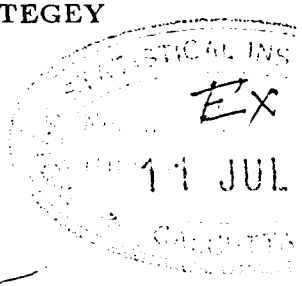
Pak. J. Statist.
1994 Vol. 10(3) pp 575-583

# ON NON-NEGATIVE UNBIASED VARIANCE ESTIMATION FOR MIDZUNO STRATEGEY



D. S. Tracy
Department of Mathematics,
University of Windsor, Canada

And

P. Mukhopadhyay[1]
Indian Statistical Institute, Calcutta, India

(Received: September, 1993 Accepted: April, 1994)

## ABSTRACT

*Vijayan (1975) and Rao (1979) obtained the necessary forms of non-negative quadratic unbiased estimators of mean square error of a linear estimator of population total. Here we consider different unbiased variance estimators which satisfy those necessity conditions for Midzuno strategy. Their properties, viz., probabilities of their taking negative values and stability have been studied empirically. The modified non-negative estimators, as in Rao and Vijayan (1977), have also been considered. The present study covers sixteen different estimators.*

## KEY WORDS

Non-negative Unbiased Variance Estimation; mean square error Midzuno strategy;

## 1. INTRODUCTION AND PRELIMINARIES

Consider a finite population $\mathcal{P}$ of $N$ identifiable units labelled $1, \ldots, i, \ldots, N$. Associated with $i$ are two real quantities $(Y_i, x_i)$, values of a main variable '$y$' and a closely related auxiliary variable '$x$' respectively $(i = 1, \ldots, N)$. In a sample survey for estimating the population total $T = \sum_1^N Y_i$ (mean $\bar{Y} = T/N$), a sample $s$ (a part of $\mathcal{P}$, with units repeated or without repetition) is selected according to a sampling plan $p$ with $p(s)$ as the probability of selecting $s(p(s) \geq 0, \sum_{s \in S} p(s) = 1, S = \{s\})$

[1]Research carried out at the University of Windsor, on leave from the Indian Statistical Institute, Calcutta.

and an estimator $e(s, \underset{\sim}{y})$, a function on $S \times R_N$ such that its value depends on $\underset{\sim}{y} = (y_1, \ldots, y_N)$ only through those $y_i$ for which $i \in s$, is employed. A combination $(p, e)$ is called a sampling strategy.

For estimating $T$, Midzuno (1950, 1952) - Lahiri (1951) - Sen (1952) proposed the following sampling strategy. The first unit in the sample (of size $n$) is chosen with probability $p_i (= x_i/X, X = \sum_1^N x_i)$ and the subsequent $(n-1)$ units by simple random sampling without replacement (srswor) from $\mathcal{P} - \{i\}$. Thus

$$p(s) = \frac{x_s}{XM_1} \tag{1.1}$$

where $x_s = \sum_{i \in s} x_i$, $M_i = \binom{N-i}{n-i}, i = 1,2$. The ratio estimator $e_R = Xy_s/x_s$, where $y_s = \sum_{i \in s} y_i$, is unbiased for $Y$ with variance

$$V(e_R) = \frac{1}{M_1}\left[\sum_1^N Y_i^2 \sum_{s \ni i} \frac{X}{x_s} + \sum_{i \neq j=1}^N \sum Y_i Y_j \sum_{s \ni i,j} \frac{X}{x_s}\right] - T^2 \tag{1.2}$$

and as an unbiased variance estimator

$$v(e_R) = e_R^2 - \frac{X}{x_s}\left\{\sum_{i \in s} y_i^2 + \frac{N-1}{n-1} \sum\sum_{\substack{i \neq j \\ \in s}} y_i y_j\right\}. \tag{1.3}$$

We shall consider the above Midzuno strategy. The estimator $v(e_R)$ can often take negative values, an undesirable property for a variance estimator. Rao (1972, 1977) and Chaudhuri (1976) considered non-negative unbiased estimator of $V(e_R)$.

Vijayan (1975) and Rao (1979) studied the necessary form of a non-negative quadratic unbiased estimator (nnque) of mean square error (MSE) of a linear unbiased estimator of $T$. Their result may be stated as follows:

**Theorem 1.** Let $\hat{Y} = \sum_{i \in s} b_{s_i} Y_i, b_{s_i} = 0$ for $i \notin s$, be a linear estimator of $T$. If MSE $(\hat{Y}) = 0$ when $Y_i = cw_i, i = 1, \ldots, N, w_i$'s being some known constants and $c$ an arbitrary constant, then

$$\text{MSE}\,(\hat{Y}) = -\sum_{i<j=1}^N \sum w_i w_j (Z_i - Z_j)^2 d_{ij} \tag{1.4}$$

where $Z_i = Y_i/w_i$, $d_{ij} = E(b_{s_i} - 1)(b_{s_j} - 1)$. Further, a nnque of MSE $(\hat{Y})$ is necessarily of the form

$$m(\hat{Y}) = -\sum_s{}' w_i w_j (z_i - z_j)^2 e_{ij}(s) \tag{1.5}$$

where

$$E(e_{ij}(s)) = d_{ij} \tag{1.6}$$

and $\sum_s'$ denotes $\sum\sum_{i<j<\in s}$. It may be noted that the equation (1.5) only provides a necessary condition for nnque of MSE $(\hat{Y})$. However, all estimators of the form (1.5) are not necessarily non-negative, i.e. the condition (1.5) is not sufficient to ensure non-negativity.

Mukhopadhyay and Vijayan (1990) investigated explicitly the different forms of nnque of $V(\hat{Y})$. When $\hat{Y} = \sum_{i\in s} b_{s_i} Y_i$ is unbiased,

$$d_{ij} = E(b_{s_i} b_{s_j}) - 1 = h_{ij} - 1 \text{ (say)} .$$

Now, from (1.4),

$$V(\hat{Y}) = \sum_{i<j=1}^{N}\sum g_{ij}(1 - h_{ij}) \tag{1.7}$$

where $g_{ij} = w_i w_j (Z_i - Z_j)^2$. One may thus get different forms of nnque of $V(\hat{Y})$ as

$$v_{k\ell} = \sum_s' g_{ij} 1^{(k)} - \sum_s' g_{ij} h_{ij}^{(\ell)}, k, \ell = 0,1,2,3, \tag{1.8}$$

where

$$1^{(0)} = \frac{b_{s_i} b_{s_j}}{E(b_{s_i} b_{s_j})}, \quad h_{ij}^{(0)} = b_{s_i} b_{s_j} \tag{1.9}$$

and

$$h_{ij}^{(1)} = \frac{h_{ij}}{M_2 p(s)}, \tag{1.10}$$

$$h_{ij}^{(2)} = \frac{h_{ij}}{\pi_{ij}}, \tag{1.11}$$

$$h_{ij}^{(3)} = \frac{h_{ij} P(s \mid i,j)}{p(s)}, \tag{1.12}$$

and similarly for $1^{(1)}, 1^{(2)}, 1^{(3)}$, when $\pi_{ij} = \sum_{s\ni i,j} p(s)$ and $P(s \mid i,j)$ denotes the conditional probability of selecting $s$ given that $i$ and $j$ were selected at the first two draws. In practice, many of these 16 estimators $v_{k\ell}$ would coincide.

It may be noted, however, that the estimators, $v_{k\ell}$ $(k, \ell = 1,\ldots,4)$ do not form an exhaustive set of estimators of the form (1.5). Padmawar (1982) has given other (more complex) estimators of this form. However, the estimators (1.8) are interesting for their simplicity.

## 2. NNU-VARIANCE ESTIMATION FOR MIDZUNO STRATEGY

Since $V(e_R) = 0$, for $y_i \alpha x_i, i = 1, \ldots, N$, it follows from (1.4) that for Midzuno strategy

$$V(e_R) = \sum_{i<j=1}^{N} \sum c_{ij} \left\{ 1 - \frac{1}{M_1} \sum_{s \ni i,j} \frac{X}{x_s} \right\} \tag{2.1}$$

where

$$c_{ij} = \left( \frac{y_i}{x_i} - \frac{y_j}{x_j} \right)^2 x_i x_j .$$

Different estimators $v_{k\ell}(k\,\ell = 0, 1, 2, 3)$ satisfying the necessarily non-negativity conditions (1.5), (1.6) are:

$$v_1 = v_{11} = v_{31} = v_{33} = \sum_s c_{ij} \frac{(N-1)X}{(n-1)x_s} \left\{ 1 - \frac{1}{M_1} \sum_{s \ni i,j} \frac{X}{x_s} \right\}$$

$$v_2 = v_{20} = \sum_s{}' c_{ij} \left( \frac{1}{\pi_{ij}} - \frac{X^2}{x_s^2} \right)$$

$$v_3 = v_{23} = v_{21} = \sum_s{}' c_{ij} \left\{ \frac{1}{\pi_{ij}} - \frac{X}{M_2 x_s} \sum_{s \ni i,j} \frac{X}{x_s} \right\}$$

$$v_4 = v_{12} = v_{13} = v_{32} = \sum_s{}' c_{ij} \left[ \frac{X}{x_s} \left\{ \frac{N-1}{n-1} - \frac{1}{M_2} \sum_{s \ni i,j} \frac{X}{x_s} \right\} \right]$$

$$v_5 = v_{00} = \sum_s{}' c_{ij} \frac{X^2}{x_s^2} \left( \frac{M_1}{X \sum\limits_{s \ni i,j} \frac{1}{x_s}} - 1 \right)$$

$$v_6 = v_{01} = v_{03} = \sum_s{}' c_{ij} \frac{X}{x_s} \left[ \frac{M_1}{x_s} \frac{1}{\sum\limits_{s \ni i,j} \frac{1}{x_s}} - \frac{X}{M_2} \sum_{s \ni i,j} \frac{1}{x_s} \right]$$

$$v_7 = v_{02} = \sum_s{}' c_{ij} X \left[ \frac{M_1}{x_s^2} \frac{1}{\sum\limits_{s \ni i,j} \frac{1}{x_s}} - X \left( \sum_{s \ni i,j} \frac{1}{x_s} \right) \frac{1}{\sum\limits_{s \ni i,j} x_s} \right]$$

$$v_8 = v_{10} = v_{30} = \sum_s{}' c_{ij} \left[ \frac{X}{x_s} \left( \frac{N-1}{n-1} - \frac{X}{x_s} \right) \right]$$

$$v_9 = v_{22} = \sum_s{}' \frac{c_{ij}}{\pi_{ij}} \left\{ 1 - \frac{1}{M-1} \sum_{s \ni i,j} \frac{X}{x_s} \right\}.$$

Of the 16 possible estimators, nine are distinct, denoted as $v_1 \ldots, v_9$.

Rao and Vijayan (1977) considered the estimators $v_s$ and $v_9$ and studied their stabilities and probability of getting a negative value empirically.

In this note we consider the performance of all the nine estimators $v_1, \ldots, v_9$ empirically. 22 populations, of which 10 are shown in table 1, were considered, including the 14 populations considered by Rao and Vijayan (1977). The cases $n =$ 3, 4 and 5 were investigated. For the cases $n = 4,5$ to save computer time, samples were drawn from modified populations, where the populations remain unchanged if $N \leq 10$, but were restricted to first 10 units if $N$ exceeded 10.

Since $v_4$ was found to have smaller variance $V(v_4) = V_4$ in many of the cases, efficiency, $e_i = V_4/V_i$ of the estimator $v_i$ was calculted with respect to $v_4$ where $V_i = V(v_i)$, $i(\neq 4) = 1, \ldots, 9$.

As in Rao and Vijayan (1977), $v_i$ was modified to a biased non-negative estimator $v_i^*$ as follows:

$$\begin{aligned} v_{is}^* &= v_{is} \text{ when } v_{is} \geq 0, \\ &= \hat{v}_s X^2 \text{ if } v_{is} < 0. \end{aligned}$$

Here $\hat{v}_s$ is the least squares estimates ($\ell se$) of $\mathcal{V}(\hat{\beta}_s) = \mathcal{E}(\hat{\beta}_s - \beta)^2$ under the model

$$\begin{aligned} &Y_i' = \beta x_i + e_i \\ &\mathcal{E}(e_i \mid x_i) = 0, \mathcal{E}(e_i^2 \mid x_i) = \sigma^2 x_i^2 \\ &\mathcal{E}(e_i e_j \mid x_i x_j) = 0, \; i \neq j \end{aligned}$$

where $Y_i'$ is a random variable whose one particular value is $Y_i, \mathcal{E}, \mathcal{V}$ denote respectively, the expectation and variance operator with respect to the model and $\hat{\beta}_s$ is the lse of $\beta$. Thus

$$\hat{v}_s = \frac{1}{n(n-1)} \sum_{i \in s} \frac{1}{x_i^2} (y_i - \hat{\beta}_s x_i)^2.$$

The model is appropriate for situations when the ratio estimator is appropriate.

The relative efficiency of $v_i^*$ with respect to $v_4^*$, denoted by $e_i^*/V_4^* = V_i^*$, (where $V_j^* = \text{MSE}\ (v_j^*), i(\neq 4) = 1, \ldots, 9$), and the relative bias $b_j^*$ where $b_j^* = \mid E(v_j^*) - V(e_R) \mid / \sqrt{\text{MSE}\ (v_j^*)}, j = 1, \ldots, 9$ were also calculated for these 22 populations.

Table 2 presents the probabilities $p_i$ of taking negative values (given by the relative frequency of number of samples yielding negative variance estimates) and the relative efficiency $e_i$ of the estimators $v_i (i = 1, 4, 5$ and 9) for samples of sizes

$n = 4$ and $n = 5$ drawn from the 10 natural populations listed in table 1. The full details for all the 22 populations for $n = 3$, 4 and 5 are available with the authors.

## 3. DISCUSSION

The following conclusions may be drawn from the detailed tables:

For $n = 3$, $v_4$ can be considered to be almost nnue of $V(e_R)$. It has got uniformly lower probability of being negative than all the other estimators considered. Then come $v_6, v_7, v_1, v_9, v_5, v_3, v_2$ in the order of decreasing desirability in terms of taking negative values more frequently (as measured by the number of populations for which they are non-negative always and the lower and upper limits of values of probabilities in case these are non-zero). The estimator $v_4$ is again, in general, the most efficient of all the estimators considered. This suggests that $v_4$ is the most preferable one, both from the point of view of non-negativity and efficiency.

For the modified estimators, relative bias of $v_4^*$ is almost always zero, $v_6^*$ takes the next position. Again $v_4^*$ is, in general, the most efficient of all the biased estimators. This suggests that $v_4^*$ is the most preferable of all the modified estimators.

For $n = 4$, $v_4$ is again seen to be taking non-negative values more frequently than the other estimators, $v_2$ is seen to be taking negative values most frequently. Except for $v_1$, $v_5$ and $v_9$, it is found to be almost always more efficient than the other estimators. The same trend is observed in respect of the modified estimators also.

For $n = 5$, $v_5$ is seen to be taking non-negative values more frequently than the others. In cases it takes negative values, the probability of taking negative values is seen to be uniformly lower (barring one case) than the others. The next desirable estimators are $v_1$ and $v_9$. Again $v_2$ (and also $v_8$) is seen to be the least preferable one in terms of non-negativity. Also, $v_5$ is most efficient of all the other estimators. Thus $v_5$ is the most desirable one both from the point of non-negativity and stability. The same trend is observed from the values of the bias ratios. For coefficient of variation of $x$ less that 15%, all the estimators are almost always non-negative for all values of $n$. There is seen to be a considerable reduction in the value of $V_i^*$ over $V_i$ throughout.

The above analyais suggests that :

(i) For $n = 3$, 4, $v_4(v_4^*)$ is the most preferable among $\{v_i(v_i^*), i = 1, \ldots, 9\}$.

(ii) For $n = 5$, $v_5(v_5^*)$ is the most preferable among $\{v_i(v_i^*), i = 1, \ldots, 9\}$.

(iii) The estimator $v_2(v_2^*)$ is the most undesirable one.

It is suggested that for large values of $n(\geq 5), v_5(v_5^*)$ should be used, while for small values of $n(\leq 4), v_4(v_4^*)$ should be used, specially if the $cv(x)$ is low, say, less than .20. However, if $cv(x)$ is moderate to large, then $v_5$ may be used for any size of the sample.

## ACKNOWLEDGEMENTS

Computational help received from Thomas Hanson is acknowledged. Partial support from NSERC Grant A 3111 is gratefully acknowledged. We are thankful to the refrees for useful comments to improve the quality of the paper.

## REFERENCES

(1) Chaudhuri, A. (1976). A non-negativity criterion for certain variance estimators. *Metrika*, **23**, 201-205.

(2) Cochran, W.G. (1977). *Sampling Techniques.* 3rd edn. Wiley, New York.

(3) Konijan, H.S. (1973). *Statistical Theory of Sampling Survey Design and Analysis.* North Holland, Amsterdam.

(4) Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimation. *Bull. Inter. Statist. Inst.*, **33**, 133-140.

(5) Midzuno, H. (1950). An outline of the theory of sampling systems. *Ann. Inst. Statist. Math.*, 1, 149-156.

(6) Midzuno H. (1952). On the sampling system with probability proportional to sum of sizes. *Annals of the Institute of Statistical Mathematics* , **3**, 99-107.

(7) Mukhopadhyay, P. and Vijayan, K. (1990). On non-negative unbiased estimator of quadratic forms in finite population sampling. *Technical Report No. 41*, University of Western Australia, Perth.

(8) Murthy, M.N. (1967). *Sampling Theory and Methods.* Statistical Publishing Society, Calcutta.

(9) Padmawar, V.R. (1982). Optimal strategies under superpopulation models. Unpublished Ph.D. thesis submitted to the Ind. Stat. Instt., Calcutta.

(10) Raj, D. (1972). *The Design of Sample Surveys.* McGraw Hill Book Company, New York.

(11) Rao, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *J. Amer. Statist. Assoc.*, **58**, 202-215.

(12) Rao, J.N.K. (1979). On designing mean square errors and their non-negative unbiased estimators. *J. Indian Soc. Agri. Statist.*, **17**, 125-136.

(13) Rao, J.N.K. and Vijayan, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *J. Amer. Statist. Assoc.*, **72**, 579-584.

(14) Rao, T.J. (1972). On the variance of the ratio estimator. *Metrika*, **18**, 209-215.

(15) Rao, T.J. (1977). Estimating the variance of the ratio estimator for the Midzuno-Sen sampling scheme. *Metrika*, **24**, 203-208.

(16) Sen, A.R. (1952). Present status of probability sampling and its use in

the estimation of a characteristic. (Abstract). *Econometrica*, **20**, 103.

(17) Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. 2nd. edn. Asian Publishing House, Bombay.

(18) Vijayan, K. (1975). On estimating the variance in unequal probability sampling. *J. Amer. Statist. Assoc.*, **70**, 713-716.

(19) Yamani, T. (1967). *Elementary Sampling Theory*. Prentice Hall, New Jersey.

Table 1
LIST OF POPULATIONS

| Population | Source | $y$ | $x$ | $N$ | $cv(x)$ | $cv(y)$ | $\rho$ |
|---|---|---|---|---|---|---|---|
| 1 | Murthy (1967), p.228 | output | number of workers | 8 | 0.056 | 0.308 | 0.822 |
| 2 | Konijn (1973), p.49 | food expenditure | total expenditure | 16 | 0.078 | 0.111 | 0.954 |
| 3 | Murthy (1967), p.178 (village 1-10) | area under paddy | geographical area | 10 | 0.085 | 0.344 | 0.254 |
| 4 | Konijn (1973), p. 389 | measurement obtained in re-interview | measurement obtained in first interview | 10 | 0.160 | 0.151 | 0.998 |
| 5 | Sukhatme & Sukhatme (1970), p. 166 | number of banana bunches | number of banana pits | 20 | 0.175 | 0.240 | 0.774 |
| 6 | Yamane (1967) p. 334 | number of vacancies | number of apartments | 10 | 0.353 | 0.344 | 0.983 |
| 7 | Murthy (1967), p. 132 (block no. 7) | timber volume | strip length | 13 | 0.368 | 0.351 | 0.945 |
| 8 | Sukhatme & Sukhateme (1970), p. 51 | area under rice | total cultivated area | 10 | 0.391 | 0.397 | 0.874 |
| 9 | Raj (1972), p. 70 | number of cattle | number of farms | 15 | 0.402 | 0.423 | 0.894 |
| 10 | Rao (1973) p. 207 | corn acreage in 1960 | corn acreage in 1958 | 14 | 0.472 | 0.379 | 0.928 |

Table 2. Probability of taking negative values and relative efficiency of the estimators $v_1, v_4, v_5, v_9$ for samples of sizes $n = 4$ and 5 for 10 natural populations.

$n = 4$

| Pop. sl. no. | $p_1$ | $p_4$ | $p_5$ | $p_9$ | relative efficiency of $v_i$ | | |
|---|---|---|---|---|---|---|---|
| | | | | | $e_1$ | $e_5$ | $e_9$ |
| 1 | .000 | .000 | .000 | .000 | 1.006 | 1.010 | 0.999 |
| 2 | .000 | .000 | .000 | .000 | 0.987 | 0.993 | 0.978 |
| 3 | .000 | .000 | .000 | .000 | 1.009 | 1.014 | 1.001 |
| 4 | .000 | .000 | .000 | .000 | 1.062 | 1.064 | 1.034 |
| 5 | .000 | .000 | .000 | .000 | 0.934 | 0.975 | 0.928 |
| 6 | .000 | .000 | .000 | .000 | 1.369 | 1.266 | 1.473 |
| 7 | .000 | .029 | .000 | .000 | 1.639 | 1.664 | 1.539 |
| 8 | .000 | .000 | .000 | .000 | 1.093 | 1.145 | 1.014 |
| 9 | .000 | .005 | .000 | .000 | 0.972 | 1.055 | 0.860 |
| 10 | .005 | .000 | .005 | .014 | 0.849 | 0.933 | 0.739 |
| | | | $n = 5$ | | | | |
| 1 | .000 | .000 | .000 | .000 | 1.173 | 1.173 | 1.170 |
| 2 | .000 | .000 | .000 | .000 | 1.005 | 1.016 | 0.992 |
| 3 | .000 | .000 | .000 | .000 | 1.081 | 1.086 | 1.072 |
| 4 | .000 | .000 | .000 | .000 | 1.240 | 1.244 | 1.226 |
| 5 | .000 | .000 | .000 | .000 | 1.031 | 1.053 | 1.001 |
| 6 | .000 | .000 | .000 | .000 | 1.847 | 1.678 | 2.031 |
| 7 | .000 | .167 | .000 | .000 | 3.536 | 3.557 | 3.333 |
| 8 | .000 | .040 | .000 | .000 | 1.783 | 1.865 | 1.658 |
| 9 | .000 | .044 | .000 | .000 | 1.676 | 1.844 | 1.456 |
| 10 | .000 | .020 | .000 | .000 | 1.208 | 1.359 | 1.024 |