Jour. Ind. Soc. Ag. Statistics
53(2), 2000 : 134-140

An Unbiased Estimator in Two-phase Sampling Using Two Auxiliary Variables

M. Dalabehera and L.N. Sahoo[1]
Orissa University of Agriculture and Technology, Bhubaneswar 751003

(Received : October, 1998)

SUMMARY

This paper presents a new unbiased estimator for the population mean when the population mean of the main auxiliary variable is unknown but the emphasis is laid on the use of an additional auxiliary variable. Empirical studies with the help of two natural populations show that there is a significant gain in efficiency of the suggested estimator over some other estimators.

Key words : Auxiliary variable, Bias, Efficiency, Ratio-type estimator, Ratio-in-ratio estimator, Two-phase sampling, Unbiased estimator.

1. Introduction

Let y and x denote the study variable and the auxiliary variable taking values y_i and x_i $(1 \le i \le N)$ respectively for the ith unit of a population of size N. When the two variables are strongly related but no information is available on the population mean $\overline{X}$ of x, we estimate the population mean $\overline{Y}$ of y, using a two-phase sampling mechanism. This requires a collection of information on x for the first-phase samples s′ of size n′ $(n' < N)$ and on y for a second-phase sample s of size n $(n < n')$ from the first-phase sample. Allowing simple random sampling (WOR) in each phase, the two-phase sampling ratio estimator will be

$$t_R = \overline{y}\frac{\overline{x}'}{\overline{x}}$$

where $$\overline{x}' = \frac{1}{n'}\sum_{i \varepsilon s'} x_i\,,\ \overline{y} = \frac{1}{n}\sum_{i \varepsilon s} y_i \text{ and } \overline{x} = \frac{1}{n}\sum_{i \varepsilon s} x_i$$

In general t_R is biased. But, an analogue of Hartley-Ross [3] unbiased ratio-type estimator was proposed by Sukhatme [11] has the form

[1] Department of Statistics, Utkal University, Bhubaneswar 751004, India

$$t_{HR} = \bar{r}\,\bar{x}' + \frac{n'-1}{n'}\,s_{rx}$$

where $$r_i = \frac{y_i}{x_i}\,,\ \bar{r} = \frac{1}{n}\sum_{i\,\varepsilon\, s} r_i \ \text{ and }\ s_{rx} = \frac{1}{n-1}\sum_{i\,\varepsilon\, s} r_i\,(x_i - \bar{x})$$

Sometimes even if $\bar{X}$ is unknown, information on a cheaply ascertainable variable z, closely related to x, is available on all units of the population. By analogy, if z has a high positive correlation with x, the ratio estimator $\frac{\bar{x}'}{\bar{z}'}\bar{Z}$ will estimate $\bar{X}$ more precisely than $\bar{x}'$. Thus, using $\frac{\bar{x}'}{\bar{z}'}\bar{Z}$ for $\bar{x}'$ in t_R, Chand [1] developed a ratio-in-ratio estimator of the form

$$t = \frac{\bar{x}'}{\bar{x}}\,\frac{\bar{Z}}{\bar{z}'}$$

where $\bar{z}' = \frac{1}{n'}\sum_{i\,\varepsilon\, s'} z_i$ and $\bar{Z}$ is the known population mean of z. Kiregyera ([4], [5]), Sahoo and Sahoo [8] also gave brief discussions of this type of situation.

The main disadvantage associated with the estimator t, is that it is seriously biased. In this paper, Chand's approach is transformed into the construction of an unbiased estimator through Hartley-Rossization.

2. Construction of the Unbiased Estimator

When x and z are positively correlated, an intuitively reasonable, but biased estimator of $\bar{Y}$ may be of the form

$$t' = \bar{r}\,\bar{g}'\bar{Z}$$

where $$\bar{g}' = \frac{1}{n'}\sum_{i\,\varepsilon\, s'} g_i \ \text{ and }\ g_i = \frac{x_i}{z_i}$$

$$\begin{aligned} \text{Now } E(t') &= E\left\{E\left(\frac{t'}{n'}\right)\right\} \\ &= E(\bar{r}'\,\bar{g}')\,\bar{Z} \\ &= \frac{N-n'}{Nn'}\,S_{rg}\,\bar{Z} + \bar{R}\,\bar{G}\,\bar{Z} \end{aligned} \qquad (2.1)$$

where $$\bar{r}' = \frac{1}{n'}\sum_{i\,\varepsilon\, s'} r_i\,,\ \bar{R} = \frac{1}{N}\sum_{i=1}^{N} r_i\,,\ \bar{G} = \frac{1}{N}\sum_{i=1}^{N} g_i$$

and $$S_{rg} = \frac{1}{N-1}\sum_{i=1}^{N} g_i\,(r_i - \overline{R})$$

Thus, the bias in t' is given by

$$B(t') = E(t') - \overline{Y}$$

$$= -\,[\,\overline{Y} - \overline{R}\,\overline{G}\,\overline{Z}\,] + \frac{N-n'}{Nn'}\,S_{rg}\,\overline{Z}'$$

which on simplification yields

$$B(t') = -\,\frac{N-1}{N}\,S_{hz} - \frac{n'-1}{n'}\,S_{rg}\,\overline{Z} \qquad (2.2)$$

where $$h_i = \frac{y_i}{z_i} \text{ and } S_{hz} = \frac{1}{N-1}\sum_{i=1}^{N} h_i\,(z_i - \overline{Z})$$

An unbiased estimate of bias in (2.2) is

$$\hat{B}(t') = -\frac{N-1}{N}\,s_{hz} - \frac{n'-1}{n'}\,s_{rg}\,\overline{Z}$$

where $$s_{hz} = \frac{1}{n-1}\sum_{i\,\varepsilon\, s} h_i\,(z_i - \overline{z}),\quad s_{rg} = \frac{1}{n-1}\sum_{i\,\varepsilon\, s} g_i\,(r - \overline{r})$$

and $\overline{z} = \frac{1}{n}\sum_{i\,\varepsilon\, s} z_i$. Hence, an unbiased estimator of $\overline{Y}$ is given by

$$t_U = \left(\overline{r}\,\overline{g}' + \frac{n'-1}{n'}\,s_{rg}\right)\overline{Z} + \frac{N-1}{N}\,s_{hz}$$

This estimator is flexible in the sense that it can be reduced to many other estimators. In particular, we now consider the following specific cases :

(a) If there is no explicit use of x and z, $t_U = \overline{y}$, the simple expansion estimator of $\overline{Y}$

(b) In case z is not used, $t_U = t_{HR}$

(c) If $n' = N$, and the emphasis is laid on the use of x only, t_U becomes exactly identical to the Hartley-Ross [3] unbiased estimator for single-phase sampling.

(d) When $n' = N$ and both x and z are involved t_U reduces to an unbiased estimator of the form

$$\bar{r}\,\bar{G}\,\bar{Z} + \frac{N-1}{N}\,(\,s_{rg}\,\bar{Z} + s_{hz}\,)$$

for single-phase sampling.

3. *Efficiency of* t_U

Using the technique developed by Tukey [12] and extended by Robson [6] for obtaining exact moments and afterwards assuming N → ∞, Sukhatme [11] derived the approximate expression for the variance of t_{HR} to a first order of approximation. The same result can also be obtained by following the work of Goodman and Hartley [2] for deriving the variance of Hartley and Ross [3] unbiased estimator. In this work following Goodman and Hartley [2] and for simplicity assuming $\frac{n'-1}{n'} \cong 1$ and $\frac{N-1}{N} \cong 1$ we obtain the variance of t_U to terms of order n^{-1} as

$$\begin{aligned} V\,(\,t_U\,) = {} & \frac{1}{n}\,[\,\sigma_y^2 + \bar{H}^2\,\sigma_z^2 + \bar{Z}^2\,\bar{R}^2\,\sigma_g^2 - 2\bar{H}\,\sigma_{yz} - 2\bar{Z}\,\bar{R}\sigma_{yg} + 2\,\bar{H}\,\bar{Z}\,\bar{R}\,\sigma_{gz}\,] \\ & + \frac{1}{n'}\,[\,\bar{Z}^2\,\bar{R}^2\,\sigma_g^2 + \bar{R}\,\bar{G}\,\bar{Z}^2\,\bar{R}\sigma_{rg} \\ & + 2\bar{Z}^2\,\bar{R}\,E\,\{(\,\Delta r\,)\,(\,\Delta g\,)^2\} + 2\bar{Z}\,\bar{R}E(\,\Delta g\,\Delta h\,\Delta z\,)\,] \end{aligned} \tag{3.1}$$

where $\bar{H} = \frac{1}{N}\sum_{i=1}^{N} h_i$, $\Delta y = y - \bar{Y}$, $\Delta z = z - \bar{Z}$, $\Delta r = r - \bar{R}$, $\Delta g = g - \bar{G}$

$\sigma_y^2 = E(\,\Delta y\,)^2$, $\sigma_z^2 = E\,(\,\Delta z\,)^2$, $\sigma_{yz} = E\,(\,\Delta y\,\Delta z\,)$, $\sigma_{rg} = E\,(\,\Delta r\,\Delta g\,)$ etc.

To the same order of approximation, we may also have

$$V\,(\,t_R\,) = \frac{1}{n}\,[\,\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{yx}\,] + \frac{1}{n'}\,[\,-R^2\sigma_x^2 + 2R\sigma_{yx}\,] \tag{3.2}$$

$$\begin{aligned} V\,(\,t_{HR}\,) = {} & \frac{1}{n}\,[\,\sigma_y^2 + \bar{R}^2\sigma_x^2 - 2\bar{R}\sigma_{yx}\,] \\ & + \frac{1}{n'}\,[\,\bar{R}^2\sigma_x^2 + 2\bar{R}\,\bar{X}\sigma_{rx} + 2\bar{R}E\,\{(\,\Delta r\,)\,(\,\Delta x\,)^2\}\,] \end{aligned} \tag{3.3}$$

$$V\,(\,t'\,) = \frac{1}{n}\bar{Z}^2\,\bar{G}^2\,\sigma_r^2 + \frac{\bar{Z}^2}{n'}\,(\,\bar{R}^2\,\sigma_g^2 + 2\,\bar{R}\,\bar{G}\sigma_{rg}\,) \tag{3.4}$$

$$V\,(t) = \frac{1}{n}\,[\,\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{yx}\,] + \frac{1}{n'}\,[\,-R^2\sigma_x^2 + 2R\sigma_{yx} + H^2\sigma_z^2 - 2H\sigma_{yz}\,] \tag{3.5}$$

where $R = \frac{\bar{\bar{Y}}}{\bar{\bar{X}}}$ and $H = \frac{\bar{\bar{Y}}}{\bar{\bar{Z}}}$. The expression in (3.3) and (3.5) are respectively given by Sukhatme [11] and Chand [1].

Now it becomes of interest to investigate the efficiency of t_U relative to others viz. t_R, t_{HR}, t' and t. But, unfortunately substraction of (3.1) from any one of the equations given in (3.2) to (3.5) leads to cumbersome mathematical expressions. Thus, it is difficult to point out the conditions under which gains in efficiency are indeed realizable with t_U. In order to facilitate assessment, an attempt has been made to investigate the efficiency of t_U compared to others empirically with the help of two natural populations.

Population I : Consists of data on yield of fiber per plant in gm (y), height of the plant in feet (x) and base diameter in cm (z) for 50 jute plants (capsularies) given by Shukla [10].

Population II : Consists of data on weight in gm (y), total length in mm (x) and body height in mm (z) for 50 fish (type catla catla) given by Sahoo [7].

Numerical values of the coefficients of $\frac{1}{n}$ and $\frac{1}{n'}$ in the variance formulae of different estimators given in (3.1) to (3.5) are presented in the Table 3.1. Coefficient of $\frac{1}{n}$ in the variance formula of $\bar{y}$ is also computed to facilitate the comparison.

Table 3.1. Numerical values of the coefficients of $\frac{1}{n}$ and $\frac{1}{n'}$

Variance	Population I		Population II	
	coefficient of		coefficient of	
	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n}$	$\frac{1}{n'}$
$V(\bar{y})$	1.8389	-	10123.3764	-
$V(t_R)$	1.0631	0.7692	2158.7761	7964.5929
$V(t_{HR})$	1.0666	0.7657	2182.0983	7921.7446
$V(t')$	1.0748	0.2785	2093.7943	3285.7578
$V(t)$	1.0631	0.2617	2158.7761	2246.0796
$V(t_U)$	0.9730	0.2551	2040.4546	2195.8614

Thus, from Table 3.1, an impressive gain in the efficiency due to t_U is seen for both the populations under consideration.

4. Variance Estimation

For estimation of variance, we resort to interpenetrating subsampling. Instead of drawing a single second-phase sample s of size n from first-phase sample s′, we draw m independent sample, $s_1, s_2, \ldots, s_m$ of equal size n/m by simple random sampling (WOR). We assume that the first sample s_1 is drawn and then replaced into s′. The second sample s_2 is then drawn by sampling again from the whole s′ and independently of s_1. Then s_2 is replaced into s′, a third independent sample s_3 is drawn, and so on until m independent samples have been drawn.

For each k = 1, 2, ... , m, if an unbiased estimator $t_U^{(k)}$ of $\overline{Y}$ is calculated using data on s_k only, then an unbiased estimator of $\overline{Y}$ based on m independent samples is

$$\overline{t}_U = \frac{1}{m} \sum_{k=1}^{m} t_U^{(k)}$$

An unbiased estimator of $V(\overline{t}_U)$, the variance of $\overline{t}_U$, is given by

$$\hat{V}(\overline{t}_U) = \frac{1}{m(m-1)} \sum_{k=1}^{m} (t_U^{(k)} - \overline{t}_U)^2$$

Very often $\hat{V}(\overline{t}_U)$ can be used for estimating $V(t_U)$, even though $\hat{V}(\overline{t}_U)$ is not unbiased for $V(t_U)$, and t_U and $\overline{t}_U$ are non-identical.

A sometimes used alternative estimator of $V(t_U)$ is

$$\hat{V}(t_U) = \frac{1}{m(m-1)} \sum_{k=1}^{m} (t_U^{(k)} - t_U)^2$$

Little seems to be known about whether $\hat{V}(\overline{t}_U)$ or $\hat{V}(t_U)$ should be preferred as an estimator of $V(t_U)$. However,

$$\hat{V}(t_U) \geq \hat{V}(\overline{t}_U)$$

showing that $\hat{V}(t_U)$ is more conservative than $\hat{V}(\bar{t}_U)$. When it comes to estimating $V(\bar{t}_U)$ is follows that $\hat{V}(t_U)$ will have a positive bias.

ACKNOWLEDGEMENT

The authors are grateful to the referee for his valuable comments which improved the paper considerably.

REFERENCES

[1] Chand, L. (1975). Some ratio-type estimators based on two or more auxiliary variables. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

[2] Goodman, L.A. and Hartley, H.O. (1958). The precision of unbiased ratio-type estimators. *Jour. Amer. Statist. Assoc.*, **53**, 491-508.

[3] Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimates. *Nature*, **174**, 270-271.

[4] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, **27**, 217-223.

[5] Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika*, **31**, 215-226.

[6] Robson, D.S. (1957). Applications of multivariate polykays to the theory of unbiased ratio-type estimation. *Jour. Amer. Statist. Assoc.*, **52**, 511-522.

[7] Sahoo, L.N. (1980). Some problems of estimation using auxiliary information in sampling from finite population. Unpublished Ph.D. dissertation. Utkal University, Bhubaneswar, India.

[8] Sahoo, J and Sahoo, L.N. (1993). A class of estimators in two- phase sampling using two auxiliary variables. *Jour. Indian Stat. Assoc.*, **31**, 107-114.

[9] Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

[10] Shukla, G.K. (1966). An alternative multivariate ratio estimate for finite population. *Bull. Calcutta Stat. Assoc.*, **15**, 127-134.

[11] Sukhatme, B.V. (1962). Some ratio-type estimators in two-phase sampling. *Jour. Amer. Statist. Assoc.*, **57**, 628-632.

[12] Tukey, J.W. (1950). Some sampling simplificated. *Jour. Amer. Statist. Assoc.*, **45**, 501-519.