Austral. J. Statist. **37**(2), 1995, 169–178

ON NON-NEGATIVE UNBIASED ESTIMATION OF QUADRATIC FORMS IN FINITE POPULATION SAMPLING

K. VIJAYAN[1], P. MUKHOPADHYAY[1,3] AND S. BHATTACHARYYA[2]

The University of Western Australia and Indian Statistical Institute

Summary

The paper investigates non-negative quadratic unbiased (NnQU) estimators of positive semi-definite quadratic forms, for use during the survey sampling of finite population values. It examines several different NnQU estimators of the variance of estimators of population total, under various sampling designs. It identifies an optimal quadratic unbiased estimator of the variance of the Horvitz–Thompson estimator of population total.

Key words: Non-negative quadratic unbiased estimator; superpopulation model; biased non-negative variance estimator.

1. Introduction

Consider a finite population of N identifiable units labelled $1,\ldots,N$. Associated with unit k is a real quantity Y_k, a realisation of a variable $\mathcal{Y}$ of interest. We want to estimate a quadratic function of $\mathbf{Y}$,

$$F(\mathbf{Y}) = \sum_i \sum_j b_{ij} Y_i Y_j = \mathbf{Y}' B \mathbf{Y}, \tag{1.1}$$

where $B = (b_{ij})$, is a symmetric $N \times N$ matrix of known elements b_{ij}, $\mathbf{Y} = (Y_1,\ldots,Y_N)'$ and $\sum_k$ denotes summation over $k = 1,\ldots,N$. To do this, we select a sample s by employing a sampling design with $\Pi_k = \sum_{s \ni k} p(s)$, $\Pi_{kk'} = \sum_{s \ni k,k'} p(s)$ as the first and second order inclusion-probabilities respectively. Here $p(s)$ is the probability of selecting sample s. Most of the quadratic forms of interest for estimation are variances of estimators of a function of $\mathbf{Y}$ or some measures of variability between the Y_i values in the population. Hence, we can assume that F is non-negative definite (NnD). Moreover, there would be some ideal populations $\mathbf{Y} = \mathbf{W}$ for which F becomes zero. For example, if F is the variance of the Horvitz–Thompson estimator (HTE),

$$e_{\mathrm{HT}} \equiv \sum_{k \in s} \frac{Y_k}{\Pi_k}$$

Received February 1991; revised December 1994; accepted January 1995.

[1]Dept of Mathematics, The University of Western Australia, Nedlands, WA 6009, Australia.
[2]Computer Science Unit, Indian Statistical Institute, 203 B T Road, Calcutta 700 035, India.
[3]On leave from Moi University, Kenya, and Indian Statistical Institute, Calcutta.

Acknowledgements. P. Mukhopadhyay thanks The University of Western Australia for a research grant for this work. The authors thank the referee for valuable suggestions which improved the paper's presentation.

of the population total $T = \sum_{i=1}^{N} Y_i$, and if samples are of equal size n, then $Y_k = \Pi_k$ makes $e_{\mathrm{HT}} = n$ for all samples. So, F is identically zero when $\mathbf{Y} = \mathbf{\Pi} = (\Pi_1, \ldots, \Pi_N)'$. Similarly, if $F = \left(\sum (Y_k - \bar{Y})^2\right)/N$, the variance of the population, $F = 0$ when all Y_is are equal, i.e. $\mathbf{Y} = a\mathbf{1}$ for some a.

In this article we consider non-negative unbiased estimation (NnUE) of $F(\mathbf{Y})$ and subsequently consider NnUE of the variance $V(\hat{T})$ for different sampling strategies with fixed sample size n. The present work generalises results of Vijayan (1975), Rao & Vijayan (1977) and Rao (1979).

2. Form of Non-negative Unbiased Estimators

Lemma 1. *Let $Q = \mathbf{X}'\mathbf{A}\mathbf{X}$ be a NnD quadratic form in $\mathbf{X} = (x_1, \ldots, x_m)'$ with $Q = 0$ for $\mathbf{X} = \mathbf{V} = (v_1, \ldots, v_m)'$, $\mathbf{A} = (a_{ij})$. Then $\mathbf{AV} = \mathbf{0}$, i.e. $\sum_j a_{ij} v_j = 0$ for all $i = 1, \ldots, m$.*

Proof. Since $\mathbf{A}$ is NnD, there exists a matrix $\mathbf{H}$ such that $\mathbf{A} = \mathbf{HH}'$. Thus

$$\mathbf{V}'\mathbf{A}\mathbf{V} = (\mathbf{H}'\mathbf{V})'(\mathbf{H}'\mathbf{V}) = 0 \implies \mathbf{H}'\mathbf{V} = \mathbf{0} \implies \mathbf{HH}'\mathbf{V} = \mathbf{0} \implies \mathbf{AV} = \mathbf{0}.$$

Theorem 1. *If $F(\mathbf{Y}) = \sum_i \sum_j b_{ij} Y_i Y_j = 0$ for $\mathbf{Y} = \mathbf{W} = (w_1, \ldots, w_N)'$, for known w_is, then*

$$F(\mathbf{Y}) = -\sum_{i<j}\sum b_{ij} w_i w_j (z_i - z_j)^2, \tag{2.1}$$

where $z_i = Y_i/w_i$. Further, a non-negative quadratic unbiased (NnQU) estimator of F is necessarily of the form

$$f_s(\mathbf{Y}) = -\sum_{i,j\in s: i<j}\sum c_{ij}(s) w_i w_j (z_i - z_j)^2, \tag{2.2}$$

where the summation is over all distinct pairs of units $\{(i,j),\ i<j\}$ contained in the sample s and

$$\mathrm{E}\big(c_{ij}(s)\big) = b_{ij} \qquad (i<j). \tag{2.3}$$

In (2.3) and the rest of the paper, E denotes expectation with respect to the sampling design.

Proof. We can write F as $F = \sum_i \sum_j b_{ij} w_i w_j z_i z_j$. Since

$$F(\mathbf{W}) = \sum_i \sum_j b_{ij} w_i w_j = 0,$$

by Lemma 1,

$$b_{ii} w_i^2 = -\sum_{j: j\neq i} b_{ij} w_i w_j \qquad (i = 1, \ldots, N). \tag{2.4}$$

Hence from (2.4),

$$F=\sum_i\left[-\sum_{j:j\neq i} z_i^2 b_{ij}w_iw_j+\sum_{j:j\neq i} z_iz_jb_{ij}w_iw_j\right]=-\sum_{i<j}\sum b_{ij}w_iw_j(z_i-z_j)^2.$$

If $f_s=\sum_{i,j\in s:i<j}\sum c_{ij}(s)y_iy_j$ is a NnQU estimator of F then

$$f_s(\mathbf{W})=0,$$

since $F(\mathbf{W})=0$ and f_s is non-negative and unbiased.

Hence, as before, $f_s(\mathbf{Y})$ reduces to the form (2.2). The condition (2.3) is necessary for unbiasedness of $f(\mathbf{Y})$.

Remark 1. Some natural choices for $c_{ij}(s)$ are

$$c_{ij}^{(1)}(s)=\frac{b_{ij}}{M_2p(s)},\qquad c_{ij}^{(2)}(s)=\frac{b_{ij}}{\pi_{ij}},\qquad c_{ij}^{(3)}(s)=\frac{b_{ij}p(s\mid i,j)}{p(s)},\tag{2.5}$$

where $M_i=\binom{N-i}{n-i}$ $(i=0,1,\ldots)$, and $p(s\mid i)\big(p(s\mid i,j)\big)$ is the conditional probability of selecting s, given that i (i and j) was (were) selected at first draw (first two draws) according to some unit-by-unit drawing sampling scheme. Recall that for every sampling design there always exists a draw-by-draw mechanism to realise the design (Hanurav, 1962).

Note 1. Following Rao & Vijayan (1977), we can show for $n=2$ that any nonnegative unbiased estimator of F is necessarily of the form (2.2). For $n>2$, we can extend the theorem to the class of all polynomial estimators of F, following Vijayan (1975).

Note 2. Theorem 1 extends the results of Vijayan (1975), Rao & Vijayan (1977) and Rao (1979) on non-negative estimation of the mean square error of $\hat{T}$ where T is the total $\sum Y_i$ written as MSE($\hat{T}$), to non-negative definite quadratic functions in survey sampling. We recall their results as follows.

Theorem 2. *Let $\hat{T}=\sum_i b_{si}Y_i$ ($b_{si}=0$ for $i\notin s$), be a linear estimator of T. If MSE($\hat{T}$) $=0$ when $Y_i=cw_i$ $(i=1,\ldots,N)$ and the w_is are some known constants and c is an arbitrary constant, then*

$$\mathrm{MSE}(\hat{T})=-\sum_{i<j}\sum w_iw_j(z_i-z_j)^2d_{ij}\,,\tag{2.6}$$

where $z_i=Y_i/w_i$, $d_{ij}=\mathrm{E}[(b_{si}-1)(b_{sj}-1)]$. Further, a non-negative quadratic unbiased estimator of MSE($\hat{T}$) is necessarily of the form

$$m(\hat{T})=-\sum_{(i,j)\in s:i<j}\sum w_iw_j(z_i-z_j)^2e_{ij}(s),\tag{2.7}$$

where

$$\mathrm{E}\big(e_{ij}(s)\big) = d_{ij} \qquad (i < j). \tag{2.8}$$

3. Different Forms of NnQU Estimators of $V(\hat{T})$

An investigation into different forms of NnQU estimators of $V(\hat{T})$, where $\hat{T}$ is defined in Theorem 2 above, enables us to choose the estimator which is most preferable in some sense, say in the sense of having maximum stability (least sampling variance) or the largest probability of being non-negative or both.

When $\hat{T} = \sum b_{si}Y_i$ is unbiased, d_{ij} from Theorem 2 equals $\mathrm{E}(b_{si}b_{sj}) - 1 = h_{ij} - 1$, say. Then

$$V(\hat{T}) = \sum_{i<j}\sum g_{ij}(1 - h_{ij}),$$

where $g_{ij} = w_i w_j (z_i - z_j)^2$. This leads to different forms of NnQU estimators of V as

$$v_{k\ell} \equiv v_{k\ell}(s) = \sum_{(i,j)\in s: i<j}\sum g_{ij}\alpha^{(k)} - \sum_{(i,j)\in s: i<j}\sum g_{ij}h_{ij}^{(\ell)} \qquad (k, \ell = 0, 1, 2, 3), \tag{3.1}$$

where $\alpha^{(k)} \equiv \alpha_{ij}^{(k)}(s)$ and $h_{ij}^{(k)} \equiv h_{ij}^{(k)}(s)$ $(k = 1, 2, 3)$ are given at (3.3) below and

$$\alpha^{(0)} = \frac{b_{si}b_{sj}}{\mathrm{E}(b_{si}b_{sj})} \quad \text{and} \quad h_{ij}^{(0)} = b_{si}b_{sj}. \tag{3.2}$$

We obtain the quantities $\alpha^{(k)}$ and $h_{ij}^{(k)}$ $(k = 1, 2, 3)$ by substituting respectively 1 and h_{ij} for b_{ij} in (2.5), yielding

$$\alpha^{(1)} = \frac{1}{M_2 p(s)}, \qquad \alpha^{(2)} = \frac{1}{\pi_{ij}}, \qquad \alpha^{(3)} = \frac{p(s \mid i, j)}{p(s)}, \tag{3.3a}$$

$$h_{ij}^{(1)} = \frac{h_{ij}}{M_2 p(s)}, \qquad h_{ij}^{(2)} = \frac{h_{ij}}{\pi_{ij}}, \qquad h_{ij}^{(3)} = \frac{h_{ij} p(s \mid i, j)}{p(s)}. \tag{3.3b}$$

Remark 2. In practice, many of the estimators $v_{k\ell}$ would coincide. Theorem 2 gives only necessary conditions for NnQU estimators of $\mathrm{MSE}(\hat{T})$. In fact, many of the estimators $v_{k\ell}$ may not be non-negative for all values of $\mathbf{Y}$.

Remark 3. Writing $d_{ij} = (h_{ij} - a) + (a - 1) \equiv m_{ij} + b$, say, i.e. $b = a - 1$ for a some real constant, leads to different other forms of NnQU estimators of $V(\hat{T})$:

$$v_{k\ell}(s; a) = b \sum_{(i,j)\in s: i<j}\sum g_{ij}\alpha^{(k)} - \sum_{(i,j)\in s: i<j}\sum g_{ij}\hat{m}_{ij}^{(\ell)} \qquad (k, \ell = 0, 1, 2, 3).$$

Optimum choice of a for a given $v_{k\ell}$ may depend on its sampling variance.

From now on, we denote $\sum\sum_{i<j}$ and $\sum\sum_{(i,j)\in s: i<j}$ by $\sum'$ and $\sum'_s$, respectively.

Example 1 (*Probability proportional to size with replacement (ppswr) sampling*). Let t_i be the number of samples $s \ni i$, and define

$$\hat{T} = \frac{1}{n}\sum \frac{t_i y_i}{p_i}, \quad b_{si} = \frac{t_i}{np_i}, \quad z_i = \frac{Y_i}{p_i}, \quad \text{and} \quad V(\hat{T}) = \frac{1}{n}\sum{}'(z_i - z_j)^2 p_i p_j.$$

We have, for example,

$$v_{00} = \frac{1}{n(n-1)}\sum{}' t_i(z_i - \bar{z}_s)^2,$$

$$v_{01} = \sum{}'_s (z_i - z_j)^2 p_i p_j \left\{\frac{t_i t_j}{n(n-1)p_i p_j} - \frac{n-1}{nM_2 p(s)}\right\},$$

$$v_{02} = \sum{}'_s (z_i - z_j)^2 p_i p_j \left\{\frac{t_i t_j}{n(n-1)p_i p_j} - \frac{n-1}{n\pi_{ij}}\right\},$$

$$v_{12} = \sum{}'_s (z_i - z_j)^2 p_i p_j \left\{\frac{1}{M_2 p(s)} - \frac{n-1}{n\pi_{ij}}\right\},$$

$$v_{22} = \sum{}'_s (z_i - z_j)^2 \frac{p_i p_j}{\pi_{ij}},$$

$$v_{32} = \sum{}'_s (z_i - z_j)^2 p_i p_j \left\{\frac{p(s \mid i,j)}{p(s)} - \frac{n-1}{n\pi_{ij}}\right\},$$

where $\pi_{ij} = 1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j)^n$. Rao (1979) considered the estimators v_{00} and v_{22}.

Example 2 (*Horvitz–Thompson estimation*).

$$e_{\mathrm{HT}} = \sum_s \frac{y_i}{\pi_i}, \qquad b_{si} = \frac{1}{\pi_i} \quad (i \in s), \qquad d_{ij} = \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}.$$

Some forms of $v(e_{\mathrm{HT}})$ are

$$v_{00} = v_{02} = v_{20} = v_{22} = \sum{}'_s \gamma_{ij}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right), \qquad \left(\gamma_{ij} \equiv \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\right),$$

$$v_{01} = \sum{}'_s \gamma_{ij}\left\{\frac{1}{\pi_{ij}} - \frac{\pi_{ij}}{\pi_i\pi_j M_2 p(s)}\right\},$$

$$v_{03} = \sum{}'_s \gamma_{ij}\left\{\frac{1}{\pi_{ij}} - \frac{\pi_{ij}}{\pi_i\pi_j}\frac{p(s \mid i,j)}{p(s)}\right\},$$

$$v_{10} = v_{23} = \sum{}'_s \gamma_{ij}\pi_i\pi_j\left\{\frac{1}{M_2 p(s)} - \frac{1}{\pi_i\pi_j}\right\},$$

$$v_{30} = v_{32} = \sum{}'_s \gamma_{ij}\pi_i\pi_j\left\{\frac{p(s \mid i,j)}{p(s)} - \frac{1}{\pi_i\pi_j}\right\}.$$

Sen (1953) and Yates & Grundy (1953) originated the well-known estimator v_{00}. Rao (1979) considered the estimators v_{10} and v_{30}.

Example 3 (*Probability proportional to size without replacement sampling* and Murthy's estimator $\hat{T}_M = \sum_s [y_i p(s \mid i)/p(s)]$). When $y_i = cp_i$ $(i = 1, \ldots, N)$, $V(\hat{T}_M) = 0$ and $b_{ij} = \sum_{s \ni i,j} [p(s \mid i)\, p(s \mid j)/p(s)] - 1$; otherwise,

$$V(\hat{T}_M) = \sum{}' c_{ij} p_i p_j \left(1 - \sum_{s \ni i,j} \frac{p(s \mid i)\, p(s \mid j)}{p(s)}\right) \qquad \left(c_{ij} \equiv \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2\right).$$

Some of the estimators in the latter case are

$$v_{00} = \sum{}'_s c_{ij} \frac{p_i p_j}{p(s)} \left\{ \frac{p(s \mid i) p(s \mid j)}{\sum_{s' \ni i,j} p(s' \mid i) p(s' \mid j)} - \frac{p(s \mid i) p(s \mid j)}{p(s)} \right\},$$

$$v_{11} = \sum{}'_s c_{ij} \frac{p_i p_j}{M_2 p(s)} \left\{ 1 - \sum_{s' \ni i,j} \frac{p(s' \mid i) p(s' \mid j)}{p(s')} \right\},$$

$$v_{21} = \sum{}'_s c_{ij} p_i p_j \left\{ \frac{1}{\pi_{ij}} - \frac{1}{M_2 p(s)} \sum_{s' \ni i,j} \frac{p(s' \mid i) p(s' \mid j)}{p(s')} \right\},$$

$$v_{30} = \sum{}'_s c_{ij} p_i p_j \left\{ \frac{p(s \mid i,j)}{p(s)} - \frac{p(s \mid i) p(s \mid j)}{p(s)} \right\},$$

$$v_{33} = \sum{}'_s c_{ij} p_i p_j \frac{p(s \mid i,j)}{p(s)} \left\{ 1 - \sum_{s' \ni i,j} \frac{p(s' \mid i) p(s' \mid j)}{p(s')} \right\}.$$

Murthy (1957) proposed v_{30}; Pathak & Sukla (1966) showed its non-negativity.

Example 4 (*Midzuno strategy*). Here samples are taken with probability proportional to the sum of the sizes of the sample units. $\hat{T} = \hat{T}_R = \left(\sum_s y_i\right) / \left(\sum_s p_i\right) = \sum_s y_i / [M_1 p(s)] = \sum_s \left(y_i p(s \mid i)/p(s)\right)$ since for the Midzuno scheme $p(s \mid i)/p(s) = 1/[M_i p(s)]$. Also $p(s \mid i,j) = (N-1)/[M_1(n-1)]$. Hence here,

$$V(\hat{T}_R) = \sum{}' c_{ij} p_i p_j \left\{ 1 - \frac{1}{M_1} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right\},$$

where $x_s = \sum_{i \in s} x_i$, $X = \sum X_i$ and c_{ij} is as in Example 3,

$$v_{10} = v_{30} = \sum{}'_s c_{ij} x_i x_j \left[\frac{1}{M_2 p(s)} - \frac{p(s \mid i) p(s \mid j)}{p(s)^2} \right]$$

$$= \sum{}'_s c_{ij} x_i x_j \left[\frac{X}{x_s} \left(\frac{N-1}{n-1} - \frac{X}{x_s} \right) \right],$$

$$v_{11} = \sum{}'_s c_{ij} \frac{x_i x_j (N-1) X}{(n-1) x_s} \left\{ 1 - \frac{1}{M_1} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right\},$$

$$v_{20} = \sum\nolimits_s' c_{ij} x_i x_j \left(\frac{1}{\pi_{ij}} - \frac{X^2}{x_s^2} \right),$$

$$v_{22} = \sum\nolimits_s' c_{ij} \frac{x_i x_j}{\pi_{ij}} \left\{ 1 - \frac{1}{M_1} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right\},$$

$$v_{23} = \sum\nolimits_s' c_{ij} x_i x_j \left\{ \frac{1}{\pi_{ij}} - \frac{X}{M_2 x_s} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right\},$$

$$v_{13} = \sum\nolimits_s' c_{ij} x_i x_j \left[\frac{X}{x_s} \left\{ \frac{N-1}{n-1} - \frac{1}{M_2} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right\} \right],$$

$$v_{12} = v_{32} = \sum\nolimits_s' c_{ij} x_i x_j \left[\frac{(N-1)X}{(n-1)x_s} - \frac{1}{\pi_{ij} M_1} \sum_{s' \ni i,j} \frac{X}{x_{s'}} \right].$$

Rao & Vijayan (1977) considered the estimators v_{10} ($= v_{30}$) and v_{22} (their estimators (2.13) and (2.11) respectively). They studied their stabilities and the probabilities of getting a negative value empirically. An investigation into the properties of some of the other estimators is in progress.

4. Empirical Study

We have investigated empirically the performances of v_{01}, v_{03} and v_{10} on 21 natural populations for Horvitz–Thompson estimation. Most of the sample survey situations are covered by these populations. Murthy (1967) described the first eight of the populations and Rao & Vijayan (1977) described the rest. For simplicity we denote the estimators by v_1, v_2 and v_3 respectively. We consider cases when the sample size $n = 3$, 4 or 5.

To save computer time, for the cases $n = 4, 5$ we drew samples from modified populations, where the populations remained unchanged if $N \leq 10$ but were restricted to the first ten units if $N > 10$. We used Sampford's (1967) procedure to draw the samples.

Tables 1, 2 and 3 give estimates of the probabilities p_i of v_i taking negative values (given by the relative frequency of number of samples yielding negative variance estimates) for different populations for $n = 3$, 4, 5 ($i = 1,2,3$). C_x denotes the coefficient of variation of x (the auxiliary variable).

The tables give the relative efficiencies of the Yates–Grundy estimator v_0 ($= v_{00}$) over v_i, denoted by E_0/E_i ($i = 1,2,3$), where $E_i = \text{var}(v_i)$ ($i = 0,\ldots,3$).

The tables also show the performances of biased non-negative variance estimators v_1^*, v_2^*, v_3^*. The estimator v_i^* is obtained by modifying v_i as in Rao & Vijayan (1977), namely

$$v_i^* = \begin{cases} v_i & \text{when } v_i \geq 0, \\ g_s X^2 & \text{when } v_i < 0. \end{cases}$$

TABLE 1

Probabilities and relative efficiencies of v_i (v_i^*), $i = 1, 2, 3$, for 21 real populations $n = 3$

Popl.				Probabilities			Relative efficiency			Relative bias			Rel. efficiency	
no.	N	C_x	ρ	p_1	p_2	p_3	E_0/E_1	E_0/E_2	E_0/E_3	b_1^*	b_2^*	b_3^*	E_1^*/E_2^*	E_1^*/E_3^*
8	8	.056	.82	0.0	0.0	0.0	1.19	1.03	1.14	0.0	0.0	0.0	.867	.954
21	16	.078	.95	0.0	0.0	0.0	1.04	1.00	1.18	0.0	0.0	0.0	.962	1.140
6	10	.085	.25	0.0	0.0	0.0	1.12	1.00	1.27	0.0	0.0	0.0	.893	1.137
20	10	.202	.76	.067	0.0	0.0	1.31	1.00	1.67	.073	0.0	0.0	.724	1.205
10	10	.248	.84	.042	0.0	.100	1.43	1.00	2.89	.049	0.0	.177	.676	1.869
17	16	.351	−.35	.155	0.0	.107	2.03	1.03	3.24	.097	0.0	.143	.528	1.635
5	13	.368	.94	.217	0.0	.077	3.18	0.96	7.75	.174	0.0	.036	.821	6.536
9	10	.392	.87	.483	0.0	.175	5.98	1.09	11.04	.416	0.0	.373	.284	2.762
15	15	.420	.77	.356	0.0	.101	3.38	1.00	6.68	.223	0.0	.289	.358	2.129
14	10	.420	.22	.467	0.0	.142	7.68	1.29	12.22	.345	0.0	.392	.266	2.303
1	8	.449	.69	.393	0.0	.143	10.49	0.91	21.97	.440	0.0	.316	.415	9.491
2	8	.449	.43	.411	0.0	.125	9.37	0.94	24.82	.472	0.0	.377	.214	5.327
19	13	.472	.52	.423	0.0	.143	5.84	1.06	10.68	.273	0.0	.252	.350	3.529
18	16	.474	.90	.348	0.0	.161	2.62	0.94	6.84	.196	0.0	.209	.485	3.350
11	12	.503	.80	.414	0.0	.150	3.32	0.86	10.17	.364	0.0	.263	.539	5.895
13	8	.562	.44	.500	.179	.179	3.61	0.78	13.33	.373	.055	.261	.617	10.35
12	9	.573	.90	.560	.012	.135	1.81	0.98	7.27	.154	.002	.140	.566	4.051
3	8	.634	.92	.589	.304	.143	18.72	1.38	40.70	.424	.086	.257	.319	9.167
4	8	.634	.89	.536	.268	.232	10.66	1.02	34.39	.417	.158	.204	.520	17.541
16	20	.723	.98	.508	0.0	.128	14.81	0.94	36.92	.279	0.0	.110	.518	20.075
7	12	.723	.93	.505	0.0	.191	7.38	0.84	27.36	.316	0.0	.131	.604	19.342

Here g_s is the least squares estimator of $\mathrm{var}(\hat{\beta}_s)$ under the model in which

$$Y_i = \beta x_i + e_i, \qquad \mathrm{E}\Big(\frac{e_i}{x_i}\Big) = 0, \quad \mathrm{E}\Big(\frac{e_i^2}{x_i}\Big) = \sigma^2 x_i^2, \quad \mathrm{E}\Big(\frac{e_i e_j}{x_i x_j}\Big) = 0 \qquad (i \neq j).$$

Thus

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

where $\hat{\beta}_s$ denotes the least squares estimator of β. This model is reasonable in situations where the Horvitz–Thompson estimator is appropriate.

In Tables 1–3, the relative efficiencies of v_2^* and v_3^* with respect to v_1^* are denoted by E_1^*/E_2^* and E_1^*/E_3^* and the relative biases are denoted by b_1^*, b_2^* and b_3^* where $E_i^* = 1/\mathrm{MSE}(v_i^*)$ and $b_i^* = |E(v_i^* - V(e_{\mathrm{HT}})|/\{\mathrm{MSE}(v_i^*)\}^{1/2}$ $(i = 1, 2, 3)$.

From Tables 1 and 2 we conclude as follows.

- For $n = 3$, v_2 can be considered to be almost a nnu estimator of $V(e_{HT})$ for populations with $C_x \leq 0.5$. It has uniformly higher probability of being non-negative than both v_1 and v_3. For v_1, p_1 increases as C_x increases for the populations considered. For populations where v_1 and v_3 both have non-zero probability of being negative, p_3 was uniformly smaller than p_1, limits of

TABLE 2

Probabilities and relative efficiencies of v_i (v_i^*), $i = 1, 2, 3$, for 19 real populations $n = 4$

Popl.				Probabilities			Relative efficiency			Relative bias			Rel. efficiency	
no.	N	C_x	ρ	p_1	p_2	p_3	E_0/E_1	E_0/E_2	E_0/E_3	b_1^*	b_2^*	b_3^*	E_1^*/E_2^*	E_1^*/E_3^*
8	8	.056	.82	.057	0.0	.114	2.962	1.023	4.122	.126	0.0	.271	.344	1.369
21	10	.060	.95	.024	0.0	.029	2.105	1.012	2.693	.039	0.0	.080	.487	1.285
6	10	.085	.25	.100	0.0	.105	2.560	.996	3.725	.158	0.0	.213	.378	1.952
5	10	.090	.62	.181	0.0	.114	2.868	.957	5.242	.314	0.0	.213	.378	1.952
20	10	.202	.77	.133	0.0	.376	3.659	1.006	5.425	.209	0.0	.055	.388	2.176
10	10	.248	.84	.443	0.0	.248	7.716	.975	15.518	.456	0.0	.388	.210	2.772
9	10	.392	.87	.695	0.0	.200	67.649	1.127	103.923	.670	0.0	.433	.058	4.419
18	10	.394	.84	.571	0.0	.229	40.684	.796	76.657	.606	0.0	.332	.082	6.908
17	10	.396	−.46	.543	0.0	.329	24.433	1.123	42.946	.461	0.0	.328	.131	4.405
15	10	.413	.63	.662	0.0	.176	69.547	1.369	113.897	.691	0.0	.468	.059	3.874
14	10	.423	.22	.710	.029	.200	127.647	1.847	182.755	.627	.77	.460	.046	3.626
1	8	.449	.69	.671	.300	.214	126.434	.717	201.336	.802	.160	.335	.050	13.915
2	8	.449	.43	.757	.014	.171	118.022	.666	200.334	.801	.019	.382	.033	9.062
11	10	.519	.79	.667	.081	.219	51.212	.813	106.398	.613	.016	.236	.114	14.047
19	10	.532	.44	.771	.181	.167	239.920	1.947	314.374	.688	.071	.359	.047	6.878
13	8	.562	.47	.757	.529	.186	69.013	.836	149.877	.714	.274	.187	.181	33.078
16	10	.562	.98	.767	.181	.181	83.434	.839	167.258	.782	.055	.301	.084	15.719
12	9	.573	.90	.786	.571	.167	45.042	1.247	105.951	.331	.108	.121	.186	15.776
7	10	.574	.91	.667	.095	.224	83.959	.666	174.223	.700	.026	.184	.122	31.198

variation of p_1 and p_3 being (0.042, 0.589) and (0.077, 0.232) respectively. We find that v_2 is, in general, the most efficient of the three, when its efficiency is compared with respect to v_0. This suggests that the estimator v_2 is the best of the three, from the viewpoints of both non-negativity and efficiency.

• For the modified estimators, relative bias of v_2^* is almost always zero, while for 16 out of 21 populations b_3^* is less than or equal to b_1^*. We find that v_2^* is uniformly more efficient than v_1^*. The modified estimator v_2^* seems to be the best of the three.

• The same trend is observed for $n = 4$. For $n = 5$, p_2 is uniformly smaller than p_1 and p_3 for populations with $C_x < 0.25$. For the remaining populations p_3 is smaller than p_2 which in its turn is smaller than p_1. The estimator v_2 has greater efficiency than both v_1 and v_3.

• The relative bias of v_i^* is uniformly higher than that of v_2^* and v_3^* (except for populations 6 and 21, where $b_1^* < b_3^*$). For populations with $C_x \leq (>) 0.25$, b_2^* is lower (higher) than b_3^*. The estimator v_2^* is more efficient than v_1^*, while v_3^* has poor efficiency.

The above analysis suggests that for the Horvitz–Thompson estimator,

(i) for $n = 3, 4$, v_2 (v_2^*) is the best of $\{v_i$ (v_i^*), $i = 1, 2, 3\}$;

(ii) for $n = 5$, for populations with $C_x \leq 0.25$, v_2 can be recommended. For the remaining types of populations, v_2 and v_3 are the better estimators, while

TABLE 3

Probabilities and relative efficiencies of v_i (v_i^*), $i = 1, 2, 3$, for 14 real populations $n = 5$

Popl.			Probabilities			Relative efficiency			Relative bias			Rel. efficiency	
no.	N	C_x	p_1	p_2	p_3	E_0/E_1	E_0/E_2	E_0/E_3	b_1^*	b_2^*	b_3^*	E_1^*/E_2^*	E_1^*/E_3^*
8	8	.056	.321	.125	.321	24.956	7.772	34.617	.684	.293	.657	.418	1.629
21·	10	.060	.250	.095	.198	9.629	8.058	12.603	.459	.315	.497	.948	1.474
6	10	.085	.317	.087	.294	13.241	4.848	18.481	.514	.347	.519	.481	1.704
5	10	.090	.341	.107	.333	17.016	3.344	26.059	.618	.236	.575	.265	2.184
20	10	.920	.222	.119	.611	12.772	2.332	17.942	.352	.109	.288	.310	2.961
10	10	.248	.544	.365	.321	53.396	5.217	82.690	.653	.351	.419	.255	4.226
9	10	.392	.746	.659	.187	482.618	41.864	633.873	.754	.732	.374	.197	6.754
18	10	.394	.663	.516	.242	370.848	8.377	541.501	.754	.615	.286	.136	11.864
15	10	.413	.746	.484	.218	765.768	12.739	1052.530	.790	.510	.358	.092	8.072
14	10	.423	.750	.373	.206	3220.821	40.286	4066.433	.761	.647	.288	.078	11.849
11	10	.519	.754	.603	.198	3434.806	12.970	5808.461	.753	.518	.073	.216	186.005
19	10	.532	.837	.794	.159	5894.681	183.662	7815.605	.817	.767	.202	.258	27.189
7	10	.574	.766	.520	.187	1547.592	14.329	2691.118	.810	.543	.091	.189	139.527

v_2^* is always the best of v_1^*, v_2^*, v_3^*.

Recently, Mukhopadhyay & Tracy (unpublished) extended similar empirical investigations to Midzuno's sampling strategy. Other sampling strategies may be investigated as well. An important problem to which the referee drew our attention is the search for a theoretical upper bound for the probability of an estimator v_{ik} taking negative values. This issue will be addressed elsewhere.

References

HANURAV, T.V. (1962). An existence theorem in sampling theory. *Sankhyā Ser. A* **24**, 327–330.

MURTHY, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā* **18**, 379–390.

— (1977). *Sampling Theory and Methods.* Second edn. Calcutta: Statistical Publishing Society.

PATHAK, P.K. & SUKLA, N.D. (1966). Non-negativity of a variance estimator. *Sankhyā Ser. A* **28**, 41–46.

RAO, J.N.K. & VIJAYAN, K. (1977). On estimating the variance in sampling with probability proportional to aggregate size. *J. Amer. Statist. Assoc.* **72**, 579–584.

— (1979). On deriving mean square errors and their non-negative unbiased estimators. *J. Amer. Statist. Assoc.* **17**, 125–136.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499–513.

SEN, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Indian Soc. Appl. Statist.* **5**, 119–127.

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

YATES, F. (1960). *Sampling Methods for Censuses and Surveys.* London: Charles Griffin and Co. Ltd.

— & GRUNDY, P.H. (1953). Selections without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* **15**, 253–261.