are also presented.

2. Modified estimator

As noted earlier in Section 1, the information about the second auxiliary variable z is available for all the units. Thus, the values of certain parameters of the variable z can be made available. Thus, utilizing the knowledge of S_z^2 and $\beta_2(z)$ and following Upadhyaya and Singh (1999) we suggest a ratio estimator for S_x^2 as

$$t_{xR} = s_{xn}^2 \left\{ \frac{S_z^2 + \beta_2(z)}{s_{zn}^2 + \beta_2(z)} \right\} \tag{2.1}$$

where $\beta_2(z) = \lambda_{004} = \mu_{004} / \mu_{002}^2$ is the known population coefficient of kurtosis of the variable z.

Replacing s_{xn}^2 by t_{xR} in (1.7) we get a modified chain ratio-type estimator for S_y^2 as

$$t_{R2} = s_{ym}^2 \frac{t_{xR}}{s_{xm}^2}$$

$$= s_{ym}^2 \left\{ \frac{s_{xn}^2}{s_{xm}^2} \right\} \left\{ \frac{S_z^2 + \beta_2(z)}{s_{zn}^2 + \beta_2(z)} \right\} \tag{2.2}$$

To obtain the bias and MSE of t_{R2} we write

$$s_{ym}^2 = S_y^2 (1 + e_0)$$

$$s_{xm}^2 = S_x^2 (1 + e_1)$$

$$s_{xn}^2 = S_x^2 (1 + e_2)$$

$$s_{zn}^2 = S_z^2 (1 + e_3)$$

such that $E(e_i)=0\,\forall\, i=0,1,2,3,4;$ and to the first degree of approximation,

$$E(e_0^2)=\left(\frac{1}{m}-\frac{1}{N}\right)C_0^2,\quad E(e_1\,e_2)=\left(\frac{1}{n}-\frac{1}{N}\right)C_1^2,$$

$$E(e_1^2)=\left(\frac{1}{m}-\frac{1}{N}\right)C_1^2,\quad E(e_1\,e_3)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{12}\,C_1C_2,$$

$$E(e_2^2)=\left(\frac{1}{n}-\frac{1}{N}\right)C_1^2,\quad E(e_2\,e_3)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{12}\,C_1C_2,$$

$$E(e_3^2)=\left(\frac{1}{n}-\frac{1}{N}\right)C_2^2,\quad E(e_4^2)=\left(\frac{1}{m}-\frac{1}{N}\right)C_2^2,$$

$$E(e_0\,e_1)=\left(\frac{1}{m}-\frac{1}{N}\right)\rho_{01}\,C_0C_1,\quad E(e_0\,e_4)=\left(\frac{1}{m}-\frac{1}{N}\right)\rho_{02}\,C_0C_2,$$

$$E(e_0\,e_2)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{01}\,C_0C_1,\quad E(e_1\,e_4)=\left(\frac{1}{m}-\frac{1}{N}\right)\rho_{12}\,C_1C_2,$$

$$E(e_0\,e_3)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{02}\,C_0C_2,\quad E(e_2\,e_4)=\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{12}\,C_1C_2,$$

$$E(e_3\,e_4)=\left(\frac{1}{n}-\frac{1}{N}\right)C_2^2.$$

Expressing (2.2) in terms of e's we have

$$t_{R2}=S_y^2(1+e_0)(1+e_2)(1+e_1)^{-1}(1+\theta\,e_3)^{-1} \tag{2.3}$$

where $\theta=S_z^2\big/\left\{s_z^2+\beta_2(z)\right\}$.

Expanding the right hand side of (2.3) up to second power of e's we have

$$t_{R2}\cong S_y^2\big[1+e_0-e_1+e_2-\theta\,e_3-e_0e_1$$

$$+e_0e_2-e_1e_2-\theta\,e_0e_3-\theta\,e_2e_3+e_1^2+\theta^2e_3^2\big]$$

or

$$\left(t_{R2}-S_y^2\right)\cong S_y^2\big[e_0-e_1+e_2-\theta\,e_3-e_0e_1+e_0e_2$$

$$-e_1e_2-\theta e_0e_3-\theta e_2e_3+\theta e_1e_3+e_1^2+\theta^2e_3^2\Big] \tag{2.4}$$

Taking expectation of both sides in (2.4) we get the bias of t_{R2} to the first degree of approximation as

$$B(t_{R2})=S_y^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)C_1^2(1-\kappa_{01})+\left(\frac{1}{n}-\frac{1}{N}\right)\theta C_2^2(\theta-\kappa_{02})\right] \tag{2.5}$$

Squaring both sides of (2.4), retaining terms containing the powers of e's up to second degree and then taking expectations of both sides of (2.4), we get the MSE of t_{R2} as

$$MSE(t_{R2})=S_y^4\left[\left(\frac{1}{m}-\frac{1}{N}\right)C_0^2+\left(\frac{1}{m}-\frac{1}{n}\right)C_1^2(1-2\kappa_{01})\right.$$

$$\left.+\left(\frac{1}{n}-\frac{1}{N}\right)\theta C_2^2(\theta-2\kappa_{02})\right] \tag{2.6}$$

3. Efficiency comparisons

From (1.5) and (2.6) we note that $MSE(t_{R2})<Var(s_{ym}^2)$ if

$$\left[\left(\frac{1}{m}-\frac{1}{n}\right)C_1^2(1-2\kappa_{01})+\left(\frac{1}{n}-\frac{1}{N}\right)\theta C_2^2(\theta-2\kappa_{02})\right]<0$$

which is always true if

$$\kappa_{01}>\frac{1}{2}\text{ and }\kappa_{02}>\frac{\theta}{2} \tag{3.1}$$

From (1.6) and (2.6) we see that $MSE(t_{R2})<MSE(t_{Rd})$ if

$$\kappa_{02}>\frac{\theta}{2} \tag{3.2}$$

It follows from (1.7) and (2.6) that $MSE(t_{R2})<MSE(t_{R1})$ if

$$\kappa_{02}<\frac{(1+\theta)}{2} \tag{3.3}$$

Thus from (3.2) and (3.3) it is observed that the suggested estimator t_{R2} is more efficient than t_{Rd} and t_{R1} if

$$\frac{\theta}{2} < \kappa_{02} < \frac{(1+\theta)}{2} \tag{3.4}$$

From (3.1) and (3.4) we see that the estimator t_{R2} dominates the estimators s_{ym}^2, t_{Rd} and t_{R1} if (3.4) holds good along with $C > 1/2$.

4. Generalized estimators

We define the following classes of estimators for S_y^2 as

$$t_g^{(1)} = s_{ym}^2 \left\{\frac{s_{xn}^2}{s_{xm}^2}\right\}^{\alpha_1} \left\{\frac{S_z^2 + \beta_2(z)}{s_{zn}^2 + \beta_2(z)}\right\}^{\alpha_2} \tag{4.1}$$

$$t_g^{(2)} = s_{ym}^2 \left\{\frac{s_{xn}^2}{s_{xm}^2}\right\}^{\alpha_1'} \left\{\frac{S_z^2 + \beta_2(z)}{s_{zn}^2 + \beta_2(z)}\right\}^{\alpha_2'} \left\{\frac{s_{zn}^2 + \beta_2(z)}{S_z^2 + \beta_2(z)}\right\}^{\alpha_3'} \tag{4.2}$$

where $s_{zm}^2 = \sum_{j \in s} (z_j - \bar{z}_m)^2 \Big/ (m-1)$, $\bar{z}_m = \sum_{j \in s} z_j \Big/ m$, α_i's and α_i''s (i =1,2,3) are suitably chosen constants. For (α_1, α_2)=(1,1) and $(\alpha_1', \alpha_2', \alpha_3')$=(1,1,0), the estimator $t_g^{(1)}$ and $t_g^{(2)}$ respectively reduce to the estimator t_{R2} while for $\alpha_i = \alpha_i' = 0 \ \forall \ i = 1,2,3$ and $t_g^{(j)}$ $(j = 1,2)$ reduces to s_{ym}^2.

To the first degree of approximation the mean squared errors of $t_g^{(1)}$ and $t_g^{(2)}$ are respectively given by

$$MSE\left(t_g^{(1)}\right) = S_y^4 \left[\left(\frac{1}{m} - \frac{1}{N}\right) C_0^2 + \left(\frac{1}{m} - \frac{1}{n}\right) \alpha_1 C_1^2 (\alpha_1 - 2\kappa_{01})\right.$$

$$\left. + \left(\frac{1}{n} - \frac{1}{N}\right) \alpha_2 \, \theta C_2^2 (\alpha_2 \theta - 2\kappa_{02})\right] \tag{4.3}$$

and

$$MSE\left(t_g^{(2)}\right) = S_y^4\left[\left(\frac{1}{m}-\frac{1}{N}\right)\left\{C_0^2+\alpha_3\theta C_2^2\left(\alpha_3\theta-2\kappa_{02}\right)\right\}\right.$$

$$+\left(\frac{1}{m}-\frac{1}{n}\right)\left\{\alpha_1\left(\alpha_1-2\kappa_{01}\right)-2\alpha_1\alpha_3\theta\kappa_{21}\right\}C_1^2$$

$$\left.+\left(\frac{1}{n}-\frac{1}{N}\right)\theta\left\{\alpha_2\left(\alpha_2\theta-2\kappa_{02}\right)-2\alpha_2\alpha_3\theta\right\}C_2^2\right] \tag{4.4}$$

which are respectively minimized for

$$\left.\begin{aligned}\alpha_{10}&=\kappa_{01}\\ \alpha_{20}&=\kappa_{02}/\theta\end{aligned}\right\} \tag{4.5}$$

and

$$\left.\begin{aligned}\alpha_{10}^{'}&=\frac{\left(\kappa_{01}-\kappa_{02}\,\kappa_{21}\right)}{\left(1-\kappa_{12}\,\kappa_{21}\right)}\\ \alpha_{20}^{'}&=\frac{\left(\kappa_{01}-\kappa_{02}\,\kappa_{21}\right)\kappa_{12}}{\left(1-\kappa_{12}\,\kappa_{21}\right)\theta}\\ \alpha_{10}^{'}&=\frac{\left(\kappa_{01}\kappa_{12}-\kappa_{02}\right)}{\left(1-\kappa_{12}\,\kappa_{21}\right)\theta}\end{aligned}\right\} \tag{4.6}$$

Thus the resulting minimum mean squared errors of $t_g^{(1)}$ and $t_g^{(2)}$ are respectively given by

$$\min.MSE\left(t_g^{(1)}\right)=S_y^4\,C_0^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{01}^2\right)+\left(\frac{1}{n}-\frac{1}{N}\right)\left(1-\rho_{02}^2\right)\right] \tag{4.7}$$

and

$$\min.MSE\left(t_g^{(2)}\right)=S_y^4\,C_0^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{0.12}^2\right)+\left(\frac{1}{n}-\frac{1}{N}\right)\left(1-\rho_{02}^2\right)\right] \tag{4.8}$$

where $\rho_{0.12}=\sqrt{\left(\rho_{01}^2+\rho_{02}^2-2\rho_{01}\rho_{02}\rho_{12}\right)/\left(1-\rho_{12}^2\right)}$.

From (1.5), (1.6), (1.7), (2.6), (4.7) and (4.8) we have

$$Var\left(s_{ym}^2\right) - \min.MSE\left(t_g^{(1)}\right) = S_y^4 C_0^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right)\rho_{01}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{02}^2\right] \tag{4.9}$$

$$\geq 0$$

$$MSE\left(t_{Rd}\right) - \min.MSE\left(t_g^{(1)}\right) = S_y^4 C_0^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{C_1}{C_0} - \rho_{01}\right)^2\right.$$

$$\geq 0$$

$$\left. + \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{02}^2\right] \tag{4.10}$$

$$MSE\left(t_{R1}\right) - \min.MSE\left(t_g^{(1)}\right) = S_y^4 C_0^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{C_1}{C_0} - \rho_{01}\right)^2\right.$$

$$\geq 0$$

$$\left. + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{C_2}{C_0} - \rho_{02}\right)^2\right] \tag{4.11}$$

$$MSE\left(t_{R2}\right) - \min.MSE\left(t_g^{(1)}\right) = S_y^4 C_0^2 \left[\left(\frac{1}{m} - \frac{1}{n}\right)\left(\frac{C_1}{C_0} - \rho_{01}\right)^2\right.$$

$$\geq 0$$

$$\left. + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\theta\frac{C_2}{C_0} - \rho_{02}\right)^2\right] \tag{4.12}$$

$$\min.MSE\left(t_g^{(1)}\right) - \min.MSE\left(t_g^{(2)}\right) = S_y^4 C_0^2 \left(\frac{1}{m} - \frac{1}{n}\right)\frac{\left(\rho_{01}\rho_{12} - \rho_{02}\right)}{\left(1 - \rho_{12}^2\right)}, \tag{4.13}$$

$$\geq 0$$

Thus, it follows from (4.9), (4.10), (4.11), (4.12) and (4.13) that the suggested estimator $t_g^{(2)}$ is the best estimator among all the estimators discussed here at its optimum conditions.

REFERENCES

CHAND L.(1975): Some ratio-type estimators based on two or more auxiliary variables. *Unpublished Ph. D. dissertation, Iowa State University, Ames, Iowa.*

ISAKI, C.T. (1983): Variance estimation using auxiliary information. *Jour. Amer. Statist. Assoc.,* 78, (381), 117–123.

UPADHYAYA, L.N. and SINGH, H.P. (1999): An estimator for population variance that utilizes the kurtosis of an auxiliary variable in sample surveys. *Vikram Mathematical Journal,* 19, 14–17.