Journal of Statistical Research *ISSN 0256 - 422 X*
2006, Vol. 40, No. 2, pp. 13–20
Bangladesh

UNBIASED ESTIMATION OF THE POPULATION VARIANCE USING MIDZUNO-SEN TYPE SAMPLING SCHEME

M. C. Agrawal
Department of Statistics,
University of Delhi, Delhi - 110007, INDIA
Email: mc_agrawal@yahoo.com

A. B. Sthapit
Department of Statistics,
Tribhuvan University, Kathmandu, NEPAL
Email: azaya_sthapit@enet.com.np

SUMMARY

We have employed ms type sampling scheme to propose two unbiased strategies for estimating the population variance. These strategies have been compared with certain known ones, and necessary and sufficient conditions have been obtained for their superior performance as compared to the known ones. An unbiased variance estimator of the population variance has also been worked out. Real-life data are shown to yield substantial gains via these strategies.

Keywords and phrases: Unbiased estimation of the population variance; Midzuno-Sen type sampling scheme

1 Introduction

By and large, the estimators of the population variance (based on auxiliary information) that have been proposed in the literature were not mooted from the point of view of statistical property of unbiasedness. Although unbiasedness should not be an obsessive property, yet it is desirable to seek unbiasedness of estimators whenever it is feasible. For the purpose of obtaining an unbiased estimator of the population variance, we, in this paper, take to Midzuno-Sen type sampling scheme.

2 Some Unbiased Estimators of the Population Variance under Midzuno-Sen Type Sampling Scheme

Consider a finite population of N units in which y_i and x_i are the measurements in respect of the study variable y and the auxiliary variable x taken on the ith unit $(i = 1, 2, \cdots, N)$ of the population from which a sample s of size n is drawn according to a certain sampling design. Let $\overline{Y}$ and $\overline{y}$ be the population and the sample means respectively of the study variable y and let $\overline{X}$ and $\overline{x}$ be the population and the sample means respectively of the auxiliary variable x. We now define the following population and sample quantities:

$$\mu_{r,s} = 1/N \sum_{i=1}^{N} (x_i - \overline{X})^r (y_i - \overline{Y})^s, \qquad m_{r,s} = 1/N \sum_{i=1}^{N} x_i^r y_i^s$$

(for any specified r and s),

$$\beta_2(y) = \frac{\mu_{04}}{\mu_{02}^2}, \qquad \theta = \frac{\mu_{22}}{\mu_{02}\mu_{20}}$$

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2 \quad \text{and} \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

We similarly define the quantities $\beta_2(x)$, S_x^2 and s_x^2 for the variable x which being based on the auxiliary information are supposed to be known. Further, later in this paper, we would, to terms of $O\left(\frac{1}{n}\right)$, use the following well-known results:

$$V(s_y^2) = \frac{\lambda}{n} S_y^4 (\beta_2(y) - 1), \qquad V(s_x^2) = \frac{\lambda}{n} S_x^4 (\beta_2(x) - 1)$$

$$\text{Cov}(s_y^2, s_x^2) = \frac{\lambda}{n} S_y^2 S_x^2 (\theta - 1),$$

where $\lambda = \frac{N-n}{N}$.

An unbiased estimator of the population variance, under the simple random sampling without replacement design, say, p_0, when no auxiliary variable is used, is given by

$$t = s_y^2 \tag{2.1}$$

Isaki (1983) proposed the ratio-type estimator of the population variance

$$t_0 = \frac{s_y^2}{s_x^2} S_x^2 \tag{2.2}$$

which is biased under the sampling design p_0. Although Agrawal and Sthapit (1995) alluded to the sampling designs which render t_0 unbiased, but, to terms of $O\left(\frac{1}{n}\right)$, the variance of t_0

under these designs remains equal to the one under the design p_0. Hence, we would continue to discuss t_0 under the design p_0.

It is known that, under the Midzuno-Sen sampling scheme, the probability of selecting a specified sample s is given by

$$p(s) = \frac{1}{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}} \sum_{i \in s} p_i$$

where p_i is the initial probability of selecting the ith unit. If we consider p_i in accordance with either of the following schemes for a suitably chosen r,

(a) $p_i \ \alpha \ x_i^r$

and

(b) $p_i \ \alpha \ (x_i - \overline{X})^r$,

then we obtain

$$p_1(s) = \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} \frac{\widehat{m}_{r,o}}{m_{r,o}}$$

for scheme (a) and

$$p_2(s) = \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} \frac{\widehat{\mu}_{r,o}}{\mu_{r,o}}$$

for scheme (b), where $\widehat{m}_{r,o}$ and $\widehat{\mu}_{r,o}$ are the sample-based quantities corresponding to $m_{r,o}$ and $\mu_{r,o}$. Now, we propose the estimators of the population variance under scheme (a) as

$$t_1 = s_y^2 \frac{m_{r,o}}{\widehat{m}_{r,o}} \tag{2.3}$$

and under scheme (b) as

$$t_2 = s_y^2 \frac{\mu_{r,o}}{\widehat{\mu}_{r,o}} \tag{2.4}$$

Both the estimators t_1 and t_2 can be verified as being unbiased. For this purpose, we note that

$$\begin{aligned} E_{p_1}(t_1) &= \sum_{s \in s} p_1(s) t_1(s) \\ &= \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} \sum_{s \in s} s_y^2 \\ &= E_{p_0}(s_y^2) \\ &= S_y^2. \end{aligned}$$

Similarly, the estimator t_2 can be shown to be unbiased.

Denoting the strategies (p_0, t), (p_0, t_0), (p_1, t_1) and (p_2, t_2) by D, D_0, D_1 and D_2 respectively, we compare them in the next section.

3 A Comparison of the Competing Strategies

The variance, to terms of $O\left(\frac{1}{n}\right)$, for the strategy D, when no auxiliary information is used, is

$$V_{p_0}(t) = \frac{\lambda}{n} S_y^4 [\beta_2(y) - 1]. \tag{3.1}$$

The mean square error (MSE), to terms of $O\left(\frac{1}{n}\right)$, of the strategy D_0 is given by

$$\mathrm{MSE}_{p_0}(t_0) = \frac{\lambda}{n} S_y^4 [\beta_2(y) + \beta_2(x) - 2\theta]. \tag{3.2}$$

Now, we proceed to obtain the variances of the proposed estimators t_1 and t_2 (defined by (2.3) and (2.4)) under the designs $p_1(s)$ and $p_2(s)$ respectively. For the strategy D_1, we can write

$$\begin{aligned} V_{p_1}(t_1) &= E_{p_1}(t_1^2) - S_y^4 \\ &= m_{r,0} \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} \sum_{s \in S} (s_y^4 / \widehat{m}_{r,0}) - S_y^4 \\ &= m_{r,0} E_{p_0}(s_y^4 / \widehat{m}_{r,0}) - S_y^4 \end{aligned}$$

which, after some algebra, is obtainable, to terms of $O\left(\frac{1}{n}\right)$, as

$$V_{p_1}(t_1) = \frac{\lambda}{n} S_y^4 \left[\beta_2(y) + \frac{\mu^*_{2r,0}}{\mu^{*2}_{r,0}} - \frac{2\mu^*_{r,2}}{\mu^*_{0,2}\mu^*_{r,0}} \right] \tag{3.3}$$

where $\mu^*_{r,s} = \dfrac{1}{N}\sum_{i=1}^{N} x_i^r (y_i - \overline{Y})^s$.

In a similar manner to the above, we can work out the variance, to terms of $O\left(\dfrac{1}{n}\right)$, for the strategy D_2 as

$$V_{p_2}(t_2) = \frac{\lambda}{n} S_y^4 \left[\beta_2(y) + \frac{\mu_{2r,0}}{\mu^2_{r,0}} - \frac{2\mu_{r,2}}{\mu_{0,2}\mu_{r,0}}\right] \tag{3.4}$$

Now, by setting $(y_i - \overline{Y})^2 = w_i$, $(x_i - \overline{X})^2 = u_i$, $x_i^r = v_i^*$ and $u_i^{r/2} = v_i$, the various variance expressions given by (3.1), (3.2), (3.3) and (3.4) can be expressed respectively as

$$V_{p_0}(t) = \frac{\lambda}{n} \overline{W}^2 C_0^2 \tag{3.5}$$

$$V_{p_0}(t_0) = \frac{\lambda}{n} \overline{W}^2 (C_0^2 + C_1^2 - 2\rho_0 C_0 C_1), \tag{3.6}$$

$$V_{p_1}(t_1) = \frac{\lambda}{n} \overline{W}^2 (C_0^2 + C_2^2 - 2\rho_1 C_0 C_2), \tag{3.7}$$

and
$$V_{p_2}(t_2) = \frac{\lambda}{n} \overline{W}^2 (C_0^2 + C_3^2 - 2\rho_2 C_0 C_3) \tag{3.8}$$

where C_0, C_1, C_2 and C_3 are the coefficients of variation of w, u, v^* and v respectively and ρ_0, ρ_1 and ρ_2 are the coefficients of correlation between w and u, w and v^*, and w and v respectively.

Needless to say, for employing the strategies D_1 and D_2, a proper choice of r has to be made. Regarding the relative performance of the competing strategies D, D_0, D_1 and D_2, we can, based on the relevant variances given by (3.5), (3.6), (3.7) and (3.8), arrive at the following conclusions:

(i) The strategy D_0 scores over the strategy D if and only if

$$\rho_0 \geq \frac{1}{2}\,\frac{C_0}{C_1};$$

(ii) The strategy D_1 performs better than D_2 if and only if

$$\frac{1}{2}\,\frac{C_3}{C_0}\left(\frac{C_2^2}{C_3^2} - 1\right) - \left(\rho_1 \frac{C_2}{C_3} - \rho_2\right) \leq 0;$$

(iii) The strategy D_1 will outperform the strategy D_0 if and only if

$$\frac{1}{2}\,\frac{C_1}{C_0}\left(\frac{C_2^2}{C_1^2} - 1\right) - \left(\rho_1 \frac{C_2}{C_1} - \rho_0\right) \leq 0;$$

while the strategy D_2 performs better than the strategy D_0 if and only if

$$\frac{1}{2}\,\frac{C_1}{C_0}\left(\frac{C_3^2}{C_1^2} - 1\right) - \left(\rho_2 \frac{C_3}{C_1} - \rho_0\right) \leq 0$$

and
(iv) The strategy D_1 fares better than the strategy D if and only if

$$\rho_1 \geq \frac{1}{2}\,\frac{C_2}{C_0};$$

while the strategy D_2 scores over the strategy D if and only if

$$\rho_2 \geq \frac{1}{2}\,\frac{C_3}{C_0}.$$

4 Unbiased Variance Estimation

To obtain an unbiased estimator, under the design p_1, of the variance of t_1, we write

$$V_{p_1}(t_1) = E_{p_1}(t_1^2) - S_y^4$$

which yields

$$\widehat{V}_{p_1}(t_1) = t_1^2 - \widehat{S_y^4}. \tag{4.1}$$

Now S_y^4 can be expressed as

$$\begin{aligned} S_y^4 &= \frac{1}{(N-1)^2}\left[\sum_{i=1}^{N} y_i^4 - 2N\overline{Y}^2 \sum_{i=1}^{N} y_i^2 + N^2\overline{Y}^4 \sum_{i\neq j}^{N} y_i^2 y_j^2\right] \\ &= \frac{1}{N^2(N-1)^2}\Bigg[(N-1)^2\sum_{i=1}^{N} y_i^4 - 4(N-1)\sum_{i\neq j}^{N} y_i^3 y_j + (N^2-2N+3)\sum_{i\neq j}^{N} y_i^2 y_j^2 \\ &\quad -2(N-3)\sum_{i\neq j\neq k}^{N} y_i^2 y_j y_k + \sum_{i\neq j\neq k\neq l}^{N} y_i y_j y_k y_l\Bigg] \end{aligned} \tag{4.2}$$

Since, under the design p_1, we have

$$E_{p_1}\left[\frac{N}{n}\sum_{i=1}^{n}\frac{m_{r,0}}{\widehat{m}_{r,0}}\,y_i^4\right] = \sum_{i=1}^{N} y_i^4,$$

$$E_{p_1}\left[\frac{N(N-1)}{n(n-1)}\sum_{i\neq 1}^{n} y_i^2 y_j^2\frac{m_{r,0}}{\widehat{m}_{r,0}}\right] = \sum_{i\neq 1}^{N} y_i^2 y_j^2$$

and so on, we can, thus, replace all the terms of the right hand side of (4.2) by the respective unbiased estimating quantities and then, after some algebra, we obtain an unbiased estimator of S_y^4 as

$$\widehat{S}_y^4 = \frac{1}{AN(N-1)}\frac{m_{r,0}}{\widehat{m}_{r,0}}\left[C\sum_{i=1}^{n}\left(y_i^2 - \sum_{i=1}^{n} y_i^2/n\right)^2 + 4C\left\{\left(\sum_{i=1}^{n} y_i^2\right)^2/n - \overline{y}\sum_{i=1}^{n} y_i^3\right\} + Bs_y^4\right] \tag{4.3}$$

where $A = (n-1)(n-2)(n-3)$

$B = n(n-1)^2(N-2)(N-3)$

$C = (N-n)(N+n+1-Nn)$

which, to terms of $O\left(\frac{1}{n}\right)$, can be expressed as

$$\widehat{S}_y^4 = \frac{m_{r,0}}{\widehat{m}_{r,0}}\left[\left(1+\frac{4\lambda}{n}\right)s_y^4 - \frac{\lambda}{n^2}\sum_{i=1}^{n}\left(y_i^2 - \sum_{i=1}^{n} y_i^2/n\right)^2 - \frac{4\lambda}{n^2}\left\{\left(\sum_{i=1}^{n} y_i^2\right)^2/n - \overline{y}\sum_{i=1}^{n} y_i^3\right\}\right], \quad (4.4)$$

and the same is then inserted in (4.1) to obtain the requisite variance estimator of t_1. In a similar manner to the above, we obtain, under the sampling design p_2, an unbiased estimator of S_y^4 if we replace $m_{r,0}$ and $\widehat{m}_{r,0}$ by $\mu_{r,0}$ and $\widehat{\mu}_{r,0}$ respectively, and hence the variance estimator of t_2.

5 Empirical Investigation

To illustrate the potential gain that might accrue from the use of the proposed strategies D_1 and D_2 over the known ones, viz., D and D_0, we consider the following Data-Sets:

Data-Set 1: We consider first fifty four (1-54) observations from Murthy (1967, p.178) and the following quantities are obtained therefrom:

$N = 54$, $\beta_2(y) = 3.799$, $\beta_2(x) = 2.012$, $\theta = 1.627$, $\frac{\text{MSE}(t_0)}{\frac{\lambda}{n}S_y^4} = 2.557$, $\frac{V(t)}{\frac{\lambda}{n}S_y^4} = 2.799$, $\frac{V(t_1)}{\frac{\lambda}{n}S_y^4} = 2.209$ (for $r = 4$) and $\frac{V(t_2)}{\frac{\lambda}{n}S_y^4} = 2.557$ (for $r = 2$).

Data-Set 2: We refer to the data available in Kish (1965, p.213, Ex.6.6). However, treating the given data as unclustered, we compute the following quantities therefrom:

$N = 17$, $\beta_2(y) = 10.078$, $\beta_2(x) = 3.979$, $\theta = 5.687$, $\frac{\text{MSE}(t_0)}{\frac{\lambda}{n}S_y^4} = 2.683$, $\frac{V(t)}{\frac{\lambda}{n}S_y^4} = 9.078$, $\frac{V(t_1)}{\frac{\lambda}{n}S_y^4} = 0.370$ (for $r = 7$) and $\frac{V(t_2)}{\frac{\lambda}{n}S_y^4} = 0.382$ (for $r = 4$).

Data-Set 3: We refer to the data available in Singh and Choudhary (1989, p.141) and have computed the following quantities:

$N = 22$, $\beta_2(y) = 13.257$, $\beta_2(x) = 5.579$, $\theta = 7.713$, $\frac{\text{MSE}(t_0)}{\frac{\lambda}{n}S_y^4} = 3.410$, $\frac{V(t)}{\frac{\lambda}{n}S_y^4} = 12.257$, $\frac{V(t_1)}{\frac{\lambda}{n}S_y^4} = 0.524$ (for $r = 7$) and $\frac{V(t_2)}{\frac{\lambda}{n}S_y^4} = 0.528$ (for $r = 6$).

In respect of the above Data-Sets, we compute the following percent gains

$$G_1 = \left[\frac{V(t)}{V(t_1)} - 1\right]\times 100$$
$$G_1' = \left[\frac{MSE(t_0)}{V(t_1)} - 1\right]\times 100$$
$$G_2 = \left[\frac{V(t)}{V(t_2)} - 1\right]\times 100$$
$$G_2' = \left[\frac{MSE(t_0)}{V(t_2)} - 1\right]\times 100$$

and presented them in the following table:

Table 1: Percent gains of t_1 and t_2 relative to t and t_0

Data-Set	G_1	G_1'	G_2	G_2'
1	26.71 (4*)	15.75 (4*)	9.46 (2*)	0 (2*)
2	2353.51 (7*)	625.14 (7*)	2276.44 (4*)	602.36 (4*)
3	2239.12 (7*)	550.76 (7*)	2221.40 (6*)	545.83 (6*)

(* indicates choice of r)

Table 1 bears it out that, for the estimating the population variance, the newly proposed strategies, D_1 and D_2 that make use of Midzuno-Sen type sampling schemes are, apart from being unbiased, capable of yielding substantial, gains in precision as compared to the known strategies D and D_0. However, between D_1 and D_2, the former is slightly better than the latter.

References

[1] Agrawal, M.C. and Sthapit, A.B. (1995). Unbiased Ratio-Type Variance Estimation. *Statistics and Probability Letters,* **25***, 361-364.*

[2] Isaki, C.T. (1983). Variance Estimation Using Auxiliary Information. *Journal of the American Statistical Association,* **76***, 117-123.*

[3] Kish, L. (1965). Survey Sampling. *John Wiley and Sons, New York.*

[4] Murthy, M.N. (1967). Sampling Theory and Method. *Statistical Publishing Society, Calcutta, India.*

[5] Singh, D. and Chaudhary, F.S. (1989). Theory and Analysis of Sample Survey Designs. *Wiley Eastern Ltd., Delhi, India.*