MULTIVARIATE RATIO-TYPE ESTIMATORS

B. V. Sukhatme, Iowa State University
Lal Chand, J. N. K. V. V.

1. Introduction

Let a finite population consist of N distinct identifiable units U_i with values $x_{0i}, x_{1i}, \ldots, x_{\lambda i}$ $i = 1, 2, \ldots, N$ of the characteristics $X_0, X_1, \ldots, X_\lambda$. Consider the problem of estimating the population mean $\bar{x}_{ON} = \frac{1}{N}\sum_{i=1}^{N} x_{0i}$ when data on two or more auxiliary characteristics X_i $i=1, 2, \ldots, \lambda$ correlated with X_0 are available or can be obtained easily. In this situation, it is customary to use data on auxiliary characteristics to obtain ratio-type estimators of $\bar{x}_{ON}$. Several authors including Olkin [3], Raj [4], Rao and Mudholkar [5], Shukla [6], Singh [7] and [8], Smith [9] and Srivastava [10, 11, 12] have proposed ratio-type estimators utilizing data on several auxiliary variables. The estimators involve unknown weights which have to be estimated and assume knowledge of the population means of the auxiliary characteristics used. Clearly, none of the estimators proposed is satisfactory from the point of view of users and there is a need to investigate the matter further. The object of this paper is to present ratio-type estimators based on two or more auxiliary characteristics which do not involve unknown weights and at the most assume knowledge of the population mean of the auxiliary characteristic least correlated with X_0 along with appropriate expressions for bias and mean square error. Almost unbiased ratio-type estimators are also developed and a discussion is given concerning the efficiency of these estimators.

2. Multivariate Ratio-type Estimators

Let ρ_{0t} denote the correlation coefficient between X_0 and X_t. We shall assume that for $i < j$ $\rho_{0j} < \rho_{0i}$. We shall first consider the case when $\lambda = 2$ and assume three phase simple random sampling without replacement in which n_2 units are drawn from N in the first phase to observe X_2, a sub-sample of n_1 units is drawn from n_2 in the second phase to observe X_1 and a sub-sample of n_0 units is drawn from n_1 in the final phase to observe X_0. Let $\bar{x}_{tm}$ denote the sample mean based on m units corresponding to the characteristic X_t.

If $\bar{x}_{2N}$, the population mean of the characteristic X_2 is unknown, the ratio-type estimator of $\bar{x}_{ON}$ based on the use of two auxiliary variables X_1 and X_2 is defined as

$$t_{2d} = \frac{\bar{x}_{On_0}}{\bar{x}_{1n_0}} \frac{\bar{x}_{1n_1}}{\bar{x}_{2n_1}} \bar{x}_{2n_2} \qquad (2.1)$$

If $\bar{x}_{2N}$ is known, the ratio-type estimator of $\bar{x}_{ON}$ is defined as

$$t_2 = \frac{\bar{x}_{0n_0}}{\bar{x}_{1n_0}} \frac{\bar{x}_{1n_1}}{\bar{x}_{2n_1}} \bar{x}_{2N} \qquad (2.2)$$

The multivariate ratio-type estimator corresponding to λ auxiliary variables is now obvious. If $\bar{x}_{\lambda N}$ is not known, the estimator is defined as

$$t_{\lambda d} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda n_\lambda} \qquad (2.3)$$

If $\bar{x}_{\lambda N}$ is known, the estimator is defined as

$$t_\lambda = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda N} \qquad (2.4)$$

It is assumed that sampling is carried out in $(\lambda + 1)$ phases with simple random sampling without replacement in each of the phases and may be diagramatically described as follows.

$$N \xrightarrow{SRS} n_\lambda (X_\lambda) \xrightarrow{SRS} n_{\lambda-1} (X_{\lambda-1}) \cdots \cdots \xrightarrow{SRS} n_1 (X_1) \xrightarrow{SRS} n_0 (X_0)$$

where at a particular phase n_t denotes the sample size to be drawn at random from n_{t+1} and X_t denotes the characteristic to be observed on n_t units.

3. Bias and Mean Square Error of the Multivariate Ratio-type Estimators

Consider first the estimator $t_{\lambda d}$. By definition

$$\left.\begin{aligned} \text{Bias}\,(t_{\lambda d}) &= E\,(t_{\lambda d}) - \bar{x}_{ON} \\ \text{and MSE}\,(t_{\lambda d}) &= E(t_{\lambda d} - \bar{x}_{ON})^2 \end{aligned}\right\} \qquad (3.1)$$

It is not possible to obtain exact expressions for the bias and mean square error. However, expressing $t_{\lambda d}$ as a power series in powers of $\delta \bar{x}_{in_i} = \dfrac{\bar{x}_{in_i} - \bar{x}_{iN}}{\bar{x}_{iN}}$, ignoring terms of order higher than two and taking expectation term by term, we obtain

$$\text{Bias}_1\,(t_{\lambda d}) = \bar{x}_{ON} \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(C_{x_i}^2 - C_{x_i x_0}\right) \qquad (3.2)$$

and

$$\text{MSE}_1\,(t_{\lambda d}) = \bar{x}_{ON}^2 \left\{\left(\frac{1}{n_0} - \frac{1}{N}\right) C_{x_0}^2 - \sum_{i=1}^{\lambda}\left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(2C_{x_0 x_i} - C_{x_i}^2\right)\right\} \qquad (3.3)$$

where $$C_{x_i}^2 = \frac{S_{x_i}^2}{\bar{x}_{iN}^2}, \quad C_{x_i x_0} = \frac{S_{x_i x_0}}{\bar{x}_{iN}\bar{x}_{0N}} \qquad (3.4)$$

with $$S_{x_i}^2 = \sum_{1}^{N} (x_{it} - \bar{x}_{iN})^2/(N-1)$$

and $$S_{x_i x_0} = \sum_{1}^{N} (x_{it} - \bar{x}_{iN})(x_{0t} - \bar{x}_{0N})/(N-1) \qquad (3.5)$$

Following the procedure of David and Sukhatme [1], it can now be shown that

$$\left.\begin{aligned} |\,\text{Bias}\,(t_{\lambda d}) - \text{Bias}_1\,(t_{\lambda d})\,| &\le \frac{A_1}{n^2} \\ |\,\text{MSE}\,(t_{\lambda d}) - \text{MSE}_1\,(t_{\lambda d})\,| &\le \frac{A_2}{n^2} \end{aligned}\right\} \qquad (3.6)$$

where A_1 and A_2 are finite. It follows that (3.2) and (3.3) provide first order approximations to the bias and mean square error of the estimator $t_{\lambda d}$. In a similar manner, it can be shown that first order approximations to the bias and mean square error of t_λ are

$$\text{Bias}_1\,(t_\lambda) = \bar{x}_{0N}\left\{\sum_{i=1}^{\lambda-1}\left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(C_{x_i}^2 - C_{x_0 x_i}\right) + \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N}\right)\left(C_{x_\lambda}^2 - C_{x_0 x_\lambda}\right)\right\}$$

and

$$\text{MSE}_1\,(t_\lambda) = \bar{x}_{0N}^2\left\{\left(\frac{1}{n_0} - \frac{1}{N}\right) C_{x_0}^2 - \sum_{i=1}^{\lambda-1}\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(2C_{x_0 x_i} - C_{x_i}^2\right) - \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N}\right)\left(2C_{x_0 x_\lambda} - C_{x_\lambda}^2\right)\right\} \qquad (3.8)$$

Higher order approximations to the bias and mean square error have been obtained by Lal Chand [2]. However, the expressions are complicated and will not be presented.

If the population is assumed to be so large that finite correction factors can be ignored and is symmetrically distributed about its means, the expressions simplify considerably. In particular it can be shown that the second order approximations to the bias and mean square error for $\lambda = 2$ are given by

$$\text{Bias}_2\,(t_{2d}) = \text{Bias}_1\,(t_{2d})\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{C_{x_2}^2}{n_1}\right] \qquad (3.9)$$

$$\text{MSE}_2\,(t_{2d}) = \text{MSE}_1\,(t_{2d})\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{3C_{x_2}^2}{n_1} + \frac{3C_{x_2}^2}{n_2}\right] \qquad (3.10)$$

and

$$\text{Bias}_2\,(t_2) = \text{Bias}_1\,(t_2)\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{C_{x_2}^2}{n_1}\right] \qquad (3.11)$$

$$\text{MSE}_2\,(t_2) = \text{MSE}_1\,(t_2)\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{3C_{x_2}^2}{n_1}\right] \qquad (3.12)$$

where the first order approximations are obtained from the expressions (3.2), (3.3), (3.7) and (3.8) by taking $\lambda = 2$.

4. Almost Unbiased Multivariate Ratio-type Estimators

In this section, we shall present multivariate analogs of the ratio-type estimators presented in section 3 which are almost unbiased in the sense that the bias to the first order of approximation is zero. The estimators corresponding to $t_{\lambda d}$ and t_λ are

$$t_{\lambda dM} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda n_\lambda} \left[1 - \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) \left\{ \frac{s^2_{x_i}}{\bar{x}^2_{in_0}} - \frac{s_{x_i x_0}}{\bar{x}_{in_0} \bar{x}_{0n_0}} \right\} \right] \quad (4.1)$$

and

$$t_{\lambda M} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda N} \left[1 - \sum_{i=1}^{\lambda-1} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) \left\{ \frac{s^2_{x_i}}{\bar{x}^2_{in_0}} - \frac{s_{x_i x_0}}{\bar{x}_{in_0} \bar{x}_{0n_0}} \right\} - \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N} \right) \left\{ \frac{s^2_{x_\lambda}}{\bar{x}^2_{\lambda n_0}} - \frac{s_{x_\lambda x_0}}{\bar{x}_{\lambda n_0} \bar{x}_{0n_0}} \right\} \right] \quad (4.2)$$

Expressing $t_{\lambda dM}$ and $t_{\lambda M}$ as power series in powers of $\delta\bar{x}_{in_i}$, ignoring powers of order higher than two and taking expectation term by term, it can be verified that to the first order of approximation $t_{\lambda dM}$ and $t_{\lambda M}$ are almost unbiased estimators of $\bar{x}_{0N}$. Proceeding in a similar manner and evaluating their mean square errors, it can be seen that to the first order of approximation $t_{\lambda dM}$ and $t_{\lambda M}$ have the same mean square errors as $t_{\lambda d}$ and t_λ respectively. We have thus proved the following result

Theorem 4.1 The estimators $t_{\lambda dM}$ and $t_{\lambda M}$ are almost unbiased estimators of $\bar{X}_{0N}$.

Further, to the first order of approximation

$$MSE_1 (t_{\lambda dM}) = MSE_1 (t_{\lambda d})$$

and

$$MSE_1 (t_{\lambda M}) = MSE_1 (t_\lambda)$$

where $MSE_1 (t_{\lambda d})$ and $MSE_1 (t_\lambda)$ are given by (3.3) and (3.8) respectively.

5. Comparison of Estimators

For the purpose of comparison, we shall consider the mean square errors of the appropriate estimators to the first order of approximation only. Since $t_{\lambda d}$ and t_λ have the same mean square errors as $t_{\lambda dM}$ and $t_{\lambda M}$ to the first order of approximation, it is enough to consider $t_{\lambda dM}$ and $t_{\lambda M}$.

We have

$$V(\bar{x}_{0n_0}) - MSE_1 (t_{\lambda dM}) = \bar{x}^2_{0N} \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) (2C_{x_0 x_i} - C^2_{x_i}) \quad (5.1)$$

and

$$V(\bar{x}_{0n_0}) - MSE_1 (t_{\lambda M}) = \bar{x}^2_{0N} \left\{ \sum_{i=1}^{\lambda-1} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) (2C_{x_0 x_i} - C^2_{x_i}) + \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N} \right) (2C_{x_0 x_\lambda} - C^2_{x_\lambda}) \right\} \quad (5.2)$$

It follows that if

$$\rho_{0i} > \frac{1}{2} \frac{C_{x_i}}{C_{x_0}} \quad \text{for } i=1, 2, \ldots, \lambda \quad (5.3)$$

then both the estimators $t_{\lambda dM}$ and $t_{\lambda M}$ will be more efficient than the simple mean estimator $\bar{x}_{0n_0}$ which does not use auxiliary data on any of the 0 variables.

Further, we have

$$MSE_1 (t_{\lambda-1\, dM}) - MSE_1 (t_{\lambda dM}) = \left(\frac{1}{n_{\lambda-1}} - \frac{1}{n_\lambda} \right) (2C_{x_0 x_\lambda} - C^2_{x_\lambda}) \quad (5.4)$$

It follows that if inequality (5.3) is true

$$MSE_1 (t_{\lambda dM}) < MSE_1 (t_{\lambda-1\, dM}) \quad (5.5)$$

for all values of λ.

It can also be seen that if inequality (5.3) is true, then

$$MSE_1 (t_{\lambda M}) < MSE_1 (t_{\lambda dM}) \quad (5.6)$$

Combining all these results, we have the following

Theorem 5.1 If

$$\rho_{0i} > \frac{1}{2} \frac{C_{x_i}}{C_{x_0}} \quad \text{for } i=1, 2, \ldots, \lambda, \text{ then}$$

$$MSE_1 (t_\lambda) < MSE_1 (t_{\lambda d}) < MSE_1 (t_{\lambda-1\, d}) \cdots$$

$$< MSE_1 (t_{1d}) < V(\bar{x}_{0n_0})$$

and

$$MSE_1 (t_{\lambda M}) < MSE_1 (t_{\lambda dM}) < MSE_1 (t_{\lambda-1\, dM}) \cdots$$

$$< MSE_1 (t_{1dM}) < V(\bar{x}_{0n_0})$$

Finally, we shall compare t_λ for $\lambda = 2$ with the ratio estimator

$\hat{\bar{x}}_{0N} = \bar{x}_{0n_0} \frac{\bar{x}_{2N}}{\bar{x}_{2n_0}}$. Then noting that

$$MSE_1(\hat{\bar{x}}_{0N}) = \left(\frac{1}{n_0} - \frac{1}{N}\right) \bar{x}_{0N}^2 \left[C_{x_0}^2 + C_{x_2}^2 - 2C_{x_0x_2}\right]$$

it can be seen that $MSE_1(t_2) < MSE_1(\hat{\bar{x}}_{0N})$ provided

$$\sum_1^N \left(x_{0i} - \frac{\bar{x}_{0N}}{\bar{x}_{2N}} x_{2i}\right)^2 > \sum_1^N \left(x_{0i} - \frac{\bar{x}_{0N}}{\bar{x}_{1N}} x_{1i}\right)^2$$

The above condition would be always true provided X_1 is a better auxiliary variable than X_2 for ratio method of estimation as assumed in this paper.

We have seen that the ratio-type estimator based on λ auxiliary variables is more efficient than the one based on $(\lambda-1)$ auxiliary variables provided

$$\rho_{0\lambda} > \frac{1}{2} \frac{C_{x_\lambda}}{C_{x_0}} .$$

Although, this result is of considerable value, what is more interesting is to know whether the reduction in variance is worth the extra cost required to observe the additional auxiliary variable. For the sake of simplicity, we shall consider the case $\lambda = 2$ and choose that estimator for which the mean square error is minimum when the total cost of collecting data cannot exceed a specified amount C_0.

Consider a simple cost function of the form

$$C = c_0 n_0 + c_1 b_1 + c_2 n_2 \qquad (5.7)$$

where c_i is the cost per unit of observing the characteristic X_i i=0, 1, 2. We shall now determine n_i such that $MSE_1(t_{2dM})$ is minimum subject to the condition that $C \leq C_0$. It can be seen that the optimal values of n_i to achieve this are given by

$$\sqrt{\frac{Q_4/c_0}{n_0}} = \sqrt{\frac{Q_5/c_1}{n_1}} = \sqrt{\frac{Q_3/c_2}{n_2}} = \frac{\sqrt{Q_4 c_0} + \sqrt{Q_5 c_1} + \sqrt{Q_3 c_2}}{C_0} \qquad (5.8)$$

where

$$Q_3 = 2C_{x_0x_2} - C_{x_2}^2$$

$$Q_4 = C_{x_0}^2 - 2C_{x_0x_1} + C_{x_1}^2 \qquad (5.9)$$

and

$$Q_5 = C_{x_2}^2 - 2C_{x_0x_2} + 2C_{x_0x_1} - C_{x_1}^2$$

For optimal choice of the n_i, the optimal mean square error of the estimator t_{2dM} is given by

$$\left[MSE_1(t_{2dM})\right]_{opt} = \frac{\left[\sqrt{Q_4 c_0} + \sqrt{Q_5 c_1} + \sqrt{Q_3 c_2}\right]^2 \bar{x}_{0N}^2}{C_0} \qquad (5.10)$$

In a similar manner, it can be seen that

$$\left[MSE_1(t_{1dM})\right]_{opt} = \frac{\left[\sqrt{Q_4 c_0} + \sqrt{Q_1 c_1}\right]^2 \bar{x}_{0N}^2}{C_0} \qquad (5.11)$$

where

$$Q_1 = 2C_{x_0x_1} - C_{x_1}^2 \qquad (5.12)$$

and

$$V(\bar{x}_{0n_0})_{opt} = \frac{C_{x_0}^2 \bar{x}_{0N}^2 c_0}{C_0} \qquad (5.13)$$

Comparing the mean square errors, it can be seen that

$$MSE_1(t_{2dM})_{opt} < MSE_1(t_{1dM})_{opt} < V(\bar{x}_{0n_0})_{opt}$$

$$\text{if } \frac{c_2}{c_1} < \frac{\left(\sqrt{Q_1} - \sqrt{Q_5}\right)^2}{Q_3} \text{ and } \frac{c_1}{c_0} < \frac{\left(C_{x_0} - \sqrt{Q_4}\right)^2}{Q_1} \qquad (5.14)$$

Since $t_{\lambda dM}$ and $t_{\lambda d}$ have the same mean square error to the first order of approximation, it follows that

$$MSE_1(t_{2d})_{opt} < MSE_1(t_{1d})_{opt} < V(\bar{x}_{0n_0})_{opt}$$

provided (5.14) is true.

6. Numerical Illustration

For the purpose of illustration, we shall consider the census data relating to 99 counties of Iowa. The three characteristics we shall consider are

X_0: Bushels of apples harvested in 1964

X_1: Apple trees of bearing age in 1964

X_2: Bushels of apples harvested in 1959

For this population, we have

$$\bar{x}_{0N} = .293458 \times 10^4 \quad \bar{x}_{1N} = .103182 \times 10^4$$

$$\bar{x}_{2N} = .365149 \times 10^4$$

$$\rho_{x_0x_1} = .93 \quad \rho_{x_0x_2} = .84 \quad \rho_{x_1x_2} = .77$$

$$C_{x_0}^2 = .402004 \times 10^1 \quad C_{x_1}^2 = .255280 \times 10^1$$

$$C_{x_2}^2 = .209379 \times 10^1$$

$$C_{x_0x_1} = .297075 \times 10^1 \quad C_{x_0x_2} = .244329 \times 10^1$$

$$C_{x_1x_2} = .177110 \times 10^1$$

For the purpose of comparing the different estimators, we shall assume that we have a large population with population parameters as given above. Further, we shall take

$n_0 = 30$ $n_1 = 60$ and $n_2 = 120$

The relevant results for comparing the different estimators are given in Table 1 below.

As is to be expected, the ratio-type estimator t_{2dM} based on two auxiliary variables is the most efficient of all the three estimators, the gain in efficiency over t_{1dM} based on one auxiliary variable being 40% while that over the mean estimator is 139%.

7. References

[1] David, I. P. and Sukhatme, B. V. 1974, On the bias and mean square error of the ratio estimator, Journal of the American Statistical Association, 69, 404-466.

[2] Lal Chand 1975, Some ratio-type estimators based on two or more auxiliary variables. Ph. D. thesis, Iowa State University, Ames, Iowa.

[3] Olkin, I. 1958, Multivariate ratio estimation for finite populations, Biometrika, 45, 154-165

[4] Raj, D. 1965, On a method of using multi-auxiliary information in sample surveys, Journal of the American Statistical Association, 60, 270-277.

[5] Rao, P. S. R. S. and Mudholkar, G. S. 1967, Generalized Multivariate estimator for the mean of finite populations, Journal of the American Statistical Association, 62, 1009-1012.

[6] Shukla, G. K. 1966, An alternative multivariate ratio estimate for finite population, Calcutta Statistical Association Bulletin, 15, 127-134.

[7] Singh, M. P. 1967, Multivariate product method of estimation for finite populations, Journal of the Indian Society of Agricultural Statistics, 19, 1-10.

[8] Singh, M. P. 1967, Ratio cum product method of estimation, Metrika, 12, 34, 42.

[9] Smith, T. M. F. 1966, Ratio of ratios and their applications, Journal of the Royal Statistical Society, Series A, 129, 531-533.

[10] Srivastava, S. K. 1965, An estimation of the mean of a finite population using several auxiliary variables, Journal of the Indian Statistical Association, 3, 189-194.

[11] Srivastava, S. K. 1967, An estimator using auxiliary information in sample surveys, Calcutta Statistical Association Bulletin, 16, 121-132.

[12] Srivastava, S. K. 1971, A generalized estimator for the mean of a finite population using multi-auxiliary information, Journal of the American Statistical Association, 66, 404-407.

Table 1

Estimator	Mean Square Error	Relative Efficiency w. r. t. $\bar{x}_{0n_0}$	Relative Efficiency w. r. t. t_{1dM}
$\bar{x}_{0n_0}$	$.115399 \times 10^7$	1	0.59
t_{1dM}	$.676577 \times 10^6$	1.70	1
t_{2dM}	$.481886 \times 10^6$	2.39	1.40