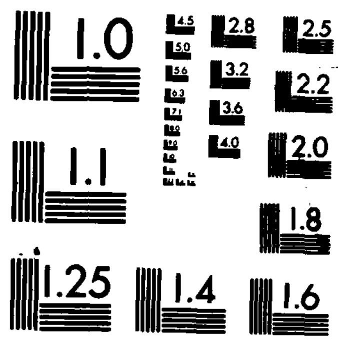
AD-A160 346 DOUBLE SAMPLING IN ESTIMATION OF A RATIO(U) PITTSBURGH 1/1
UNIV PA CENTER FOR MULTIVARIATE ANALYSIS
M AHMAD ET AL. JUN 85 TR-85-24 AFOSR-TR-85-0750
UNCLASSIFIED F49620-85-C-0008 F/G 12/1 NL

END



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR- 85-0750

AD-A160 346

DOUBLE SAMPLING IN ESTIMATION OF A RATIO*

Manzoor Ahmad
Claire Marcotte

Universite du Quebéc a Montréal

Center for Multivarite Analysis
University of Pittsburgh

and

R.S. Singh

University of Guelph

Center for Multivariate Analysis

University of Pittsburgh

DTIC FILE COPY

85 10 11 146

DOUBLE SAMPLING IN ESTIMATION OF A RATIO*

Manzoor Ahmad
Claire Marcotte

Universite du Quebéc a Montréal

Center for Multivarite Analysis
University of Pittsburgh

and

R.S. Singh

University of Guelph

June 1985

Technical Report No. 85-24

Center for Multivariate Analysis
515 Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

*Part of the work was supported by the Air Force Office of Scientific Research under Contract F49620-85-C-0008. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.
Authors are thankful to NSERC of Canada for research support.

- A -

DOUBLE SAMPLING IN ESTIMATION OF A RATIO

Manzoor Ahmad*
Claire Marcotte
Universite du Quebéc a Montréal

and

R.S. Singh*
University of Guelph

ABSTRACT

Problems connected with the estimation of the ratio of the means of a finite bivariate population have been considered in this report. The usual estimator of the ratio, based on the means of a bivariate simple random sample drawn without replacement (s.r.s. (w.o.r.)), has been compared with estimators based on alternative double sampling design. Under this design a very large s.r.s. (w.o.r.) is drawn for measuring only one of the variables and a subsample (s.r.r. (w.o.r.)) is drawn out of the first phase units for measuring the other variable. Efficiency and bias comparisons have been made by subjecting each of the competitors to the same budgetary constraint. It turns out that deviation from the usual set-up sometimes leads to better sampling strategies.

Additional keywords: mean square errors; statistical inference; multivariate analysis; approximation (mathematics).

1. INTRODUCTION

An inference problem, not uncommon in practice, is the estimation of the ratio $R = \bar{Y}/\bar{X}$ of two means of a finite bivariate population; $[(X_i, Y_i): i = 1, \ldots, N]$, where X_i and Y_i are measurements of two characteristics X and Y for a member of the population identified by the integer i. Addressing to this problem, we have confined ourselves to estimators of the form $\hat{R} = \hat{\bar{Y}}/\hat{\bar{X}}$ based on some suitably chosen random sampling scheme. A commonly used 'strategy' (see Des Raj [1963]) is to draw a bivariate simple random sample without replacement (s.r.s. (w.o.r.)), $(x_1, y_1), \ldots, (x_n, y_n)$ and to choose $\hat{\bar{X}} = \bar{x}$ and $\hat{\bar{Y}} = \bar{y}$, the sample means. We propose an alternative strategy in which the units are drawn by 'Double Sampling' scheme (see Cochran [1977], pp. 327-355, Sukhatme [1954], Murthy [1977]).
We have made comparisons by subjecting the competitors to a common constraint. Ensuing conclusions are relevant to situations in which the availability of resources (e.g. budget or time) is a dominant factor, it is a given constant B.

Let q_0 be the overhead per-unit cost of selecting a unit from a given population and, q_x and q_y be the per-unit costs of measuring respectively the characteristics X and Y for the selected unit. The sample size for the usual estimator $\hat{R} = \bar{y}/\bar{x}$ is then determined by

$$n(q_0 + q_x + q_y) = B. \tag{1.1}$$

In many practical situations q_x and q_y differ significantly, sometimes to an extent that a sacrifice of one observation on the costlier component may allow considerably large number of observations on the cheaper component. In situations of this kind it is possible to obtain a net gain in precision (over the usual estimator $\hat{R} = \bar{y}/\bar{x}$) using a modified estimator. Such as estimator is based on Double sampling (also called two phase sampling) in which one takes the advantage of some prior knowledge about the correlation between the component characteristics and the relative costs of measurements $q = (q_0 + q_x)^{-1} q_y$ and $q^* = q_x(q_0 + q_y)^{-1}$ by allowing a considerably large sample on the cheaper component (X or Y) at the expense of a small reduction in the size of the joint sample. The idea of double sampling was first proposedby Neyman (1938) in the context of stratified sampling. Since then many authors have used it for different problems, see for example, Robson and King (1953), Sukhatme (1962), Raj (1963), Sendransk [(1965,1967)], Khan and Tripathi (1967), Sen (1972).

We have proposed two different modified estimators of the ratio $R = \bar{Y}/\bar{X}$ depending of whether X or Y is cheaper and compared them with the usual estimator $\hat{R} = \bar{y}/\bar{x}$ in terms of bias and MSE. Accordingly two different cases are considered in sections 2 and 3

case (i): $q = q_y(q_0 + q_x)^{-1} > 1$ (X cheaper)

case (ii): $q^* = q_x(q_0 + q_y)^{-1} > 1$ (Y cheaper).

In section 3 the method of approximation of Bias and MSE has been discussed. This is in the same spirit as in David and Sukhatme (1974) except that asymptotic in the context of double sampling has been clarified and unified with the usual asymptotic for $\hat{R}$.

2. THE CASE: $q = q_y(q_0 + q_x)^{-1} > 1$

As mentioned earlier, the modified estimator, proposed in this section, will be compared with the usual estimator $\hat{R} = \bar{y}/\bar{x}$ based on a s.r.s. (w.o.r.) whose size n is pre-determined by the budgetary constraint: $n = B(q_0 + q_x + q_y)^{-1}$.

Our modified estimator is based on a double sampling scheme (or a two phase scheme) in which s.r. sampling (w.o.r.) is used in both phases. If we allow a reduction of order αn, $\alpha \in (0,1)$, in the size of the joint sample (that we use for $\hat{R}$) i.e. if we allow a sacrifice of αn observations on the costlier component Y, we can take $q\alpha n$ additional observations on X; we implicitly assume αn and $q\alpha n$ to be positive integers. We shall denote by n_s and n_ℓ respectively the size of the smaller (than n) sample on Y and the size of larger sample on X; we shall frequently use the identities

$$n_s = n - \alpha n = n(1-\alpha)$$

$$n_\ell = n + q\alpha n = n(1+q\alpha). \tag{2.1}$$

Let $x_1, x_2, \ldots, x_\ell$ be the sample of X-measurements for the units selected in the first phase leading to

$$\bar{x}_\ell = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} x_j, \quad \text{and} \quad s^2_{\ell x} = \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}(x_j - \bar{x}_\ell)^2 \tag{2.2}$$

which are known to be unbiased estimators of $\bar{X}$ and S^2_X (Cochran [1977] pp. 19-27), where

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad S^2_X = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2. \tag{2.3}$$

It is also well known that

$$E(\bar{x}_\ell - \bar{X})^2 = (1-f_\ell)\ S^2_X/n_\ell \tag{2.4}$$

where $f_\ell = n_\ell N^{-1}$.

Let $\{j_1, j_2, \ldots, j_{n_s}\}$ be the sample of units drawn from the set of first phase units, $j_k \in \{1,2,\ldots,n\}$, $k = 1,2,\ldots,n_s$, and let their Y-values be $y_{i_1}, y_{i_2}, \ldots, y_{i_s}$, leading to

$$\bar{y}_s = \frac{1}{n_s}\sum_{k=1}^{n_s} y_{j_k} . \tag{2.5}$$

Denoting the unobserved mean and the variance of Y - values of the first phase units by

$$\bar{y}_\ell = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} y_j \quad \text{and} \quad s_{\ell y}^2 = \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell} (y_j - \bar{y}_\ell)^2, \tag{2.6}$$

following properties will be required in the sequel:

$$E(\bar{y}_s | n_s) = \bar{y}_\ell$$

$$E[(\bar{y}_s - \bar{Y})^2 | n_s] = (\frac{1}{n_s} - \frac{1}{n_\ell}) s_{\ell y}^2 + (\bar{y}_\ell - \bar{Y})^2$$

$$E(\bar{y}_\ell) = \bar{Y} = \sum_{i=1}^{N} Y_i / N$$

$$E(s_{\ell y}^2) = S_Y^2 = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 / (N-1)$$

$$E(\bar{y}_\ell - \bar{Y})^2 = (1-f_\ell) S_Y^2 / n_\ell$$

and

$$E(\bar{y}_\ell - \bar{Y})(\bar{x}_\ell - \bar{X}) = \rho(1-f_\ell) S_X S_Y / n_\ell \tag{2.7}$$

where

$$\rho = [(N-1)S_X S_Y]^{-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

is the population correlation coefficient between the characteristics X and Y and, $E(\cdot|n_\ell)$ denotes the conditional expectation, given the Y - values of units selected in the first-phase.

Now we propose the following modified estimator of the ratio $R = \bar{Y}/\bar{X}$:

$$\hat{R}_1 = \bar{y}_s / \bar{x}_\ell . \tag{2.8}$$

Instead of the Bias $B(\hat{R}_1) = E(\hat{R}_1 - R)$ and the mean squared error $MSE(\hat{R}_1) = E(\hat{R}_1 - R)^2$ we will consider the relative bias and the relative mean squared error of $\hat{R}_1$ as well as of any other estimator. Denoting them respectively by RB and RMSE, we have

$$RB(\hat{R}_1) = E(\frac{\hat{R}_1}{R} - 1) = \frac{1}{R} B(\hat{R}_1)$$

$$RMSE(\hat{R}_1) = E(\frac{\hat{R}_1}{R} - 1)^2 = \frac{1}{R^2} MSE(\hat{R}_1) \qquad (2.9)$$

and denoting their 1^{st} order approximations respectively by $RB_1(\hat{R}_1)$ and $RMSE_1(\hat{R}_1)$ they turn out to be

$$RB_1(\hat{R}_1) = (\frac{1}{n_\ell} - \frac{1}{N})(C_X^2 - \rho C_X C_Y)$$

$$= \frac{1}{n}(\frac{1}{1+q\alpha} - f)\ (C_X^2 - \rho C_X C_Y)$$

$$RMSE_1(\hat{R}_1) = (\frac{1}{n_s} - \frac{1}{N})C_Y^2 + (\frac{1}{n_\ell} - \frac{1}{N})(C_X^2 - 2\rho C_X C_Y)$$

$$= \frac{1}{n}\{(\frac{1}{1-\alpha} - \frac{1}{1+q\alpha})C_Y^2 + (\frac{1}{1+q\alpha} - f)C_{XY}^2\}$$

where

$$f = n|N$$

$$C_X = S_X/\bar{X},\ C_Y = S_Y/\bar{Y}$$

$$C_{XY}^2 = C_Y^2 + C_X^2 - 2\rho C_X C_Y. \qquad (2.10)$$

The errors in these approximations satisfy

$$|RB(\hat{R}_1) - RB_1(\hat{R}_1)| = 0(n^{-2}) \qquad (2.11)$$

and

$$|RMSE(\hat{R}_1) - RMSE_1(\hat{R}_1)| = O(n^{-2}). \tag{2.12}$$

A proof in a slightly general setting and with some mild assumptions appears in section 3. Asymptotics for the context of double sampling has been clarified and unified with the usual asymptotics for $\hat{R}$.

Comparison with $\hat{R}$

It is well known (see Sukhatme and David, 1974) that a first order approximation of the relative Bias $B(\hat{R})$ and the relative $MSE(\hat{R})$ of $\hat{R}$ are

$$\begin{aligned} RB_1(R) &= (\frac{1}{n} - \frac{1}{N})(C_X^2 - \rho C_X C_Y) \\ &= \frac{1}{n}(1-f)(C_X^2 - \rho C_X C_Y) \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} RMSE_1(\hat{R}) &= (\frac{1}{n} - \frac{1}{N}) C_{XY}^2 \\ &= \frac{1}{n}(1-f) C_{XY}^2 \quad . \end{aligned} \tag{2.14}$$

The errors satisfy

$$|RB(\hat{R}) - B_1(\hat{R})| = O(\frac{1}{n^2}) \tag{2.15}$$

and

$$RMSE(\hat{R}) - MSE_1(\hat{R}) = O(\frac{1}{n^2}). \tag{2.16}$$

Expressing the $RMSE_1(\hat{R}_1)$ in terms of $RMSE_1(\hat{R})$, using (2.10) and (2.14), we obtain

$$\begin{aligned} RMSE_1(\hat{R}_1) &= RMSE_1(\hat{R}) \\ &+ \frac{C_{XY}^2}{n} \{\frac{q\alpha}{(1-\alpha)(1+q\alpha)} (\alpha - \frac{qC-1}{q+qC})\} \end{aligned} \tag{2.17}$$

where

$$C = \frac{C_{XY}^2 - C_Y^2}{C_Y^2} = C_{X|Y}^2 - 2\rho C_{X|Y}$$

with

$$C_{X|Y} = C_X/C_Y \quad \text{and} \quad C_{Y|X} = C_Y/C_X.$$

In case C is positive and

$$q > \frac{1}{C} \tag{2.18}$$

$\hat{R}_1$ would perform better than $\hat{R}$ as long as α, $0 < \alpha < 1$, satisfies

$$\alpha < \frac{C - \frac{1}{q}}{C+1} . \tag{2.19}$$

In terms of the correlation coefficient ρ, $C_{X|Y} = C_X/C_Y$ and $C_{Y|X} = 1/C_{X|Y}$ conditions (2.18) and (2.19) can be expressed as

$$2\rho C_Y < C_X - C_Y^2/qC_X \tag{2.20}$$

and

$$\alpha < \frac{C_{X|Y}^2 - 2\, C_{X|Y} - \frac{1}{q}}{C_{X|Y}^2 - 2\, C_{X|Y} + 1} . \tag{2.21}$$

Observe that if coefficient of variation of the variable X is twice as much as the coefficient of variation of the variable Y, and the cost ratio q is moderately large (say $q \geq 5$), the condition (2.20) holds. In terms of ρ, whenever q is large and $\rho \leq 0$, (2.20) is clearly satisfied.

Assuming that we are in a practical situation in which (2.18) is satisfied, the best choice of α can be found by minimizing (2.17) w.r.t. α. Taking the first derivative of the bracketted expression in (2.17) and equating it to zero, we are led to solve the quadratic equation:

$$(q-C)\alpha^2 + 2(1+C)\alpha + (1-qC)q^{-1} = 0. \tag{2.21}$$

The roots of (2.21) turn out to be

$$\frac{-2(1+C) \pm \Delta}{2(q-C)} \tag{2.22}$$

where

$$\Delta = 2\sqrt{C}\,(1+q)/\sqrt{q}\,.$$

Denoting the critical points by α_0 and α_1, we get

$$\alpha_0 = \frac{\sqrt{q}\sqrt{C} - 1}{q + \sqrt{q}\sqrt{C}} = \frac{\sqrt{C} - (\sqrt{q})^{-1}}{\sqrt{C} + \sqrt{q}}$$

and

$$\alpha_1 = \frac{\sqrt{q}\sqrt{C} + 1}{\sqrt{q}\sqrt{C} - q}\,. \tag{2.23}$$

Due to the assumption $q > 1$ and $qC > 1$ (vide (2.18)), it follows that $\alpha_1 \notin (0,1)$, $\alpha_0 \in (0,1)$. The second derivative of (2.17) with respect to α, evaluated at α_0, turns out to be

$$\frac{\partial^2}{\partial\alpha^2}(\mathrm{RMSE}_1(\hat{R}_1))\,|\,\alpha = \alpha_0$$

$$= \frac{2q^2C_Y^2}{n(1+q\alpha_0)^3}\{C + (\frac{1+q\alpha_0}{1-\alpha_0})^3 \frac{1}{q^2}\} \tag{2.24}$$

which is positive for $C \geq 0$.

Hence, under the condition (2.18), the optimum size of the joint sample is

$$n_C = n(1-\alpha_0) = n\,\frac{1+\frac{1}{q}}{1+\sqrt{C/q}} \tag{2.25}$$

and the size of the additional sample on X is

$$nq_0 = n \frac{\sqrt{q}\ \sqrt{C} - 1}{1 + \sqrt{C/q}} . \tag{2.26}$$

The $\mathrm{RMSE}_1(\hat{R}_1)$ corresponding to the optimum choice of α turns out to be

$$[\mathrm{RMSE}_1(\hat{R}_1)]_0 = \mathrm{RMSE}_1(\hat{R}) - \frac{C_Y^2}{n(1+\frac{1}{q})} (\sqrt{C} - \frac{1}{\sqrt{q}})^2 . \tag{2.27}$$

Ignoring the terms of order N^{-1},

$$[\mathrm{MSE}_1(\hat{R}_1)]_0 = R^2 \frac{C_{XY}^2}{n} - \frac{C_Y^2}{n} \frac{(\sqrt{C} - \frac{1}{\sqrt{q}})^2}{(1+\frac{1}{q})} \tag{2.28}$$

and when q is very large

$$\simeq \frac{S_Y^2}{n\bar{X}^2} \tag{2.29}$$

implying that for a small reduction in the joint sample (see (2.25)) a very large sample of additional observations (see (2.26)) enables the modified estimator behave as if its denominator were the population mean $\bar{X}$. In terms of the parameters $R = \bar{Y}/\bar{X}$, S_X^2, S_Y^2 and ρ, the optimum $[\mathrm{MSE}_1(\hat{R}_1)]_0$ is

$$[\mathrm{MSE}_1(\hat{R}_1)]_0 = \frac{1}{n(1+q)\bar{X}^2} [\sqrt{q}\ S_Y + \sqrt{R^2 S_X^2 - 2R\rho S_X S_Y}\ ]^2 - \frac{R^2}{N} C_{XY}^2 . \tag{2.30}$$

The relative efficiency of the optimum modified estimator $\hat{R}_1$ (say $(\hat{R}_1)_0$ w.r.t. $\hat{R}$ is

$$\text{R.E.}(\hat{R}_1)_0 = \frac{\text{MSE}_1(\hat{R})}{[\text{MSE}_1(\hat{R}_1)]_0} = \frac{\text{RMSE}_1(\hat{R})}{[\text{RMSE}_1(\hat{R}_1)]_0}$$

$$= \frac{(1+q)[1+C^2_{X|Y} - 2\ C_{X|Y}] - \frac{n}{N}[1+C^2_{X|Y} - 2\ C_{X|Y}](1+q)}{[\sqrt{q} + \sqrt{C^2_{X|Y} - 2\rho C_{X|Y}}\ ]^2 - \frac{n}{N}\ [1 + C^2_{X|Y} - 2\rho C_{X|Y}](1+q)} \tag{2.31}$$

In situations in which $\frac{(1+q)n}{N} \cong 0$, we have

$$[\text{MSE}_1(\hat{R}_1)]_0 = \text{MSE}_1(\hat{R}_1)_0$$

$$\cong \frac{1}{n(1+q)}\,\bar{X}^{-2}\ [\sqrt{q}\ S_Y + \sqrt{R^2 S_X^2 - 2R\ S_X S_Y}\ ]^2 \tag{2.32}$$

and

$$\text{RE}(\hat{R}_1)_0 \cong \frac{(1+q)[1+C^2_{X|Y} - 2\ C_{X|Y}]}{[\sqrt{q} + \sqrt{C^2_{X|Y} - 2\rho C_{X|Y}}\ ]^2}\ . \tag{2.33}$$

Further if $C_{X\ Y} = C_X\ C_Y = 1$

$$\text{RE}(\hat{R}_1) \cong \frac{2(1+q)(1-\rho)}{[\sqrt{q} + \sqrt{1-2\rho}\ ]^2} \tag{2.34}$$

3. APPROXIMATION OF BIAS AND MSE

Let $\hat{\hat{X}}$ and $\hat{\hat{Y}}$ denote, respectively, unbiased estimators of $\bar{X}$ and $\bar{Y}$, based on a finite number of randomly chosen units. We assume taht

$$\hat{\hat{X}} > 0$$

for all possible realizations of the sampling scheme, and we propose the ratio

$$\hat{R} = \hat{\hat{Y}}/\hat{\hat{X}} \tag{3.1}$$

as the estimator of the population ratio

$$R = \bar{Y}/\bar{X}. \tag{3.2}$$

The bias of this estimator, in terms of the random relative deviations

$$\delta = (\hat{\bar{Y}} - \bar{Y})/\bar{Y}$$

and

$$\epsilon = (\hat{\bar{X}} - \bar{X})/\bar{X} \tag{3.3}$$

is

$$B(\hat{R}) = E(\hat{R}-R) = RE[(\delta-\epsilon)(1+\epsilon)^{-1}]. \tag{3.4}$$

If we define a k^{th} order approximation of $B(\hat{R})$ as

$$B_k(\hat{R}) = RE[(\delta-\epsilon)\sum_{i=1}^{2k}(-\epsilon)^{i-1}], \tag{3.5}$$

then the error of this approximation turns out to be

$$\begin{aligned} e_k B &= B(\hat{R}) - B_k(\hat{R}) \\ &= \bar{Y}\,E[(1/\hat{\bar{X}})(\delta-\epsilon)\epsilon^{2k}]. \end{aligned} \tag{3.6}$$

The mean squared error of $\hat{R}$, $MSE(\hat{R})$, in terms of δ and ϵ is

$$\begin{aligned} MSE(\hat{R}) &= E(\hat{R}-R)^2 \\ &= R^2E[(\delta-\epsilon)^2(1+\epsilon)^{-2}]. \end{aligned} \tag{3.7}$$

If we define a k^{th} order approximation of $MSE(\hat{R})$ as

$$MSE_k(\hat{R}) = R^2E[(\delta-\epsilon)^2\sum_{i=1}^{2k-1} i(-\epsilon)^{i-1}] \tag{3.8}$$

then the error of this approximation turns out to be

$$e_k MSE = MSE(\hat{R}) - MSE_k(\hat{R})$$
$$= -\bar{Y}^2 E[(1/\hat{\bar{X}})^2(\delta-\epsilon)^2(2k\epsilon^{2k-1}+(2k-1)\epsilon^{2k})] \quad . \tag{3.9}$$

In the sequel we give bounds for $e_k B$ and $e_k MSE$ for each of the estimators considered in this report. In doing so we follow the techniques and arguments of David and Sukhatme [1974]. We also make use of the following order relation the proof of which can be found in the work of these authors; for non-negative integers r and s,

$$\bar{Y}^r \bar{X}^s E(\delta^r \epsilon^s) = \begin{cases} 0\{1/n^{(r+s)/2}\} & \text{if } r+s \text{ is even} \\ 0\{1/n^{(t+s+1)/2}\} & \text{if } r+s \text{ is odd} \end{cases} \tag{3.10}$$

where the random 'relative' deviations

$$\epsilon = (\hat{\bar{X}} - \bar{X})/\bar{X} = (\bar{x}-\bar{X})/\bar{X} \ ,$$

and

$$\delta = (\hat{\bar{Y}} - \bar{Y})/\bar{Y} = (\bar{y}-\bar{Y})/\bar{Y},$$

are based on the sample means $\bar{x}$ and $\bar{y}$ of a simple random sample (without replacement) of size n. However, when the sample means $\hat{\bar{X}}$ and $\hat{\bar{Y}}$ are based on a two-phase sampling design in which a large sample of n_ℓ units are drawn in the first phase to compute one of the means and a subsample of n_s units are drawn out of first phase units to compute the other, a similar order relation follows.

<u>REMARK 1</u>. In the case of $\hat{R}_1 = \bar{y}_s/\bar{x}_\ell$,

$$\delta = \delta_\ell = (\bar{y} - \bar{Y})/\bar{Y} = E(\delta_s | n_\ell)$$
$$\epsilon = \epsilon_\ell = (\bar{x} - \bar{X})/\bar{X} \tag{3.11}$$

where

$$\delta_s = (\bar{y}_s-\bar{Y})/\bar{Y} = \delta_\ell + \{(\bar{y}_s-\bar{y}_\ell)/\bar{Y}\}$$

and $E(\cdot|n_\ell)$ denotes conditional expectation given the Y - values of the first phase units.

REMARK 2. In the case of $\hat{R}_2 = \bar{y}_\ell/\bar{x}_s$,

$$\epsilon = \epsilon_\ell = (\bar{x} - \bar{X})/\bar{X} = E(\epsilon_s|n_\ell) \tag{3.12}$$

where

$$\epsilon_s = (\bar{x}_s-\bar{X})/\bar{X} = \epsilon_\ell + \{(\bar{x}_s-\bar{x}_\ell)/\bar{X}\}.$$

We now study the order relation for the estimator $\hat{R}_1$. The same for $\hat{R}_2$ can be studied similarly. For the positive and even integers r and s the computation of the order of $E(\bar{y}_s-\bar{Y})^r(\bar{x}_\ell-\bar{X})^s$ proceeds as follows;

$$C\,E[\delta_\ell^r\epsilon_\ell^s] = E[(\bar{y}_s-\bar{Y})^{2a}(\bar{x}_\ell-\bar{X})^{2b}]$$

with $r = 2a4s = 2b$, where a and b are non-negative integers and $C = \bar{X}^s\bar{Y}^r$. Let

$$\Delta\bar{x}_\ell = \bar{x}_\ell-\bar{X}, \quad \Delta\bar{y}_{s|\ell} = \bar{y}_s - \bar{y}_\ell$$

and

$$\Delta\bar{y}_\ell = \bar{y}_\ell - \bar{Y}. \tag{3.13}$$

We note that

$$C\delta_s^r\epsilon_\ell^s = (\Delta\bar{x}_\ell)^{2b}(\Delta\bar{y}_{s|\ell} + \Delta\bar{y}_\ell)^{2a},$$

which in turn can be expressed as

$$C\delta_s^r\epsilon_\ell^s = (\Delta\bar{x}_\ell)^{2b}\Big[\sum_{i=0}^{a}\binom{2a}{2i}(\Delta\bar{y}_{s|\ell})^{2i}(\Delta\bar{y}_\ell)^{2a-2i}$$

$$+\sum_{i=0}^{a-1}\binom{2a}{2i+1}(\Delta\bar{y}_{s|\ell})^{2i+1}(\Delta\bar{y}_\ell)^{2a-2i-1}\Big]. \qquad (3.14)$$

Now taking the expectation, conditionally given the first phase units and adopting the order relation (3.10) also for this situation, dispensing momentarily with the justification for doing so, we have

$$(\Delta\bar{x}_\ell)^{2b}\Big[\sum_{i=0}^{a}\binom{2a}{2i}\ O\Big(\frac{1}{\sqrt{n_s}}\Big)^{2i}(\Delta\bar{y}_\ell)^{2a-2i}$$

$$+\sum_{i=0}^{a-1}\binom{2a}{2i+1}\ O\Big(\frac{1}{\sqrt{n_s}}\Big)^{2i+2}(\Delta\bar{y}_\ell)^{2a-(2i+1)}\Big]. \qquad (3.15)$$

Next, the expectation with respect to the first phase units yields the unconditional expectation as

$$\sum_{i=0}^{a}\binom{2a}{2i}\ O\Big(\frac{1}{\sqrt{n_s}}\Big)^{2i}\ O\Big(\frac{1}{\sqrt{n_\ell}}\Big)^{2a+2b-2i}$$

$$+\sum_{i=0}^{a-1}\binom{2a}{2i+1}\ O\Big(\frac{1}{\sqrt{n_s}}\Big)^{2i+2}\ O\Big(\frac{1}{\sqrt{n_\ell}}\Big)^{2a+2b-(2i-1)+1},$$

which in turn is of order $O(1/\sqrt{n})^{2a+2b}$, yielding

$$C\ E[\delta_s^r\epsilon_\ell^s] = O\Big(\frac{1}{\sqrt{n}}\Big)^{r+s},\quad r+s \text{ even};$$

since $n_s = (1-\alpha)n$ and $n_\ell = (1-q\alpha)n$ with $0<\alpha<1$. Similarly we get the order relation for the other cases of $r+s$. The justification for these order relations are as follows.

In order to compare $\hat{R}$ with $\hat{R}_i$, their respective settings for asymptotics must be interlinked and this in both situations $i=1,2$. For doing so we implicitly follow the setting of David and Sukhatme [1974] for the asymptotic study of $\hat{R}$ according to which a sequence of finite populations $S_1, S_2, \ldots, S_n$ of sizes $N_1, N_2, \ldots, N_n$ is drawn from a bivariate super population $S = \{(u_1,v_1),(u_2,v_2),\ldots\}$. The sequence of sizes is strictly increasing so that $n = t_n N_n$, $\lim_{n\to\infty} t_n = t$, $0 < t < 1$ and $0 < t_n < 1$ for every n.

The sampling scheme adopted in the case of $\hat{R}_i$ $(i = 1,2)$ is two phase sampling. In the first phase, a simple random sample (w.o.r.) of size $n_\ell = (1+q\alpha)n$, $q\alpha > 1$, is drawn from S_n. We assume that $n_\ell \le N_n$ for every n, so that

$$n_\ell = (1+q\alpha)t_n N_n = t_n^* N_n$$

The condition $n \to \infty$ implies $n_\ell \to \infty$ and also that $t_n^* \to t^* = (1+q\alpha)t$ with $0 < t^* < 1$ and $0 < t_n^* < 1$ for every n. In the second phase, a simple random sample (without replacement) of size

$$n_s = (1-\alpha)n = \frac{1-\alpha}{1+qs} n_\ell = \lambda_n n_\ell, \quad 0 < \alpha < 1,$$

$\lambda_n = \lambda = (1-\alpha)/1+qs$, is drawn from a sub-population consisting of n_ℓ units. The sizes of these sub-populations are strictly increasing. The condition $n_s \to \infty$ is implied by the condition $n \to \infty$ and we also have $\lim_{n\to\infty} \lambda_n = \lambda$ with $0 < \lambda = \lambda_n < 1$ for every n. The settings for asymptotics being linked in this manner the content of Sukhatme and David [1974] applies in verbatim.

<u>Bias and MSE of $\hat{R}_1$</u>

Following (3.4), (3.6), and (3.7) and the Remark 1,

$$B(\hat{R}_1) = R\, E[(\delta_s - \epsilon_\ell)(1+\epsilon_\ell)^{-1}], \tag{3.16}$$

$$B_k(\hat{R}_1) = R\, E[(\delta_s - \epsilon_\ell) \sum_{i=1}^{2k} (-\epsilon_\ell)^{i-1}], \tag{3.17}$$

and

$$e_k B = \bar{Y}\, E[(1/\bar{x}_\ell)(\delta_s - \epsilon_\ell)\epsilon_\ell^{2k}. \tag{3.18}$$

As mentioned earlier, we assume that X-measurements are positive for all the units of the population. Let x_0 be the lower bound for $\bar{x}_\ell$, we have then

$$e_k B \leq \frac{\bar{Y}}{x_0} [E(\delta_s \epsilon_\ell^{2k}) - E(\epsilon_\ell)^{2k+1}] \tag{3.19}$$

and using the order relation,

$$|e_k B| \leq 0(1/n)^k . \tag{3.20}$$

This bound over the error of paproximation of order k shows that it suffices to approximate $B(\hat{R}_1)$ by $B_1(\hat{R}_1)$ for sufficiently large sample size; letting $k=1$ in (3.5), we get

$$B_1(\hat{R}_1) = RE\,[(\delta_s - \epsilon_\ell)(1-\epsilon_\ell)]$$

$$= RE\,[\delta_s - \epsilon_\ell + \epsilon_\ell^2 - \delta_s \epsilon_\ell] \tag{3.21}$$

where

$$E(\delta_s) = EE(\delta_s | n_\ell) = E(\delta_\ell)$$

$$= \frac{1}{\bar{Y}} E(\bar{y}_\ell - \bar{Y}) = 0.$$

Similarly

$$E(\delta_s \epsilon_\ell) = \frac{1}{\bar{X}\bar{Y}} E[(\bar{y}_\ell - \bar{Y})(\bar{x}_\ell - \bar{X})]$$

$$= \rho\, C_x C_y$$

and

$$E(\epsilon_\ell^2) = \frac{1}{\bar{X}^2} E(\bar{x}_\ell - \bar{X})^2$$

$$= (\frac{1}{n_\ell} - \frac{1}{N}) C_x^2 .$$

We have therefore

$$B_1(\hat{R}_1|R) = (\frac{1}{n_\ell} - \frac{1}{N})(C_x^2 - \rho\, C_x C_y). \tag{3.22}$$

As to the $MSE(\hat{R}_1)$ we have

$$MSE(\hat{R}_1) = R^2 E[(\delta_s - \epsilon_\ell)^2 (1+\epsilon_\ell)^{-2}] \tag{3.23}$$

$$MSE_k(\hat{R}_1) = R^2 E[(\delta_s - \epsilon_\ell)^2 \sum_{i=1}^{2k-1} i(-\epsilon)^{i-1}] \tag{3.24}$$

and

$$e_k MSE = - \bar{Y}^2\; E[k/\bar{x}_\ell^2)(\delta_s - \epsilon_\ell)^2 (2k\epsilon_\ell^{2k-1} + (2k-1)\epsilon_\ell^{2k})]. \tag{3.25}$$

Noting that

$$|e_k MSE| \le \frac{\bar{Y}^2}{x_0^2} |2k\; E\{(\delta_s^2 - 2\delta_s \epsilon_\ell + \epsilon_\ell^2)\epsilon_\ell^{2k-1}\}$$

$$+ (2k-1)\; E\{(\delta_s^2 - 2\delta_s \epsilon_\ell + \epsilon_\ell^2)\epsilon_\ell^{2k}\}| \tag{3.26}$$

we observe that

$$|e_k MSE| \le O(1/n)^k \tag{3.27}$$

and consequently approximate $MSE(\hat{R}_1|R)$ by $MSE_1(\hat{R}_1|R)$ which turns out to be

$$MSE_1(\hat{R}_1|R) = E[(\delta_s - \epsilon_\ell)^2]$$

$$= (\frac{1}{n_s} - \frac{1}{N})C_y^2 + (\frac{1}{n_\ell} - \frac{1}{N})(C_x^2 - 2\rho C_x C_y). \tag{3.28}$$

Bias and MSE of $\hat{R}_2$

Computations similar to those involved in the case of $\hat{R}_1$ lead to the following

$$|e_k B| = O(1/n)^k \ , \tag{3.29}$$

$$|e_k MSE| = O(1/n)^k \ , \tag{3.30}$$

$$B_1(\hat{R}_2) = (\frac{1}{n_s} - \frac{1}{N})C_x^2 - (\frac{1}{n_\ell} - \frac{1}{N})\rho \ C_x C_y \tag{3.31}$$

and

$$MSE_1(\hat{R}_2|R) = (\frac{1}{n_s} - \frac{1}{N})C_x^2 + (\frac{1}{n_\ell} - \frac{1}{N})(C_y^2 - 2\rho \ C_x C_y). \tag{3.32}$$

REFERENCES

[1] Cochran, W.G. [1977]. Sampling Techniques. Wiley Eastern Limited.

[2] David, I.P. and Sukhatme, B.V. [1974]. On the bias the mean square error of the ratio-estimate, Jour. Amer. Stat. Assoc., 69, 464-466.

[3] Khan, S. and Tripathi, T.P. [1967]. The use of multivariate auxiliary information in double sampling. J. Ind. Stat. Assoc., 5, 42-48.

[4] Murthy, M.N. [1977]. Sampling Theory and Methods. Statistical Publishing Society, Calcutta, India.

[5] Neyman, J. [1938]. Contributions to the theory of sampling human populations, Jour. Amer. Stat. Assoc. 33, 101-116.

[6] Raj, Des [1964]. On double sampling for pps estimation. Ann. Math. Stat., 35.

[7] Raj, Des [1979]. Sampling Theory. Tata McGraw-Hill Publ. Co. Ltd.

[8] Robson, D.S. and King, A.J. [1953]. Double sampling and the curtis impact survey, Cornell Univ. Agr. Exp. Sta. Mem., 231.

[9] Sen, A.R. [1972]. Successive sampling with p (p21) auxiliary variables, Ann. Math. Stat. 43, 2031-2034.

[10] Sendransk, [1965]. A double sampling scheme for analytical surveys. Jour. Amer. Stat. Assoc., 60, 985-1004.

[11] Sendransk, [1972]. Designing some multifactor analytical studies, Jour. Amer. Stat. Assoc., 62, 1121-1139.

[12] Sukhatme, P.V. [1954]. Sampling Theory of Surveys, with Applications. Iowa State College Press, Ames, Iowa.

[13] Sukhatme, P.V. [1962]. Some ratio type estimators in two-phase sampling, Jour. Amer. Stat. Assoc. 57, 628-632.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 85-0750	2. GOVT ACCESSION NO. AD-A160	3. RECIPIENT'S CATALOG NUMBER 346
4. TITLE (and Subtitle) Double Sampling in Estimation of a Ratio		5. TYPE OF REPORT & PERIOD COVERED Technical - June, 1985
		6. PERFORMING ORG. REPORT NUMBER 85-24
7. AUTHOR(s) Manzoor Ahmad, Claire Marcotte, and R. S. Singh		8. CONTRACT OR GRANT NUMBER(s) F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Multivariate Analysis 515 Thackeray Hall Univ.of Pittsburgh, Pittsburgh, PA 15260		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Department of the Air Force Bolling Air Force Base DC 20332		12. REPORT DATE June, 1985
		13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Double Sampling, Ratio Estimator, Simple Random Sample, Bias, Mean Square Error

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

Problems connected with the estimation of the ratio of the means of a finite bivariate population have been considered in this report. The usual estimator of the ratio, based on the means of a bivariate simple random sample drawn without replacement (s.r.s. (w.o.r.)), has been compared with the estimators based on alternative double sampling design. Under this design a very large s.r.s. (w.o.r.) is drawn for measuring only one of the variables and a subsample (s.r.r. (w.o.r.)) is drawn out of the first phase of units for measuring the other variable. Continued

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Efficiency and bias comparisons have been made by subjecting each of the competitors to the same budgetary constraint. It turns out that deviation from the usual set-up sometimes leads to better sampling 'strategies'.

END

FILMED

11-85

DTIC