

MEDIAN ESTIMATION USING DOUBLE SAMPLING

SARJINDER SINGH^{1*}, ANWAR H. JOARDER² AND D.S. TRACY³

University of Saskatchewan, K.F.U. of Petroleum and Minerals and University of Windsor

Summary

This paper proposes a general class of estimators for estimating the median in double sampling. The position estimator, stratification estimator and regression type estimator attain the minimum variance of the general class of estimators. The optimum values of the first-phase and second-phase sample sizes are also obtained for the fixed cost and the fixed variance cases. An empirical study examines the performance of the double sampling strategies for median estimation. Finally, an extension of the methods of Chen & Qin (1993) and Kuk & Mak (1994) is considered for the double sampling strategy.

Key words: auxiliary information; cost aspects; empirical study; median estimation; two-phase sampling; variance.

1. Introduction

In survey sampling, statisticians often find that they are dealing with variables, such as income, expenditure, etc., that have highly skewed distributions. In such situations, the median is regarded as a more appropriate measure of location than the mean. Kuk & Mak (1989) suggest estimators of the median using auxiliary information in survey sampling. Francisco & Fuller (1991) also deal with the problem of estimation of the median as part of the estimation of a finite population distribution function. A brief description of the motivation of Kuk & Mak (1989) is given below.

Let Y_i and X_i , $i = 1, 2, \dots, N$, denote the values of the population units for the study variable Y and the auxiliary variable X , respectively. Further let y_i and x_i , $i = 1, 2, \dots, n$, denote the values of the units included in a sample S_n of size n drawn by simple random sampling without replacement (SRSWOR). Assuming that the median M_X of variable X is known, Kuk & Mak (1989) proposed a ratio estimator

$$\hat{M}_Y^{(r)} = \hat{M}_Y \frac{M_X}{\hat{M}_X}, \quad (1)$$

where $\hat{M}_Y$ and $\hat{M}_X$ are the sample estimators of M_Y and M_X , respectively. Suppose that $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ are the y values of sample units in ascending order. Further, let t be an

Received May 1998; revised January 2000; accepted January 2000.

* Author to whom correspondence should be addressed.

¹ Dept of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK S7N 5E6, Canada.
e-mail: sarjinder@yahoo.com

² Dept of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia.

³ Dept of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada N9B 3P4.

Acknowledgments. The authors thank an Associate Editor, a referee, the Editor and the Technical Editor for valuable comments that improved this paper. Professor D.S. Tracy passed away on 27 December 1998 at the airport, New Delhi, India. The opinions and results discussed in this paper are the authors' and not necessarily of their institute(s). This work was done while the first author was at the Australian Bureau of Statistics, Canberra, Australia.

integer such that $Y_{(t)} \leq M_Y < Y_{(t+1)}$ and let $p = t/n$ be the proportion of Y values in the sample that are less than or equal to the median value M_Y , an unknown population parameter. If $Q_Y(t)$ denotes the t -quantile of Y , and $\hat{Q}_Y(t)$ denotes the sample t -quantile of Y , then $\hat{M}_Y = \hat{Q}_Y(0.5)$. Kuk & Mak (1989) defined a matrix of proportions (p_{ij}) as

	$Y \leq M_Y$	$Y > M_Y$	Total
$X \leq M_X$	p_{11}	p_{21}	$p_{\bullet 1}$
$X > M_X$	p_{12}	p_{22}	$p_{\bullet 2}$
Total	$p_{1\bullet}$	$p_{2\bullet}$	1

and a position estimator of M_Y given by

$$\hat{M}_Y^{(p)} = \hat{Q}_Y(\hat{p}_Y), \quad (2)$$

where
$$\hat{p}_Y = \frac{1}{n} \left(\frac{n_X \hat{p}_{11}}{\hat{p}_{\bullet 1}} + \frac{(n - n_X) \hat{p}_{12}}{\hat{p}_{\bullet 2}} \right) \approx 2 \left(\frac{n_X \hat{p}_{11} + (n - n_X) \hat{p}_{12}}{n} \right),$$

with the $\hat{p}_{ij}$ being the sample analogues of the p_{ij} obtained from the population and n_X the number of units in S_n with $X \leq M_X$.

Kuk & Mak (1989) also proposed an alternative estimator which they called a stratification estimator. Let $\tilde{F}_{YA}(y)$ and $\tilde{F}_{YB}(y)$ denote the proportion of units in the sample with $X \leq M_X$ and $X > M_X$, respectively, that have Y values less than or equal to y . The stratification estimator proposed by them is given by

$$\hat{M}_Y^{(st)} = \inf\{y: \tilde{F}_Y(y) \geq 0.5\}, \quad (3)$$

where $\tilde{F}_Y(y) \approx 0.5(\tilde{F}_{YA}(y) + \tilde{F}_{YB}(y))$. We have seen that the estimators defined by (1), (2) and (3) are based on prior knowledge of the median M_X of the auxiliary character. In many cases M_X may not be known. This necessitates an analogue of the above estimators in double sampling.

In the first phase of double sampling, we select a preliminary large sample S' of n' units by SRSWOR and only the auxiliary character X is measured. Let $\hat{M}'_X$ be the estimator of M_X based on the first-phase sample. In the second phase, a sub-sample S of n units is drawn from the preliminary large sample by SRSWOR and both the study variable Y and the auxiliary variable X are measured. Let $\hat{M}_X$ and $\hat{M}_Y$ denote the estimators of M_X and M_Y , respectively, based on the sample drawn in the second phase.

2. General class of estimators

Following Srivastava (1971) and Srivastava & Jhaji (1981), we propose a general class of estimators for estimating the median as

$$\mathcal{H} = \left\{ \hat{M}_Y^{(H)}: \hat{M}_Y^{(H)} = H\left(\hat{M}_Y, \frac{\hat{M}_X}{\hat{M}'_X}\right) \right\},$$

where the function H assumes a value in a bounded closed convex subset $W \subset \mathbb{R}_2$, which contains the point $(M_Y, 1)$ and is such that $H(M_Y, 1) = M_Y$. Using a first order Taylor's series expansion around the point $(M_Y, 1)$, we get

$$\hat{M}_Y^{(H)} = H(M_Y, 1) + (\hat{M}_Y - M_Y)H_{10}(M_Y, 1) + (U - 1)H_{01}(M_Y, 1) + o(n^{-1}), \quad (4)$$

where $H_{10}(M_Y, 1)$ and $H_{01}(M_Y, 1)$ denote the first order partial derivatives of $H(\hat{M}_Y, U)$ with respect to $\hat{M}_Y$ and U respectively. Under the assumption that $H_{10}(M_Y, 1) = 1$, we have the following theorems.

Theorem 2.1. Any estimator in $\mathcal{H}$ is asymptotically unbiased and normal.

Proof. Following Kuk & Mak (1989), let p_Y and p_X denote the proportion of Y and X values in the sample for which $Y \leq M_Y$ and $X \leq M_X$, respectively; then we have

$$\hat{M}_Y - M_Y = \frac{1}{f_Y(M_Y)} \left(\frac{1}{2} - p_Y \right) + o_p(n^{-1/2}),$$

$$\hat{M}_X - M_X = \frac{1}{f_X(M_X)} \left(\frac{1}{2} - p_X \right) + o_p(n^{-1/2}),$$

and

$$\hat{M}'_X - M_X = \frac{1}{f_X(M_X)} \left(\frac{1}{2} - p_X \right) + o_p(n'^{-1/2}).$$

On using these results in (4), we have the theorem.

Theorem 2.2. The minimum variance of the proposed class of estimators $\hat{M}_Y^{(H)}$, up to terms of order $o(n^{-1})$, is given by

$$\text{var}(\hat{M}_Y^{(H)}) \geq \frac{1}{4f_Y(M_Y)^2} \left(\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{n'} \right) (4p_{11} - 1)^2 \right), \quad (5)$$

where f_Y denotes the pdf of Y .

Proof. The proof is straightforward following Srivastava & Jhajj (1981).

In these two sections, we have studied the properties of some estimators, which are in fact special cases of the general class of estimators.

3. Regression type estimator

We also consider the following regression type estimator

$$\hat{M}_Y^{(\ell)}(d) = \hat{M}_Y + d(\hat{M}'_X - \hat{M}_X),$$

where d is a suitably chosen constant such that the variance of the proposed estimator $\hat{M}_Y^{(\ell)}(d)$ is a minimum. The minimum variance of the estimator $\hat{M}_Y^{(\ell)}(d)$, up to terms of order $o(n^{-1})$, is given by

$$\text{var}(\hat{M}_Y^{(\ell)}(d)) \geq \frac{1}{4f_Y(M_Y)^2} \left(\left(\frac{1}{n} - \frac{1}{N} \right) \left(\frac{1}{n} - \frac{1}{n'} \right) (4p_{11} - 1)^2 \right). \quad (6)$$

The minimum value is achieved for

$$d = \frac{f_X(M_X)}{f_Y(M_Y)} (4p_{11} - 1).$$

The proof is straightforward.

An estimator of this optimum value of d is given by

$$\hat{d} = \frac{\hat{f}_X(\hat{M}_X)}{\hat{f}_Y(\hat{M}_Y)} (4\hat{p}_{11} - 1),$$

where $\hat{f}_Y(\hat{M}_Y)$ and $\hat{f}_X(\hat{M}_X)$ can be obtained by following Silverman (1986).

The regression type estimator, after replacing d with $\hat{d}$, becomes

$$\hat{M}_Y^{(\ell)}(\hat{d}) = \hat{M}_Y + \hat{d}(\hat{M}'_X - \hat{M}_X).$$

Lemma 3.1. *The estimators $\hat{M}_Y^{(\ell)}(d)$ and $\hat{M}_Y^{(\ell)}(\hat{d})$ have the same asymptotic variance because $E(\hat{d}) = d + O(n^{-1})$.*

Proof. Define

$$\delta_0 = \frac{\hat{f}_X(\hat{M}_X)}{f_X(M_X)} - 1, \quad \delta_1 = \frac{\hat{f}_Y(\hat{M}_Y)}{f_Y(M_Y)} - 1 \quad \text{and} \quad \delta_2 = \frac{\hat{p}_{11} - 1}{p_{11} - 1} - 1,$$

so that $E(\delta_0) = E(\delta_1) = E(\delta_2) \approx 0$. It can be seen that the asymptotic variances of the estimators $\hat{M}_Y^{(\ell)}(d)$ and $\hat{M}_Y^{(\ell)}(\hat{d})$ are the same. This proves the lemma.

The estimators of the form

$$\hat{M}_Y^{(a)} = \hat{M}_Y \frac{\hat{M}'_X}{\hat{M}_X}, \quad \hat{M}_Y^{(b)} = \hat{M}_Y \left(\frac{\hat{M}_X}{\hat{M}'_X} \right)^b \quad \text{and} \quad \hat{M}_Y^{(c)} = \hat{M}_Y \left(\frac{\hat{M}'_X}{c\hat{M}_X + (1-c)\hat{M}'_X} \right),$$

which are in fact analogous to the ratio estimator and the estimators proposed by Srivastava (1967) and Walsh (1970) in double sampling, respectively, belong to the class of estimators, $\mathcal{H}$.

Lemma 3.2. *The general class of estimators, $\mathcal{H}$ is always more efficient than the ratio estimator $\hat{M}_Y^{(a)}$.*

Proof. Since $\hat{M}_Y^{(a)} \in \mathcal{H}$, we must have $\text{var}(\hat{M}_Y^{(a)}) \geq \min_{\mathcal{H}} \text{var}(\hat{M}_Y^{(H)})$. It can be seen that

$$\text{var}(\hat{M}_Y^{(a)}) - \min_{\mathcal{H}} \text{var}(\hat{M}_Y^{(H)}) = \left(\frac{M_Y}{M_X} - \frac{4(p_{11} - 0.25)f_X(M_X)}{f_Y(M_Y)} \right)^2 > 0.$$

Hence the lemma.

4. Position estimator

Suppose n_x^* is the number of units in the second-phase sample S_n with $X \leq \hat{M}'_X$. Thus, if the p_{ij} ($i, j = 1, 2$) are known, we can predict p by

$$\hat{p}_Y^{(0)} = \frac{1}{n} \left(\frac{n_x^* p_{11}}{p_{\bullet 1}} + \frac{(n - n_x^*) p_{12}}{p_{\bullet 2}} \right) \approx \frac{2}{n} (n_x^* p_{11} + (n - n_x^*) (0.5 - p_{11})), \quad (7)$$

because $p_{\bullet j} = p_{1j} + p_{2j} \approx 0.5$, $j = 1, 2$. If we replace p_{ij} by $\hat{p}_{ij}$ from the sample in (7), we get an estimator of p_Y of the form

$$\hat{p}_Y^{(1)} = \frac{1}{n} \left(\frac{n_x^* \hat{p}_{11}}{\hat{p}_{\bullet 1}} + \frac{(n - n_x^*) \hat{p}_{12}}{\hat{p}_{\bullet 2}} \right) \approx \frac{2}{n} (n_x^* \hat{p}_{11} + (n - n_x^*) (0.5 - \hat{p}_{11})).$$

Then we define a position estimator in two-phase sampling as

$$\hat{M}_Y^{(p)} = \hat{Q}_Y(\hat{p}_Y^{(1)}).$$

We have the following theorems.

Theorem 4.1. *The estimator $\hat{M}_Y^{(p)}$ is asymptotically unbiased and normal.*

Proof. We have

$$\hat{M}_Y^{(p)} - M_y = \frac{1}{f_y(M_y)}(\hat{p}_Y^{(1)} - p_Y) + o_p(n^{-1/2}).$$

Since $\hat{p}_Y^{(1)}$ is asymptotically normal with mean p_Y , the result follows.

Theorem 4.2. *The variance of the estimator $\hat{M}_Y^{(p)}$, up to terms of order $o(n^{-1})$, is*

$$\text{var}(\hat{M}_Y^{(p)}) \geq \frac{1}{4f_Y(M_Y)^2} \left(\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{n'} \right) (4p_{11} - 1)^2 \right). \quad (8)$$

Proof. See Appendix 1.

5. Stratification estimator

Suppose $\tilde{F}_{YA}^*(y)$ and $\tilde{F}_{YB}^*(y)$ denote the proportion of units in the second-phase sample for which $X \leq \hat{M}_X'$ and $X > \hat{M}_X'$, respectively, that have Y values less than or equal to y . Then $F_Y(y)$ can be estimated by

$$\tilde{F}_Y(y) = \frac{n'_X \tilde{F}_{YA}^*(y)}{n'} + \frac{(n - n'_X) \tilde{F}_{YB}^*(y)}{n'} \approx 0.5(\tilde{F}_{YA}^*(y) + \tilde{F}_{YB}^*(y)),$$

where n'_X is the number of units in the preliminary large sample with $X \leq \hat{M}_X'$. We consider an analogue of the stratification estimator in two-phase sampling of the form

$$\hat{M}_Y^{(st)} = \inf\{y: \tilde{F}_Y(y) \geq 0.5\}.$$

The estimator $\hat{M}_Y^{(st)}$ is asymptotically unbiased and normal. Theorem 5.1 gives its variance.

Theorem 5.1. *The variance of the estimator $\hat{M}_Y^{(st)}$, up to terms of order $o(n^{-1})$, is*

$$\text{var}(\hat{M}_Y^{(st)}) \geq \frac{1}{4f_Y(M_Y)^2} \left(\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{n'} \right) (4p_{11} - 1)^2 \right). \quad (9)$$

Proof. See Appendix 2.

Relations (5), (6), (8) and (9) show that the general class of estimators, regression type estimator, position estimator and stratification estimator attain the same minimum variance to the first order of approximation. It is difficult to distinguish the best estimator from among the proposed estimators using first order asymptotic properties, so we have resorted to an empirical study for comparing the different estimators.

6. Estimators of variances and optimum sample sizes

An estimator to estimate the variance of the estimators, $\hat{M}_Y^{(g)}$ (except $\hat{M}_Y^{(a)}$) is

$$\hat{v}(\hat{M}_Y^{(g)}) = \frac{1}{4\hat{f}_Y(\hat{M}_Y)^2} \left(\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{n'} \right) (4\hat{p}_{11} - 1)^2 \right).$$

We obtain the optimum first-phase and second-phase sample sizes for the fixed cost as well as for the fixed variance cases.

Let c and c' denote the cost per unit in the second phase and first phase, respectively, then the fixed cost C_0 is given by

$$C_0 = nc + n'c'.$$

Let $p_{11} = \frac{1}{4}(1 + \varepsilon)$. The variance of $\hat{M}_Y^{(g)}$ is the minimum for the fixed cost C_0 if

$$n = \frac{C_0\sqrt{1-\varepsilon^2}}{\sqrt{c}(\sqrt{c(1-\varepsilon^2)} + \sqrt{c'\varepsilon^2})} \quad \text{and} \quad n' = \frac{C_0\sqrt{\varepsilon^2}}{\sqrt{c'}(\sqrt{c(1-\varepsilon^2)} + \sqrt{c'\varepsilon^2})}.$$

The variance var_0 can be achieved for minimum cost with

$$n = \frac{Nv_Y\sqrt{1-\varepsilon^2}(\sqrt{c(1-\varepsilon^2)} + \sqrt{c'\varepsilon^2})}{\sqrt{c}(\text{var}_0 + v_Y)} \quad \text{and} \quad n' = \frac{Nv_Y\sqrt{\varepsilon^2}(\sqrt{c(1-\varepsilon^2)} + \sqrt{c'\varepsilon^2})}{\sqrt{c'}(\text{var}_0 + v_Y)},$$

where $v_Y = 1/(4Nf_Y(M_Y)^2)$.

7. Empirical study

We have made an empirical study of two types of populations, namely natural populations and artificial populations, by following Singh, Horn & Yu (1998). The natural populations were taken from books or journals, whereas the artificial populations were generated according to actual practical situations using standard subroutines available in FORTRAN. We first explain the steps taken for simulation using natural populations.

Natural populations

Consider the Horvitz & Thompson (1952) population consisting of $N = 20$ units. The study variable y represents the number of households on the i th block and the auxiliary character x denotes the eye estimate of the number of households on the i th block. First we selected all 77 520 possible preliminary large samples of size n' units. From every given preliminary large sample of $n' = 7$ units, the last two units in the sample were dropped to obtain second-phase samples of $n = 5$ units. The mean and mean square error of the method g estimator $\hat{M}_Y^{(g)}$ are given by

$$\text{av}(\hat{M}_Y^{(g)}) = \frac{1}{k} \sum_{s=1}^k \hat{M}_{Ys}^{(g)} \quad \text{and} \quad \text{mse}(\hat{M}_Y^{(g)}) = \frac{1}{k} \sum_{s=1}^k (\hat{M}_{Ys}^{(g)} - M_Y)^2,$$

where $k = \binom{N}{n'}$. Then the relative bias and relative efficiency are given by

$$\text{RB}(\hat{M}_Y^{(g)}) = \frac{\text{av}(\hat{M}_Y^{(g)})}{\sqrt{\text{mse}(\hat{M}_Y^{(g)})}} \quad \text{and} \quad \text{RE}(\hat{M}_Y^{(g)}) = \frac{\text{mse}(\hat{M}_Y)}{\text{mse}(\hat{M}_Y^{(g)})}.$$

A similar procedure was repeated for a few more natural populations. The results obtained are presented in Table 1. The value of RE for the ratio estimator $\hat{M}_Y^{(a)}$ is only 0.59 for

TABLE 1

Relative bias (a), and relative efficiency (b), of proposed estimators with respect to usual estimator of median from natural populations for $n' = 7$, $n = 5$ and various values of N

Source	Description of populations	d	$\hat{M}_Y^{(l)}(d)$	$\hat{M}_Y^{(b)}$	$\hat{M}_Y^{(c)}$	$\hat{M}_Y^{(a)}$	$\hat{M}_Y^{(p)}$	$\hat{M}_Y^{(st)}$
(a) Horvitz & Thompson (1952)	$N = 20$	0.1	0.012	0.013	0.023			
	y: no. of households	0.3	0.012	0.013	0.025			
	x: eye estimate of no. of households on i th block	0.5	0.013	0.015	0.034	0.006	0.004	0.008
		0.7	0.015	0.014	0.026			
		0.9	0.025	0.019	0.024			
Wang & Chow (1994 p. 349)	$N = 31$	0.1	0.039	0.027	0.034			
	y: volume	0.3	0.024	0.026	0.030			
	x: diameter	0.5	0.014	0.021	0.028	0.008	0.006	0.009
		0.7	0.012	0.026	0.029			
		0.9	0.011	0.028	0.037			
Sadasivan & Aggarwal (1978)	$N = 20$	0.1	0.024	0.023	0.025			
	y: buffaloes in milk	0.3	0.026	0.025	0.026			
	x: cattle in milk	0.5	0.034	0.031	0.034	0.003	0.003	0.009
		0.7	0.029	0.027	0.028			
		0.9	0.028	0.026	0.029			
Sanyal & Sinha (1977)	$N = 16$	0.1	0.016	0.014	0.024			
	y: % of NSS estimate over Agr. census during 1971	0.3	0.018	0.012	0.020			
	x: % of NSS estimate over Agr. census during 1970	0.5	0.018	0.016	0.019	0.001	0.002	0.005
		0.7	0.014	0.013	0.026			
		0.9	0.013	0.012	0.031			
(b) Horvitz & Thompson (1952)	$N = 20$	0.1	1.14	1.13	1.11			
	y: no. of households	0.3	1.25	1.25	1.22			
	x: eye estimate of no. of households on i th block	0.5	1.34	1.34	1.31	0.98	1.06	1.07
		0.7	1.37	1.37	1.36			
		0.9	1.34	1.32	1.32			
Wang & Chow (1994 p. 349)	$N = 31$	0.1	1.04	1.04	1.03			
	y: volume	0.3	1.07	1.06	1.06			
	x: diameter	0.5	1.09	1.09	1.08	1.08	1.12	1.15
		0.7	1.11	1.10	1.10			
		0.9	1.12	1.10	1.10			
Sadasivan & Aggarwal (1978)	$N = 20$	0.1	1.04	1.05	1.05			
	y: buffaloes in milk	0.3	1.09	1.10	1.10			
	x: cattle in milk	0.5	1.13	1.15	1.15	1.55	1.26	1.37
		0.7	1.17	1.20	1.20			
		0.9	1.21	1.25	1.25			
Sanyal & Sinha (1977)	$N = 16$	0.1	1.08	1.08	1.07			
	y: % of NSS estimate over Agr. census during 1971	0.3	1.12	1.13	1.12			
	x: % of NSS estimate over Agr. census during 1970	0.5	1.13	1.13	1.14	0.59	1.04	1.03
		0.7	1.09	1.08	1.11			
		0.9	1.03	0.97	1.00			

the fourth population. The ratio estimator also remains less efficient in the case of the first population. Most of the other estimators are more efficient than the simple median estimator.

Artificial populations

To study the performance of the proposed estimators in practice, we generated five auxiliary characters having different levels of correlation with the study variable. The transfor-

TABLE 2

Relative bias (a), and relative efficiency (b), of proposed estimators with respect to usual estimator of median from artificial populations generated through beta distributions for $n' = 7, n = 5, N = 20$

Correlation (ρ) between Y and X	d	$\hat{M}_Y^{(\ell)}(d)$	$\hat{M}_Y^{(b)}$	$\hat{M}_Y^{(c)}$	$\hat{M}_Y^{(a)}$	$\hat{M}_Y^{(p)}$	$\hat{M}_Y^{(st)}$
(a)	0.938	0.1	0.002	0.005	0.003		
		0.3	0.003	0.005	0.004		
		0.5	0.002	0.006	0.003	0.005	0.006
		0.7	0.002	0.005	0.004		0.009
		0.9	0.002	0.004	0.005		
	0.850	0.1	0.003	0.008	0.007		
		0.3	0.003	0.006	0.004		
		0.5	0.002	0.007	0.005	0.006	0.004
		0.7	0.003	0.007	0.007		0.008
		0.9	0.002	0.008	0.004		
	0.619	0.1	0.003	0.007	0.008		
		0.3	0.003	0.006	0.007		
		0.5	0.002	0.008	0.002	0.005	0.005
		0.7	0.002	0.007	0.006		0.008
		0.9	0.001	0.006	0.004		
	0.387	0.1	0.009	0.006	0.005		
		0.3	0.008	0.007	0.005		
		0.5	0.007	0.008	0.004	0.004	0.005
		0.7	0.008	0.007	0.005		0.009
		0.9	0.008	0.008	0.004		
	0.236	0.1	0.012	0.006	0.005		
		0.3	0.014	0.008	0.006		
		0.5	0.013	0.007	0.008	0.005	0.006
		0.7	0.012	0.006	0.006		0.009
		0.9	0.012	0.005	0.005		
(b)	0.938	0.1	1.04	1.15	1.36		
		0.3	1.09	1.19	1.38		
		0.5	1.16	1.26	1.40	1.15	1.25
		0.7	1.26	1.36	1.33		1.37
		0.9	1.06	1.18	1.26		
	0.850	0.1	1.03	1.16	1.30		
		0.3	1.08	1.19	1.29		
		0.5	1.13	1.18	1.25	1.12	1.24
		0.7	1.23	1.06	1.30		1.31
		0.9	1.06	1.18	1.26		
	0.619	0.1	1.03	1.15	1.25		
		0.3	1.07	1.18	1.23		
		0.5	1.12	1.15	1.23	1.10	1.18
		0.7	1.22	1.05	1.23		1.26
		0.9	1.05	1.16	1.25		
	0.387	0.1	1.01	1.02	1.02		
		0.3	1.03	1.02	1.03		
		0.5	1.06	1.03	1.02	1.00	1.12
		0.7	1.04	1.04	1.03		1.10
		0.9	1.04	1.02	1.03		
	0.236	0.1	1.01	1.00	1.01		
		0.3	1.00	1.00	1.01		
		0.5	1.01	1.00	1.00	0.86	0.99
		0.7	1.00	0.98	1.00		1.00
		0.9	1.00	0.97	1.00		

mations used to generate the various variables are given by

$$X_{ki} = 20 + \sqrt{(1 - \rho^2)}X_{ki}^* + \rho \frac{\sigma_X}{\sigma_Y} Y_i^* \quad \text{and} \quad Y_i = 10 + Y_i^* \quad (k = 1, \dots, 5),$$

where X_{ki}^* and Y_i^* are the independent beta variates generated by the subroutine BETACH for $a = 1.6$, $b = 1.1$, seed1 = 13 031 963, seed2 = 19 631 303, $\sigma_Y = 5.5$, $\sigma_X = 1.5$ and $N = 20$ following Bratley, Fox & Schrage (1983) in FORTRAN 77 for different values of the correlation coefficient: $\rho_1 = 0.938$, $\rho_2 = 0.850$, $\rho_3 = 0.619$, $\rho_4 = 0.387$ and $\rho_5 = 0.236$. The rest of the procedure was repeated as for the natural populations and the results are reported in Table 2. Similar results were observed, with slight modifications in the above transformations where needed, with the Normal and Gamma populations. As the value of the correlation coefficient between x and y decreases, the magnitude of the percent relative efficiency also decreases, although this fact is not clear from the variance formulae of the estimators. In the simulation study we did not estimate the optimum value of d from sample information, therefore a relative efficiency was obtained for the expected range within the interval $(0, 1)$ for the estimators based on it. We increased the size of the artificial population to $N = 1500$ units. We selected 10 000 preliminary large samples, each of size $n' = 100$. From a given preliminary sample, we selected second-phase samples of different sizes, namely $n = 10, 20, 30, 35, 50, 70$, etc., to compute the ratio of the exact variance to approximate variances, defined as

$$VR = \frac{\text{var}(\hat{M}_Y^{(g)})}{\text{mse}(\hat{M}_Y^{(g)})},$$

for the estimators considered in the present investigation. For calculating the approximate variance, we assumed that X and Y had a beta density function given by

$$f(z) = \frac{1}{B(a, b)} z^{a-1} (1 - z)^{b-1} \quad (0 < z < 1),$$

with $a = 1.6$ and $b = 1.1$. We used an estimator of the optimum value of d , given by $\hat{d} = (4\hat{p}_{11} - 1)\hat{f}_X(\hat{M}_X)/\hat{f}_Y(\hat{M}_Y)$, from each sample while studying the estimators $\hat{M}_Y^{(g)}$. As the sample size increases beyond $n = 30$, the ratio converges to 1 as shown in Table 3. This part of the simulation study was suggested by one of the referees. Following the Associate Editor's comments, we have also extended the paper to include the more recent work of Kuk & Mak (1994) and Chen & Qin (1993) in the following sections.

8. The Kuk and Mak method

Kuk & Mak (1994) proposed another ingenious method, which in double sampling can also be extended for estimating the finite population distribution function of Y , defined as

$$F_Y(y) = \frac{1}{N} \sum_{i=1}^N \Delta(y - y_i),$$

where $\Delta(a) = 1$ when $a \geq 0$ and $\Delta(a) = 0$ otherwise. Note that F_Y puts probability $1/N$ at each y_i , $i = 1, 2, \dots, N$, assuming them to be distinct. The naïve estimator of F_Y is the sample distribution function defined as

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i \in S} \Delta(y - y_i),$$

TABLE 3
Ratio (VR) of exact variance to approximate variance for several sample sizes

ρ	Estimator	Sample size (n)					
		10	20	30	35	50	70
0.95	$\hat{M}_Y^{(\ell)}(d)$	0.86	0.89	0.90	0.93	0.97	0.98
	$\hat{M}_Y^{(a)}$	0.93	0.95	0.98	0.99	0.99	0.99
	$\hat{M}_Y^{(b)}$	0.89	0.89	0.91	0.92	0.95	0.96
	$\hat{M}_Y^{(c)}$	0.92	0.95	0.96	0.98	0.99	0.99
	$\hat{M}_Y^{(p)}$	0.94	0.96	0.98	0.99	0.99	0.99
	$\hat{M}_Y^{(st)}$	0.93	0.95	0.96	0.97	0.98	0.98
0.50	$\hat{M}_Y^{(\ell)}(d)$	0.85	0.90	0.93	0.95	0.98	0.99
	$\hat{M}_Y^{(a)}$	0.91	0.94	0.97	0.98	0.98	0.99
	$\hat{M}_Y^{(b)}$	0.89	0.88	0.94	0.96	0.97	0.97
	$\hat{M}_Y^{(c)}$	0.93	0.96	0.97	0.99	0.98	0.99
	$\hat{M}_Y^{(p)}$	0.95	0.94	0.96	0.98	0.98	0.99
	$\hat{M}_Y^{(st)}$	0.92	0.96	0.97	0.99	0.98	0.99
0.25	$\hat{M}_Y^{(\ell)}(d)$	0.90	0.91	0.92	0.93	0.96	0.97
	$\hat{M}_Y^{(a)}$	0.94	0.96	0.97	0.98	0.98	0.98
	$\hat{M}_Y^{(b)}$	0.89	0.90	0.93	0.94	0.96	0.98
	$\hat{M}_Y^{(c)}$	0.91	0.92	0.95	0.97	0.99	0.98
	$\hat{M}_Y^{(p)}$	0.93	0.97	0.98	0.98	0.98	0.99
	$\hat{M}_Y^{(st)}$	0.94	0.96	0.97	0.98	0.99	0.98

which puts the probability mass $1/n$ at each sample point. Before we introduce a new estimator in double sampling, we review the definition of the customary ratio estimator of the population mean of Y in double sampling given by

$$\hat{y}_r = \left(\frac{\bar{x}'}{\bar{x}}\right)\bar{y} = h(\bar{y}),$$

where $\bar{y}$ and $\bar{x}$ are the second-phase sample means of y and x , respectively, and $\bar{x}'$ is the first-phase sample mean of the x variable. Also $h(a) = (\bar{x}'/\bar{x})a$ is a function satisfying $h(\bar{x}) = \bar{x}'$ for the given first-phase sample. Let F_X denote the population distribution function of the auxiliary variable X . Also let $\hat{F}'_X$ and $\hat{F}_X$ be the sample distribution functions of the auxiliary variable X based on the first- and second-phase sample information, respectively. We have an obvious choice for h :

$$h(u) = \hat{F}'_X(\hat{F}_X^{-1}(u)) \quad (0 < u < 1),$$

where $\hat{F}_X^{-1}(u)$ denotes the inverse of the function F_X . Based on this information, we define an estimator of $F_Y(y)$ in double sampling as

$$\hat{F}_X^{(t)}(y) = \hat{F}'_X(\hat{F}_X^{-1}(F_Y(y))). \tag{10}$$

Let the variables y_i and x_i be related to each other through a model defined as

$$\hat{y}_{ij} = \hat{B}x_i + v(x_i)\hat{u}_j, \quad j \in S, \quad i \in S' \subset \Omega,$$

where S' and S denote the first-phase and second-phase samples from the population Ω . The $\hat{u}_j$ are the residuals. Here our aim is to redistribute the point mass at each x_i of the given first-phase sample evenly to the n points. Also

$$\hat{B} = \frac{\sum_{i \in S} y_i x_i / v^2(x_i)}{\sum_{i \in S} x_i^2 / v^2(x_i)}$$

is the weighted least squares estimator of the regression coefficient. After the redistribution of mass based on a given first-phase sample, we obtain $\hat{Q}'$ and $\hat{Q}$ as

$$\begin{aligned} \hat{Q}'(y) &= \frac{1}{n'n} \sum_{i \in S'} \sum_{j \in S} \Delta(y - \hat{y}_{ij}) = \frac{1}{n'} \sum_{i \in S'} \hat{P}_i(y), \\ \hat{Q}(y) &= \frac{1}{n^2} \sum_{i \in S} \sum_{j \in S} \Delta(y - \hat{y}_{ij}) = \frac{1}{n} \sum_{i \in S} \hat{P}_i(y), \end{aligned}$$

where $\hat{P}_i(y) = 1/n \sum_{j \in S} \Delta(y - \hat{y}_{ij})$ is an estimator of $\Pr(y' \leq y \mid x' = x)$ under the model $y_i = \beta x_i + v(x_i)u_i$ for a given first-phase sample. Replacing $\hat{F}'_X$ and $\hat{F}_X^{-1}$ with $\hat{Q}'$ and $\hat{Q}^{-1}$ in (10) we get another estimator of $F_Y(y)$ as

$$\hat{F}_Y^{(q)}(y) = \hat{Q}'(\hat{Q}^{-1}(\hat{F}_Y(y))).$$

Now we have the following theorem.

Theorem 8.1. *The estimator $\hat{F}_Y^{(t)}(y)$ is asymptotically normal with mean $F(y)$ and variance given by*

$$\begin{aligned} \text{var}(\hat{F}_Y^{(t)}(y)) &= \left(\frac{1}{n'} - \frac{1}{N}\right) F_Y(y)(1 - F_Y(y)) \\ &\quad + \left(\frac{1}{n} - \frac{1}{n'}\right) \left(F_Y(y)(1 - F_Y(y)) + F_X(x)(1 - F_X(x)) - 2(H_{XY}(x, y) - F_Y(y)F_X(x)) \right), \end{aligned}$$

where $x = F_X^{-1}(F(y))$ and $H_{XY}(x, y)$ denotes the finite population joint distribution of X and Y .

Proof. The proof follows from the relation

$$\hat{F}_Y^{(t)}(y) - F_Y(y) \approx (\hat{F}_Y(y) - F_Y(y)) - (\hat{F}_X(x) - F_X(x)) + (\hat{F}'_X(x) - F_X(x)).$$

The estimator $\hat{F}_Y^{(q)}(y)$ can be written as

$$\hat{F}_Y^{(q)}(y) - F_Y(y) \approx (\hat{F}_Y(y) - F_Y(y)) - (\hat{Q}(y^*) - \hat{Q}'(y^*)),$$

where $y^* = Q^{-1}(F_Y(y))$. Thus the expression for the variance of the estimator $\hat{F}_Y^{(q)}$ can be developed by following Kuk & Mak (1994).

The estimator of the median can be obtained by inverting $\hat{F}_Y^{(t)}$ or $\hat{F}_Y^{(q)}$ at 0.5. In fact, this is equivalent to the inversion of the sample cdf, but at a position other than 0.5.

9. The Chen and Qin method

Chen & Qin (1993) suggested an empirical likelihood method for quantile estimation in the presence of auxiliary information. We follow their notation. They considered the problem of estimation of $\theta = E(T(X, Y))$, subject to the constraint $E(w(X)) = 0$. An estimator of θ based on the method of empirical likelihood function is given by

$$\hat{\theta}_n = \frac{1}{n} \sum_{i \in S} \frac{T(x_i, y_i)}{1 + \hat{\lambda} w(x_i)} \quad (11)$$

with $\hat{\lambda}$ satisfying

$$\sum_{i \in S} \frac{w(x_i)}{1 + \hat{\lambda} w(x_i)} = 0$$

for some suitable weights $w(x_i)$, $i = 1, 2, \dots, n$. We consider some special cases of the estimator (11).

1. If $w(x) = \bar{x} - \bar{x}'$, where $\bar{x}$ and $\bar{x}'$ are the second- and first-phase sample means of the auxiliary character X , then $\hat{\theta}$ is an analogue of the estimator of Hartley & Rao (1968) in double sampling.
2. If $w(x) = I_{(x \leq \hat{M}_X)} - 0.5$ and θ is the median of X , then the estimator (11) leads to the position estimator in double sampling.
3. If $w(x_i) = x_i - \bar{x}'$ and x_i is a 0–1 variable and $\bar{x}$ is the estimator of the population proportion $\bar{X}$ (say) based on the first-phase sample information, then the estimator (11) leads to an analogue of the post-stratified estimator of Silva & Skinner (1995) in double sampling.
4. The ratio type estimator proposed by Garcia & Cebrian (1998) is also an estimator of the form (11).

Appendix 1

A.1. Proof of Theorem 4.2

We have $\hat{M}_Y^{(p)} = \hat{Q}_Y(\hat{p}_Y^{(1)})$. Following Kuk & Mak (1989), we have

$$F_Y(\hat{M}_Y^{(p)}) = F_Y(M_Y) + f_Y(M_Y)(\hat{M}_Y^{(p)} - M_Y) + o_p(n^{-1/2}).$$

This implies

$$\hat{M}_Y^{(p)} - M_Y = \frac{1}{f_Y(M_Y)} (F_Y(\hat{M}_Y^{(p)}) - F_Y(M_Y)) + o_p(n^{-1/2}).$$

Now we have

$$\hat{p}_Y^{(0)} - \hat{p}_Y = 2(\hat{p}_X^* p_{11} + (1 - \hat{p}_X^*)(0.5 - p_{11})) - \hat{p}_Y = (4p_{11} - 1)(\hat{p}_X^* - 0.5) - (\hat{p}_Y - 0.5).$$

Also we have

$$\begin{aligned}\hat{p}_Y^{(1)} - \hat{p}_Y^{(0)} &= 2(\hat{p}_X^* \hat{p}_{11} + (1 - \hat{p}_X^*)(0.5 - \hat{p}_{11})) - 2(\hat{p}_X^* p_{11} + (1 - \hat{p}_X^*)(0.5 - p_{11})) \\ &= 2(\hat{p}_{11} - p_{11})(2\hat{p}_X^* - 1).\end{aligned}$$

Now $(\hat{p}_Y^{(1)} - \hat{p}_Y^{(0)})\sqrt{n} \rightarrow 0$ in probability, as $\hat{p}_X^* \rightarrow \frac{1}{2}$ in probability and $(\hat{p}_{11} - p_{11})$ is of order $o_p(n^{-1/2})$ and we may replace $\hat{p}_Y^{(1)}$ with $\hat{p}_Y^{(0)}$. The details of these assumptions are available in Kuk & Mak (1989), and we have

$$\hat{M}_Y^{(p)} - M_Y = \frac{1}{f_Y(M_Y)}(\hat{p}_Y^{(0)} - \hat{p}_Y) + o_p(n^{-1/2}).$$

Hence $\hat{M}_Y^{(p)}$ is approximately normal with mean M_Y . To find the variance of the estimator $\hat{M}_Y^{(p)}$, let E_2 and E_1 denote the expected values for the given first-phase sample and over all possible first-phase samples, respectively. Similarly defining the variances var_2 and var_1 , we have

$$\text{var}(\hat{M}_Y^{(p)}) = E_1 \text{var}_2(\hat{M}_Y^{(p)}) + \text{var}_1 E_2(\hat{M}_Y^{(p)}). \quad (12)$$

$$\text{Now } \text{var}_2(\hat{M}_Y^{(p)}) = \text{var}_2(\hat{M}_Y^{(p)} - M_Y) = \frac{1}{f_Y(M_Y)^2} \text{var}_2(\hat{p}_Y^{(0)} - \hat{p}_Y). \quad (13)$$

Again

$$\begin{aligned}\text{var}_2(\hat{p}_Y^{(0)} - \hat{p}_Y) &= \text{var}_2((4p_{11} - 1)(\hat{p}_X^* - 0.5) - (\hat{p}_Y - 0.5)) \\ &= (4\hat{p}_{11}^* - 1)^2 \text{var}(\hat{p}_X^*) + \text{var}(\hat{p}_Y) - 2(4\hat{p}_{11}^* - 1) \text{cov}_2(\hat{p}_X^*, \hat{p}_Y) \\ &= \frac{1}{4} \left(\frac{1}{n} - \frac{1}{n'} \right) (4\hat{p}_{11}^* - 1)^2 + \frac{1}{4} \left(\frac{1}{n} - \frac{1}{N} \right) - \frac{1}{2} \left(\frac{1}{n} - \frac{1}{n'} \right) (4\hat{p}_{11}^* - 1)^2 \\ &= \frac{1}{4} \left(\frac{1}{n} - \frac{1}{N} \right) - \frac{1}{4} \left(\frac{1}{n} - \frac{1}{n'} \right) (4\hat{p}_{11}^* - 1)^2,\end{aligned} \quad (14)$$

where $\hat{p}_{11}^*$ is defined by a cross-classification of the phase one sample using $\hat{M}_X'$ and $\hat{M}_Y'$ as the cut-offs; cov_2 denotes the covariance term for the given first-phase sample. Using (14) in (13), and then using the result in (12), we get Theorem 4.2.

Appendix 2

To prove Theorem 5.1, we have

$$F_Y(\hat{M}_Y^{(st)}) = F_Y(M_Y + (\hat{M}_Y^{(st)} - M_Y)) = F_Y(M_Y) + f_Y(M_Y)(\hat{M}_Y^{(st)} - M_Y) + o_p(n^{-1/2}).$$

This implies

$$\begin{aligned}\hat{M}_Y^{(st)} - M_Y &= \frac{1}{f_Y(M_Y)}(F_Y(\hat{M}_Y^{(st)}) - F_Y(M_Y)) + o_p(n^{-1/2}) \\ &= \frac{1}{f_Y(M_Y)}(0.5 - \hat{p}_Y) + o_p(n^{-1/2}).\end{aligned}$$

For large N , we have

$$F_Y(M_Y) \approx F_{YA}^*(M_Y) + F_{YB}^*(M_Y)$$

$$\text{var}(F_Y(M_Y)) = E_1 \text{var}_2(F_Y(M_Y)) + \text{var}_1 E_2(F_Y(M_Y)). \quad (15)$$

Now

$$\begin{aligned} \text{var}_2(F_Y(M_Y)) &\approx \text{var}_2(F_{YA}^*(M_Y)) + \text{var}_2(F_{YB}^*(M_Y)) \\ &= \frac{1}{4} \left(\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{n} - \frac{1}{n'} \right) (4\hat{p}_{11}^* - 1)^2 \right). \end{aligned}$$

On substituting in (15), we get Theorem 5.1.

References

- BRATLEY, P., FOX, B.L. & SCHRAGE, L.E. (1983). *A Guide to Simulation*. New York: Springer-Verlag.
- CHEN, J. & QIN, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107–116.
- FRANCISCO, C.A. & FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Ann. Statist.* **19**, 454–469.
- GARCIA, M.R. & CEBRIAN, A.A. (1998). Quantile interval estimation in finite populations using a multivariate ratio estimator. *Metrika* **47**, 203–213.
- HARTLEY, H.O. & RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika* **55**, 547–557.
- HORVITZ, D.G. & THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.
- KUK, A.Y.C. & MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **51**, 261–269.
- KUK, A.Y.C. & MAK, T.K. (1994). A functional approach to estimating finite population distribution functions. *Comm. Statist. Theory Methods* **23**, 883–896.
- SADASIVAN, G. & AGGARWAL, R. (1978). Optimum points of stratification in bi-variate populations. *Sankhyā Ser. C* **40**, 84–97.
- SANYAL, S.K. & SINHA, S.K. (1977). Methodological problems in large scale sample surveys — experiences from national sample survey. *Sankhyā Ser. C* **39**, 47–70.
- SILVA, P.L.D.N. & SKINNER, C.J. (1995). Estimating distribution functions with auxiliary information using poststratification. *J. Official Statist.* **11**, 277–294.
- SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- SINGH, S., HORN, S. & YU, F. (1998). Estimation of variance of the regression estimator: higher level calibration approach. *Survey Meth.* **24**, 41–50.
- SRIVASTAVA, S.K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statist. Assoc. Bull.* **16**, 121–132.
- SRIVASTAVA, S.K. (1971). A generalized estimator for the mean of a finite population using multi-auxiliary information. *J. Amer. Statist. Assoc.* **66**, 404–407.
- SRIVASTAVA, S.K. & JHAJJ, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika* **68**, 341–343.
- WALSH, J.E. (1970). Generalization of ratio estimator for population total. *Sankhyā Ser. A* **42**, 99–106.
- WANG, S.G. & CHOW, S.C. (1994). *Advanced Linear Models — Theory and Applications*. New York: Marcel Dekker, Inc.