COMPARISON OF RATIO ESTIMATORS IN TWO-PHASE SAMPLING

K. T. DeGraft-Johnson and J. Sedransk

(Received Jan. 5, 1972)

Summary

To estimate the finite population mean, $\bar{Y}$, a two-phase sample may be selected. A simple random sample of size n' is chosen, and a concomitant variable, X, is measured for all units. Then, a simple random subsample of size n $(0<n\leqq n')$ is chosen, and Y is measured. Seven ratio-type estimators of $\bar{Y}$ are given, and their biases and mean square errors determined to $O((n')^{-2})$. Then, the estimators are compared (a) without any assumptions about the relation between Y and X, and (b) assuming that Y and X are linearly related.

1. Introduction

Given a sample of size n from a finite population of N units, it is often possible to estimate the finite population mean with greater precision by utilizing known information about a concomitant variable, X, related to the variable under study, Y. When simple random sampling is used, a common approach is to employ a ratio estimator. Denote by (Y_j, X_j) the value of the variable (Y, X) for the jth unit in the population $(j=1, 2, \cdots, N)$, and by (y_i, x_i) the corresponding value of the variable for the ith draw in the sample $(i=1, 2, \cdots, n)$. Then, assuming simple random sampling, the ordinary ratio estimator of $\bar{Y}=\sum_{j=1}^{N} Y_j/N$ is given by $\hat{\bar{Y}}=\bar{X}\sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i$ where $\bar{X}$ is the (assumed known) population mean of the X variable.

When the per unit cost of measuring the concomitant variable, X, is non-negligible, it may not be feasible to utilize estimators such as $\hat{\bar{Y}}$. However, if the per unit cost of measuring X is considerably smaller than the per unit cost of measuring Y, one may use a two-phase sampling procedure. Here, a simple random sample of n' units is selected and the value of X determined for all units in this "first-phase" sample. Then, a simple random subsample of size n is selected from the n' units, and the value of Y is determined for these n units*. An estimator analogous to $\hat{Y}$ is given by $\hat{Y}_R=\bar{x}_{n'}\sum_{i=1}^{n} y_i\Big/\sum_{i=1}^{n} x_i$ where $\bar{x}_{n'}=\sum_{i=1}^{n'} x_i/n'$ denotes the sample mean of X from the first-phase sample.

With extensive use of two-phase sampling, ratio-type estimators analogous to those developed for use with single-phase sampling are of interest. Seven of these estimators are presented and discussed in Section 2 together with a brief description of the methods used to derive the moments of the estimators. Then, in Section 3, the biases and mean square errors of these estimators are compared (1) without any assumptions about (Y, X), and (2) assuming a linear relationship between Y and X, but without any distribution assumptions about X. (Note that these comparisons are summarized in Section 3.4.) The expected values and mean square errors of the ratio-type estimators are given in Section 4. Finally, the results of a numerical analysis of the efficacy of the approximate formulas for the biases and mean square errors of the estimators (Section 4) are summarized in Section 5.

2. The ratio estimators and some preliminary results

To aid in choosing the estimators to be compared, there are available both theoretical and numerical studies in which various ratio-type estimators are compared assuming *single-phase* sampling. Both Fraudendorfer [3] and Rao [5] summarize these studies; and, in addition, both carry out numerical investigations of the properties of various estimators using real finite populations. While some specific conclusions about the relative merits of the estimators may be drawn from the theoretical investigations, the numerical studies suggest that a ranking of the estimators may depend substantially on the actual finite population being sampled. Thus, it was decided to choose for comparison (two-phase sampling counterparts of) estimators which are "representative" of some of the *kinds* of (single-phase sampling) ratio-type estimators which have been suggested in the literature.

Assuming (unless otherwise specified) the two-phase sample design given in the second paragraph of Section 1, the first estimator is $\hat{Y}_R$ defined in Section 1:

$$\hat{Y}_R=\bar{y}_n\bar{x}_{n'}/\bar{x}_n \tag{2.1}$$

where $\bar{y}_n=\sum_{i=1}^{n} y_i/n$, $\bar{x}_n=\sum_{i=1}^{n} x_i/n$ and $\bar{x}_{n'}=\sum_{i=1}^{n'} x_i/n'$.

* For example, to estimate the total cocoa yield in a given locality in Ghana, one might use a sample of n' farms to estimate the total area under cocoa cultivation and a subsample of n farms to determine the actual yields.

The next two estimators are two-phase sampling counterparts of two estimators ("Beale" and "Tin") described and compared by Tin [6]. The constants in $\hat{\bar{Y}}_B$ and $\hat{\bar{Y}}_T$ given below have been chosen so that the biases of these estimators are, as in single-phase sampling, of $O(n^{-2})$.

$$\hat{\bar{Y}}_B=\hat{\bar{Y}}_R[1+\{(n'-n)s_{xy}/nn'\bar{x}_n\bar{y}_n\}][1+\{(n'-n)s_x^2/nn'\bar{x}_n^2\}]^{-1} \tag{2.2}$$

and

$$\hat{\bar{Y}}_T=\hat{\bar{Y}}_R\{1+(n'-n)[(s_{xy}/\bar{x}_n\bar{y}_n)-(s_x^2/\bar{x}_n^2)]/nn'\} \tag{2.3}$$

where $(n-1)s_{xy}=\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)$, $(n-1)s_x^2=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$. It may be noted that $\hat{\bar{Y}}_B$ and $\hat{\bar{Y}}_T$ are somewhat similar in form, and they have similar properties (Section 3). This is also true for their counterparts in single-phase sampling.

To apply the procedure (first suggested by Quenouille [4]) for modifying an estimator having a bias of $O(n^{-1})$ to produce an estimator having a bias of $O(n^{-2})$, the following sample design is assumed: The preliminary (simple random) sample of size n' is divided at random into two groups of equal size (*i.e.*, $n_1'=n_2'=n'/2$), and a (simple random) subsample of $n/2$ units is selected from each half of the preliminary sample. Letting $(\bar{x}_{n_1'}, \bar{x}_{n_2'})$ denote the sample means of X for the two halves of the preliminary sample, and $(\hat{R}_1, \hat{R}_2)$ the sample ratios $(\hat{R}=\bar{y}/\bar{x})$ calculated from the two second-phase samples,

$$\hat{\bar{Y}}_{Q1}=2\hat{\bar{Y}}_R-(\hat{R}_1\bar{x}_{n_1'}+\hat{R}_2\bar{x}_{n_2'})/2 \tag{2.4}$$

is an estimator of $\bar{Y}$ which has a bias of $O(n^{-2})$.

For *each* of the (remaining) three estimators, the sample design described in the second paragraph of Section 1 is assumed with the added specification that the (second-phase) simple random sample of size n be divided (at random) into halves. Denoting, as above, the two sample ratios (each based on $n/2$ units) by $(\hat{R}_1, \hat{R}_2)$

$$\hat{\bar{Y}}_{Q2}=\{[2\bar{y}_n/\bar{x}_n]-[(\hat{R}_1+\hat{R}_2)/2]\}\bar{x}_{n'} \tag{2.5}$$

is a two-phase sampling counterpart of the "Quenouille" estimator (see Cochran [1], p. 180). Since the bias of $\hat{\bar{Y}}_{Q2}$ is of $O((n')^{-1})$ (Section 4), a modification of $\hat{\bar{Y}}_{Q2}$ is also to be considered:

$$\hat{\bar{Y}}_{Q3}=\{[(2n'-n)(\bar{y}_n/n'\bar{x}_n)]-[(n'-n)(\hat{R}_1+\hat{R}_2)/2n']\}\bar{x}_{n'} \tag{2.6}$$

where the bias of $\hat{\bar{Y}}_{Q3}$ is of $O(n^{-2})$ (Section 4).

The last estimator is a two-phase sampling counterpart of "Mickey's" single-phase (see Rao [5], p. 214) unbiased estimator:

$$\hat{\bar{Y}}_M=\{(\hat{R}_1+\hat{R}_2)\bar{x}_{n'}/2\}+\{(2n'-n)/n'\}\,\{\bar{y}_n-(\hat{R}_1+\hat{R}_2)\bar{x}_n/2\}\ . \tag{2.7}$$

To derive the (approximate) expected values and variances of the alternative estimators, we shall employ series expansions of quantities such as $(1+\delta\bar{x}_n)^{-1}$ where $\delta\bar{x}_n=(\bar{x}_n-\bar{X})/\bar{X}$. While such series expansions have been extensively used in similar investigations (see, for example, Tin [6]), their use does not, of course, guarantee adequate approximations for the desired expected values and variances. However, the results of a numerical investigation of the efficacy of such approximations suggest that at least the approximations for the mean square errors are quite satisfactory (see Section 5).

To *illustrate* the derivations used for *each* of the estimators, consider $\hat{\bar{Y}}_R$ and define $\delta\bar{y}_n=(\bar{y}_n-\bar{Y})\bar{Y}^{-1}$, $\delta\bar{x}_n=(\bar{x}_n-\bar{X})\bar{X}^{-1}$ and $\delta\bar{x}_{n'}=(\bar{x}_{n'}-\bar{X})\bar{X}^{-1}$. Then,

$$\hat{\bar{Y}}_R=\bar{Y}(1+\delta\bar{y}_n)(1+\delta\bar{x}_{n'})(1+\delta\bar{x}_n)^{-1}\ , \tag{2.8}$$

and the right-hand side of (2.8) can be expanded to permit each of $\mathrm{E}(\hat{\bar{Y}}_R)$ and $\mathrm{V}(\hat{\bar{Y}}_R)$ to be evaluated to terms of $O((n')^{-2})$. To facilitate the evaluation of $\mathrm{E}(\hat{\bar{Y}}_R)$ and $\mathrm{V}(\hat{\bar{Y}}_R)$, note that the two-phase sample design (Section 1) is identical to one defined by an initial selection of n units by simple random sampling (from the population of N units) followed by a simple random sample of $n'-n$ units from the $N-n$ units not selected initially. Then,

$$\mathrm{E}\,(\delta\bar{x}_{n'}\,|\,x_1,\cdots,x_n;\,y_1,\cdots,y_n)=n(N-n')\delta\bar{x}_n/n'(N-n) \tag{2.9}$$

and

$$\begin{aligned}\mathrm{E}\,[(\delta\bar{x}_{n'})^2\,|\,x_1,\cdots,x_n;\,y_1,\cdots,y_n]&\\ =[n(N-n')\delta\bar{x}_n/n'(N-n)]^2&+(N-n')(n'-n)\\ \cdot\,[(n'\bar{X})^2(N-n)(N-n-1)]^{-1}&\\ \cdot\,[(N-1)S_x^2-(n-1)s_x^2-nN\bar{X}^2(\delta\bar{x}_{n'})^2(N-n)^{-1}]&\end{aligned} \tag{2.10}$$

where, in (2.9) and (2.10), the expectation is conditional on the values of X and Y in the second-phase sample. To determine $\mathrm{E}(\hat{\bar{Y}}_R)$ and $\mathrm{V}(\hat{\bar{Y}}_R)$, it is helpful to use (2.9) and (2.10) because all of the statistics then pertain to the sample of size n (see (2.8), (2.9) and (2.10)).

Formula (2.10) can be simplified considerably if it is assumed that N is sufficiently large that the terms (n'/N) and (n/N) can be neglected. Then,

$$\mathrm{E}\,(\delta\bar{x}_{n'}\,|\,x_1,\cdots,x_n;\,y_1,\cdots,y_n)\doteq n\delta\bar{x}_n/n' \tag{2.11}$$

and

$$(2.12) \qquad \mathrm{E}\,[(\delta\bar{x}_{n'})^2 \,|\, x_1, \cdots, x_n; y_1, \cdots, y_n] \doteq (n\delta\bar{x}_n/n')^2 + (n'-n)C_{20}/(n')^2$$

where

$$C_{20} = \sum_{i=1}^{N} (X_i - \bar{X})^2 / N\bar{X}^2 .$$

To simplify the derivations and to facilitate the comparisons it will be assumed throughout that N is sufficiently large that (n'/N) and (n/N) can be neglected. Of course, if the per unit cost of observing X is almost negligible, (n'/N) may be too large to be neglected. However, if this occurs, each two-phase sampling procedure should be quite similar to its single-phase sampling counterpart and, thus, this case is of little interest because we wish to compare ratio-type estimators in *two-phase* sampling.

Finally, note that (1) the subsampling fraction (n/n') is not neglected; (2) for all of the estimators to be well-defined, it is assumed that $X > 0$; and (3) to simplify the comparisons (Section 3) it is assumed that $Y > 0$.

3. Comparisons

3.1. *Introduction*

In this Section we compare the alternative estimators using both bias and mean square error as criteria. To ensure meaningful and, hopefully, accurate comparisons even for fairly small sample sizes, both the expected value and mean square error for each estimator have been determined up to (and including) terms of $O((n')^{-2})$. Thus the effect of the bias of $\hat{Y}_R$ (which is of $O(n^{-1})$) on its mean square error will be considered whereas the bias will have no effect on the mean square error (determined to terms of $O((n')^{-2})$) for most of the alternative ratio-type estimators. This is important because the purpose of most of the alternative estimation methods is to provide an estimator which is either unbiased or has a smaller (order) bias than $\hat{Y}_R$.

Some general comparisons of the estimators are given in Sections 3.2 and 3.3; and they are summarized in Section 3.4. These comparisons give a general picture of the relative merits of the estimators. However, the complexity of some of the formulas (for both the mean square errors and biases) precludes the possibility of ranking all of the estimators (under specified *simple* conditions). Thus, an investigator who repeatedly conducts surveys on the "same" (or a similar) population may wish to choose an estimator for a current survey by evaluating the bias and mean square error for each estimator (using estimates, from prior surveys, of the necessary parameters). To meet this need we give, in Section 4, the bias and mean square error (to terms of $O((n')^{-2})$) for each estimator.

Finally, when comparing the alternative estimators, the differential costs of calculating the estimates from a large-scale survey should be considered. However, since no information about comparative costs appears to be available (for either single- or two-phase sampling), such cost comparisons are not feasible at present.

Throughout Section 3 it is assumed that ρ, the population correlation coefficient, is positive. If $\rho<0$, one would ordinarily use a "product-type" rather than a "ratio-type" estimator. Further, define

$$C_{uv}=\sum_{k=1}^{N}(X_k-\bar{X})^u(Y_k-\bar{Y})^v/N\bar{X}^u\bar{Y}^v\,, \tag{3.1}$$

$$C_X=\sqrt{C_{20}}=\text{coefficient of variation of } X\,, \tag{3.2}$$

$$R=\bar{Y}/\bar{X}\,. \tag{3.3}$$

Note that $\rho>0$ implies that $C_{11}>0$ (because it has been assumed that $X>0$ and $Y>0$).

3.2. *General comparisons*

Using the expressions given in Section 4 for the biases of the estimators, it may be shown that

(a$_1$) $\mathrm{E}(\hat{\bar{Y}}_M)=\bar{Y}$.

(b$_1$) The biases of $\hat{\bar{Y}}_B$, $\hat{\bar{Y}}_T$, $\hat{\bar{Y}}_{Q1}$ and $\hat{\bar{Y}}_{Q3}$ are of $O(n^{-2})$ whereas the bias of $\hat{\bar{Y}}_R$ is of $O(n^{-1})$, and that of $\hat{\bar{Y}}_{Q2}$ is of $O((n')^{-1})$.

(c$_1$) If $C_{21}=C_{30}$ (as in a bivariate normal population), to terms* of $O((n')^{-2})$,

(i) $|B(\hat{\bar{Y}}_B)|\leqq|B(\hat{\bar{Y}}_T)|\leqq|B(\hat{\bar{Y}}_{Q3})|\leqq|B(\hat{\bar{Y}}_{Q1})|\leqq|B(\hat{\bar{Y}}_{Q2})|$

(ii) $|B(\hat{\bar{Y}}_{Q1})|\leqq|B(\hat{\bar{Y}}_R)|$ if $n\geqq 3C_{20}$ (which is likely to hold). Here, $B(Y)$ denotes the bias of Y.

(d$_1$) To terms of $O((n')^{-1})$, $|B(\hat{\bar{Y}}_{Q2})|\leqq|B(\hat{\bar{Y}}_R)|$ if and only if, $(n/n')\leqq(1/2)$.

To facilitate comparing the mean square errors of the estimators (see Section 4), it is usually possible to write (for two estimators $\hat{\bar{Y}}_i$, $\hat{\bar{Y}}_j$):

$$\mathrm{MSE}\,(\hat{\bar{Y}}_i)-\mathrm{MSE}\,(\hat{\bar{Y}}_j)=\bar{Y}^2\{(n'-n)/nn'\}\,\{[K_1n^{-1}]+[K_2(n')^{-1}]\} \tag{3.4}$$

where K_1 and K_2 are simple quadratic functions of the C_{uv}. Then, using relationships such as $(1-\rho^2)C_{20}C_{02}=C_{20}C_{02}-C_{11}^2$, the results summarized below can be obtained.

(a$_2$) To terms of $O((n')^{-1})$, $\hat{\bar{Y}}_R$, $\hat{\bar{Y}}_B$, $\hat{\bar{Y}}_T$, $\hat{\bar{Y}}_{Q1}$, $\hat{\bar{Y}}_{Q2}$, $\hat{\bar{Y}}_{Q3}$ and $\hat{\bar{Y}}_M$ have the

* "Terms of $O(\cdot)$" means "up to and *including* terms of $O(\cdot)$" (when such terms exist).

same mean square error.

(b$_2$) To terms of $O(n^{-2})$,

(i) MSE $(\hat{\bar{Y}}_T)=$MSE $(\hat{\bar{Y}}_B)\leqq$MSE $(\hat{\bar{Y}}_{Q1})=$MSE $(\hat{\bar{Y}}_{Q2})=$MSE $(\hat{\bar{Y}}_{Q3})\leqq$ MSE $(\hat{\bar{Y}}_M)$.

(ii) If $C_{21}=C_{30}=C_{12}$, MSE $(\hat{\bar{Y}}_M)\leqq$MSE $(\hat{\bar{Y}}_R)$.

Note that when n' is sufficiently large that the terms of $O((nn')^{-1})$ and of $O((n')^{-2})$ can be neglected, the comparisons in (b$_2$) are definitive.

(c$_2$) To terms of $O((n')^{-2})$,

(i) MSE $(\hat{\bar{Y}}_T)=$MSE $(\hat{\bar{Y}}_B)\leqq$MSE $(\hat{\bar{Y}}_{Q3})\leqq$MSE $(\hat{\bar{Y}}_M)$.

(ii) If $(n/n')<(2/3)$, MSE $(\hat{\bar{Y}}_T)<$MSE $(\hat{\bar{Y}}_{Q1})$.

(d$_2$) Comparisons involving $\hat{\bar{Y}}_R$. If $C_{21}=C_{12}=C_{30}$, to terms of $O((n')^{-2})$,

(i) If 1) $\rho>(C_X/C_Y)$, or 2) $\rho<(C_X/C_Y)$ *and* $[n/(n+n')]<[1-(\rho C_Y/C_X)]/4$, MSE $(\hat{\bar{Y}}_R)>$MSE $(\hat{\bar{Y}}_M)$.

(ii) If 1) $\rho>(C_X/C_Y)$, or 2) $\rho<(C_X/C_Y)$ *and* $[n/n']<[1-(\rho C_Y/C_X)]$, MSE $(\hat{\bar{Y}}_R)>$MSE $(\hat{\bar{Y}}_{Q3})$ and MSE $(\hat{\bar{Y}}_R)>$MSE $(\hat{\bar{Y}}_{Q1})$.

(iii) If 1) $\rho>(C_X/C_Y)$, or 2) $\rho<(C_X/C_Y)$ *and* $[n/n']<5[1-(\rho C_Y/C_X)]/8$, MSE $(\hat{\bar{Y}}_R)>$MSE $(\hat{\bar{Y}}_{Q2})$.

(e$_2$) Comparisons involving $\hat{\bar{Y}}_{Q2}$. If $C_{21}=C_{12}=C_{30}$, to terms of $O((n')^{-2})$,

(i) MSE $(\hat{\bar{Y}}_{Q2})<$MSE $(\hat{\bar{Y}}_{Q1})$ if 1) $\rho>(C_X/C_Y)$, or 2) $\rho<(C_X/C_Y)$ *and* $[n/n']<[1-(\rho C_Y/C_X)]/4$.

(ii) MSE $(\hat{\bar{Y}}_{Q2})<$MSE $(\hat{\bar{Y}}_{Q3})$ if 1) $(C_X/C_Y)<\rho<(5C_X/C_Y)$, or 2) $\rho<(C_X/C_Y)$ *and* $[n/n']<2[1-(\rho C_Y/C_X)]/5$.

(iii) MSE $(\hat{\bar{Y}}_{Q2})>$MSE $(\hat{\bar{Y}}_T)$ if $\rho<(C_X/C_Y)$ *and* $(n/n')<(1/3)$.

3.3. *General comparisons: linear model*

To further compare the alternative estimators it is postulated that

$$Y_i=\beta X_i+\varepsilon_i \tag{3.5}$$

where $\mathrm{E}(\varepsilon_i|x_i)=0$, $\mathrm{V}(\varepsilon_i|x_i)=\sigma^2X_i^g$ $(g\geqq 0)$ and $(\varepsilon_i, \varepsilon_j)$ are assumed to be independent. Then, it is easily shown that each of the ratio-type estimators is an unbiased estimator of $\bar{Y}$. Further, to terms of $O((n')^{-2})$:

(a$_3$) For any value of $g\geqq 0$, we have

$$\text{MSE}\,(\hat{\bar{Y}}_B)=\text{MSE}\,(\hat{\bar{Y}}_T)\leqq\text{MSE}\,(\hat{\bar{Y}}_M)=\text{MSE}\,(\hat{\bar{Y}}_{Q3})\leqq\text{MSE}\,(\hat{\bar{Y}}_{Q1})\,.$$

(b$_3$) For $g=0$, we have

$$\begin{aligned}\text{MSE}\,(\hat{\bar{Y}}_B)&=\text{MSE}\,(\hat{\bar{Y}}_T)\leqq\text{MSE}\,(\hat{\bar{Y}}_{Q2})\leqq\text{MSE}\,(\hat{\bar{Y}}_M)\\&=\text{MSE}\,(\hat{\bar{Y}}_{Q3})\leqq\text{MSE}\,(\hat{\bar{Y}}_{Q1})\leqq\text{MSE}\,(\hat{\bar{Y}}_R)\end{aligned}$$

where MSE $(\hat{\bar{Y}}_T)\leqq$MSE $(\hat{\bar{Y}}_{Q2})$ if $(n/n')\leqq(1/2)$. (Note that the other relationships in (b_3) do *not* depend on the condition $(n/n')\leqq(1/2)$.)

(c_3) For $g=1$, we have

$$\text{MSE}(\hat{\bar{Y}}_B)=\text{MSE}(\hat{\bar{Y}}_T)=\text{MSE}(\hat{\bar{Y}}_R)\leqq\text{MSE}(\hat{\bar{Y}}_M)$$
$$=\text{MSE}(\hat{\bar{Y}}_{Q3})\leqq\text{MSE}(\hat{\bar{Y}}_{Q1})\leqq\text{MSE}(\hat{\bar{Y}}_{Q2})\ .$$

3.4. *Summary of comparisons*

The use of a ratio estimator is typically associated with a relationship between Y and X which is, roughly, of the form (3.5). Thus, the comparisons in Section 3.3 may serve as a recommendation for using either the "Beale" or "Tin" estimator in preference to the others. Such a recommendation is strengthened by the comparisons of the mean square errors given in Section 3.2: (b_2) and (c_2). In particular, the comparisons in (b_2) may be indicative since, in some applications, the terms of $O((nn')^{-1})$ and $O((n')^{-2})$ will be negligible. Finally, apart from the unbiased estimator, $\hat{\bar{Y}}_M$, the biases of $\hat{\bar{Y}}_B$ and $\hat{\bar{Y}}_T$ are smaller (under the conditions stated in Section 3.2 (c_1)) than those of the alternatives. The prospective user should, however, note that most of the comparisons (Section 3.2) involving $\hat{\bar{Y}}_R$ and $\hat{\bar{Y}}_{Q2}$ depend on the assumption that $C_{12}=C_{21}=C_{30}$.

4. Expected values and mean square errors

To facilitate comparisons of the ratio-type estimators for specific finite populations, the expectations and mean square errors of the seven estimators are given (up to and including terms of $O((n')^{-2})$) below. A practitioner who conducts repeated surveys of the same (or a similar) population may wish to choose an estimator for a current survey by evaluating the bias and mean square error of each ratio-type estimator using estimates of the C_{uv} obtained from a previously conducted survey (or, from a pilot sample survey).

$$\text{(4.1)}\qquad \text{E}(\hat{\bar{Y}}_R)=\bar{Y}\Big[1+\Big(\frac{1}{n}-\frac{1}{n'}\Big)(C_{20}-C_{11})+\frac{1}{n^2}\{(C_{21}-C_{30})+3C_{20}(C_{20}-C_{11})\}$$
$$+\frac{1}{nn'}\{(C_{30}-C_{21})-3C_{20}(C_{20}-C_{11})\}\Big]\ ,$$

$$\text{(4.2)}\qquad \text{E}(\hat{\bar{Y}}_B)=\bar{Y}\Big[1+\frac{1}{n^2}\{2(C_{30}-C_{21})-2C_{20}(C_{20}-C_{11})\}$$
$$+\frac{1}{nn'}\{3(C_{21}-C_{30})+4C_{20}(C_{20}-C_{11})\}$$
$$+\frac{1}{(n')^2}\{(C_{30}-C_{21})-2C_{20}(C_{20}-C_{11})\}\Big]\ ,$$

$$\begin{aligned}\mathrm{E}(\hat{\bar{Y}}_T)=\bar{Y}\Bigg[1&+\frac{1}{n^2}\{2(C_{30}-C_{21})-3C_{20}(C_{20}-C_{11})\}\\&+\frac{1}{nn'}\{3(C_{21}-C_{30})+6C_{20}(C_{20}-C_{11})\}\\&+\frac{1}{(n')^2}\{(C_{30}-C_{21})-3C_{20}(C_{20}-C_{11})\}\Bigg]\,,\end{aligned}\tag{4.3}$$

$$\begin{aligned}\mathrm{E}(\hat{\bar{Y}}_{Q1})=\bar{Y}\Bigg[1&-\frac{2}{n^2}\{(C_{21}-C_{30})+3C_{20}(C_{20}-C_{11})\}\\&+\frac{2}{nn'}\{(C_{21}-C_{30})+3C_{20}(C_{20}-C_{11})\}\Bigg]\,,\end{aligned}\tag{4.4}$$

$$\mathrm{E}(\hat{\bar{Y}}_{Q2})=\bar{Y}\Bigg[1+\frac{1}{n'}(C_{11}-C_{20})+\frac{1}{n^2}\{-2(C_{21}-C_{30})+6C_{20}(C_{11}-C_{20})\}\Bigg]\,,\tag{4.5}$$

$$\begin{aligned}\mathrm{E}(\hat{\bar{Y}}_{Q3})=\bar{Y}\Bigg[1&+\frac{1}{n^2}\{-2(C_{21}-C_{30})+6C_{20}(C_{11}-C_{20})\}\\&+\frac{1}{nn'}\{3(C_{21}-C_{30})+9C_{20}(C_{20}-C_{11})\}\\&+\frac{1}{(n')^2}\{(C_{30}-C_{21})+3C_{20}(C_{11}-C_{20})\}\Bigg]\,,\end{aligned}\tag{4.6}$$

$$\mathrm{E}(\hat{\bar{Y}}_M)=\bar{Y}\,.\tag{4.7}$$

$$\begin{aligned}\mathrm{MSE}\,(\hat{\bar{Y}}_R)=\bar{Y}^2\Bigg[&\frac{1}{n}(C_{02}-2C_{11}+C_{20})+\frac{1}{n'}(2C_{11}-C_{20})+\frac{1}{n^2}(4C_{21}-2C_{12}\\&+6C_{11}^2+3C_{20}C_{02}-2C_{30}-18C_{20}C_{11}+9C_{20}^2)\\&+\frac{1}{nn'}(4C_{30}-6C_{21}+26C_{20}C_{11}+2C_{12}-8C_{11}^2-3C_{20}C_{02}\\&-15C_{20}^2)+\frac{1}{(n')^2}(2C_{21}+2C_{11}^2-2C_{30}-8C_{20}C_{11}+6C_{20}^2)\Bigg]\,,\end{aligned}\tag{4.8}$$

$$\begin{aligned}\mathrm{MSE}\,(\hat{\bar{Y}}_B)&=\mathrm{MSE}\,(\hat{\bar{Y}}_T)\\&=\bar{Y}^2\Bigg[\frac{1}{n}(C_{02}-2C_{11}+C_{20})+\frac{1}{n'}(2C_{11}-C_{20})\\&\quad+\frac{1}{n^2}\{2(C_{20}-C_{11})^2+(C_{20}C_{02}-C_{11}^2)\}+\frac{1}{nn'}\{-(C_{20}-C_{11})^2\\&\quad-(C_{20}C_{02}-C_{11}^2)\}+\frac{1}{(n')^2}\{-(C_{20}-C_{11})^2\}\Bigg]\,,\end{aligned}\tag{4.9}$$

$$\mathrm{MSE}\,(\hat{\bar{Y}}_{Q1})=\bar{Y}^2\Bigg[\frac{1}{n}(C_{20}-2C_{11}+C_{02})+\frac{1}{n'}(2C_{11}-C_{20})\tag{4.10}$$

$$+\frac{1}{n^2}(2C_{11}^2+2C_{20}C_{02}-8C_{20}C_{11}+4C_{20}^2)$$
$$+\frac{1}{nn'}(12C_{20}C_{11}-4C_{11}^2-2C_{20}C_{02}-6C_{20}^2)$$
$$\left.+\frac{1}{(n')^2}(2C_{11}^2-4C_{20}C_{11}+2C_{20}^2)\right],$$

$$\text{(4.11)}\quad \text{MSE}\,(\hat{\bar{Y}}_{Q2})=\bar{Y}^2\left[\frac{1}{n}(C_{02}+C_{20}-2C_{11})+\frac{1}{n'}(2C_{11}-C_{20})\right.$$
$$+\frac{1}{n^2}(4C_{20}^2+2C_{20}C_{02}+2C_{11}^2-8C_{11}C_{20})$$
$$+\frac{1}{nn'}(2C_{12}+2C_{30}-4C_{21}-3C_{20}C_{02}-4C_{11}^2+14C_{11}C_{20}$$
$$\left.-7C_{20}^2)+\frac{1}{(n')^2}(-2C_{30}+2C_{21}+2C_{11}^2+6C_{20}^2-8C_{11}C_{20})\right],$$

$$\text{(4.12)}\quad \text{MSE}\,(\hat{\bar{Y}}_{Q3})=\bar{Y}^2\left[\frac{1}{n}(C_{02}+C_{20}-2C_{11})+\frac{1}{n'}(2C_{11}-C_{20})\right.$$
$$+\frac{1}{n^2}(4C_{20}^2+2C_{20}C_{02}+2C_{11}^2-8C_{11}C_{20})$$
$$+\frac{1}{nn'}(10C_{20}C_{11}-5C_{20}^2-2C_{11}^2-3C_{20}C_{02})$$
$$\left.+\frac{1}{(n')^2}(C_{20}^2-2C_{11}C_{20}+C_{20}C_{02})\right],$$

$$\text{(4.13)}\quad \text{MSE}\,(\hat{\bar{Y}}_{M})=\bar{Y}^2\left[\frac{1}{n}(C_{02}+C_{20}-2C_{11})+\frac{1}{n'}(2C_{11}-C_{20})\right.$$
$$+\frac{1}{n^2}(6C_{11}^2+2C_{20}C_{02}-16C_{20}C_{11}+8C_{20}^2)$$
$$+\frac{1}{nn'}(22C_{20}C_{11}-8C_{11}^2-3C_{20}C_{02}-11C_{20}^2)$$
$$\left.+\frac{1}{(n')^2}(2C_{11}^2-6C_{20}C_{11}+3C_{20}^2+C_{20}C_{02})\right].$$

5. Numerical study of the approximations

To investigate how closely the expressions given in Section 4 for the biases and mean square errors approximate the true values, a small Monte Carlo study was conducted. Two thousand two-phase samples each with ($n'=30$, $n=10$) were selected from a finite population of $N=3{,}164$ units according to the sample design described in Section 1 (*i.e.*, simple random sampling at each phase). For each sample, $\hat{\bar{Y}}_R$, $\hat{\bar{Y}}_B$ and

$\hat{Y}_T$ were evaluated. Then, from the two thousand replications, unbiased estimates of $\mathrm{E}(\hat{Y})$ and $\mathrm{Var}(\hat{Y})$ were calculated for each of the three ratio-type estimators. These "Monte Carlo" estimates were then compared with the corresponding asymptotic expressions given in Section 4. In addition to ($n'=30$, $n=10$), the sample sizes ($n'=80$, $n=20$) and ($n'=300$, $n=100$) were used. The finite population consisted of $N=3, 164$ trees*, but four different X variables were used in conjunction with Y=gross volume: X_1=diameter, breast-high; X_2=height; $X_3=X_1^2$; $X_4=X_1^2X_2$. Thus, four populations of (Y, X) values and three sets of sample sizes were considered. Both the relationship between Y and X_1 and that between Y and X_2 can be characterized, roughly, as "quadratic". The relation between Y and X_3 and that between Y and X_4 are both "strongly" linear ($\rho=0.982$ and $\rho=0.997$, respectively). For further details about the variables and finite populations see deGraft-Johnson [2].

Denote by MSE(A) and $B(A)$ the asymptotic expressions for the

Table 1. Comparison of the asymptotic expressions for the bias and mean square error with the (Monte Carlo) estimated values of these parameters

	(10, 30)		(20, 80)		(100, 300)	
	$\frac{\mathrm{MSE}(A)}{\mathrm{MSE}(M)}$	$B(M)$, $B(A)$	$\frac{\mathrm{MSE}(A)}{\mathrm{MSE}(M)}$	$B(M)$, $B(A)$	$\frac{\mathrm{MSE}(A)}{\mathrm{MSE}(M)}$	$B(M)$, $B(A)$
(Y, X_1)						
$\hat{Y}_R$	1.025	−0.317, −0.261	1.037	−0.158, −0.148	1.042	0.008, −0.0263
$\hat{Y}_B$	1.026	−0.090, −0.030	1.038	−0.019, −0.009	1.041	0.034, −0.0003
$\hat{Y}_T$	1.024	−0.085, −0.025	1.037	−0.017, −0.007	1.041	0.034, −0.0003
(Y, X_2)						
$\hat{Y}_R$	1.075	−0.162, −0.192	0.958	−0.026, −0.107	0.969	0.005, −0.0188
$\hat{Y}_B$	1.069	0.015, −0.013	0.953	0.077, −0.004	0.968	0.024, −0.0001
$\hat{Y}_T$	1.068	0.020, −0.009	0.953	0.078, −0.002	0.968	0.024, −0.0001
(Y, X_3)						
$\hat{Y}_R$	1.079	−0.247, −0.244	0.967	−0.105, −0.132	1.032	0.004, −0.0227
$\hat{Y}_B$	1.072	−0.078, −0.066	0.960	0.007, −0.019	1.033	0.026, −0.0007
$\hat{Y}_T$	1.069	−0.063, −0.048	0.959	0.012, −0.013	1.033	0.026, −0.0005
(Y, X_4)						
$\hat{Y}_R$	1.119	−0.024, −0.007	1.018	0.014, −0.005	1.003	0.001, −0.0010
$\hat{Y}_B$	1.114	−0.029, −0.014	1.020	0.015, −0.004	1.002	0.002, −0.0001
$\hat{Y}_T$	1.113	−0.030, −0.012	1.019	0.015, −0.004	1.002	0.002, −0.0001

* Note that in forest inventory surveys, sampling fractions of less than 0.01 are typical.

mean square error and bias of any ratio-type estimator. Similarly, MSE(M) and $B(M)$ are the "Monte Carlo" estimates of the mean square error and bias. For each population and set of sample sizes, the values of [MSE(A)/MSE(M)] and [$B(M)$, $B(A)$] are tabled for $\hat{\hat{Y}}_R$, $\hat{\hat{Y}}_B$ and $\hat{\hat{Y}}_T$. In Table 1, the maximum and median values of [|MSE$(A)-$MSE(M)|/MSE(M)] are 0.119 and 0.038, respectively. Thus, at least for these populations, the asymptotic expressions for the mean square errors appear to provide good approximations.

Noting that $\bar{Y}=8.9633$, the biases appear to be very small (in absolute value) for each of these four populations. Thus, it is difficult to judge the adequacy of the asymptotic expressions for the biases.

The authors wish to thank Miss Greta Ljung for checking many of the formulas presented in Section 4, Mr. Eric West for writing the necessary computer programs, and Dr. Kenneth Ware for providing the data used in Section 5. This work was partially supported by the Population Council of America, the Government of Ghana, and the U.S. Office of Education under contract No. OEC-3-6-002041-2041.

CENTRAL BUREAU OF STATISTICS, GHANA
UNIVERSITY OF WISCONSIN

REFERENCES

[1] Cochran, W. G. (1963). *Sampling Techniques*, John Wiley and Sons, New York, Second edition.

[2] deGraft-Johnson, K. T. (1969). Some contributions to the theory of two-phase sampling, Ph.D. dissertation, Iowa State University, Ames, Iowa.

[3] Frauendorfer, R. (1967). Numerical analysis of some unbiased and approximately unbiased ratio type estimators, M.S. thesis, Iowa State University.

[4] Quenouille, M. H. (1956). Notes on bias in estimation, *Biometrika*, **43**, 353-360.

[5] Rao, J. N. K. (1969). Ratio and regression estimators, *New Developments in Survey Sampling*, N. L. Johnson and H. Smith, editors, New York: Wiley, 213-234.

[6] Tin, M. (1965). Comparison of some ratio estimators, *J. Amer. Statist. Ass.*, **60**, 294-307.